Scale-free growth in regional scientific capacity building explains long-term scientific dominance

Servedio, V.D.P.; Ferreira, M.R.; Reisz, N.; Costas Comesana R.; Thurner, S.

Chaos, Solitons and Fractals 167 (2023) 113020

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

journal homepage: www.elsevier.com/locate/chaos

Scale-free growth in regional scientific capacity building explains long-term scientific dominance

Vito D.P. Servedio [a,1], Márcia R. Ferreira [a,1], Niklas Reisz [a], Rodrigo Costas [b,c], Stefan Thurner [a,d,e,*]

[a] Complexity Science Hub Vienna, Josefstädter Str. 39, 1080 Vienna, Austria
[b] Centre for Science and Technology Studies (CWTS), Leiden University, Willem Einthoven Building, Kolffpad 1, 2333 BN Leiden, The Netherlands
[c] DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, RW Wilcocks Building, Stellenbosch, Western Cape, South Africa
[d] Section for Science of Complex Systems, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, A-1090 Vienna, Austria
[e] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

ARTICLE INFO

Keywords:
Science of science
Preferential attachment
Critical mass
Scaling

ABSTRACT

The regional capability of performing front-running research and technological development has been identified as a necessary condition for future wealth creation. The quality and amount of scientific capabilities in specific fields vary dramatically across different world regions. Capabilities, knowledge, and skills are embodied by scientists working in research institutions or companies. The conditions for the emergence of a leading regional scientific environment – and the resulting early technological leadership – are poorly understood. The existence of a critical mass – the threshold above which a region can build comparably strong scientific capabilities – of scientists is often assumed. Using a unique dataset of global scientific activity and researcher mobility over several decades, we present empirical evidence in three scientific areas (semiconductor research, embryonic stem cells, and Internet research) that the process of scientific knowledge accumulation is remarkably general and applies to practically all regions. Regional knowledge accumulation data obtained from an analysis of scientists' geolocated trajectories follow a preferential attachment mechanism characterized by a sub-linear kernel with a robust growth exponent. Scale-free growth patterns suggest that regions that move early into new technologies tend to dominate the corresponding scientific fields. We find no evidence that critical mass is required to achieve prolonged scientific dominance. We propose a simple preferential attachment model that explains the empirical data and allows us to understand deviations from the growth exponent as focused interventions to strategically attract scientists at regional level. We demonstrate this explicitly for China in the three scientific fields examined.

1. Introduction

It has long been known that a region's ability to attract scientists is key to its future scientific and economic development [1]. Far less is known about how new fields of science are created and how a leading edge is achieved, established, and maintained. Are there optimal strategies for regional institutions and stakeholders to achieve global leadership and thus economic benefits? In this context, the success story of Silicon Valley comes to mind, in particular, what led to the development of semiconductor science that later made the region the largely unchallenged leader in these technologies for many decades [2]. Why could not other regions do the same? To explain what it takes for regions to accumulate scientific capacity and sustain decades of dominance, it is often argued that a minimum number of researchers is required to grow and create a self-sustaining capacity-a *critical mass* [3,4].

Global scientific mobility and international collaboration [5] further reinforce the spatial concentration of knowledge, favoring regions that are becoming major scientific hubs [6]. Silicon Valley is often cited as a canonical example of the regional growth and dominance of high-tech industry and scientific knowledge. It developed in the 1950s when the U.S. military nurtured companies in the state of California with multibillion-dollar contracts. Electrical and electronics companies then settled in Santa Clara to take advantage of the city's location near defense-related aircraft, missile, and space markets [2]. Before this,

* Corresponding author at: Section for Science of Complex Systems, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, A-1090 Vienna, Austria.
E-mail address: stefan.thurner@meduniwien.ac.at (S. Thurner).
[1] These authors contributed equally to the paper.

https://doi.org/10.1016/j.chaos.2022.113020
Received 20 July 2022; Received in revised form 8 November 2022; Accepted 10 December 2022
Available online 21 December 2022

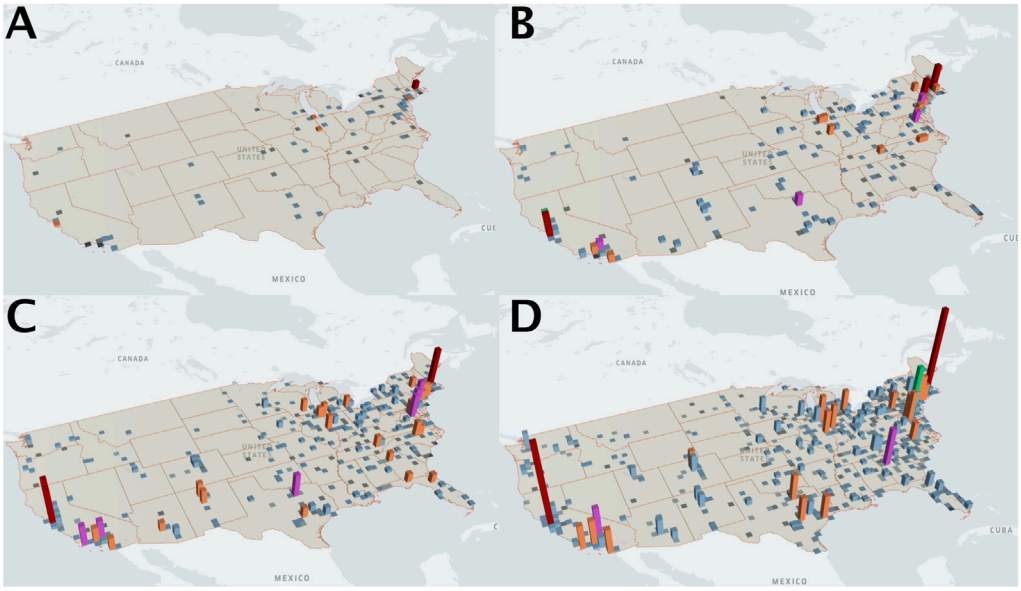

Fig. 1. Regional evolution of the number of in-flowing researchers into the field of semiconductors in the US (height of bars) Panels A–D correspond to time periods 1954–1970, 1970–1986, 1986–2002, and 2002–2020, respectively. In the fifties and sixties, only the Boston area plays a role, some researchers are present in Chicago and Silicon Valley. From the seventies on, the Boston area and Silicon valley are dominant; other regions are building up some expertise but never get into a position of becoming a challenger. The five colors (blue, orange, purple, green, and red) indicate steps of 20% of the aggregated sum of authors' flows value of the inflows at each time period. The respective maximum values for panels A–D are: max = 57, 207, 1086, and 8270. For the definition of a region, see methods Section 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the San Francisco Bay Area was already a place where the electronics industry existed, which shows that the Bay Area needed some time to build momentum [7]. While there is little doubt that Silicon Valley has evolved from a productive environment to one of the leading areas for high-tech industries and scientific knowledge, producing important new technologies and industrial clusters, the underlying mechanisms and critical parameters for this development are still unclear.

With ever more complete bibliographic databases available, it is possible to reconstruct the historical buildup of regional capacities in specific fields in all regions of the world. This is possible by reconstructing the mobility of individual scientists and their productivity in terms of individual papers in specific scientific fields [5,8]. Researchers can be reliably tracked in time and space by the publication date of articles and the history (temporal sequence) of the affiliations where they produced their scientific output. Recent advances in bibliographic databases have also greatly improved the consistency and quality of metadata extracted from publications, with particular attention to author-affiliation links [5,8].

Here, we use the *Dimensions* database, see Appendix A. It contains millions of publications and data spanning several decades. The database includes about 20 million disambiguated researchers associated with a researcher ID. The database contains links to the Global Research Identifier Database, which currently covers more than 98,000 research institutions worldwide [9,10]. *Dimensions* currently covers more local journals than any other large-scale bibliographic database such as *Web of Science* or *Scopus*. Its broader scope allows us to include more organizations with better local resolution [9].

With the *Dimensions* dataset, we can reliably reconstruct the changes in affiliations of all researchers in a specific field and infer how they move between geographic regions. We can see which region attracts researchers with certain scientific skills in a time-resolved manner. We see how quickly different regions accumulate specific types of skills over time, which allows us to define regional growth rates of scientific skills. A schematic picture of semiconductor researchers moving into different regions over seven decades is shown in Fig. 1. The reasons why researchers switch locations are plentiful. In addition to the attractiveness of a region and the everyday surroundings, the scientific environment and the innovation potential also play an important role [6,11]. This also includes the number of scientists already present in a region [12].

This paper studies the growth mechanism leading to the observed temporal and regional distributions of accumulated scientific skills in three specific scientific fields: Semiconductor Research, Embryonic Stem Cells, and Internet Research. We identify the underlying mechanism as the so-called "rich-get-richer" effect and find no evidence of a critical mass of scientists. In particular, in our analysis of empirical flows of scientists (and their skills), we find strong evidence for a sub-linear preferential attachment (PA) growth mechanism, meaning that a region becomes increasingly attractive to scientists as the number of existing scientists in a field increases.

Preferential attachment mechanisms (or rich-get-richer or Mathew effect) have been identified in a wide variety of social [13–16] and network phenomena [17–19]. For a given quantity of interest (integer), $x(t)$, at time t, growth following a PA mechanism means that the probability of the quantity gaining one unit within the next timestep is

$$\mathcal{P}[x(t+1) = x(t) + 1] \propto x(t)^{\alpha} . \tag{1}$$

Preferential attachment can appear in linear, sub- and super-linear versions, depending on whether the growth exponent $\alpha = 1$, $\alpha < 1$, or $\alpha > 1$, respectively. Sub-linear PA growth has been reported, for example, in the actor collaboration network and scientific co-authorship networks [18,19], as well as in the friendship network formation in a massive multiplayer online computer game [20]. The microscopic mechanisms underlying a sub-linear PA dynamics may be of various types, for example it could be an ageing effect, where scientists prefer newborn institutions, or it could be a fitness-related attachment, where researchers prefer institutions with a high impact, or it could be hiring selection rules, where institutions may lower the entry requirements for the candidates. A few publications consider these effects in their modeling efforts [21–24] but do not explicitly relate them to a sub-linear *effective* PA kernel. In our context we observe that a PA mechanism does not need the knowledge of the entire network for the researchers to move to a region. Picking random scientists (colleagues) and copying their behavior leads to a PA dynamics. In our study, we do not consider

Table 1
Descriptive features of the data set used in the three studied scientific fields.

	Semiconductors	ESC	Internet
Starting year	1941	1941	1956
End year	2019	2019	2019
Researchers	2,011,170	752,575	109,098
Articles	5,062,639	1,083,100	246,953
Regions	1633	1161	1032
Heaps' exp. γ	0.39	0.37	0.48
PA exponent α	0.79	0.84	0.79

the details of the causes leading to an effective power-law sub-linear PA. We focus on how to measure it and demonstrate its importance for explaining the regional growth of scientific fields.

To test if the conclusion of an underlying effective sub-linear PA mechanism is valid, we design a simple global model of the regional attractiveness of regions to scientists. The model not only replicates the empirical frequency distribution of researchers in different areas but also helps us understand the peculiar behavior of regions that are late adopters of new scientific fields. We illustrate the case of late adopters in the context of semiconductor research, where Silicon Valley intellectually dominated the scene for a long period, from the end of the 1940s on [25]. Only very late China entered the scene, starting in the early 1980s.

From 2006 on, China became the most dominant player in semiconductor research, also thanks to the Chinese "Medium- and Long-Term Plan for the Development of Science and Technology" [26]. We find that the cumulative number of scientists moving to China and publishing scientific papers exceeds those of California from 2007 on. Understanding how this take-over was possible needs special attention — a naive explanation with PA growth will not be sufficient. Change of leadership is only possible if special action is taken to locally change the PA mechanism with the help of strategic interventions that make regions radically more attractive by investments in technology [27] or research and development [28]. We show that our model can implement these interventions and we quantify how large these efforts have to be to allow regional latecomers to become dominant.

2. Results

We first characterize the mobility of scientists, in particular the growth (and change) of numbers of scientists of a given field working in different world regions. This is based on analyzing the temporal sequences of their publications and affiliations. We define a region as the first-level administrative division of a country. In the case of the United States, such a region corresponds to a state. For the data collection procedure, see Appendix A; for an overview of the data set, consult Table 1.

To arrive at a practical data structure for the analysis, we define an event, $E(T)$, as a publication of an article in a given field of research at time T (date of publication) by a scientist. It consists of four entries, $E(T) = (S, R, T, F)$, the scientist, S, the region, R, with the condition that S has published in the region R for the first time (i.e., the pair (S, R) has never occurred before, and the scientific field is F. We skip an event for which this condition is not fulfilled. We ensure that we only record the scientists that move into a new region or who start their careers. Our work focuses on academic mobility from the perspective of receiving countries and regions. Since we mainly focus on how regions receive new people in specific scientific domains, we do not account for returned mobility (i.e., people who return to a location where they previously have become "attached" to). Post-migration retention, attrition, and returning are highly significant dimensions of the knowledge transfer equation and are ideal for follow-up studies focusing on questions of *brain circulation* and strategies to mitigate the effects of *brain drain*. We do not deal with these themes here.

Table 2
Sequence of publication events, E, in the field of physics. Intrinsic time increases by one each time a scientist publishes an article in a given region for the first time. From this data, all necessary sequences can be extracted. In the specific example, Bob has already published in Arizona at intrinsic time $t = 2$, so all his future publications in Arizona no longer appear in the stream. Using initials for regions, this example generates the regional stream, $\mathcal{R} = (A, A, A, C, A, D, C)$.

Paper	Real time	intr. time	Scientist	Region	Field
Article 1	1960	1	Alice	Arizona	phys.
Article 1	1960	2	Bob	Arizona	phys.
Article 1	1960	3	Charlie	Arizona	phys.
Article 2	1961	4	David	California	phys.
Article 3	1961	–	Bob	Arizona	phys.
Article 3	1961	5	David	Arizona	phys.
Article 4	1962	6	Alice	Delaware	phys.
Article 5	1963	–	Bob	Arizona	phys.
Article 5	1963	–	David	Arizona	phys.
Article 5	1963	–	Alice	Arizona	phys.
Article 6	1964	7	Bob	California	phys.

One publication may generate multiple events according to the number of scientists in the author list. Here, we consider a full counting of all authors and their affiliations (and regions) recorded in the publications. Note that a "fractionalization" of researchers' and regions' contributions to papers, where researchers and regions are assigned a relative weight, makes little sense in the present context since we build sequences of events. Moreover, fractional counting is usually not considered when the analysis is conducted for a single discipline, nor when studying mobility flows, as in our case [29–31]. After sorting events, $E(T)$, ascending in time, we get a sequence, E_t, where t is an index that indicates the event's position in the sequence. We call t the *intrinsic time*. For every index, t, there is an associated time, T, given by a non-decreasing function $T(t)$. Intrinsic time is a convenient way for treating sequences of events whenever their rate of appearance in real-time is not constant. This is indeed the case since, in most fields, the number of published articles increases exponentially in real-time [21], and, correspondingly, the real-time difference between two consecutive articles decreases exponentially. We can consider the derivative of intrinsic time with respect to real-time, as a proxy for attractiveness of new locations. We increase intrinsic time whenever someone moves to a new location in their life. This means that the field has to be sufficiently attractive to justify a person with their family to move or to justify the addition of a new affiliation to one's list. From the sequence, E_t, we extract the corresponding sequences of regions, R_t, which we call *regional stream*, $\mathcal{R}$. For example, see Table 2.

Given a regional stream, one defines two useful quantities. First, the cumulative number of new scientists who moved into the region i (or started their publishing career there), before time t

$$k_i(t) = \sum_{\tau=1}^{t} \delta(R_\tau, i)\,, \tag{2}$$

where $\delta(x, y)$ is the Kronecker symbol, $\delta(x, y) = 1$ if $x = y$ and zero, otherwise. The second quantity is the number of different regions appearing in the regional stream before time t

$$D(t) = \sum_{\tau=1}^{t} \delta(k_{R_\tau}(\tau), 1)\,. \tag{3}$$

$D(t)$ is the number of regions appearing at least once and resembles the "regional diversity" of a field. We only focus on the number of papers published within a region and ignore their impact and quality.

The sub-linear PA kernel. From the time evolution of the quantities, $k_i(t)$, we can estimate the probability for a scientist to move to a new location of work (and publish there for the first time) that has already had k scientists up to time $t - 1$:

$$P_k(t) = \mathcal{P}\left[k_i(t) = k + 1 \mid k_i(t-1) = k\right]\,. \tag{4}$$

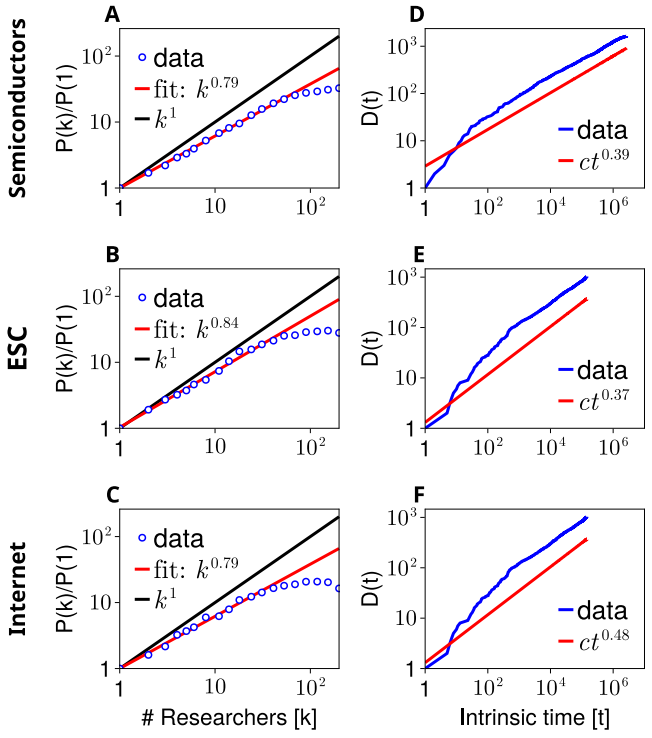

Fig. 2. Growth statistics of the three fields. Panels A-C show the probability for a scientist to move into a region where k researchers are present. The attachment probability, $P(k)/P(1)$ (PA kernel), (blue) shows a sub-linear power-law increase as a function of k. The least-square fits are shown in red, the black line indicates the linear exponent, $\alpha = 1$. (A) shows the case for semiconductor science, B for ESC, and (C) for Internet research. For the sake of readability, in panels A-C we show only data for small k where we find an approximately linear relation in the log–log plot. We do not show data for $k > 200$ where curves drop due to insufficient statistics. Panels D–F shows the increase of regional diversity, $D(t)$, occurring in the regional stream, $\mathcal{R}$, as a function of intrinsic time, t (blue). After an initial transient, $D(t)$ approaches an approximate power-law (Heaps' law). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

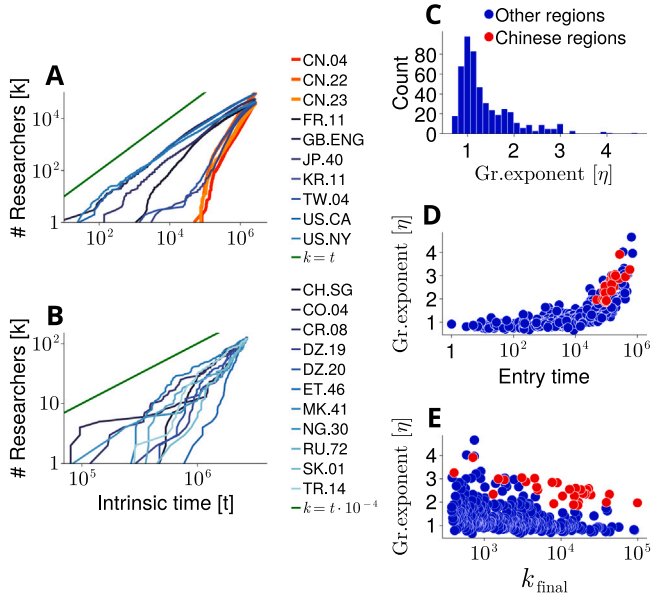

Fig. 3. Panels A and B show the cumulative number of scientists coming into a region (lines) in intrinsic time in semiconductor research; (A) shows the 10 most visited regions. Most regions grow sub-linearly. Chinese regions are shown with red thick lines and increase super-linearly. In (B), we see 11 poorly visited regions ranked from 700 to 710 regarding the number of scientists. Curves are (additively) shifted to the left so that their first point is placed at coordinates (1,1); In both (A) and (B), the green straight lines on the top represent the linear regime. (C) Histogram of growth exponents, η, for all regions; (D) Growth exponents, η, of all regions, i, as a function of their entry time, t_i (intrinsic time). Note that Chinese regions (red) come in late (after timestep 10^5) and have higher exponents than other regions (blue). (E) Growth exponents, η, as a function of the cumulative number of scientists in 2019, k_{final}. The high exponents of Chinese regions are visible (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For simplicity, we assume that $P_k(t)$ does not explicitly depend on time, t, but on the number of occurrences, k, of the regions in the regional stream, and we write $P_k(t) \approx P(k)$. We will see that this approximation already explains the experimental data well. We estimate $P(k)$ with the running histogram method [18], briefly described in the methods Section 4.

In Fig. 2 panels A, B, and C, we show the empirical settlement probabilities, $P(k)$, for a scientist (in the fields of semiconductor research, embryonic stem cells, and Internet research, respectively) to move to a new location of work and publish there for the first time — as a function of the number of scientists who already work in that region, k. The probability is normalized by $P(1)$; we call $P(k)/P(1)$ the effective *attachment kernel* [32].

Clearly, with increasing numbers of already present scientists, a power-law increase of the attachment kernel is visible. The attachment exponents, defined in Eq. (1), are $\alpha \sim 0.79$ for semiconductor and Internet research, and 0.84 for ECS (see Table 1). For reference, the black lines indicate the linear PA mechanism, i.e., $\alpha = 1$.

Panels D, E, and F of Fig. 2 show the number of different regions appearing in the stream $D(t)$ as a function of intrinsic time. After a brief linear transient that ends around $t \approx 100$ publications in all three fields, D increases as a power-law, $D(t) \approx t^{\gamma}$, with $\gamma < 1$. The initial linear growth, which is the fastest growth possible in intrinsic time, reflects the fast diffusion of the three scientific fields across the world at their onset. As time goes on, the initial constant rate, $\frac{d}{dt}D(t)$, turns into a time decreasing regime with an exponent $\gamma - 1$. We observe no saturation effects at high intrinsic times despite the number of different regions in the three cases has crossed (semiconductors, embryonic stem cells), or is close to (Internet), the 50% of the total number of regions in the world, i.e., 2092 in the year 2019.

A power law increase in novel entries in a stream has been associated with the existence of the so-called *Heaps' law* that often appears in evolutionary time series [33–37]. In the present case, Heaps' law shows how fast a new scientific field spreads in regions around the globe.

Growth of regional capacity. In Fig. 3, we show the cumulative number of semiconductor scientists working in several selected regions in intrinsic time. We do not distinguish those scientists leaving a region for a new one from those who start in a region for the first time in their career in the semiconductor field. We fit with a least square procedure the regional growth curves with $k(t) \sim t^{\eta}$, where η is the *growth exponent*, that can be interpreted as the *regional growth capability*. In all the three scientific fields we find $R^2 > 0.9$ for the η fits, with very few values between 0.85 and 0.90.

In panel A, every curve represents a region that belongs to one of the 10 most successful ones (cumulative number of scientists in the order of 10^5). Most curves grow sub-linearly with the exception of the three regions in China (denoted by CN.{04,22,23}); these grow faster than linear (left-bent). Both California (US.CA) and England (GB.ENG) have an exponent of approximately 0.8 (yellow and green curves practically overlap), while the Chinese region around Beijing (CN.22) shows an exponent of about 1.8. This means that Chinese regions follow a very different growth pattern. Note that Chinese regions entered the scene only after about 90,000 papers were published in the field, corresponding to the year 1977.

In panel B, we show the growth curves of 11 low-ranked regions (rank 700–710). These all have a final cumulative number of scientists of about 120. All of them grow super-linearly with an exponent of

approximately 1.6. The general trend is that well-established regions in a field generally grow sub-linearly (curves bent to the right), while late adopters grow super-linearly.

Fig. 3C shows the histogram of the growth exponents, η, for the semiconductor case. We find a distribution with mean $m = 1.55$, standard deviation $s = 0.69$ and skewness $b = 1.52$. For the other fields the situation is similar: Internet: $m = 1.36$, $s = 0.59$ and $b = 1.48$; embryonic stem cells: $m = 1.56$, $s = 0.86$ and $b = 1.73$.

In Fig. 3D we show the growth exponents, η, of all world regions, i, as a function of their entry time, t_i, defined as the time (measured in intrinsic time) at which a region appears in the affiliation list of an article for the first time. Exponents are larger the later a region enters. Late adopter regions seem to bring in scientists faster than regions where the field started. Yet, in most cases, higher exponents are not enough to catch up and challenge the leaders in the field. Chinese regions are marked in red. They come in late (10^5) and have exponents higher than other regions (blue).

Finally, in Fig. 3E we see the regional growth exponents, η, as a function of the cumulative number of scientists up to year 2019 in a region, k_{final}. The visual general decay means that the more scientists are present in a region, the lower is its growth capability. Note again the exceptionally high growth exponents for the Chinese regions.

Relations between the three exponents. The exponents α, γ, and η are related. The corresponding functional expressions can be estimated with simple reasoning, see Appendix B, or are obtained through an approximate analytic solution of the model, see Appendix C.

A simple generative model. To understand the observed statistical features presented in Figs. 2 and 3 we devise a simple model. We first specify a scientific field, e.g., semiconductor research. We use the empirical regional stream, $\mathcal{R}$, as the baseline to build a new synthetic stream, $\mathcal{S}$. We start at $t = 0$ with the first region, R_0, where a scientist's affiliation in an article appeared for the first time. We insert this region in $\mathcal{S}$ as its first element S_0. For each following intrinsic time, $t > 0$, we add an element S_t in $\mathcal{S}$ in the following two ways according to whether R_t has already appeared in $\mathcal{R}$ or not:

- if region R_t has never appeared in $\mathcal{R}$ before, i.e., $k_{R_t}(t) = 1$, we also insert it in $\mathcal{S}$, so that $S_t \equiv R_t$;
- if region R_t has already appeared in $\mathcal{R}$, i.e., $k_{R_t}(t) > 1$, we randomly select a region s, from those regions already in the stream $\mathcal{S}$, with a preferential attachment probability

$$\begin{aligned} P_k(t) &= \mathcal{P}\left[k_s(t) = k+1 \mid k_s(t-1) = k\right] \\ &= Z^{-1}k^{\alpha}, \end{aligned} \tag{5}$$

where $Z(t) = \sum_{s=1}^{D(t)} k_s^{\alpha}(t)$ is a normalization term. In other words, we choose regions with a probability proportional to a power, α, of their number of occurrences in $\mathcal{S}$ until time $t-1$.

Since the entry times of new regions in the two streams, $\mathcal{R}$ and $\mathcal{S}$, are the same by construction, the number of different regions in time, $D(t)$, coincides in the two streams, meaning that also Heaps' law is the same (Fig. 4B). The model has one free parameter, the PA attachment exponent, α, of the sub-linear PA mechanism. Note that this model is similar to the one presented in [38] where, however, the PA was linear. The model can be approximately solved analytically; see Appendix C.

For the bulk of the model simulations, we chose α as measured in the data, i.e., $\alpha = 0.79$, 0.79, and 0.84 for the Internet, semiconductors, and stem cell areas, respectively; see Table 1. For this choice, a run of the model for the semiconductor research followed by a simple linear regression yields $\gamma \approx 0.39$ and $\eta \approx 0.74$.

Understanding exceptional super-linear regions. To model the exceptional super-linear growth of Chinese regions in the field of semiconductors, visible as the red curves in Fig. 3A, we use as an input a slightly larger PA exponent $\alpha_{\text{CN}} = 0.915$ for these regions only, while for all the other regions we keep $\alpha = 0.79$, see the methods Section 4.

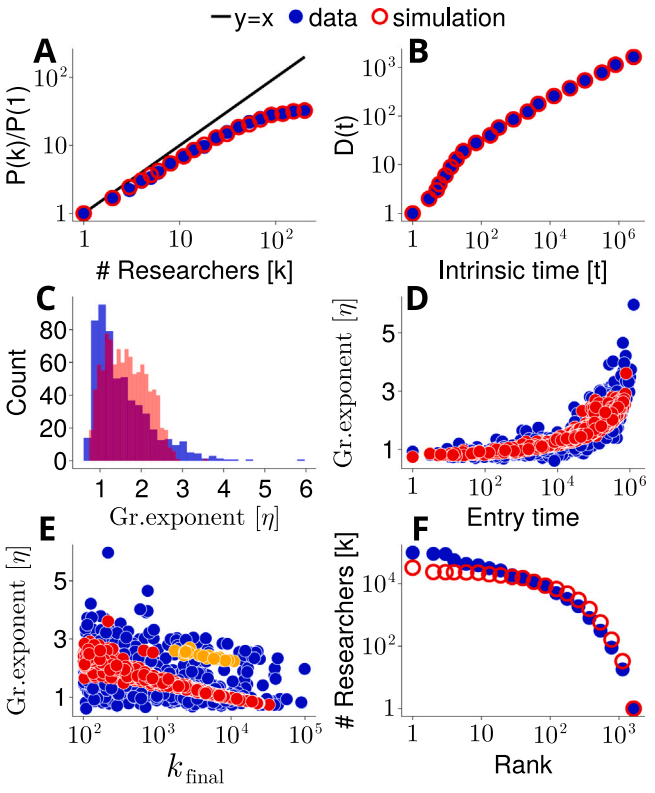

Fig. 4. Comparison of model results (red) with experimental data (blue) for the field of Semiconductors. (A) PA kernel; data and simulation practically coincide; the black line depicts a linear kernel. (B) Heaps's law; the two curves coincide by construction. (C) Histograms of the regional growth exponent, η. Simulations show a slightly higher mean. (D) Exponents, η, reproduce the empirical increase as a function of entry time. (E) The decrease of η as a function of the cumulative number k_{final} of scientists in the year 2019 is well captured by the model. Note the fact that Chinese regions (orange circles) are predicted with higher than usual values. (F) Also, the frequency-rank distributions of the regions appearing in the two streams practically coincide except for very small ranks. In panels C, D, and E, only regions with $k_{\text{final}} > 100$ are taken into account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 4, we compare the model results (red) with the empirical data (blue). In panel A, we show the PA kernel, as in Fig. 2, panels A, B, and C. In the simulation, we use the empirical value of the PA exponent of $\alpha = 0.79$ for all the 1633 regions (except the 31 Chinese regions for which we use $\alpha_{\text{CN}} = 0.915$). The two curves almost coincide even for $k > 60$, i.e., outside the interval used to fit α. This confirms the goodness of the fit.

In Fig. 4B, we plot the number of distinct regions, $D(t)$, as a function of intrinsic time, in the regional, $\mathcal{R}$, and the synthetic stream, $\mathcal{S}$. Since the entry time of new regions is the same in both streams, the two curves are identical.

Fig. 4C compares the distribution of the growth exponents, η. In the simulation, the distribution has a somewhat higher mean than the data and lower skewness; we get mean $m = 1.63$, standard deviation $s = 0.49$ and skewness $b = 0.28$. For the data we found $m = 1.55$, $s = 0.69$ and $b = 1.52$. The model under-populates the large exponents and underestimates the small ones.

In Fig. 4D, the growth exponent at the entry time of a given region is plotted against the entry time in intrinsic time. The simulation results explain the data, particularly that latecomers tend to have higher growth exponents.

Fig. 4E shows the comparison for the growth exponent, η, as a function of the cumulative number of researchers in a region in 2019, k_{final}. The more scientists are present, the smaller the growth exponent. Fewer scientists correspond to higher entry times and, thus, to higher

exponents. The red circle marks the Chinese regions in the model to which we assigned higher α values. Also, for these, the model practically reproduces the data.

Finally, Fig. 4F gives the frequency rank distribution of the regions appearing in the two streams. The model also reproduces the distribution obtained from data well. A consequence of a sub-linear PA is that the corresponding frequency-rank of the cumulative number of distinct scientists who worked in a region is not an exact power-law [17] – a fact that we hereby verify.

In Appendices E and F we show the results for the embryonic stem cells and Internet research in the same manner as in Figs. 3 and 4.

3. Discussion

Using metadata from scientific publication databases, we reconstructed the regional cumulative number of publishing scientists across the world regions. We find that researchers tend to move to regions according to a sub-linear PA mechanism, where the settling probability for a region is proportional to a power of the cumulative number of scientists already working in that region. In other words: We find evidence for the rich-get-richer phenomenon of scientific mobility of a sub-linear type. This means that regions that move early-moving are likely to dominate the field (or become one of the few dominant regions) and then retain that dominance for a long time.

We find power-law-like regional growth curves that indicate the absence of any specific scale that would reflect the presence of a critical mass. We find no signs of a critical mass (or density) of scientists in a particular topic necessary before the field takes off regionally; a simple PA model is sufficient to explain the main features of the data.

The number of researchers in early-moving regions tends to grow sub-linearly in intrinsic time. The more researchers there are, the lower the regional growth rates. Latecomer regions, which start attracting researchers at a later stage (after the first several hundred articles have been published), are characterized by a number of researchers that can be several orders of magnitude smaller than that of pioneer regions. We find evidence that latecomer regions tend to grow faster, typically even superlinearly, in their initial phase, but with numbers so small that they can practically never even approach dominance. The exponents of regional growth increase with the delay of entry into the field relative to the age of the discipline. Since regions that have been in research for a long time have higher cumulative numbers of scientists than regions that enter late, regional growth exponents tend to decrease as more scientists populate the region. We demonstrate the existence of the same generic mechanism in three different scientific domains over many decades.

For these three domains, we find that Chinese regions behave differently from all the others. They do not follow the sub-linear growth of the pioneering regions but grow super-linearly (in intrinsic time), starting after 500,000 articles have been produced in the field. This suggests that Chinese regions follow a different pattern. The simplest way to explain the differences is to assign them a higher PA exponent — still being sub-linear. This means that as compared to the other regions, Chinese ones attract scientists with higher probability. We speculate that Chinese regions become more attractive thanks to strategic efforts, including proactive hiring policies, massive funding, and by training a huge reservoir of scientists abroad and bringing them back. We emphasize that the PA exponent does not depend on the multiplicity of Chinese scientists (population effect) since it determines how researchers attach to a region. In the first approximation, the PA mechanism does not depend on the number of attaching nodes. Of course, there could be an indirect effect by which the high number of scientists requires the introduction of *ad hoc* policies. Fig. 3E shows how the growth exponents decrease with the final number of scientists, with Chinese regions being the outlier red circles. Introducing a higher PA exponent for Chinese regions in our model yields the correct massively larger growth exponents, η, that are observed. The deeper reasons why Chinese regions behave differently by attracting researchers more strongly than others – still under a PA scheme – have to be investigated in future research.

Note that the study considers all regions as having the same PA exponent, except for the Chinese regions that have a larger one. This is a gross simplification. However, it still explains the data reasonably well. Nevertheless, the shortcoming of identical exponents should be addressed in future work by a more precise description that assigns different exponents to different regions, ideally on an empirical basis. How this can be done is not so clear.

Finally, we mention that, remarkably, we find that the rate at which regions join a specific field follows an approximate power-law. This feature has been referred to as Heaps' law that occurs in numerous evolutionary processes [33–35,37]. Heaps' law is intimately connected to the idea of the adjacent possible [36], where the discovery of new things or ideas leads to other discoveries and results that were impossible to achieve before [39].

In summary, our simple PA model captures the essence of the underlying preferential attachment process. To a large extent, it explains the empirical data of the historical evolution of the three studied disciplines. It captures both the growth of scientific capacity in individual regions and the observed super-linear initial growth of latecomer regions. The model is simple enough to be analytically tractable, which allows for a detailed understanding of the relations between the three power-law exponents involved in the process. In particular, it explains why faster growth of latecomer regions is observed.

The popular explanation that attaining (or catching up to) leadership in a scientific discipline requires building a critical mass of scientists is not supported by data on scientific mobility. We conclude that there are two ways to secure leadership and scientific dominance in a field: One is to have a strong presence among the earliest and earliest players, which is entirely consistent with the theory of increasing returns [40]. The other is to ensure the continuation of exceptionally high, superlinear growth rates through strategic interventions that must be sustained over long periods of time (decades). The latter has been most evident in some areas of Chinese science: The beginning of the catch-up process in the late 1970s led to a dominant role today.

4. Methods

Estimating the PA kernel. We estimate the PA attachment probability, $P(k)$, by means of a method originally devised for co-authorship networks [18]. Following this method, we go through the regions in the time-sorted regional stream, $\mathcal{R}$, one by one and build a histogram from which we estimate the attachment kernel. If a PA process is realized, at each point in time t, the probability of a region already occurred k times to appear again is given by

$$\Pi(k,t) = P(k)\frac{n(k,t)}{D(t)}, \tag{6}$$

where $n(k,t)$ is the number of regions already appeared k times at time t, and $D(t)$ is the total number of regions already appeared irrespective of their number of occurrences. Then, the attachment probability, $P(k)$, can be estimated from a histogram where each contribution is weighted with the inverse frequency $\frac{D(t)}{n(k,t)}$. The resulting histogram starts to deviate from a straight line (in double logarithmic scale) at around $k \approx 60$ (Fig. 2 panels A–C). This is a well-known behavior of the method. Therefore, we resort to the first 60 points to estimate the exponents with a least squares method in all three scientific fields studied. In all the three fields we find $R^2 > 0.99$.

Determining the PA exponent for Chinese regions. We obtain the special value of $\alpha_{\mathrm{CN}} = 0.915$ in the following way. We first determine from the data the average growth exponent, η, exclusively for the Chinese regions obtaining $\langle\eta\rangle_{\mathrm{CN}} \approx 2.53 \pm 0.09$. Then, from a starting value of $\alpha_{\mathrm{CN}} = 0.79$ – the value of all the remaining regions – we increment α_{CN} in the simulation in steps of 0.005 until we find an

agreement between the empirical $\langle \eta \rangle_{\mathrm{CN}}$ and the one estimated from the stream of regions generated by the simulation.

Figs. 3C–E and 4 panels C–E are obtained by first removing all regions with a cumulative number of scientists at the year 2019 less than 100. For those regions with a lower number of scientists, the curves of their growth in intrinsic time become noisy, and the estimated exponent is unreliable.

CRediT authorship contribution statement

Vito D.P. Servedio: Analyzed the data, Devised the model, Studied it analytically, Writing – original draft. **Márcia R. Ferreira:** Conceptualized the idea, Wrangled and curated the data, Analyzed the data, Writing – original draft. **Niklas Reisz:** Analyzed the data. **Rodrigo Costas:** Wrangled and curated the data. **Stefan Thurner:** Conceptualized the idea, Devised the model, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge the Centre for Science and Technology Studies (CWTS) of Leiden University for providing access to their in-house version of the Dimensions database and support by the Austrian Research Promotion Agency FFG under grant 882184. All authors contributed to the final version.

Appendix A. Establishing the data set

The Dimensions data. We access the *Dimensions* database through the Centre for Science and Technology Studies (CWTS), Leiden University. Dimensions covers millions of research publications connected by more than 1.6 billion citations, supporting grants, datasets, clinical trials, patents, and policy documents. In this work, we only use research publications. The publications database of Dimensions contains publication data spanning several decades. The database comprises about 20 million disambiguated researchers assigned to a researcher ID. It also comprises affiliation linkages to the Global Research Identifier Database that currently covers more than 98,000 research institutions worldwide [9,10]. *Dimensions* is produced by Digital Science and launched in January 2018. For further references, see the Dimensions.ai website.[2]

Preparing the data. We start by collecting documents (i.e., all document types) in three different topics, i.e., Semiconductors, Embryonic Stem Cell (ESC) research, and Internet research. We chose these areas since they are sufficiently large and important to provide a meaningful statistical analysis. Delineating research areas is crucial to study their growth. Different types of data and techniques have been proposed to delineate research areas such as document co-citation [41], author co-citation [42], co-word analysis [43], and journal-based mapping [44]. The semiconductors, ESC, and Internet research areas are defined by collecting three sets of documents and associated groups of disambiguated authors, their affiliated institutions, and the geographical regions where the institutions are located. We use a combination of term matching, citing and cited links, and a selection of specific Fields of Research (FOR) to identify documents that belong to those areas of research. The method we employ for defining the three research areas in this paper can be divided into three steps:

Step 1: Creating a lexicon of key technical terms. We started by extracting keywords from the literature and validating them with domain experts. Terms identified as relevant within each topic by experts were kept for further analysis. This step involved identifying the core technical jargon that authors use in each of the areas under study. As a result, we prioritized technical terminology over popular terms, as the former were more likely to be used and recognized by our target research community. The vocabulary was expanded further using the VOSviewer software [45] to identify very frequently co-occurring concepts in addition to the initial terms, and the newly extracted concepts were validated by domain experts during a second round of consultation. The list of the lexicon used can be found in Appendix D.

Step 2: Retrieving items using term matching and Dimensions' concepts. In the second step of our approach, we searched the Dimensions database for all document types that matched the terminology established in the previous step using a combination of exact and fuzzy term matching. We matched the phrases to pre-extracted *concepts* from Dimensions publications that had an abstract between 1941 and 2019. According to the Dimensions documentation, concepts are normalized noun phrases that describe the core concepts of a document and are derived automatically from the publication's abstracts [46]. Additionally, we used Dimensions' four-digit FOR categories to ensure that the concepts correspond to a rather consistent collection of documents. For example, several concept matches can yield publication items associated with FOR categories such as "Anthropology" or "Law". We excluded these domains from our document search due to our strong focus on technical knowledge in semiconductors, ESC, and Internet research. In Appendix D, we list the whole vocabulary that was used to retrieve the documents.

Step 3: Extracting cited and citing items. Cited works stand for or symbolize works that have been inspired by past publications, while citing works symbolize works that have inspired subsequent items. Therefore, we also collected all citing and cited items to the publication sets retrieved in step 2. Both cited and citing items are restricted to the overall time spans in 1 and FOR categories.

Algorithms for author name disambiguation and institutional registries have been implemented and linked to the majority of large bibliometric databases, including Scopus from Elsevier and Dimensions from Digital Science. Most of these algorithms leverage open systems for uniquely identifying scholars, such as ORCID, or for identifying institutional affiliations, such as GRID [8].

Using this method, we can extract a representation of research areas that includes core documents in the international scientific literature in a period between 1941 and 2019. We use the affiliation of researchers to assign an article to one or more world regions. Regions are considered at the granularity of the first level of administrative regions, equivalent to provinces (e.g., US.MA [Massachusetts, USA], GB.ENG [England, Great Britain]). We then select a chosen scientific field and sort all the corresponding publications in ascending temporal order, whereas articles sharing the same year of publication are listed randomly. We checked that the reshuffling of the publications inside the same year did not change any of the results. Since we are interested in the cumulative growth of the number of scientists in a region, we discard all those events where scientists publish a paper with one of their old affiliations under which they had already published in the past (see Table 2 for an example of the procedure). Finally, we build the sequence of regions in the order they appear in time. The position in the sequence defines an intrinsic time that runs faster than real-time. One tick of intrinsic time corresponds to a region in the sequence. The relationship between intrinsic and real-time is an inverse stretched exponential, i.e., for all the three scientific fields considered, the relation between the two is approximately of the type $t \approx \exp(\sqrt{\frac{T-T_0}{\tau}})$ with t intrinsic time, T real time, T_0 starting year reported in Table 1, and τ representing a sort of characteristic time that gets values around 130–160 days in all three cases.

[2] https://www.dimensions.ai/.

Appendix B. Scaling relations between exponents

The important quantities we consider in this study are, asymptotically, for large values of intrinsic time t

$$D(t) \approx t^{\gamma} \qquad k_i(t) \approx t^{\eta}$$
$$P_k \propto k^{\alpha} \qquad Z(t) = \sum_{i=1}^{D} k_i^{\alpha} \approx t^{\sigma}, \tag{B.1}$$

where $D(t)$ is the number of distinct regions (Eq. (3) in the main text); $k_i(t)$ is the cumulative number of scientists in region i (Eq. (2) in the main text); P_k is the preferential attachment probability (Eq. (5) in the main text); $Z(t)$ is the normalization term in the PA probability definition.

One can deduce some identities with simple approximate reasoning. Since $\sum_{i=1}^{D} k_i = t$ and we have D terms in the sum, we can write $t^{\gamma+\eta} \approx t$, hence $\eta = 1-\gamma$. Similarly, $Z(t) \approx t^{\gamma} t^{\alpha\eta} = t^{\gamma+\alpha(1-\gamma)}$, hence $\sigma = \alpha + \gamma(1-\alpha)$ (we rearranged the terms to highlight the role of α). These crude approximations are confirmed by a less straightforward analytic solution, which can also provide the behavior of $k_i(t)$ when t is close to the first introduction of region i, i.e., $k_i(t \approx t_{0,i}) \approx (1+c(t-t_{0,i}))^{\frac{1}{1-\alpha}}$. Since $0 < \alpha < 1$, the exponent $1/(1-\alpha)$ is larger than 1, in accordance with the observed super-linear growth of latecomer regions. Practically, the number of occurrences of regions that enter late in the system starts to grow super-linearly and eventually reaches the asymptotic sub-linear regime.

Appendix C. Approximate analytical solution of the model

We have a stream of geographical regions from empirical data and from it, we build a new synthetic stream. We shall generically refer to geographical regions as *tokens* in the following since the reasoning below can be applied to any kind of stream with sub-linear PA. Each time a brand new token appears in the real stream, we put it as it is in our own created stream; each time an already occurred token appears in the real stream, we extract a token from our synthetic stream with a probability that is sublinear in the number of its occurrences k so far, i.e., according to the probability P_k defined in the main text in Eq. (5).

Definition of parameters

Our model contains one free parameter: the exponent α of the sublinear rich-get-richer mechanism. We take into account Heaps' exponent automatically by adopting the real stream of new tokens. We need to introduce two parameters more, which eventually will be connected to the PA and Heaps' exponents, i.e., the asymptotic exponent σ of the PA normalization term and the asymptotic exponent η of the growth of tokens in intrinsic time. The definition of the exponents is that of Eq. (B.1). We call the normalization $Z(t)$ as partition function in the following. A rough estimate of the parameters based on one run of the model in the semiconductor field and simple linear regressions gives:

$$\gamma \approx 0.39 \qquad \eta \approx 0.74 \qquad \alpha \approx 0.79 \qquad \sigma \approx 0.85. \tag{C.1}$$

The value of η was inferred from the most populated token.

Occurrence of token i in intrinsic time

In the approximation of continuous time we can write

$$\frac{dk_i}{dt} = \frac{k_i^{\alpha}}{Z(t)} \tag{C.2}$$

where we know that asymptotically $Z(t) \approx ct^{\sigma}$ (with c a multiplicative constant). We solve the previous equation

$$\int_1^k \frac{dk_i}{k_i^{\alpha}} = c \int_{t_{0,i}}^{t} \frac{d\tau}{\tau^{\sigma}} \tag{C.3}$$

with t_0 the entry time of the token, to get

$$k = \left(1 + c\frac{1-\alpha}{1-\sigma}(t^{1-\sigma} - t_{0,i}^{1-\sigma})\right)^{\frac{1}{1-\alpha}}. \tag{C.4}$$

For $t \gg t_0$ we get (we drop the index i for simplicity)

$$k \approx t^{\frac{1-\sigma}{1-\alpha}} \quad \text{i.e.} \quad \boldsymbol{\eta = \frac{1-\sigma}{1-\alpha}}. \tag{C.5}$$

If $t = t_0 + x$ we get by expanding around t_0

$$k \approx \left(1 + c\frac{1-\alpha}{t_0^{\sigma}}x\right)^{\frac{1}{1-\alpha}}, \tag{C.6}$$

with a super-linear exponent $\frac{1}{1-\alpha}$.

Probability density function of occurrences

Let us call $N_k(t)$ the number of tokens having already occurred k-times at time t. After dropping the index i we can write the following master equation for $k > 1$:

$$N_k(t+1) = N_k(t) + N_{k-1}\frac{(k-1)^{\alpha}}{Z(t)} - N_k\frac{k^{\alpha}}{Z(t)}, \tag{C.7}$$

which can be written in a continuous approximation as the PDE

$$\frac{\partial N_k}{\partial t} = -\frac{1}{Z(t)}\frac{\partial(k^{\alpha}N_k)}{\partial k}. \tag{C.8}$$

Some important relations to keep in mind:

$$\sum_{i=1}^{D} k_i = \sum_{k=1}^{k_M} kN_k = t \tag{C.9}$$

$$\sum_{i=1}^{D} 1 = \sum_{k=1}^{k_M} N_k = D \approx ct^{\gamma} \tag{C.10}$$

with k_M representing the maximum value of the k_i. In order to normalize N_k and get the probability ρ_k of finding tokens with k occurrences we divide by D

$$N_k \approx ct^{\gamma}\rho_k. \tag{C.11}$$

We substitute the previous relation and $Z(t) \approx st^{\sigma}$ into the PDE. After simplifying the derivatives explicitly, we get

$$c\gamma t^{\gamma-1}\rho_k \approx -\frac{ct^{\gamma-\sigma}}{s}\left(\alpha k^{\alpha-1}\rho_k + k^{\alpha}\frac{d}{dk}\rho_k\right) \tag{C.12}$$

and eventually, obtain the following ODE

$$\frac{d}{dk}\rho_k = -s\gamma k^{-\alpha}t^{\sigma-1}\rho_k - \frac{\alpha}{k}\rho_k. \tag{C.13}$$

Its solution is proportional to

$$\rho_k \propto k^{-\alpha}e^{-\frac{s\gamma t^{\sigma-1}}{1-\alpha}k^{1-\alpha}} \tag{C.14}$$

which asymptotically at large t, gives

$$\rho_k \approx k^{-\alpha}. \tag{C.15}$$

Since the decay in time is pretty slow ($\sigma \approx 0.85$ in our systems), the limit distribution is attained over long times. Moreover, we recall that $0 < \alpha < 1$ so that ρ_k would have no thermodynamic limit without the stretching exponential term. For determining the scaling relations between the exponents, we stick to the power-law limit distribution and call k_M the maximum value of k at a certain large value of time t. By normalizing the ρ_k we find

$$\rho_k = \frac{1-\alpha}{k_M^{1-\alpha} - 1}k^{-\alpha}. \tag{C.16}$$

Partition function in time

We recall the partition function

$$Z(t) = \sum_{i=1}^{D} k_i^{\alpha} = \sum_{k=1}^{k_M} N_k k^{\alpha} \approx t^{\sigma}. \tag{C.17}$$

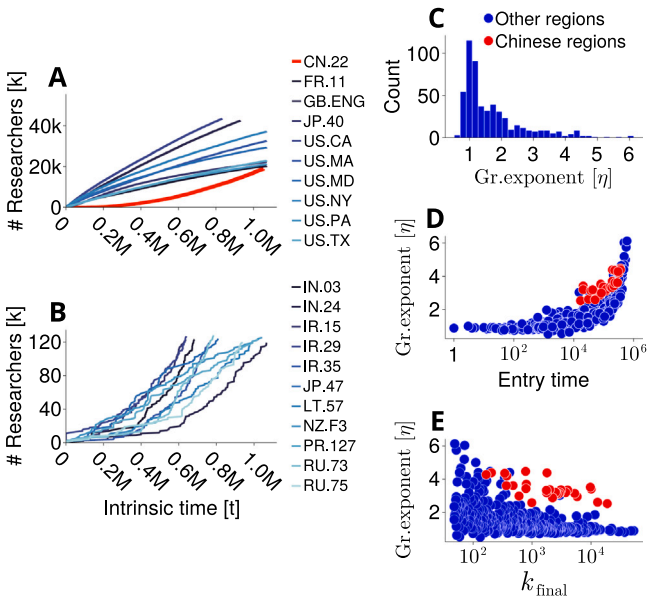

Fig. E.5. Panels A and B show the cumulative number of scientists coming into a region (lines) in intrinsic time in embryonic stem cell research; with respect to Fig. 2A and B, the axes are in linear scale to appreciate the different curvature of the growths at high intrinsic times; (A) shows the 10 most visited regions. Most regions grow sub-linearly. Chinese regions are shown with red thick lines and clearly increase super-linearly. In (B), we see 11 poorly visited regions ranked from 700 to 710 in terms of the number of scientists. Curves are (additively) shifted to the left so that their first point is placed at coordinates (1,1); (C) Histogram of growth exponents, η, for all regions; (D) Growth exponents, η, of all regions, i, as a function of their entry time, t_i (intrinsic time). Note that Chinese regions (red) come in late (after timestep 10^5) and have higher exponents than other regions (blue). (E) Growth exponents, η, as a function of the cumulative number of scientists in the year 2019, k_{final}. The high exponents of Chinese regions are visible (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

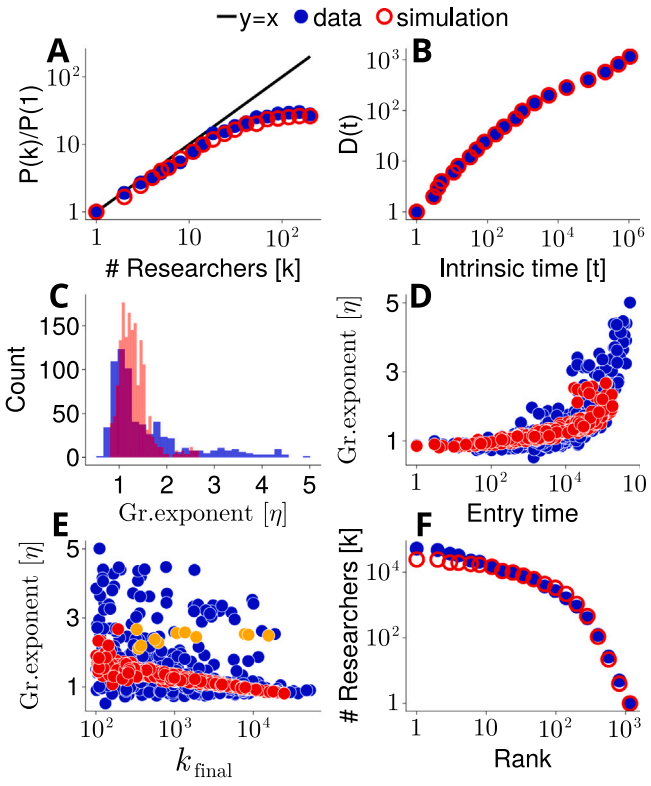

Fig. E.6. Comparison of model results (red) with experimental data (blue) for the field of embryonic stem cells. (A) PA kernel; data and simulation practically coincide; the black line depicts a linear kernel. (B) Heaps's law; the two curves coincide by construction. (C) Histograms of the regional growth exponent, η. The distributions of the empirical values have average $m = 1.56$, standard deviation $s = 0.86$ and skewness $b = 1.73$, while the distribution of the simulated values has $m = 1.31$, $s = 0.31$ and $b = 1.70$. (D) Exponents, η, reproduce the empirical increase as a function of entry time. (E) The decrease of η as a function of the cumulative number k_{final} of scientists in the year 2019 is well captured by the model. Note the fact that Chinese regions (orange circles) are predicted with a higher than usual values. (F) Also the frequency-rank distributions of the regions appearing in the two streams practically coincide except for very small ranks. In panels C, D, and E only regions with $k_{\text{final}} > 100$ are taken into account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

If $\alpha \to 0$ then $Z(t) = D \approx t^{\gamma}$; if $\alpha = 1$ then $Z(t) = t$. Therefore we expect $\gamma < \sigma < 1$. We can write

$$\begin{aligned} Z(t) &\approx \sum_{k=1}^{k_M} t^{\gamma} \rho_k k^{\alpha} \\ &\approx \sum_{k=1}^{k_M} t^{\gamma} \frac{1-\alpha}{k_M^{1-\alpha}-1} k^{-\alpha} k^{\alpha} \\ &\approx t^{\gamma} \frac{1-\alpha}{k_M^{1-\alpha}-1} \sum_{k=1}^{k_M} 1 \approx t^{\gamma} k_M^{\alpha}. \end{aligned} \tag{C.18}$$

If we now recall that the occurrences of a token grow as t^{η}, and therefore also k_M does, we find

$$\sigma = \gamma + \alpha\eta = \gamma + \alpha\frac{1-\sigma}{1-\alpha}, \tag{C.19}$$

which after solving for σ gives:

$$\boldsymbol{\sigma = \alpha + \gamma(1-\alpha)}. \tag{C.20}$$

By inserting the above expression of σ in the expression of η of Eq. (C.5) we also get

$$\boldsymbol{\eta = 1 - \gamma}. \tag{C.21}$$

Recap of results

After substituting the result for σ we get:

- $Z(t \gg 1) \approx t^{\alpha+\gamma(1-\alpha)}$
- $k_i(t \gg t_{0,i}) \approx t^{1-\gamma}$
- $k_i(t \approx t_{0,i}) \approx (1 + c(t - t_{0,i}))^{\frac{1}{1-\alpha}}$.

Note how the number of occurrences of tokens that enter late in the system starts to grow super-linear and eventually reaches the asymptotic sublinear regime.

Appendix D. List of terms

In the following, we list the lexicon of terms used to select the articles according to their scientific fields.

Semiconductor research transistor, analog circuits, bipolar junction transistor, bipolar transistor, carbide, carborundum, cat s-whisker detector, conductivity, crystal diode, darlington transistor, discrete device, doped monocrystalline silicon grid, electrical conduction, electrical conductivity, electron–hole pairs, electronic band structure, field effect junction transistor, field-effect transistor, four-terminal devices, gallium arsenide, germanium, gunn diode, hall effect sensor, heterojunctions, impatt diode, insulated-gate bipolar transistor, integrated circuit, iotatron, laser diode, light-emitting diode, light-emitting diode, metal rectifier, metal–oxide–semiconductor, metal–oxide–semiconductor field-effect transistor, microchip, microprocessor, mixed-signal circuits, mos transistor, mosfet, n-type semiconductor, optocoupler, photocoupler, photodiode, phototransistor, photovoltaic solar cells, pin diode, p–n junctions, p–n–p point-contact germanium,

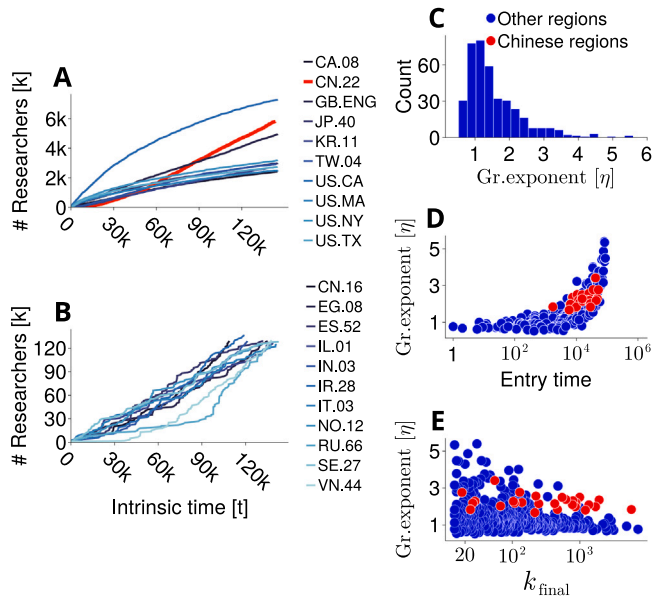

Fig. F.7. Panels A and B show the cumulative number of scientists coming into a region (lines) in intrinsic time in Internet research; with respect to Fig. 2A and B, the axes are in linear scale to appreciate the different curvature of the growths at higher intrinsic times; (A) shows the 10 most visited regions. Most regions grow sub-linearly. Chinese regions are shown with red thick lines and clearly increase super-linearly. In (B) we see 11 poorly visited regions ranked from 700 to 710 in terms of the number of scientists. Curves are (additively) shifted to the left so that their first point is placed at coordinates (1,1); (C) Histogram of growth exponents, η, for all regions; (D) Growth exponents, η, of all regions, i, as a function of their entry time, t_i (intrinsic time). Note that Chinese regions (red) come in late (after timestep 10^5) and have higher exponents than other regions (blue). (E) Growth exponents, η, as a function of the cumulative number of scientists in the year 2019, k_{final}. The high exponents of Chinese regions are visible (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

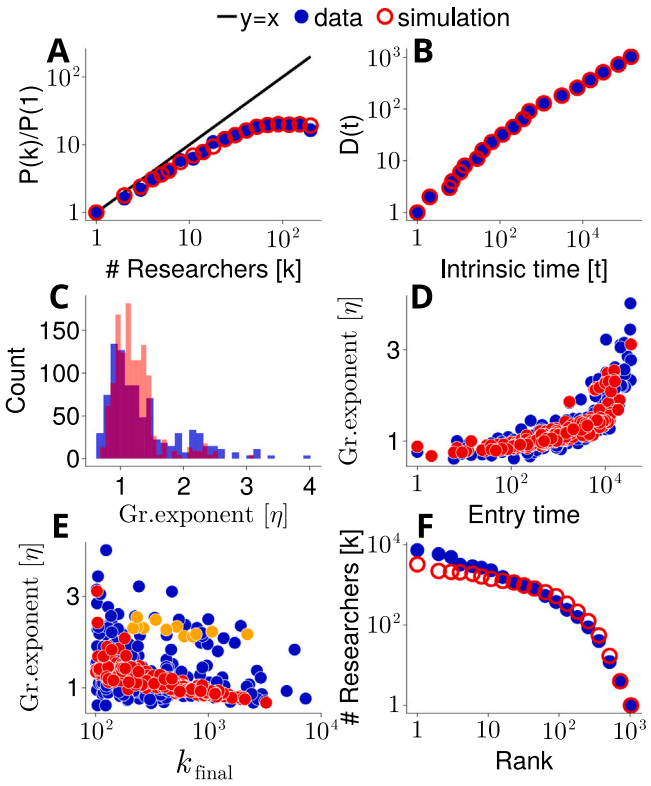

Fig. F.8. Comparison of model results (red) with experimental data (blue) for the field of Internet. (A) PA kernel; data and simulation practically coincide; the black line depicts a linear kernel. (B) Heaps's law; the two curves coincide by construction. (C) Histograms of the regional growth exponent, η. The distributions of the empirical values has average $m = 1.36$, standard deviation $s = 0.59$ and skewness $b = 1.48$, while the distribution of the simulated values has $m = 1.22$, $s = 0.35$ and $b = 1.93$. (D) Exponents, η, reproduce the empirical increase as a function of entry time. (E) The decrease of η as a function of the cumulative number k_{final} of scientists in the year 2019 is well captured by the model. Note the fact that Chinese regions (orange circles) are predicted with higher than usual values. (F) Also, the frequency-rank distributions of the regions appearing in the two streams practically coincide except for very small ranks. In panels C, D, and E, only regions with $k_{\text{final}} > 100$ are taken into account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

polysilicon, p-type semiconductor, rectifier diode, reverse-biased p–n junction, schottky diode, selenium sulfide, semiconductor, semi-insulators, silicon-controlled rectifier, solid triode, three-terminal devices, thyristor, transconductance, transient-voltage-suppression diode, transistron, triode, tunnel diode, two-electrode vacuum tube, unijunction transistor, vaccum tube, varistor, vcsel, zen diode, zener diode.

Embryonic Stem Cells blastocyst, blastoid, embryonic AND stem AND cell, embryonic fibroblast, embryonic germ layers, hescs AND stem AND cell, mescs AND stem AND cell, pluripotency, pluripotent AND stem AND cell, undifferentiated pluripotent state.

Internet research arpanet, atm networks, atm switch, bitnet, broad-based electronic communication, classless interdomain routing, csnet, darpa internet, distributed network protocol, domain name system, end-to-end circuit, end-to-end transmission, esnet, exterior gateway protocol, federal internet exchanges, gateway protocol, gateways router, global information infrastructure, gopher protocol, ground-based packet, hop wireless network, host-to-host protocol, hypertext transfer protocol, ibm rscs protocol, input-queued packet, interior gateway protocol, internet infrastructure, internet protocol, internet router, internet routing, internetting, internetwork, internetworking architecture, ip service, ipv4, ipv6, lan technology, local area network technology, message switching AND internet, minitel, mmdf protocol, multihop networks, networked computers, network control protocol, network emulator, network traffic protocol, npl network, open systems interconnection, open-architecture network, optical packet, osi reference model, packet radio, packet satellite, packet switch, packet traffic, packetization, packet-switching, radio transmitting system, relay network, remote machines, simple mail transfer protocol, store-and-forward switching, switched network, tcp performance, tcp/ip, tcp-based applications, telecommunications infrastructure, telenet, time-sharing internet, time-sharing network, transmission control protocol, trans-oceanic circuits, transport layer protocol, user datagram protocol, virtual circuit model, virtual circuits, wireless atm AND internet, wireless sensor network AND internet, x.25 protocol.

Appendix E. Results for embryonic stem cells

Figs. E.5 and E.6 show the results for the field of embryonic stem cells and share the same design of Figs. 3 and 4.

Appendix F. Results for internet research

Figs. F.7 and F.8 show the results for the field of Internet and share the same design of Figs. 3 and 4.

References

[1] Gibbons M, Johnston R. The roles of science in technological innovation. Res Policy 1974;3(3):220–42.
[2] Saxenian A. The genesis of silicon valley. Built Environ 1983;(1978-):7–17.
[3] Harrison M. Does high quality research require critical mass. Question R&D Spec: Pers Policy Implic. 2009;57–9.
[4] Johnston R. Effects of resource concentration on research performance. Higher Educ. 1994;28(1):25–37.

[5] Sugimoto CR, Robinson-García N, Murray DS, Yegros-Yegros A, Costas R, Larivière V. Scientists have most impact when they're free to move. Nat News 2017;550(7674):29.
[6] Trippl M. Islands of innovation and internationally networked labor markets. magnetic centers for star scientists?. 2009.
[7] Sturgeon TJ. How silicon valley came to be. Underst Silicon Val: Anat Entrepreneurial Reg 2000;15–47.
[8] Macháček V, Srholec M, Ferreira MR, Robinson-Garcia N, Costas R. Researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization. Sci Public Policy 2021.
[9] Hook DW, Porter SJ, Herzog C. Dimensions: Building context for search and evaluation. Front Res Metr Anal 2018;3:23.
[10] Herzog C, Hook D, Konkiel S. Dimensions: Bringing down barriers between scientometricians and data. Quant Sci Stud 2020;1(1):387–95.
[11] Menter M, Lehmann EE, Klarl T. In search of excellence: A case study of the first excellence initiative of Germany. J. Bus Econ 2018;88(9):1105–32.
[12] Zucker LG, Darby MR. Star scientists, innovation and regional and national immigration. Working paper 13547, National Bureau of Economic Research; 2007.
[13] de Solla Price D. A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 1976;27(5):292–306.
[14] Simkin MV, Roychowdhury VP. Re-inventing willis. Phys Rep 2011;502:1–35.
[15] Bol T, de Vaan M, van de Rijt A. The matthew effect in science funding. Proc Natl Acad Sci 2018;115(19):4887–90.
[16] Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. Phys Rev E 2006;74:036116.
[17] Krapivsky PL, Redner S, Leyvraz F. Connectivity of growing random networks. Phys Rev Lett 2000;85:4629–32.
[18] Newman MEJ. Clustering and preferential attachment in growing networks. Phys Rev E 2001;64:025102.
[19] Jeong H, Néda Z, Barabási AL. Measuring preferential attachment in evolving networks. Europhys Lett (EPL) 2003;61(4):567–72.
[20] Szell M, Lambiotte R, Thurner S. Multirelational organization of large-scale social network in an online world. Proc Natl Acad Sci 2010;107(31):13636–41.
[21] Reisz N, Servedio VDP, Loreto V, Schueller W, Ferreira MR, Thurner S. Loss of sustainability in scientific work. New J Phys 2022;24(5):053041.
[22] Wang D, Song C, Barabási A-L. Quantifying long-term scientific impact. Science 2013;342(6154):127–32.
[23] Jin C, Song C, Bjelland J, Canright G, Wang D. Emergence of scaling in complex substitutive systems. Nat Hum Behav 2019;3(8):837–46.
[24] Gleeson JP, Cellai D, Onnela J-P, Porter MA, Reed-Tsochas F. A simple generative model of collective online behavior. Proc Natl Acad Sci 2014;111(29):10411–5.
[25] Rao A, Scaruffi P. A history of silicon valley: the greatest creation of wealth in the history of the planet. 2nd ed.. Omniware Group; 2013.
[26] Lazonick W, Li Y. China's path to indigenous innovation. 2012.
[27] Li Y. The semiconductor industry: A strategic look at China's supply chain. In: The new Chinese dream: industrial transition in the post-pandemic era. Cham: Springer International Publishing; 2021, p. 121–36, Chapter 8.
[28] Tollefson J. China declared world's largest producer of scientific articles. Nature 2018;553:390.
[29] Robinson-Garcia N, Sugimoto CR, Murray D, Yegros-Yegros A, Larivière V, Costas R. The many faces of mobility: Using bibliometric data to measure the movement of scientists. J Informetr 2019;13(1):50–63.
[30] Waltman L, van Eck NJ. Field-normalized citation impact indicators and the choice of an appropriate counting method. J Informetr 2015;9(4):872–94.
[31] Moed HF. Citation analysis in research evaluation. Springer Dordrecht; 2005.
[32] Pham T, Sheridan P, Shimodaira H. PAFit: A statistical method for measuring preferential attachment in temporal complex networks. PLoS One 2015;10(9):e0137796.
[33] Serrano MÁ, Flammini A, Menczer F. Modeling statistical properties of written text. PLoS One 2009;4(4):1–8.
[34] Tria F, Loreto V, Servedio VDP, Strogatz SH. The dynamics of correlated novelties. Sci Rep 2014;4:5890.
[35] Mazzolini A, Grilli J, De Lazzari E, Osella M, Lagomarsino MC, Gherardi M. Zipf and heaps laws from dependency structures in component systems. Phys Rev E 2018;98:012315.
[36] Tria F, Loreto V, Servedio VDP. Zipf's, Heaps' and Taylor's laws are determined by the expansion into the adjacent possible. Entropy 2018;20(10).
[37] Simini F, James C. Testing Heaps' law for cities using administrative and gridded population data sets. EPJ Data Sci 2019;8(1):24.
[38] Zanette D, Montemurro M. Dynamics of text generation with realistic Zipf's distribution. J Quant Linguist 2005;12(1):29–40.
[39] Kauffman SA. Investigations. New York/Oxford: Oxford University Press; 2000.
[40] Arthur WB. Competing technologies, increasing returns, and lock-in by historical events. Econ J 1989;99(394):116–31.
[41] Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inf Sci 1973;24(4):265–9.
[42] White HD, Griffith BC. Author cocitation: A literature measure of intellectual structure. J Am Soc Inf Sci 1981;32(3):163–71.
[43] Callon M, Courtial J-P, Turner WA, Bauin S. From translations to problematic networks: An introduction to co-word analysis. Soc Sci Inf 1983;22(2):191–235.
[44] Leydesdorff L. Clusters and maps of science journals based on bi-connected graphs in journal citation reports. J Doc 2004.
[45] Van Eck NJ, Waltman L. Software survey: Vosviewer, a computer program for bibliometric mapping. Scientometrics 2010;84(2):523–38.
[46] Working with concepts in the dimensions API. 2022, https://api-lab.dimensions.ai/cookbooks/1-getting-started/7-Working-with-concepts.html, accessed 2022-01-26.