Accepted Manuscript

The dynamics of innovations and citations

Christian Ghiglino, Nicole Tabasso

PII: S0165-1765(15)00146-9
DOI: http://dx.doi.org/10.1016/j.econlet.2015.04.004
Reference: ECOLET 6725

To appear in: *Economics Letters*

Received date: 20 January 2015
Revised date: 26 March 2015
Accepted date: 3 April 2015

Please cite this article as: Ghiglino, C., Tabasso, N., The dynamics of innovations and citations. *Economics Letters* (2015), http://dx.doi.org/10.1016/j.econlet.2015.04.004

*Highlights (for review)

Highlights

- Patent citations occur as new ideas are produced by combining existing ideas.
- Ideas are intervals in a variety space.
- Interval lengths determine the likelihood of citation.
- The model derives exponential aging of patents, which fits the data very well.
- Endogenous aging sets the model apart from preferential attachment models.

The Dynamics of Innovations and Citations

Christian Ghiglino[a,b,*], Nicole Tabasso[c]

[a] *University of Essex, Department of Economics, Colchester, CO4 3SQ, UK*
[b] *GSEM, Geneva, Switzerland*
[c] *University of Surrey - School of Economics, Guildford, GU2 7XH, UK*

Abstract

We present a model in which patent citations occur as new ideas are produced from combinations of existing ideas. An idea's usability in this process is represented as an interval in a variety space of ideas, whose length determines the likelihood of citation. This process endogenously derives exponential aging of patents, which is consistent with empirical observations. The endogeneity of aging sets our process apart from the standard preferential attachment literature.

Keywords: Citation Dynamics, Citation Network, Innovations, Idea Applicability

JEL Classification: O30; O31; D85

1. Introduction

Citation patterns between academic papers or patents have received considerable attention. They have been analyzed predominantly with network models based on preferential attachment or intrinsic fitness in which papers/patents are nodes, and citations directed links between them (Atalay (2013), Barabási et al. (1999), Jackson and Rogers (2007), Peterson et al. (2010), Valverde et al. (2007)). In these models the probability to link to an existing node is proportional to a scalar quantity, e.g., intrinsic fitness or the number of past citations. However, without the introduction of a specific aging function they are not

[*] Corresponding author
Email addresses: cghig@essex.ac.uk (Christian Ghiglino), n.tabasso@surrey.ac.uk (Nicole Tabasso)

able to match the observed aging of patents, see Figure 1 or Marco (2007). In particular, with pure preferential attachment citation rates are only negatively affected by the total network size which provides a weak form of aging.

[**Figure 1 here**]

Figure 1: **Citation Dynamics.** Non-parametrically estimated population hazard rates of being cited for USPTO patents that have received 1, 5, and 10 citations respectively.

We rather leave preferential attachment and model the attachment process between patents as random, guided by the applicability of patents, representing intrinsic heterogeneity. We model applicability in a way that leads to aging very naturally through a behavioral choice of innovators. In our model, patents represent ideas, which are built from combination of older ideas as in Auerswald et al. (2000), Ghiglino (2012), or Weitzman (1998). Innovations arrive to innovators, who decide which ideas to combine to realize the innovation. This choice is largely driven by technological identity of ideas. In particular, we model a patent as an interval in the variety space of ideas, which represents its applicability range. More broadly applicable patents thus are more likely to be cited. The choice of the innovator that leads to aging is simple: If there exist multiple ideas that may be used as inputs for his innovation, he chooses the youngest. Such behavior might be justied if innovators do not know perfectly which input idea is best for them to use.

2. The model

We model patents/ideas as nodes in a network, and citations among them as directed links. A link from node i to j thus signifies that j is an input idea to i. Ideas are of different varieties, and the support of the variety space is a circle of circumference 1. At time t there are $N(t)$ nodes. Time is continuous and new nodes arrive sequentially, as a Poisson process with arrival rate of 1. Each node i is characterized by an interval $I_i \subset (0,1]$, a set of m scalars, $\mu_k^i \in (0,1]$ with

$k = 1, ..., m$, and its birth date t_i. We assume that each μ_k^i is extracted from a uniform distribution on $(0, 1]$, and I_i is extracted from a distribution Ψ such that the position (middle point) of the (connected) interval is extracted from a uniform distribution on $(0, 1]$, and $|I_i| \in (0, \frac{1}{m})$.

A necessary condition for the existence of a link from i to j, is that for at least one $k = 1, ..., m$, $\mu_k \in I_j$ and $t_j \leq t_i$, in which case j is a feasible input for i. However, there might be several nodes that satisfy this condition. Let the set of these nodes be $\hat{I}_k$. A sufficient and necessary condition for a link from i to j is that $j \in \hat{I}_k$ and $t_j > t_{j'}$ for all $j' \in \hat{I}_k$. As nodes are added sequentially there is, at most, only one such node. The attachment process is illustrated in Figure 2 for $m = 3$.

[Figure 2 here]

Figure 2: **Attachment process.** Node i enters the system, with $\mu^i = \{\mu_j, \mu_k, \mu_l\}$. Nodes c, e, and g will receive links.

We proceed to obtain the probability for a node to receive a link. First, note that a given node j has m chances to receive an additional edge from a newly entered node i. For each of these, as μ_k is uniformly extracted from $(0, 1]$, the probability that j fulfills the necessary condition that $j \in \hat{I}_k$ is equal to

$$Pr(\mu_k \in I_j) = |I_j|. \tag{1}$$

In turn, the probability that $t_j > t_{j'}$ for all $j' \in \hat{I}_k$, is the probability that between t_j and t, no other node j' has entered the system for which $\mu_k \in I_{j'}$. Given the Poisson arrival process of ideas, the mean arrival rate of such nodes j' is given by the average interval length, denoted $\bar{I}$,

$$Pr(t_j > t_{j'} | j, j' \in \hat{I}_k) = e^{-\bar{I}(t - t_j)}. \tag{2}$$

Let $k_j(t)$ be the number of edges that node j has received up to t (its *in-degree*). The expected change in $k_j(t)$ is computed assuming that $k_j(t)$ is continuous and that the mean-field approximation holds (see, e.g., Barabási et al. (1999) or Jackson and Rogers (2007)), which allows us to denote the expected rate of change as the actual one. Consequently, the probability that node j's in-degree will increase by at least one at t, $\Pi(I_j, t_j, t)$ (the *hazard rate* of node j),[1] can be expressed as the continuous rate of change of $k_j(t)$:

$$\Pi(I_j, t_j, t) = \frac{\partial k_j(t)}{\partial t} = m \cdot |I_j| \cdot e^{-\bar{I}(t-t_j)}. \tag{3}$$

The predictions of (3) are in line with stylized facts of patent citations: citation rates vary across patents and older patents are less likely to be cited. In contrast to preferential attachment models the likelihood to receive an edge depends on the *age* of a node, rather than on time t itself, and is independent of the current number of edges. Note that the model does not attempt to describe the increase in the likelihood of receiving an edge that is observed early in the life of patents. Instead, it predicts exponential aging of patents, which we now test against available patent citation data.

3. Results

Our dataset consists of a random sample of $N = 214,071$ patents granted by the United States Patent and Trademark Office (USPTO) between 1975 and 1999, made available by the National Bureau of Economic Research. We observe each patent whenever it gets cited and at the end of the 25 year period, which provides a total of $n = 1,059,475$ observations. We measure time as the number of patents granted, i.e., it coincides with the number of patents in the system.

We test our model by comparing the predicted hazard rates from (3) against the empirically estimated citation rates. To do so, we first integrate (3) subject to $k_j(t_j) = 0$, which yields

[1] While theoretically, node j can be cited multiple times by i, we find that the distribution of $|I_j|$ makes this probability negligible.

$$k_j(t) = m\frac{|I_j|}{\bar{I}}\left(1 - e^{-\bar{I}(t-t_j)}\right). \tag{4}$$

This allows us to calculate $|I_j|$ as

$$|I_j| = k_j(t) \cdot \bar{I} \cdot \frac{1}{m\left(1 - e^{-\bar{I}(t-t_j)}\right)}. \tag{5}$$

With the exception of $\bar{I}$, the variables in (5) are directly observable in the data, e.g., $m = 7.72$. As shown in Appendix 1, $\bar{I}$ can be obtained numerically as a fixed point associated to (5). We find $\bar{I} = 9.22 \cdot 10^{-7}$ and use this value to derive individual $|I_j|$'s.[2] Figure 3 shows that the hazard rates from (3) follow the citation rates very closely:

[Figure 3 here]

Figure 3*:* **Fit of the Model.** The hazard rates from Figure 1 (solid, dotted, and dashed lines) are plotted against the calculated hazard rates of our model (triangles).

Individual, rather than population, hazard rates can be estimated with Survival Analysis, using the specification delivered by our model:

$$\ln h_j(t) = \beta_1 \ln(m) + \beta_2 \ln(|I_j|) + \beta_3(t - t_j). \tag{6}$$

The results of the estimation are given in Table 1, column 1. Our model predicts $\beta_1 = \beta_2 = 1$, and $\beta_3 = -\bar{I}$. Column 2 imposes the restriction $\beta_3 = -\bar{I}$. Under both specifications, the coefficient estimates are extremely close to the predicted ones.

To appreciate the success of the model, we re-do the same analysis for a preferential attachment citation process as in Jackson and Rogers (2007). The

[2] Equation (4) also allows us to derive the distribution of in-degrees as $t \to \infty$, which is $f(k) = g\left(\frac{k\bar{I}}{m}\right)$, where $g(\cdot)$ denotes the distribution of interval lengths.

equivalent to (6) is

$$\ln h_{j|JR}(t) = \beta_1 \ln(t) + \beta_2 \ln(1+r) + \beta_3 \ln(rm + k_j(t)), \tag{7}$$

where r is the ratio of random to network-based meetings.[3]

Table 1: Survival Analysis estimates

	(1)	(2)	(3)	(4)
$\ln(m)$	1.259^{***} (0.007)	1.262^{***} (0.007)		
$\ln(I_j)$	1.021^{***} (0.001)	1.029^{***} (0.001)		
$t - t_j$	$-1.04 \cdot 10^{-6***}$ $(2.58 \cdot 10^{-9})$	$-9.22 \cdot 10^{-7}$ (n/a)		
$\ln(t)$			-6.026^{***} (0.011)	-1 (n/a)
$\ln(1+r)$			39.808^{***} (0.096)	-5.384^{***} (0.007)
$\ln(k_j(t) + m \cdot r)$			3.523^{***} (0.003)	3.334^{***} (0.003)
n	$1,053,738$	$1,053,557$	$1,053,557$	$1,053,557$
N	$214,071$	$214,071$	$214,071$	$214,071$
AIC	$-1,345,045$	$-1,342,796$	$-758,793$	$-556,400$
BIC	$-1,345,009$	$-1,342,772$	$-758,757$	$-556,376$

N denotes patents, n observations.

Standard errors in parentheses, significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Regressions were run without a constant, only observations where $t - t_j > 250,000$ are included.

The estimation is reported in column 3, and the predicted coefficients are $\beta_1 = \beta_2 = -1$, and $\beta_3 = 1$. Column 4 imposes $\beta_3 = 1$. We find a statistically worse fit to the data, and coefficient estimates that are incompatible with their

[3] Matching the distribution of in-degrees, we find $r \approx 4.5$.

predicted values.

4. Conclusion

The proposed model of patent citations introduces endogenous aging of patents and provides an excellent fit to empirical citation rates. Indeed, it fits the data better than power law aging, as predicted by standard preferential attachment models.

Acknowledgements

The authors thank Andrea Galeotti and an anonymous referee for useful comments and discussions.

References

Atalay, E., 2013. Sources of variation in social networks. Games and Economic Behavior 79, 106–131.

Auerswald, P., Kauffman, S., Lobo, J., Shell, K., 2000. The production recipes approach to modeling technological innovation: An application to learning by doing. Journal of Economic Dynamics and Control 24, 389–450.

Barabási, A.-L., Albert, R., Jeong, H., 1999. Mean-field theory for scale-free random networks. Physica A 272, 173–187.

Ghiglino, C., 2012. Random walk to innovation: Why productivity follows a power law. Journal of Economic Theory 147 (2), 713 – 737.

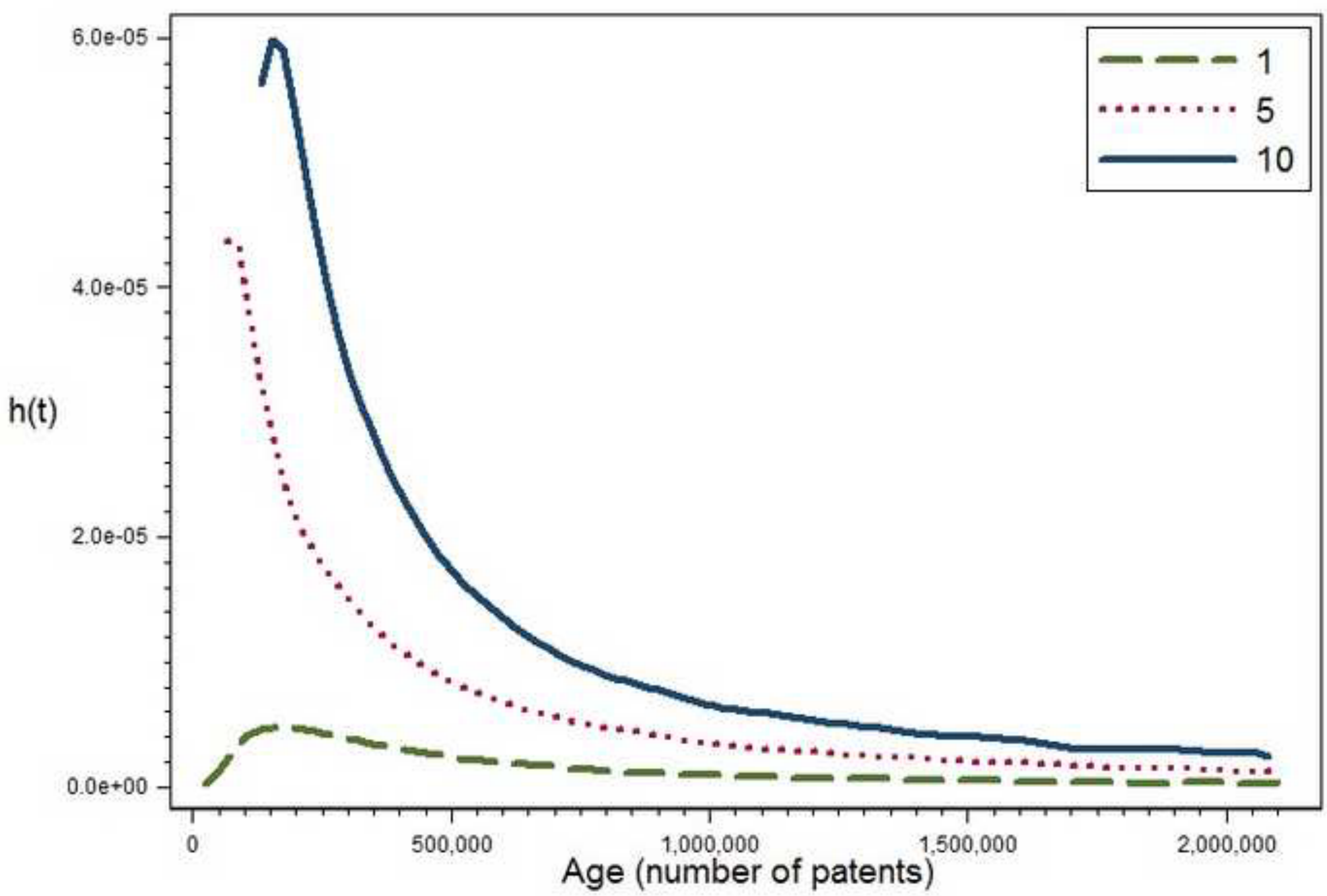
Jackson, M. O., Rogers, B. W., June 2007. Meeting strangers and friends of friends: How random are social networks? American Economic Review 97 (3), 890–915.

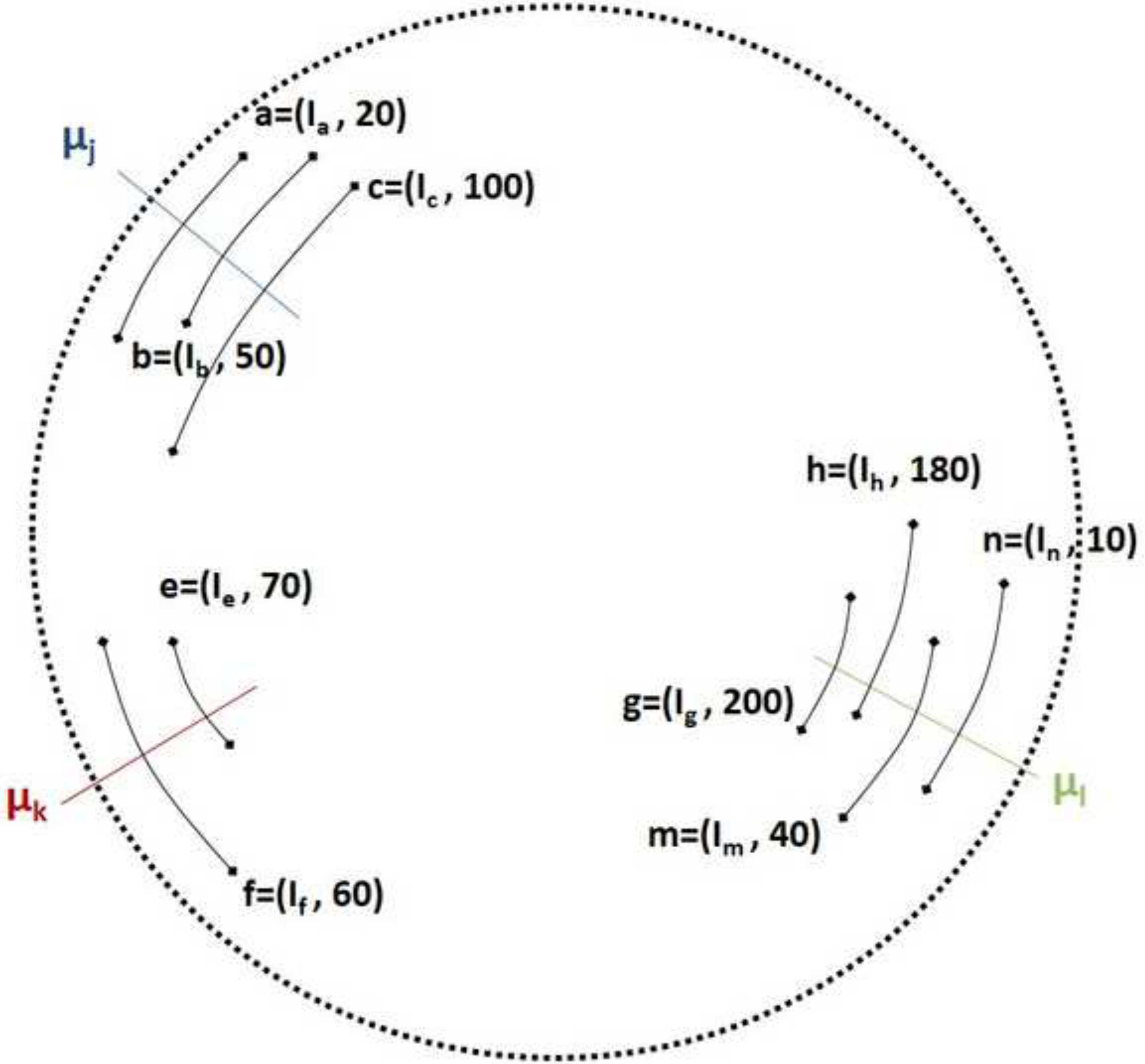
Marco, A. C., 2007. The dynamics of patent citations. Economics Letters 94, 290–296.

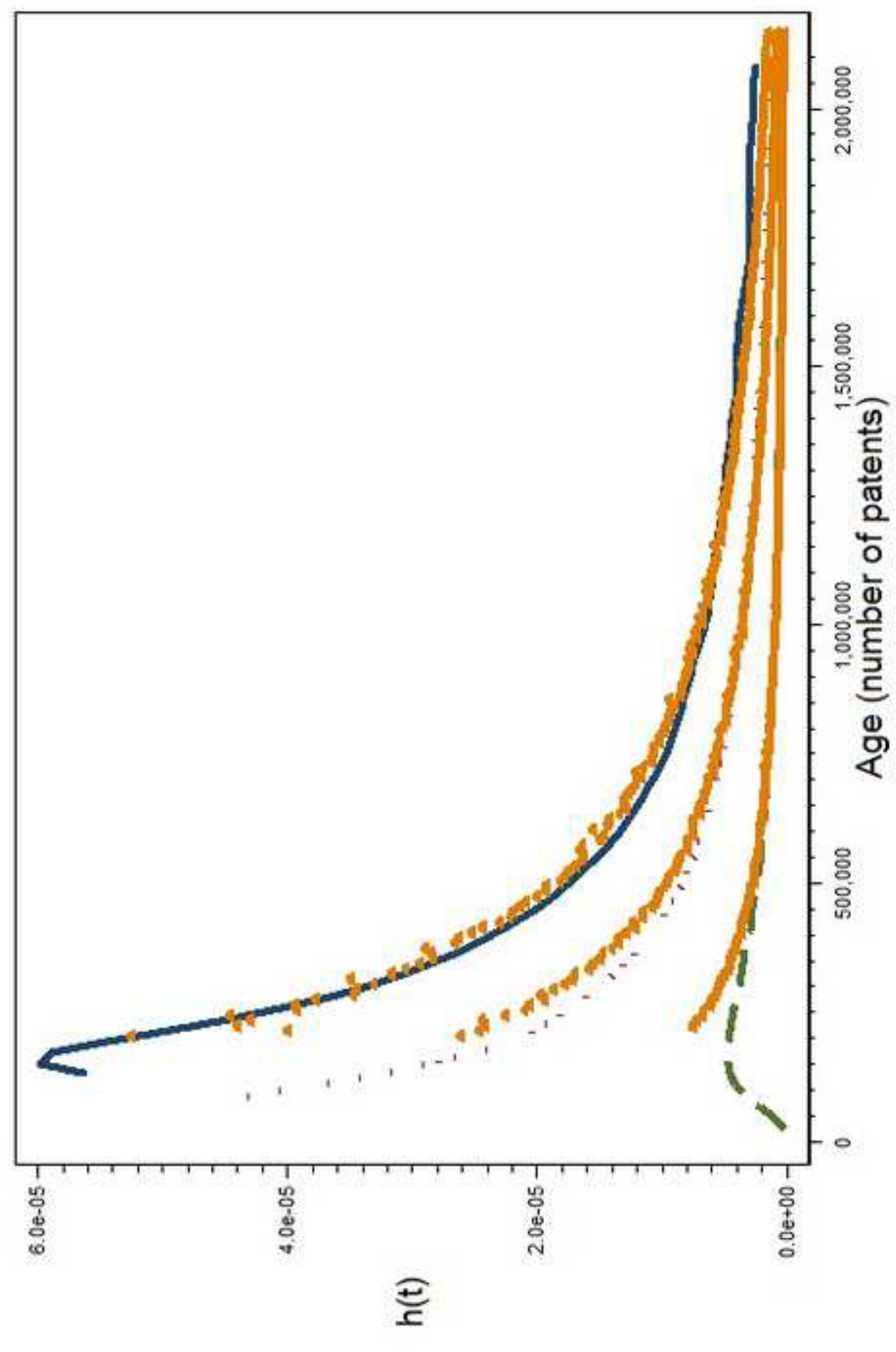
Peterson, G. J., Pressé, S., Dill, K. A., 2010. Nonuniversal power law scaling in the probability distribution of scientific citations. Proceedings of the National Academy of Sciences 107 (37), 16023–16027.

Valverde, S., Solé, R. V., Bedau, M. A., Packard, N., Nov 2007. Topology and evolution of technology innovation networks. Phys. Rev. E 76.

Weitzman, M. L., May 1998. Recombinant growth. The Quarterly Journal of Economics CXIII (2), 331–360.


6.0e-05
4.0e-05
2.0e-05
0.0e+00
h(t)
0
500,000
1,000,000
1,500,000
2,000,000
Age (number of patents)
1
5
10


μ_j
$a=(I_a, 20)$
$c=(I_c, 100)$
$b=(I_b, 50)$
$h=(I_h, 180)$
$n=(I_n, 10)$
$e=(I_e, 70)$
$g=(I_g, 200)$
μ_k
μ_l
$m=(I_m, 40)$
$f=(I_f, 60)$


h(t)
6.0e-05
4.0e-05
2.0e-05
0.0e+00
0
500,000
1,000,000
1,500,000
2,000,000
Age (number of patents)

Appendix 1

To numerically solve for the value of $\bar{I}$ that can be seen as a fixed point associated to equation (5), we use the following algorithm:

1. Guess a value of $\bar{I}$. Our initial guess is $\bar{I} = 9 \cdot 10^{-7}$.
2. Use the guessed value of $\bar{I}$ in equation (5) to calculate for each patent the implied $|I_j|$. For patents that have not been cited at all, set $k_j(t) = 0.0001$.
3. Calculate the average $|I_j|$ in the dataset at the end of the observed period, which is $t = 6,009,554$.
4. Calculate the difference between the guessed $\bar{I}$ and the one calculated in step 3, denoted x.
5. Replace the guess of $\bar{I}$ with $\bar{I}^{new} = \bar{I}^{old} - \frac{x}{2}$.
6. Repeat steps 2-5 until $|x| \leq 10^{-10}$.

Equation (5) assumes that the citations a patent receives is a continuous variable. However, in the data it is discrete, which is why for each patent we observe one value of $|I_j|$ for each t at which we observe the patent, which is at each citation and at the end of the dataset. We calculate the average $|I_j|$ from the data at the very end of our observed period as we believe that at this point, the data provide the most accurate estimate of $|I_j|$, as patents have been observed for the longest time.

We can instead look at the empirical average of (5) over all observations in our data in step 3 of our algorithm. That is, for each patent we take the average value of $|I_j|$ calculated from (5), and then average these across patents. In this case, we find that $\bar{I} = 9.07 \cdot 10^{-7}$. None of our qualitative results are sensitive to our choice between these values of $\bar{I}$.