Learning and the structure of citation networks

Working Paper Series

#2012-071

Learning and the structure of citation networks

Francois Lafond

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Maastricht Graduate School of Governance (MGSoG)
email: info-governance@maastrichtuniversity.nl | website: http://mgsog.merit.unu.edu

Keizer Karelplein 19, 6211 TC Maastricht, The Netherlands
Tel: (31) (43) 388 4400, Fax: (31) (43) 388 4499

UNU-MERIT Working Papers

ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology, UNU-MERIT

Maastricht Graduate School of Governance MGSoG

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT and MGSoG to stimulate discussion on the issues raised.

Learning and the structure of citation networks*

François Lafond ‡

October 22, 2012

Abstract

The distribution of citations received by scientific publications can be approximated by a power law, a finding that has been explained by "cumulative advantage". This paper argues that socially embedded learning is a plausible mechanism behind this cumulative advantage. A model assuming that scientists face a time trade-off between learning and writing papers, that they learn the papers known by their peers, and that they cite papers they know, generates a power law distribution of popularity, and a shifted power law for the distribution of citations received. The two distributions flatten if there is relatively more learning. The predicted exponent for the distribution of citations is independent of the average in-(or out-) degree, contrary to an untested prediction of the reference model (Price, 1976). Using publicly available citation networks, an estimate of the share of time devoted to learning (against producing) is given around two thirds.

Keywords: shifted power law, scale free networks, two-mode networks, cumulative advantage, polynomial attachment kernel, innovation, diffusion.

JEL codes: D83, D85, O31, O33

1 Introduction

In a paradigmatic model of the evolution of science, Price (1976) assumed that new papers tend to cite preferentially the papers already well cited. But do scientists cite the most cited papers or do they read the most well known? This article argues that the former is an artefact of the latter. In other words, the so-called preferential attachment of new papers to well-cited existing papers comes from a preferential learning mechanism by which scientists learn preferentially the popular papers and cite uniformly at random among the papers they know. While the literature proposes that the mechanism generating preferential citing is that scientists cite papers found in the bibliography of other papers, It is suggested here that the mechanism behind preferential learning is that scientists learn papers known by their peers. It is found that these assumptions predict a distribution of citations received similar

*I have benefited from comments of and/or discussions with Marco Capasso, Robin Cowan, Luiz Izquierdo, Alejandro Lavopa, Hibret Maemir, Daniel Opolot, Bulat Sanditov, Gerald Silverberg and Michael Verba. I am also grateful to audiences at the tenth workshop on networks in economics and sociology (Utrecht, 2011) and at UNU-MERIT internal seminars. This research was supported by METEOR, Maastricht University. All errors, omissions and approximations are mine.

‡Address: UNU-MERIT, Keizer Karelplein 19, 6211 TC, Maastricht, The Netherlands. E-mail *lafond@merit.unu.edu*.

to that of Price's model with generalized initial attractiveness (Dorogovtsev et al., 2000), but with a different effect of the main parameter (average bibliography size).

Since the prominent study of Price (1965), a large literature has repeatedly found that the number of citations received by scientific papers is approximately distributed according to a power law[1], that is, the share of scientific papers having received k citations is $p(k) = Ck^{-\gamma}$, where C is a normalizing constant and with the power law exponent γ generally lying between 2 and 5. This finding has been explained by a mechanism that was originally pointed out to explain another well known power law in scientometrics called Lotka's law of scientific productivity, which states that the number of authors having published k papers follows again a power law. Simon (1955) proposed a model to derive Lotka's law as follows. He assumed that at each period one new paper appears, and it has a new author with probability v, otherwise this new paper is authored by an author who is chosen from the pool of existing authors with probability (called attachment kernel) proportional to his publication record, thus creating a "rich get richer" effect. These two principles (growth and positive feedback) are enough to give rise to Lotka's law. Price (1976), however, was interested in the *citation* network. Arguing for the existence of what he dubbed "cumulative advantage", Price derived a power law distribution of citations received from the following assumption: a paper receives citations at a rate proportional to the number of citations that it already has, that is, there is a process of preferential citing by which papers cite preferentially the papers that have been already well cited. The purpose of the present article is to show that a power law-like distribution can emerge if we assume that scientists *cite uniformly at random*, but *learn preferentially* the papers which are already well known. A formal model is built using a growing two-mode[2] (author-paper) network in which there is a link between an author and a paper if the author has produced or has learned the paper. Thus, the focus shifts from the dynamics of the epistemic network (the citation network alone) towards the co-evolution of the socio-epistemic network (the social network of authors, the citation network, and the author-paper network). The key feature of the model is that the number of citations received by a paper is correlated to its popularity, defined as the number of people knowing the paper. This departs from the literature, which typically assumes that increasing returns are generated by people finding new papers by reading the papers in the bibliography of other papers. In the model, the source of cumulative advantage comes from the "social" side: agents learn papers known by their peers. The argument is that scientists like to know what fellow scientists already know, and ask one another relevant literature or pick it up on various occasions, such as seminars, workshops and conferences. This is not exactly the same as saying that they like to cite what peer scientists have cited nor that they like to know what peer scientists have cited. Clearly, the argument captures only partly the truth, but such a social aspect behind citation networks has been understudied, and the present paper is aimed at filling this gap.

To see how socially embedded learning can create a cumulative advantage, observe that if an agent chooses uniformly at random a paper known by another agent, then papers are chosen with probability proportional to their popularity. By reference to the landmark model of scale-free networks (Barabási

[1] Or a relatively similar law (to be discussed below). In all cases, the distribution is very (right) skewed, indicating a high level of inequality, where very few papers receive a very high number of citations, while the large majority of them receive none or a few.

[2] Two-mode networks are networks with two types of nodes, and links only between nodes of different types. By definition, they are bipartite.

and Albert, 1999), one may call "preferential learning" this "preferential attachment" (of people to papers). The mechanism which generates preferential learning is a two-mode version of what has been proposed for one-mode networks, a form of copying or referral (an early reference is Kleinberg et al. (1999), but see also Vázquez (2003) and Jackson and Rogers, 2007). Based on different analytical methods and simulations, the results indicate that preferential learning generates a power law distribution of paper's popularity and a *shifted* power law (Mandelbrot) distribution of citations received, the latter being in agreement with latest empirical evidence (Eom and Fortunato, 2011). In both cases, the power law exponent is related to the trade-off between learning and producing papers. However, a surprising outcome is that the exponent of the (shifted) power law for citations received does *not* depend on the average bibliography size (which, in the model, is both the mean in- and out-degree). This seems more in line with empirical evidence than Price's (1976) model which predicts a *negative* relationship, although evidence is scant.

The paper is structured as follows. Section 2 describes the fundamental assumptions of the model. Section 3 reviews related branches of literature. Section 4 describes the model and gives preliminary results. Sections 5 and 6 contain the results, i.e. the explicit formulas describing the evolution of popularity and citations received by individual papers, as well as the distributions, for the cases of an infinite and finite population respectively. Section 7 estimates the parameters of the model from publicly available data on citations among high energy physics papers posted on *arXiv.org,* and compares the model with the Price-Newman model. Finally, section 8 elaborates on some limitations of the model and possible extensions. The last section concludes. Analytical results that have been found using a master equation method are in the appendix, together with an extension of the model where productivity is parametrized, a brief discussion of the absence of Lotka's law in the model, and some robustness checks where sparse rather than full social networks are assumed.

2 A minimalist model of an author-paper system with learning

The beauty of Price's model is simplicity. It captures the essence of the phenomenon, and constitutes a framework that can be modified to accommodate more realistic features, for instance papers' aging (obsolescence), papers' quality, or accelerating growth. The learning-based model proposed here also aims at being as simple, general and flexible as possible. To this end, one needs to start by stating clearly what is taken to be the set of fundamental assumptions for modeling an author-paper network.

First of all, we will need to have a clear concept of "knowing":

Assumption 1 (Knowledge). *Papers are either known or unknown by any particular agent.*

In particular, the model will not distinguish between knowing a paper because one has learned it or because one has actually written it. This assumption brings in lots of analytical power, since it implies that we can construct a two-mode network where a link between i and j exists if and only if agent i "knows" paper j.

Second, direct empirical observation unequivocally shows that

Assumption 2 (Growth). *The total number of papers increases.*

Assumption 2 does not appear controversial, but it will also be assumed that growth takes place at a constant rate, which is against empirical evidence, but is a useful first approximation.

So far, we have a two-mode network with one of its two sets of nodes growing. A mechanism producing the well known "Matthew effect" can now be introduced:

Assumption 3 (Embeddedness)**.** *Learning is socially embedded in the sense that agents learn what their peers know.*

Clearly, science is a collective endeavor shaped and constrained by social processes, and often scientists learn from others, in seminars, conferences, and interactions with colleagues. For instance, it is common in the literature to assume that since a large part of knowledge is tacit, it flows through face-to-face interactions (Cowan, 2005). In the model, papers are pieces of codified knowledge which exist independently of the people who know them, and flows of tacit knowledge are not explicitly modeled. It is simply assumed that the choice of which paper to learn is socially embedded in the sense that agents randomly choose a paper known by a randomly chosen friend. In what follows, it will not be possible to account for all the details of the ways in which the structure of the social network influences learning. For simplicity (and mathematical tractability), a full network, i.e. where everybody is friend with everyone else, will be assumed[3]. The point of focus here is that if agents learn what is well known, there are increasing returns in the dynamics of papers' popularity. In other words, assumption 3 leads to a "rich get richer" effect:

Lemma 1 (Cumulative advantage)**.** *Past accumulation of success increases the probability of success.*

The original insight of Simon's model is that assumption 2 and lemma 1 (which is an assumption in his case) generate a power law. Simon showed that growth of the number of papers and cumulative advantage of authors regarding new authorship generates Lotka's law. Price (1976) showed that growth of the number of papers and cumulative advantage of papers regarding new citations generates scale-free citations networks[4]. Ramasco et al. (2004) showed that growth of the number of authors and papers and the cumulative advantage of authors regarding authorship generates Lotka's law and a scale-free co-authorship network. I will show that growth of the number of papers and cumulative advantage of papers regarding new diffusion generates a scale-free distribution of papers popularity, which, associated with a random citing mechanism, produces a citation network characterized by a shifted power law in-degree distribution.

Note that, as compared to Simon's model, the two sets (authors and papers) are inverted. In Simon's model, a new authorship relationship arrives and with some probability the author is new. In our context, a new "knowing" relationship appears and with some probability the paper is new. To justify putting Simon's model upside down, observe the following premises:

Assumption 4 (Knowledge creation)**.** *Knowledge is created by the actors, using effort and their own knowledge.*

[3] In appendix G, it is argued that the full network assumption is enough to capture the essence of embeddedness for the purpose of this paper. When the social network structure is a circle the shape of the distributions changes a bit, but it is very similar when the social network is a Erdős-Renyi random graph.

[4] "Scale-free" is used loosely here and in the literature in general. In fact Price's model features a shifted power law degree distribution, which is not scale-free except for (very) large degrees.

One common way to deal with assumption 4 is to consider the dynamics of knowledge accumulation alone: knowledge builds out of itself. However, this would leave aside an important feature of the real world, which is that agents need to know what *others* know. In other words,

Assumption 5 (Learning)**.** *Agents are learning*[5]*.*

Assumption 1, 2, 4 and 5 together imply that

Lemma 2 (Learning/Creating trade-off)**.** *Agents divide their time between learning existing papers and creating new ones.*

This is equivalent to saying that it is impossible to read a paper while writing one's own paper. Since time is limited, the two options are traded-off (it does *not* imply that they are not complementary activities). The easiest then is to assume that this trade-off is solved by an exogenous, fixed probability.[6]

Finally, assumptions 1, 2 and 5 imply two lemmas

Lemma 3. *An agent does not learn the same paper twice.*

Lemma 4. *Creating a paper or learning a paper produce an equivalent result: the actor knows the paper.*

Note that lemma 4 is not too unrealistic; after all, reproducing the work of others can lead to a quite equivalent "understanding".

In summary, the view expressed here is that to understand the evolution of science, it is absolutely necessary to incorporate actors, and the facts that they are learning, that they rely on their peers for awareness of relevant knowledge items to be learned, and that they are trading-off learning and producing.[7] This is a necessary consequence of the view that humans are social beings, and that knowledge is created by humans using their existing knowledge, subject to a time constraint. Furthermore, these elements do not add much complication. Thanks to these assumptions, the system can be summarized in a simple growing matrix with binary entries called the incidence matrix of the two-mode network.

Before presenting the detailed description of the model, let us see how it relates to several strands of literature.

3 Related Literature

The model is related to a long tradition of modeling science, and in particular citation networks (section 3.1). The availability of very large datasets has revived a debate regarding the best functional form for empirical citation distribution (section 3.2). Meanwhile, many have argued that two-mode networks are

[5]Throughout the paper, "learning" is used to refer to the act of increasing one's knowledge by learning a piece of *already existing* knowledge, that is, already known by somebody else. Learning thus characterizes knowledge "diffusion" or "imitation". By contrast, increasing one's knowledge with original, new-to-the-world knowledge will be referred to as "innovation".

[6]Alternatively, one can explicitly model choice, e.g. by maximizing a certain utility function which weighs the benefits of learning v.s. producing. This 'microfoundation' of the model is not necessary for the main argument presented here.

[7]In the model, it is implicitly assumed that the productivity of learning is equal to the productivity of innovating, and it is equal to one new paper (invented or learned) per period in which an agent is selected. Parametrizing productivity leads to qualitatively similar results, see appendix F.

an insightful way of representing science (section 3.3), a modelling standpoint which is easily backed up by social theories of knowledge such as actor-network theory (section 3.4). A more general background discussion of the importance of social networks in economics is omitted and can be found in Goyal (2007), Vega-Redondo (2007), and Jackson (2008).

3.1 Models of citation networks

The literature on citation networks has become prominent following the seminal contribution of Price (1965) who explicitly conceived scientific publications and their bibliographies as a citation network, and explored the power law nature of the citation distribution. Power laws are ubiquitous in natural and artificial systems, and in the 1950's they had already been found for city sizes by Zipf (1949) and for biological genera sizes by Yule (1925), which motivated the paper by Simon (1955). Simon's model is in fact a growing two-mode network, but Price (1976) showed that it could be used to analyse a growing citation network. He assumed that new papers constantly arrive and cite previous papers with probability proportional to their already accumulated stock of citations.

One problem with this heuristic is that since papers start their life with zero citations, they have no chance to receive their first citation. In his original article, Price simply counts publication as the first citation, such that everything happens as if papers were arriving with an in-degree equal to one. Dorogovtsev et al. (2000) have generalized Price's (1976) model[8] to allow for an initial attractiveness (call it a), and showed that the exponent of the (shifted) power law is given by $2 + a/h$. Hence, in this theory, the slope of the shifted power law depends on two parameters: the average bibliography size (h) and the initial attractiveness (a).

The scale-free network model (Barabási and Albert, 1999) has triggered a huge amount of research on growing networks, providing important new inputs for the understanding of citation networks. For instance, super linear preferential attachment, the assumption that existing nodes receive new links at a rate more than proportional to their degree (Krapivsky et al., 2000), and initial attractiveness (Dorogovtsev et al., 2000) have been used to explain the in-degree distribution of the patent citation network by Sanditov (2005). Super linear preferential attachment and aging (Dorogovtsev and Mendes, 2000) have been used in the same context by Csárdi et al. (2007) and Valverde et al. (2007).

Another strand of literature has focused on the idea that authors typically cite recent literature, and then cite papers cited by this first article. An ingenious empirical study partly justifying this assumption is that of Simkin and Roychowdhury (2011), who have shown that misprints in citations tend to repeat (in fact, the number of misprint repetitions follows a power law). They propose a model which allows them to estimate that 70 to 90% percent of the citations are copy/pasted from other papers[9]. Another recent paper, by Peterson et al. (2010), explicitly distinguishes between two

[8] Newman (2010) proposed to simplify the in-degree distribution of Dorogovtsev and Mendes's (2000) model as a ratio of Beta functions (by multiplying Dorogovtsev and Mendes's (2000) equation (9) by $1 = \Gamma(2 + a)/[(1 + a)\Gamma(1 + a)]$). Golosovsky and Solomon (2012) recently called it a Waring distribution (this can be checked from its generating function defined in Irwin (1963), appendix II p.29). Throughout the paper I will refer to "Price's model", but credit is due to many authors for extending and clarifying Price's original contribution.

[9] According to them, copying a misprinted citation constitutes evidence that the faulty author has not read the paper. One may dispute this interpretation, but their data at least suggests that sometimes authors know, among their references, which one cites another, and use this information to save time in preparing their reference list. From this, it is tempting to infer that an author became aware of a certain paper by seeing it referenced in another paper. Nevertheless, this

mechanisms underlying citations. In the *direct mechanism*, which happens with probability $(1 - c)$, a paper simply gets to be known and cited (the new paper cites every old paper with equal probability). In the *indirect mechanism*, which happens with complementary probability c, the paper cited is taken from the bibliography of the known paper (thus the new paper cites every old paper with probability proportional to the in-degree of the old paper). They find a power law exponent equal to $1 + \frac{1}{c}$.

The difference between the two mechanisms above is related to Jackson and Rogers's (2007) model for social networks, where newborn agents choose to link to random existing agents ("random meetings"), and to random neighbors of these first chosen agents ("search")[10]. Atalay (2011) argued that this simple model cannot account for differences in citations received by papers of similar age. He modified the model by introducing a fitness function, as in Bianconi and Barabási (2001), to account for the differentiated quality of the papers.

In the context of patents, Ghiglino and Kuschy (2011) have proposed that the origin of the in-degree distribution is to be found in the heterogeneous nature of patent's applicability. They assume that there is a fixed number of patent classes, each of which sees new patents arriving following a Poisson process. Each patent i has a type μ_i and a broadness a_i which determines its support $F_i = [\mu_i \pm a_i]$. When a new patent arrives, it cites the youngest patents which are in the support of technologically related but non already existing patents. As a result, the probability of patent i being cited is proportional to its broadness, a_i, so the distribution of citations received also depends on the distribution of the a_i's.

Bramoullé et al. (2012) also propose a model to explain the pattern of citations, though their main focus is on "long-run integration", which occurs when the distribution of types of a node's friends is the same as in the whole population. Again based on Jackson and Rogers's (2007) idea of an initial random choice followed by neighbor-picking, Bramoullé et al. (2012) also add paper's type, and a homophilous process: the initial random choice of papers is biased towards same type (same subfield) papers. Their key insight is that, as time passes, the search process becomes more prevalent than the random meetings. Since in-degree is correlated with time, if the search process is unbiased we should observe a form of integration, which they confirm empirically on the American Institute of Physics dataset (1985-2003, 207,912 papers): as in-degree increases, the share of citations from same-type papers decreases.

This literature recognizes that authors must know the existence of a paper before they cite it, but does not explicitly keep track of who knows what. To my knowledge, the only model that studies citation networks and (partly) does so is the TARL (Topic, Aging and Recursive Linking) model of Börner et al. (2004). They proposed that authors and papers co-evolve, but additional features of their model (papers' topics and aging) prevent them from finding closed-form expressions for the degree distributions. Notwithstanding details, the main difference with the model presented here is that they assume that authors, at each period, choose a number of items to read among all existing papers, and

obviously does not constitute a proof, because one can for instance know of a paper, see it subsequently cited by another paper, and copy the citation from this last paper.

[10]In fact Peterson et al.'s (2010) model is identical to Jackson and Rogers's (2007), but they obtain the Waring distribution from the master equation (as Atalay, 2011) instead of the shifted power law from the mean-field approximation. As a result the exponent of the Peterson et al.'s (2010) model can be rewritten $\gamma = 2+(1-c)/c$ to be compared to Jackson and Rogers's (2007) $\gamma = 2 + m_r/m_n$. In both cases $\gamma - 2$ is equal to ratio of "random" over "preferential" meetings. In the learning-based model presented here, the exact same citation distribution is obtained, but the models' primitives are different. The exponent reflect the fact that learning is "preferential" (because agents learn random ideas of random friends) but innovating is not (a written paper always arrive with a one author). See *infra*.

then scan the bibliographies of these papers to find additional literature, as in Bramoullé et al. (2012) and Peterson et al. (2010). By contrast, in the learning-based model below, authors do not find new references by looking in the reference list of existing papers; they find new papers to learn by choosing at random a paper known by a friend. Another important difference is that the TARL formulation does not explicitly model the trade-off between learning existing papers and producing new ones, whereas this distinction is central in the learning-based model where it allows to study the consequences of the allocation of attention between knowledge consumption and production.

3.2 Degree distribution of citation networks

The statistical analysis of power laws is a rather technical and controversial topic (Perline, 2005; Stumpf and Porter, 2012) nicely summarized in Clauset et al. (2009) and Gabaix (2009). Regarding citations data, the early contribution of Price (1965) and Seglen (1992) have been followed since the turn of the century by a number of studies on large datasets, following the influential note of Redner (1998), who estimated that the tail of the distribution for the two datasets he studied could roughly be fitted by a power law with an exponent around 3. Arguing that a stretched exponential function (Laherrere and Sornette, 1998) is helpful in fitting the left side (papers with few citations) but fails in the tail, Tsallis and De Albuquerque (2000) introduced a form, which can fit the whole distribution, namely $p(k) \propto [1+(g-1)\lambda k]^{\frac{-g}{g-1}}$, where g and λ are parameters. It can be rewritten $p(k) \propto (k+\frac{1}{(g-1)\lambda})^{\frac{-g}{g-1}}$ to show that it is a particular case of the shifted power law, or Mandelbrot's (1953) law (or Pareto type II or Lomax distribution) $p(k) \propto (k+r)^{-\gamma}$, with a particular correlation of its parameters.

Several recent papers have used massive datasets to discriminate between the three major candidates: power law, shifted power law, and log-normal. Stringer et al. (2010), looking at the ultimate number of citations received (i.e considering only papers which are not cited anymore), found that the distribution is a discrete version of the log-normal. Albarrán and Ruiz-Castillo (2011), with a huge dataset and considering citations received after 5 years, found that in most fields it is not possible to reject the hypothesis of a power law for the tails. Eom and Fortunato (2011), using a citation network constructed from papers published in American Physical Society journals, compared the fits given by the power law, the log-normal, and the shifted power law, finding that the last one is the best. The details of those studies shall be considered in section 7 where an empirical discussion of the implications of the model is provided.

3.3 Two mode networks

Sociologists have long been using two-mode networks to analyze social phenomena (Freeman, 2003). For instance, the classical paper of Breiger (1974) studies the dual relationship between individual people and groups using a two-mode network where one mode is the actors, and the other is the groups to which they belong or the events that they attend. The intuition here is that actors are linked through their common affiliations or participation in certain activities. Indeed, although many classical concepts for one-mode networks have been extended to two-mode networks (Faust, 1997; Latapy et al., 2008), two-mode networks are often projected into one-mode networks, in which two actors are linked if they have a common affiliation. However, many have argued that in most applications it is better to stick with two-mode networks, which allow, among other things, to study the co-evolution of the two sets of nodes. For instance, Roth and Cointet (2010) studied empirically a scientific collaboration network and the concepts used in the associated articles, effectively characterizing socio-semantic co-evolutionary dynamics.

The model below is close to this literature in that it adapts Price's (1976) and Barabási and Albert's (1999) models to a two-mode network framework, following the pioneering work of Ramasco et al. (2004) which was focused on co-authorship networks. Ramasco et al. (2004) consider both the set of actors and the set of papers, but only the *production* of papers. They propose that at each time step a new paper arrives, this paper having n_a authors. The authors are partly new, and partly taken from the existing set of authors *proportionately to their existing number of authored papers*, as in Simon's original model. Thus, a difference with the model proposed below is that actors create new papers but they never learn existing ones. Another difference is that Ramasco et al. (2004) are interested in the distribution of the number of papers per author (whereas the interest here is in the number of agents knowing a given paper, which is the other side of the two-mode network), and on the projection of the two-mode network into its "co-authorship" one-mode projection (whereas projections are not considered here).

Ramasco et al.'s (2004) model is an extension of Simon's (1955) original model. These two models are growing two-mode networks where the two sides are growing. However, many situations involve that one set of nodes is fixed (below, the number of authors is kept constant, and population size is shown to have important implications). Peruani et al. (2007) have provided detailed results regarding the degree distribution of one set of nodes of a two-mode network when the other set grows, but to the best of my knowledge, nobody has analyzed the degree distribution of a set of nodes of a two-mode network when this set grows and the other one is fixed.

Finally, there is a larger, rapidly expanding literature on self-organizing multi-mode networks which is impossible to review here. A few recent contributions can be mentioned, starting with the network rewiring model of Evans and Plato (2008). Their model is a two-mode network with *both* sets of nodes fixed, and only a rewiring process. Actors are linked to one and only one artefact, and the distribution of artefacts' popularity is studied. In an extension of the model, actors imitate the artefact of their friends, which generates a preferential attachment of actors to artefacts. Their model applies in particular in anthropology where one is interested in the transmission of cultural artefacts. The model proposed below also applies to this context, but assumes that new artefacts appear over time, and that actors accumulate artefacts over time. There is also an upsurge of interest in two-mode or multi-mode networks in which all sets of nodes are growing. For instance, Beguerisse-Diaz et al. (2010) studied a system where users rate videos. Liu et al. (2011) study a social tagging system, which can be seen as a three-mode network (users, resources tagged, and tags). Zeng et al. (2012) show that certain recommender systems produce more unequal popularity distribution than others. The model proposed below contributes to this literature by providing a detailed analysis of the artefact degree distribution under the assumption of a non-growing population of actors, and by using the two-mode network in conjunction with the two associated one-mode networks.

3.4 Science in the making: some arguments for actor-paper systems with learning

Perhaps one of the most general, least controversial statement about scientists is that they are learning. But what do they learn? Once we assume that they learn knowledge items nicely encoded into "papers", the question becomes: what is the structure of this actor-paper network?[11] The point made here is that the social embeddedness of learning creates a bias towards popular items which greatly determines the shape of this network. It is proposed to model the self-organization of science as a morphogenesis of the actor-paper system driven by heuristics: agents learn ideas of their friends, and, when they write a paper, they cite what they have learned before. This approach, which significantly complicates the one-mode approach to citation networks, is justified by a large literature in science studies where it was highlighted that many "social" processes affect the self-organization of science (Merton, 1968; Latour and Woolgar, 1979). Furthermore, not only there are arguments to introduce humans in citation network models, but there are also arguments to introduce non-humans in social network models of science. Proponents of actor-network theory (Latour, 2005) contend that artefacts – material or not, but "non-humans" – indeed populate the net in which human agency takes place. One may argue that the most important such objects in science are the publications. Hence, one can choose as a fundamental unit for the description of science the relationship between an actor and a piece of knowledge. The dynamics of the system follows from the constitution of this relation (the learning event), because it is not independent of the rest of the network: it is this feedback loop which leads to self-organization. A proposal of formal modelling of these ideas appears in the next section.

4 The model

4.1 Description

Consider a two mode network with n agents and w papers. Papers are either known or unknown by any given agent, which is represented by the presence or absence of a link between an agent and a paper. The number of agents is kept fixed, but the number of papers grows. Time is discrete and indexed by t, and, because there is exactly one event (learning or creating) per period, it does not measure clock time, but rather system time. Denote by E_t the total number of actor-papers relationships, i.e. the number of edges of the two-mode network. At the beginning , there is one paper known by one randomly chosen (r.c.[12]) agent ($w_0 = 1$ and $E_0 = 1$). Then at each time period, the following algorithm[13] is applied:

I/ pick an agent i at random.

II/ with probability b, the agent i creates a new paper (a new node is added to the set of papers, and

[11] Throughout the paper, I use the words agent, author and actor interchangeably, but notice that as compared to the literature considering only authorship, the agency of actors is extended by the fact that they can learn. Although this is modeled in a minimal way (a constant probability), it captures a crucial behavioral aspect, and can be refined to include more subtle rules or, if need be, strategic considerations.

[12] Throughout the paper, "random" refers to a uniform distribution

[13] The model is firstly a computer model, because this allows one to get results that are not possible to find analytically, and it forces one to specify every single detail of the procedure. The mathematical analysis presented below is always a "model of the model", with different degrees of approximation giving different degrees of accuracy.

an edge is added to the two-mode network, between i and the new node). If so, pick h papers at random in i's portfolio, and add directed links between the new paper and these h existing papers.[14]

III/ otherwise (i.e. if the r.c. agent does not create a new paper), pick another agent i' at random.[15] Then pick at random a paper s among those papers known by i' and unknown by i. Then i learns s (an edge is added to the two-mode network, between i and s)

Each step of the algorithm is commented successively. The first step (I) ensures symmetry among agents, which prevents the emergence of Lotka's law (see appendix E), but will greatly facilitates the derivations because we will need to know some agent-level properties, in particular the average number of degrees of the ideas known by i (see the discussion on the normalization factor of the attachment kernel, or "partition factor", *infra*).

The second step (II) states that with some probability b, a new node will be added to the set of papers. This new paper is linked to its inventor/writer, i. The parameter b is assumed constant. b can be interpreted as the share of attentional time devoted to the creation of a paper, as compared to learning existing papers.[16] An interesting feature of the model is that diffusion and creation are integrated. To be sure, this is simply because the time unit of the model is the appearance of a new paper, pre-existing elsewhere or not, in the portfolio of an agent. Thus b is an important parameter, and we will see that it determines, for instance, the steepness of the power law tail of the distribution of papers' popularity and of papers' number of citations received.

Moreover, the citation network is also modelled. That is, this new paper will make some citations to other (pre-existing) papers. However, the new paper can cite only papers that are known by its writer.[17] It is assumed that, among this set of papers known by i, the h papers that will eventually be cited are chosen uniformly at random.

[14] One might address to this description a critique similar to the one addressed by Bollobás et al. (2001) to Barabási and Albert's (1999) model regarding the specification of initial conditions: At the start of the algorithm, the agent does not know h papers. In the computer code, it is assumed that the agent cites all the papers that she knows (this can be none). In the mathematical model, one can rely on a classic argument from the growing networks literature that the frequency of this problematic event goes to zero as t goes to infinity (and, in this case, as h tends to 0).
Another boundary condition issue is that if both b and n are small, there are not enough new papers to satisfy the number of required learning events. This problematic configuration always happen with non negative probability, and to ensure that the model always run, the computer code is as follows: when a r.c. agent is supposed to learn but there is no papers that she does not already know, she creates a new one. To get a glimpse of how important is this boundary condition issue, and for which values of the pair (b, n) it becomes negligible, a corner-of-the-table computation may be done as follows: For any given agent, we need that when she has to learn (an event which happens with probability $(1 - b)(1/n)$, enough other agents have created new papers (events which happen with probability $b(1-1/n)$. Solving $(1-b)(1/n) < b(1-1/n)$ gives the condition $b > (1/n)$.
Finally, also note that the upper bound on b is 1, but it is uninteresting: in this case each paper is known one time.

[15] The assumption is that a friend of i is chosen, but for simplicity the social network is assumed to be a full graph. This assumption is relaxed in appendix G.

[16] Alternatively, one can think of b as the result of the following requirement: agents need to learn z papers before they can create a new one. As a result, they innovate every $z+1$ periods, or on average $b = \frac{1}{z+1}$.

[17] This sounds rather obvious in scientific research: an author has to know (at least the existence of) a paper before she cites it when writing her own paper. But it is not valid for patents, which have at least part of their reference list added by the patent assignee.

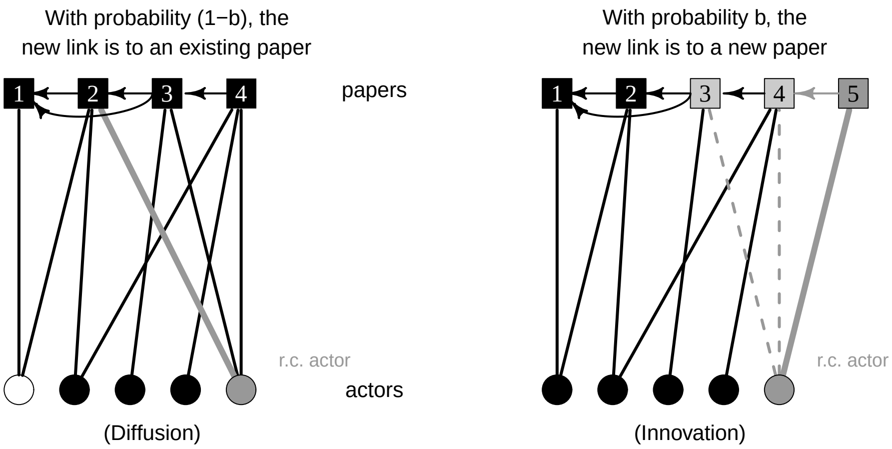

Figure 1: **Schematic description of the model**. At each time step, one and only one of the two events represented above happens. In both cases, a link is added (the thick grey one). Note that since the grey link is the ninth (of the two-mode network), we are looking at what happens during period $t = 8$. There are three different degree distributions. The main focus here is on the distribution of (undirected) degrees of the top nodes (papers' popularity), $p(k)$. The distribution of degrees of the bottom nodes (actors' knowledge), $p_a(k_a)$ is discussed in appendix E. The directed arrows and the top nodes constitute the citation network, used to compute the distribution of citations received (in-degree distribution) $p^C(k^C)$. On the left panel, the r.c. agent is learning. In this case, a r.c. neighbor has been chosen and turns out to be the leftmost (in white). There are only two papers unknown by the r.c. actor and known by the r.c. neighbor (1 and 2). She chooses one of them at random – in the example above she turned out to choose the second paper. On the right panel, the r.c. agent has created a new paper. The dashed grey edges represent the papers that she knew previously. A number of them (here $h = 1$), are chosen (uniformly at random) to receive a citation from the new paper (dark grey, 5). Papers 3 (grey) and 4 (grey) had equal chance to be cited; eventually only paper 4 is cited. The social network between bottom nodes, not depicted here, is assumed to be full throughout the paper except in appendix G where robustness checks are presented.

The third step (III) leads to *preferential learning*. The simplest argument to justify step III is "referrals": scientists suggest one another references. Although common and intuitive, one may think of other arguments leading to preferential learning. For instance, this is a fairly legitimate assumption if one considers the theory of attention. In this literature (e.g. Styles, 2006), it is emphasized that the main driver of attention is *saliency* (defined from context). In the model, what makes a paper salient is the number of people who know it. For instance, in a perhaps more interactionist view, one can say that the more it is known, the more it is discussed, so the higher the chance that it is heard by or come to the attention of the agent.[18]

[18] In the two justifications above, and in the way the model is described and computed, the agent does not know how popular a paper is. However, one may think that agents actively choose to learn what is well known, perhaps because they see popularity as an indicator of quality, or because in general agents need to have some common knowledge if they want to communicate (a "coordination" argument). In this case, papers have a role of focal points in the sense that people learn them because they expect others to know them and think they are expected to know them. In a more sociological parlance, they would be called "boundary objects" (Star and Griesemer, 1989): of course they may be understood slightly differently by different learners, yet they are understood homogenously enough that people can refer to them when building collective understanding. It should be stressed that to a large extent, the results on the degree distributions do not depend on the origin of the attachment kernel, but simply on its final form. In other words, for the degree distributions, it does not matter if agents learn preferentially because they like popular items or because they are

The formulation of step III has very important consequences for the resulting network. Since agents learn what their friends know, they have higher chances to learn what is already well-known. More precisely (but not exactly), the chance of a paper to be learned is proportional to how many times it is already known. In the complex networks literature, the rates (probabilities) at which nodes having a certain degree receive additional links is called *attachment kernel*. The famous "preferential attachment" case is the one in which a node receives new links with probability proportional to how many links it already has, and is also called linear attachment.[19] A classical explanation for linear attachment in one-mode networks is "copying". For instance, in social networks (Jackson and Rogers, 2007), choosing friends of a node means choosing nodes with probability proportional to their degree. In citation networks (Peterson et al., 2010), choosing a paper cited by another paper means choosing a paper with probability proportional to its (in)degree. The model specified above uses the same intuition, but in a two-mode fashion: choosing a paper known by an agent means choosing a paper with probability proportional to its degree (popularity, or number of other agents who know it).

The model does not have, however, a linear attachment kernel, but a polynomial one, a case which does not seem to have been studied before. The reason is that agents do not learn twice the same paper. It implies that the number of times that a paper is known is bounded by the total number of agents. It also implies that a paper can be learned only by agents who do not already know it, so the larger the number of agents who know it, the less the chances that the r.c. agent does not know it, so the less the chances of that paper being learned. This effect creates a factor in the attachment kernel which is similar to the carrying capacity in population growth models. The resulting attachment kernel is thus a polynomial with a negative coefficient for the second order term, as will be discussed in more details below.

The following section clarifies the setup of the model by deriving key mathematical relationships implied by the algorithm (I-III).

4.2 Preliminary results

To understand the model, let us consider its matrix representation. Consider a matrix Q which has a fixed number of rows, a number of columns w_t that depends on time, and entries Q_{ij} equal to one if agent i knows paper j, and zero otherwise. Start with a column vector filled with a one and $(n-1)$ zeros. At each period, with probability b, a column is added (a new paper is created). Then, with probability 1, one entry of Q is changed from zero to one (if a new column has been added, this modified entry must be in that new column). Letting E_t be the total number of ones in Q, we have, for large t, $E(t) = t + 1 \approx t$, and $w(t) = b(t + 1) \approx bt$. Then, it is direct to see that the density of the system, denoted D, is stable because b is stable:

$$D(t) = D = \frac{E_t}{nw_t} = \frac{t}{nbt} = \frac{1}{nb}$$

embedded in a (full) social network and learn the papers known by their friends. For more detailed topological quantities, such as clustering or path length, and for convergence properties, this would make a difference.

[19] The term linear has been introduced to distinguish this case from the cases in which nodes gain links with probability proportional to a certain power of their degree. Those cases are called sub-linear (super-linear) if the exponent is less (more) than 1.

Hence the density of the two-mode network is constant (independent of system time t). We see that an increased rate of innovation b will make the system sparser (since more papers have to be learned), whereas a high rate of learning $(1-b)$ will make it denser. In this model, growth (assumption 2) corresponds to the increment of a structural dimension (a column). The learning component (assumption 5) ensures that the density of the system stays stable.

The main objective is to derive $p_t(k)$, the probability that a r.c. paper in t is known k times (i.e. has degree k). To this end, for reasons to be clarified below, we need to know the rate at which papers are learned (i.e. become more popular). Consider the degree of paper j, denoted k_j. How does k_j increase? That is, under which conditions paper j will be learned at time t? First, the r.c. agent i must be learning, which happens with probability $(1-b)$. Second, paper j must be unknown to i. Since j is known k_j times by definition, the probability that i knows j is k_j/n, and thus the probability that the paper j is unknown to i is $(1-k_j/n)$. Third, the paper j must be chosen, and because of the preferential learning mechanism described above, this happens with a probability *proportional to* k_j.

Putting these together, the evolution of paper j's popularity is

$$\frac{dk_j(t)}{dt} = (1-b)A_t(k_j) \tag{1}$$

where, since one and only one idea must be selected, the attachment kernel $A_t(k_j)$ has been defined such that it is equal to unity if summed up over j.

$$A_t(k_j) = \frac{k_j}{\sum_{l \notin N(i)} k_l}\left(1-\frac{k_j}{n}\right) = \frac{k_j(n-k_j)}{\mu n t} \underset{n\to\infty}{\to} \frac{k_j}{t} \tag{2}$$

where $N(i)$ denotes the set of papers known by i. By definition, one degree is added at each period so $\sum_{j=1}^{w_t} k_j = E(t) \approx t$. Thus we see that the restricted sum in the denominator must be a fraction of t - let us call this share μ.[20] It is the fraction of actor-papers relationships that do not involve a paper known by an average r.c. agent (i), and it tends to 1 as $n \to \infty$, because if the population is infinite, a r.c. agent has been previously chosen a number of times which tends to 0, so the number of papers that she knows (and the sum of their degrees) tends to 0. Note that if all ideas were to have a chance of being learned by the r.c. agent of the current period, the normalization factor would just be the total number of degrees, t. But it would be assuming that agents can learn twice (or more) the same paper, a scenario that was explicitly excluded in step III of the algorithm (see lemma 3). In fact, the correct normalization factor μt is the *total number of degrees of the papers unknown by the r.c. agent* ($\sum_{l \notin N(i)} k_l$), or equivalently the total number of degrees of all papers (t), minus the total number of degrees of the papers *known* by the r.c. agent. One can obtain the total number of degrees of the papers known by i by summing up for all papers their degree times the probability that i knows them.

[20] In fact, μ should be indexed by i and by t, and ideally one should be studying the evolution of the properties of the distribution of the μ_i. In this paper, I will only claim that the distribution of papers' popularity can be found assuming that $\mu_{it} = \mu$. A detailed analysis is certainly informative of the learning trajectory of individual agents, but it is not necessary for deriving the key results of the paper. I couldn't find a closed-form expression for μ in terms of b and n. Using equation 3 and the fact that that if $p(k)$ is uniform (implying the minimum possible $\langle k^2 \rangle$), $\langle k^2 \rangle = 1/b^2$, one finds that μ is bounded above by $1-1/(bn) = 1-D$. So the upper bound of μ increases with b and with n. To find a lower bound, assume that all the mass goes to the least number of papers (so we find the maximum possible $\langle k^2 \rangle$). Assume that t is a multiple of n for simplicity ($t = cn$), so the first c papers are known n times, giving $\langle k^2 \rangle = n/b$, and so a lower bound on μ is simply zero. Simulations show that the actual value of μ increases with b and n.

Since the probability that a r.c. agent knows idea j is simply k_j/n, we can write

$$\mu t = t - \sum_{l=1}^{w_t} k_l \frac{k_l}{n} = t - \frac{1}{n} \sum_{l=1}^{w_t} k_l^2$$

or, dividing both sides by t

$$\mu = 1 - \frac{w_t}{nt} \frac{1}{w_t} \sum_{l=1}^{w_t} k_l^2 = 1 - \frac{b}{n} \langle k^2 \rangle \tag{3}$$

where $\langle k^2 \rangle$ is the mean square degree. To be sure, notice that by definition of the attachment kernel, to ensure that exactly one paper will be learned if the current period event is "learning", the following must hold

$$\sum_{j=1}^{w_t} A_t(k_j) = 1 \tag{4}$$

As a consistency check, one can put equation 2 into equation 4, use the fact that the sum of all degrees is t, and solve for μ, which gives equation 3. Equation 3 shows a non trivial relationship between degree heterogeneity (or the variance of papers' popularity) and the normalization factor of the attachment kernel. This can be summarized in the following proposition.

Proposition 4.1. *The more the papers are heterogenous in terms of popularity, the larger the share of the total popularity of the papers that are known by a r.c. agent.*

Another way to understand proposition 4.1 is to ask: how many times can we find an agent who knows a paper that i knows? Proposition 4.1 tells us that this number increases if the variance of the popularity distribution increases.

It can be seen from equation 2 that when we consider a finite population, the attachment kernel is non linear. In fact, it is a polynomial. To the best of my knowledge, this case has not appeared in the literature, and finding exact analytical solutions is cumbersome. On the other hand, mean-field continuous approximations are relatively straightforward, and provide a good intuition. In addition, assuming a infinite population, we can get rid of the factor μ, and obtain particularly simple expressions. Thus, I start below with the simplest case (mean field continuous approximation of the infinite population case) and then I turn to the case of a finite population. More detailed analytical results, obtained by master equation methods, are mentioned when available and their derivation appears in the appendix; a summary of which results have been obtained and which ones are missing can be found in table 1. All cases are compared with simulations.

5 Infinite population: Simon's model

To get an expression for the distribution of papers' popularity and citations received, I used the method from Barabási and Albert (1999). It consists in writing a differential equation[21] for the evolution of a paper's popularity, and then using the fact that in expected terms there is a monotonic relationship between paper's index (its order of arrival) and paper's popularity. Regarding the distribution of

[21] This is only a crude approximation, because this assumes continuous time, and because it is deterministic.

		$n \to \infty$		$n = n_0$	
		Proposition	Equation	Proposition	Equation
Popularity	Continuous	5.2 A	8 *Power law*	6.2 A	15
	Discrete	5.2 B	9 *Yule-Simon*	6.2 B	16
Citations	Continuous	5.4 A	12 *Shifted power law*		
	Discrete	5.4 B	(13) *Waring*		

Table 1: **Summary of the results**. The table displays the number and name (if there is one) of the equation of the corresponding result. The number of the corresponding proposition is also reported. These equations are plotted in figure 2 against simulations. Equation 13 is put in parenthesis because it is not formally proven (see appendix D). Equation 13 is the distribution of citations which is used to test the model empirically (section 7).

citations received, it is possible to exploit the fact that we know the expected popularity of each paper at each point in time to derive the probability that a given paper is known, and thus cited, by a r.c. agent.

5.1 Distribution of popularity

Let us consider that there is a large number of agents ($n \to \infty$). In this limit, we can get rid of the non-linearity of equation 2, because one finds from equation 3 that $\mu \to 1$. In this case, the following proposition describes the evolution of the popularity of paper j born at time t_j

Proposition 5.1. *When $n \to \infty$, the continuous approximation of the popularity of paper j is given by*

$$k_j(t) = \left(\frac{t}{t_j}\right)^{1-b} \tag{5}$$

In discrete time, it is given by

$$k_j(t) = \frac{B(1-b,t_j)}{B(1-b,t)} \tag{6}$$

where $B()$ is the Beta function[22].

Proof. In the continuous approximation, from equations 1 and 2 we have

$$\frac{dk_j(t)}{dt} = (1-b)\frac{k_j(t)}{t} \tag{7}$$

Solving this differential equation, with the initial condition that the paper j is born at time t_j with degree one ($k_j(t_j) = 1$) gives 5. Equation 6 is found by solving the discrete version of 7, see appendix A $\square$

Clearly, the popularity of each paper is non decreasing with time. Thus popularity is unbounded. Remarkably - this is a key insight from Simon's paper-, this does not prevent the system to self-organize such that the distribution of paper's popularity is stable, and has a scale-free tail. Introducing $\hat{b} = \frac{1}{1-b}$, we have[23]

[22]The Beta function is defined in terms of the Gamma function: $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. The Gamma function generalizes the factorial function for non integer values, such that when x is an integer $\Gamma(x+1) = x\Gamma(x) = x!$, but x can also take non-integer values. The Beta and Gamma functions admit integral representations due to L. Euler.

[23]$\hat{b}$ can be interpreted in the same way as b since it varies in the same direction, but takes values on $[1,\infty)$ instead of $[0,1]$. It is interesting to notice that $\hat{b}$ is related to b by the formula $\hat{b} = \sum_{n=0}^{\infty} b^n$, which is the simple geometric series.

Proposition 5.2. *A) When $n \to \infty$, the continuous approximation of the distribution of papers' popularity is a power law*

$$p(k) = \hat{b}k^{-(1+\hat{b})} \tag{8}$$

B) Using a master equation method, the distribution of papers' popularity is the Yule-Simon distribution

$$p(k) = \hat{b}B(1+\hat{b}, k) \tag{9}$$

Proof. When $n \to \infty$ the calculations are the same as Simon's model. Equation 9 is the Yule-Simon distribution and is derived in appendix B. The continuous distribution is computed thus (using equation 5)

$$p(k_j \leq k) = p\left(\left(\frac{t}{t_j}\right)^{1-b} \leq k\right) = 1 - p\left(t_j \leq \frac{t}{k^{\frac{1}{1-b}}}\right)$$

Since the t_j's are uniformly distributed[24] their probability mass function is $prob(t_j = Y) = 1/t$ for Y from 1 to t, so $prob(t_j \leq Y) = \sum_1^Y \frac{1}{t} = \frac{Y}{t}$. This leads to $p(k_j \leq k) = 1 - k^{\frac{-1}{1-b}}$. Applying $p(k) = \frac{dp(k_j \leq k)}{dk}$ leads to the probability distribution of papers popularity, equation 8. It is easy to check that this is a proper distribution function, $\int_1^\infty p(k)dk = 1$ □

Equations 8 and 9 show how the innovation/diffusion trade-off determines the shape of the popularity distribution: the more innovation (i.e. the less diffusion), the more papers known only a few times (the less papers known many times). Equation 8 is plotted against simulation results in the left panel of figure 2, showing only limited agreement. Whereas the error for low degrees is due to the continuous approximation (this can be checked by comparison with the master equation solution, equation 9), the misfit of the tail comes from the assumption of an infinite population, which will be relaxed in section 6. But for the time being, let us turn to the derivation of the main result: the distribution of citations received.

5.2 Distribution of citations received

Denote by $p_t^C(k^C)$ the probability that a r.c. paper has already received k^C citations (has in-degree k^C) at time t. To derive this probability, let us start by expressing the in-degree of each paper j. What is the probability that paper j gets one new citation at time t? It is h times the probability that it gets a particular citation, and it gets a particular citation when several events are simultaneously realized. First, the r.c. agent is producing (probability b); Second, j is known by the r.c. agent ; Third, j is chosen. The joint probability that j is known and chosen is the probability that j is chosen conditional on being known, times the probability that it is known. At time t, the r.c. agent knows on average t/n

[24] Contrary to one mode scale free network models, this is not exactly true, since there is not one new paper per period, but only one at each period with probability b. The uniform distribution is, nevertheless, an appropriate approximation since the t_js of many independent realizations of the stochastic process are uniformly distributed over $[1, t]$. To see this, consider the original scale free network model (Barabási and Albert, 1999). There the t_js are simply the natural numbers, $1, 2, 3, ...t$, so they are obviously uniformly distributed on $[1, t]$. In the learning based model, however, there are some periods for which no paper is created. Assume for simplicity that $b = 0.5$, and the first simulation gives the following sequence of birth dates (omitting the first one, and until $t = 10$): $3, 5, 7, 10$. Another simulation may give $2, 4, 6, 8, 9$. This is a simple, "lucky" example to show that indeed each integer appears in a simulation with equal probability, $1/t$.

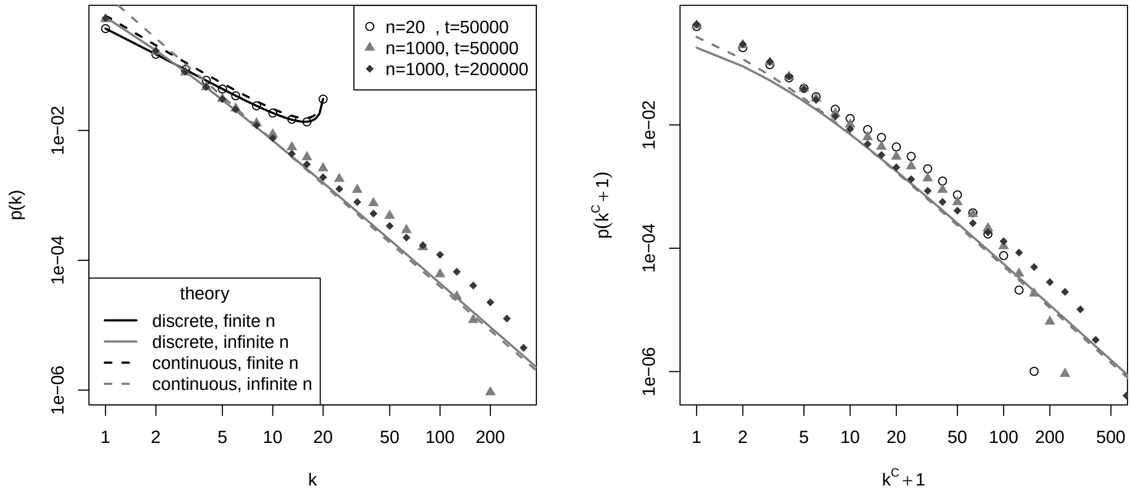

Figure 2: **Comparison of different approximations**. The simulation setup is $b = 0.2$, $h = 5$. The equations for finite n (i.e. $n = 20$) are plotted using the recorded value $\mu = 0.45$. 100 simulations runs are averaged out (only 60 for the case $n = 1000, t = 200000$), and only some points are shown to increase readability. The discrepancy between the points for large n (i.e. $n = 1000$) and the theoretical curves for infinite n can be reduced by increasing n and/or by running the process for longer times (for $n = 1000$, the process does not reach a steady state after only 50000 periods).

papers, so if she knows j, for each citation to be made j will be picked with probability $1/(t/n)$. The probability that j is known by the r.c. agent is easily found by its definition: since j is known k_j times by n agents, it is known by a r.c. agent with probability k_j/n. This leaves us with[25]

$$\frac{dk_j^C(t)}{dt} = bh\frac{1}{t/n}\frac{k_j(t)}{n} = bh\frac{k_j(t)}{t} \tag{10}$$

In fact, choosing randomly a paper among those known by an agent amounts to choosing each paper with a probability proportional to its popularity (recall that agents are similar, so on average they have the same papers' portfolio. So it is like sampling uniformly at random one link of the two-mode network). Since $k_j(t)$ is given by equation 5, we can substitute it into equation 10 and solve the differential equation with the initial condition that $k_j^C(t_j) = 0$ to obtain

Proposition 5.3. *When $n \to \infty$, the continuous approximation of the number of citations received by paper j is given by*

$$k_j^C(t) = \hat{b}bh\left(\left(\frac{t}{t_j}\right)^{1-b} - 1\right) \tag{11}$$

Equation 11 and the same method than for proposition 5.2A leads to:

Proposition 5.4. *A) When $n \to \infty$, the continuous approximation of the number of citations received*

[25]As is usual in the literature, I do not explicitly take into account that a paper is never cited more than once by a newly arriving paper (a draw without replacement is approximated by one with replacement). This is reasonable because the probability that a paper would be chosen twice in a row tends to zero as the number of papers in the network tends to infinity. As a footnote, one may notice that in reality some papers do cite other papers several times- just as this paper cites Simon (1955) many times - but they appear only once in the reference list.

is a shifted power law

$$p^C(k^C) = \hat{b}(\hat{b}bh)^{\hat{b}} \left(k^C + \hat{b}bh\right)^{-(1+\hat{b})} \tag{12}$$

B) A discrete version of (12) *is*

$$p^C(k^C) = \frac{B(k^C + \hat{b}bh, 1 + \hat{b})}{B(\hat{b}bh, \hat{b})} \tag{13}$$

The proof of equation 12 is similar to that of equation 8 and is omitted. Equation 13 is derived in appendix D by comparing equation 12 with the mean-field continuous approximation of a model for which the exact solution is known (Price's model). Again, it holds for equation 12 that $\int_0^\infty p^C(k^C)dk^C = 1$. Equation 12 deserves two remarks.

First, note that k^C does not appear alone with its exponent, but together with a constant (denoted $r = \hat{b}bh$). This form has been called a shifted power law (or Zipf-Mandelbrot's law) in the literature, and its discrete equivalent equation 13 is sometimes called the Waring distribution (Irwin (1963) and e.g. Golosovsky and Solomon, 2012). This constant creates a "shoulder" on the left of the distribution, but does not change the tails much since it converges to the classical power law for large k (i.e. for $k \gg r$). Interestingly, some recent publications have convincingly fit this form to appropriately constructed datasets (Newman, 2009; Eom and Fortunato, 2011; Golosovsky and Solomon, 2012).

Second, the exponent is the same as that of the distribution of papers' popularity (equation 8), and is independent of h (the mean in- or out-degree). This last fact is crucial, because Price's model, which also predicts a shifted power law (appendix D), features a negative relationship between h and the power law exponent. I shall come back to this when discussing empirical matters. Before that, let me present more rigorous results on the distribution of popularity.

6 Distribution of popularity in finite population

To facilitate the derivations above, I have assumed that $n \to \infty$. When the population is infinite, the probability that an agent is r.c tends to zero, and we arrive at the mathematically useful but somewhat empty result that, at a given time step, agents tend to have learned a number of papers that tends to zero. This avoided a mathematically annoying but theoretically fundamental feature: if agents cannot learn twice the same paper, the paper that they choose to learn at any given period depends on its popularity relative to the papers that the agent does not already know. In other words, the normalization factor μt in equation 1 cannot be reduced to t (recall that μt is the total number of degrees of the papers unknown by a r.c. agent). Inserting the quadratic attachment kernel in (1), the dynamics of individual paper's popularity is now given by a non autonomous logistic differential equation. Its solution gives the following proposition

Proposition 6.1. *When the size of the population is finite, the continuous approximation of the popularity of paper j is given by*

$$k_j(t) = n \left[1 + (n-1)\left(\frac{t_j}{t}\right)^{\frac{1-b}{\mu}}\right]^{-1} \tag{14}$$

We can easily see that for $n \to \infty$ (also implying $\mu \to 1$), we recover equation 5. As in the infinite population case, we can use the expected value of the popularity of paper j (14) to find the continuous approximation of the popularity distribution:

Proposition 6.2. *A) If n is finite, the continuous approximation of the distribution of papers popularity is given by*

$$p(k) = \hat{b}n\mu(n-1)^{-\hat{b}\mu}\ (n-k)^{-1+\hat{b}\mu}\ k^{-1-\hat{b}\mu} \tag{15}$$

B) Using the master equation method, the distribution of papers popularity is given by

$$p(k) = p(1)\frac{(1)_{k-1}(1-n)_{k-1}}{(-r_1)_{k-1}(-r_2)_{k-1}} \tag{16}$$

where

$$\{r_1, r_2\} = \frac{n-4 \pm \sqrt{n(n+4\hat{b}\mu)}}{2}$$

and

$$p(1) = \frac{n\mu}{b+n-bn+n\mu-1} \tag{17}$$

Equation 16 is proven in appendix C and $(x)_y$ is the Pochhammer symbol.[26] The usual check for equation 15 is that $\int_1^n p(k)dk = 1$, and by taking the limit for $n \to \infty$ and setting $\mu = 1$ we recover equation 8.

Equation 15 is plotted against simulations in the left panel of figure 2, showing good agreement, although the last value ($k = n$) cannot be plotted. Equation 16 shows excellent agreement. The most striking feature of these equations is that, contrary to all other power-law like distributions encountered here, it is not monotonically decreasing. Towards the end of the distribution, it goes up. This means that the number of papers known n times is higher than the number of papers known $n-1$ times. This effect might be relevant in all contexts where one set of nodes of a two-mode network growing by preferential attachment is fixed or small as compared to the other set (that is, growing much more slowly).[27] The finiteness of n implies that papers cannot diffuse infinitely, so the bins of the histogram for high k keep getting papers from the lower bins without having so many papers leaving them towards higher bins. Remarkably, this phenomenon does not prevent the distribution from featuring a steady-state.

I couldn't find analytically the effect of population finiteness on the distribution of citations received. Simulations reveal that small populations create a cut-off on the citation distribution, because the most

[26] Sometimes called the rising factorial, it is defined as

$$(x)_y = x(x+1)(x+2)\dots(x+y-1) = \frac{\Gamma(x+y)}{\Gamma(x)} \tag{18}$$

[27] Interestingly, a similar phenomenon was found by Peruani et al. (2007) on the degree distribution of the *fixed* set of nodes, in a growing two-mode network with mixed (random and preferential) attachment, and a high value of the parameter tuning the relative amount of preferential versus random attachment. In their case, the distribution is a beta distribution. The model of Evans and Plato (2008), which is a fixed two-mode network with rewiring, produces a U-shaped distribution too, when the relative amount of preferential v.s. random attachment is high.

popular papers are not as popular as they would be in a large population, and thus are less cited than they would be in a larger population (see the right panel of figure 2)[28]

In what follows, it will be assumed that the difference between the citation distributions of finite and infinite population is not too important, such that equation 13 is approximately valid for a context such as those generally encountered in empirical work, where $n \in [10^2; 10^5]$ and $t \in [10^3; 10^7]$.

7 A quick look at the data

In this section, the parameters of the learning-based and of Price's model are estimated for two citation networks of high energy physics papers, and related empirical work is discussed.

7.1 Comparing Price's and the learning-based model

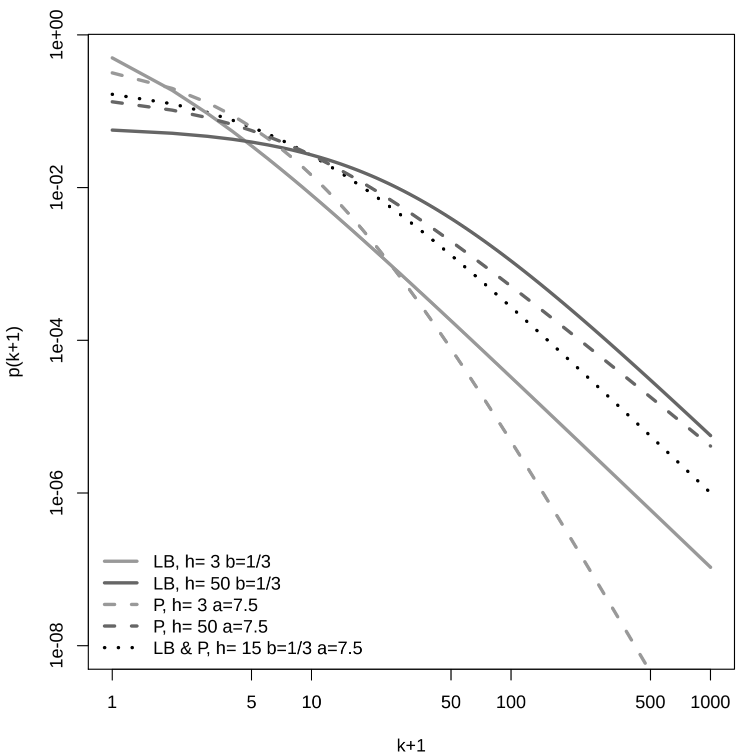

Figure 3: **Effect of h on the shape of citation distribution**. The figure shows theoretical predictions according to Price's (P) and to the Learning Based (LB) models. The curves are calculated using the values in the legend. The base case, in black, shows a Waring distribution with $\gamma = 2.5$ and $r = 7.5$, which, assuming $h = 15$, corresponds both to the LB model with $b = 1/3$ and the P model with $a = 7.5$. The other curves depict what the two models would predict under a change of h, *ceteris paribus*, that is, with a constant in the P model and with b constant in the LB model. The slope of the tail is invariant with h in the LB model, whereas it is not in the P model.

[28]This cut-off does not seem to be in agreement with recent empirical evidence. Golosovsky and Solomon (2012) found that in very large datasets, the tail of the citation distribution is "superheavy", exhibiting a "runaway": the most cited papers are even more cited than what linear preferential attachment would predict.

Both models predict that the citation distribution is a Waring distribution (which could also be called discrete shifted power law)

$$p(k) = \frac{B(k+r,\gamma)}{B(r,\gamma-1)} \tag{19}$$

where the superscript C used previously to distinguish popularity and citation distributions is omitted from now on. By assuming that the data is distributed according to this law, it is possible to obtain the empirical values of the parameters, r and γ, by maximum likelihood.[29]

To see what are the similarities and differences between the learning-based and Price's model, let us start by noting that from equation 19 we find that the first moment of the distribution is $\langle k \rangle = \sum_{k=0}^{\infty} k \frac{B(k+r,\gamma)}{B(r,\gamma-1)} = \frac{r}{\gamma-2}$, which is also valid with the continuous form, i.e. $\int_0^\infty k(\gamma-1)r^{(\gamma-1)}(k+r)^{-\gamma} = r/(\gamma-2)$. In both the learning-based and Price's model, by definition, we add h edges each time that we add a node, so the mean in- or out-degree is h, hence

$$\langle k \rangle = h = \frac{r}{(\gamma-2)} \tag{20}$$

Now suppose that h increases. This can be compensated on the RHS of (20) by either an increase of r or a decrease of γ, or a mix of both. Remarkably, the learning based model predicts that an increase of h leads to an increase of r but keeps γ constant, whereas Price's model predicts that an increase of h leads to a decrease of γ, but keeps r constant. This can be seen by comparing the general form (19) with the equation from the model for $n \to \infty$ (13). We have

$$\gamma = 2 + \frac{b}{1-b} \tag{21}$$

which is independent of h and

$$r = \frac{b}{1-b}h \tag{22}$$

which depends linearly on h. In Price's model (see appendix D) the in-degree distribution is equation D.1, so we have

$$\gamma = 2 + \frac{a}{h} \tag{23}$$

which decreases and converges to 2 when $h \to \infty$ and, the shoulder parameter is just equal to initial attractiveness:

$$r = a \tag{24}$$

[29]Other approaches are possible. One can estimate the shifted power law equation 12, but it is not a discrete law, and we would have either to discretize it, by finding the normalization factor that makes it *summing up* (not integrate) to one (this would be the inverse of a Hurwitz zeta function), or to accept that we fit a continuous law to discrete data. One could estimate the distribution assuming that it is valid only after a certain value k_{min}, but that would not qualify as a fit of the model, which predicts that *all* values are drawn from a Waring. Yet another approach is that of Jackson and Rogers (2007), who also fit a shifted power law, and who input the mean in-degree as the true value of h, and derive their second parameter by iterated least squares. The approach followed here *predicts* a value of h from the estimated parameters of the distribution (see table 2). When fitting the distributions of the networks used to construct the right panel of figure 2, disappointing results are obtained in the case $n = 20, t = 50000$ where the Maximum Likelihood Estimation (MLE) gives $b = 0.13 \neq 0.2$ and $h = 12 \neq 5$. Much more encouraging results are found in the case $n = 1000, t = 200000$, with $b = 0.21 \approx 0.2$ and $h = 6 \approx 5$. This is quite good taking into account that $n = 1000$ is not that large, and so the distribution is actually not a Waring but a Waring with a cut-off. I conclude that the MLE is indeed reliable. Clearly, the research needed on these estimation issues is out of the scope of this paper.

The difference between the two models in terms of the predicted change of the distribution when h changes can be seen in figure 3. In the learning-based model, h is increasing the shoulder (the flat part on the left is flat until a higher value of k) but does not impact the exponent (the slope of the curve on the rightmost part does not change), whereas the reverse is true for Price's model.[30].

Once we have estimated γ and r, equations 21 and 22 form a system of two equations with two unknowns, b (the learning/writing trade-off) and h. In the case of Price's model, the system (23)-(24) has two unknowns, a (the initial attractiveness) and h. The first empirical exercise is thus to estimate these parameters. Second, since we know that h differs across fields, and that within fields h has increased, it is possible to discriminate between the two models using both cross-section and time-series evidence. This task will be only superficially attempted here, because of the lack of appropriate freely available data. Nevertheless, it is possible to discuss the differences between the two models using the estimates provided below on two citation networks from *arXiv*, and using heuristic arguments to interpret figures published in the literature that do not correspond to an exact test of the models but may be close to it.

7.2 The data

Two publicly available datasets[31] are analysed below. Both are citation networks of a subset of *arXiv* publications: High Energy Physics Theory (HepTh) and High Energy Physics Phenomenology (HepPh)[32]. They cover almost completely the history of these subsets of *arXiv*, until 2003. In both cases, a citation network is constructed for each year, where year T network includes all the papers published on *arXiv* before December 31st of year T.[33] Descriptive statistics can be found in the leading columns of table 2.

7.3 Results

The objective of this section is to obtain values for the parameters r and γ, which can be done by maximum likelihood. In practice, a Nelder-Mead algorithm was used to maximize the logarithm of the

[30] The distribution predicted by the learning-based model is exactly the same as that predicted by Jackson and Rogers (2007) and Peterson et al. (2010). However, the interpretation of the parameters differ. What is the ratio of random over preferential meetings (i.e. citations) in their case in the ratio of producing over learning papers here. The intuition behind this connection is that citation dynamics is directly correlated to popularity dynamics, and in the popularity dynamics learning is done "preferentially" whereas writing papers corresponds to a random attachment. See also appendix F.

[31] Downloaded from http://snap.stanford.edu/data/index.html#citnets. Credit is due to Gehrke et al. (2003) and Leskovec et al. (2005).

[32] There is an overlap between the two. *arXiv* is a server where scientists, mainly physicists, can freely post pre- and post-prints of their work. See *arXiv.org*.

[33] An important problem arising if one tries to test the prediction of the learning-based or Price's model is that they model the "complete" system. In other words, the distribution that they predict includes all papers and all citations to those papers. Most available studies use "cohort" datasets. For instance, Golosovsky and Solomon (2012) use physics papers published between 1980 and 1989 and cited by 2008, omitting all the papers published between 1990 and 2008 and all the citations they make and receive. Note that in his own empirical test of Price's model, Mark Newman constructs the network by taking all articles citing five key papers, which is much closer to what the model actually models. Another way to think about this question is to realize that in both Price's and the learning based model, no citation comes from outside the network or is made to a paper outside the network (the number of references is equal to the number of citations). An ideal test of these models would require data on science as a whole, which still omits citations to and references by non scientific literature.

1	2	3	4	5	6	7	8	9	10
Year	Number of edges	Number of nodes	Empirical in/out degree	Predicted in/out degree	Minimum degree for PL fit	PL exponent	"Shoulder" param. for shifted PL	Shifted PL exponent	Estimated innovation probability
T		w_t	$\langle k \rangle$	h	k_{min}	γ_{PL}	r	γ_{SPL}	b
High Energy Physics - Theory									
1994	12879	4924	2.62	2.81	6	2.53	2.79	2.99	0.50
1995	28131	7078	3.97	4.31	24	3.11	3.65	2.85	0.46
1996	53091	9541	5.56	6.06	25	2.79	4.37	2.72	0.42
1997	87713	12124	7.23	8.08	28	2.63	4.80	2.59	0.37
1998	125163	14766	8.48	9.75	29	2.57	5.11	2.52	0.34
1999	167110	17485	9.56	10.99	36	2.61	5.66	2.52	0.34
2000	217790	20523	10.61	12.26	58	2.77	6.27	2.51	0.34
2001	271863	23587	11.53	13.30	100	2.98	6.90	2.52	0.34
2002	333973	26792	12.47	14.45	62	2.75	7.53	2.52	0.34
2003	352807	27770	12.70	14.83	55	2.70	7.60	2.51	0.34
High Energy Physics - Phenomenology									
1994	11387	4423	2.57	3.29	22	3.56	1.68	2.51	0.34
1995	29802	7276	4.10	4.91	20	3.04	3.05	2.62	0.38
1996	58904	10478	5.62	6.86	16	2.74	4.07	2.59	0.37
1997	98307	14013	7.02	8.44	25	2.94	5.19	2.61	0.38
1998	142934	17736	8.06	9.80	76	3.74	5.82	2.59	0.37
1999	201300	21739	9.26	11.54	56	3.37	6.34	2.55	0.35
2000	263313	25764	10.22	12.90	64	3.38	6.75	2.52	0.34
2001	333942	29877	11.18	14.17	72	3.40	7.21	2.51	0.34
2002	411527	34014	12.10	15.42	94	3.49	7.66	2.50	0.33
2003	421578	34546	12.20	15.61	94	3.47	7.69	2.49	0.33

Table 2: **High Energy Physics Theory and Phenomenology citation network statistics**. The fourth column gives the empirical average in-degree. Columns 8 and 9 give the estimated parameters of a Waring distribution. The fifth column gives the average in-degree as predicted by the model, using the estimated parameters from columns 8 and 9 and equations 21 and 22. The last column gives b as computed using column 9 and equation 21. Columns 6 and 7 give the parameters estimated under the assumption of a Power Law (PL). See the main text for additional details.

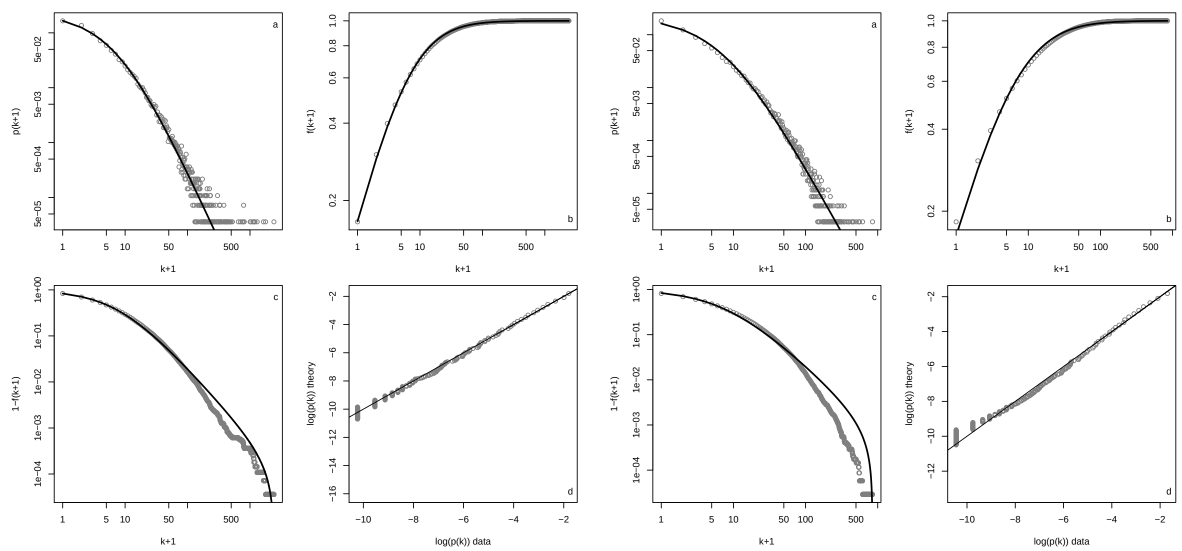

Figure 4: **Maximum likelihood fit for the citation distributions of the Hep-Th (left) and Hep-Ph (right) networks** at the end of 2003. In panels a, b and c, the grey points are empirical values, and the black lines are the theoretical results constructed by using equation 13 with values of the fitted parameters (columns 8 and 9 in table 2) from $k = 0$ to the maximum observed degree, and renormalizing such that it sums up to one. All subpanels of a given panel display the same raw data. a) Probability distribution. b) Cumulative distribution. c) Counter cumulative distribution. Panel d) is a Quantile-Quantile plot, where the black line materializes perfect fit.

likelihood function associated to equation 13. For comparison purposes, the fit of a power law is also added.[34]

Results are reported in table 2. The sixth and seventh columns show the estimated parameters assuming a power law (for the tail), and the eighth and ninth columns give the estimated parameters assuming a shifted power law (for the whole distribution). Early networks (1992-93) are not reported. The resulting look of the predicted laws can be seen in figure 4. The last column gives the corresponding value of b, computed using the estimated shifted power law exponents and equation 21. The fourth and fifth columns compare the empirical mean in-degree (equal to the mean out-degree) and the value of h deduced from the estimated shifted power law parameters and the system 21-22 (or, equivalently, 23-24).

First, it seems that the fits are reasonably good, although a deviation can be observed when the counter cumulative distribution is plotted, especially for the HepPh network (subpanel c of the right panel of figure 4). Second, one sees that the predicted mean in-degree is close to the observed one (columns 4 and 5 of table 2). Finally, the shifted power law exponent is strikingly stable over time and across the two networks, and imply $b \approx 1/3$. If one is to take this exercise seriously, it says that high energy physicists spend a third of their time working on their own paper and two thirds learning the work of others.

[34]The power law fit was done using the procedure and code of Clauset et al. (2009) (translated to R by Laurent Dubroca)

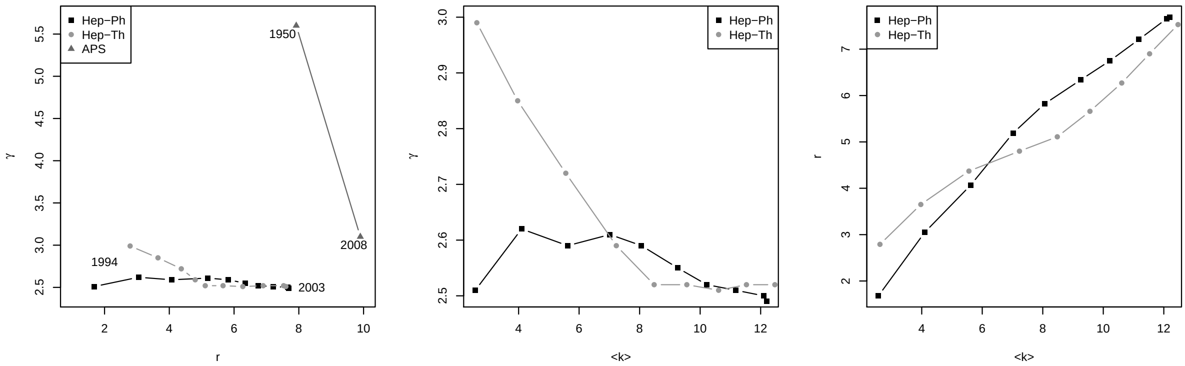

Figure 5: **Temporal relationship between γ and h.** The data is taken from table 2, columns 8 and 9 (left panel), 4 and 9 (middle) and 4 and 8 (right). On the left panel, the two points available from Eom and Fortunato's (2011) study of the American Physical Society data are added. On the left, a flat curve favors the LB model and a horizontal one favors the P model. On the middle panel, a flat curve favors the LB model and a decreasing ones favors the P model. On the right panel, a flat curve favors the P model whereas. an increasing curve favors the LB model.

7.4 Other results from the literature

One can also compare the estimates above with those reported in the literature. Jackson and Rogers (2007) estimated a shifted power law on a dataset containing all papers which either cite Milgram's small world paper or contain the phrase "small world". They found $\gamma = 2.63$, which imply $b = 0.39$, and report a mean degree $h = 5$. Tsallis and De Albuquerque (2000) estimated the Tsallis distribution, from which we can deduce the shifted power law parameters. For the Physical Review D dataset originally studied by Redner (1998), that includes citations to papers published in this journal between 1975 and 1994, one finds from Tsallis's parameters that $\gamma = 2.56$ and $r = 12.02$, implying $b = 0.36$ and $h = 21.46$. Peterson et al. (2010) estimated a Waring distribution on the same data[35] and found $\gamma = 3.1$ and $r = 31$ leading to $b = 0.52$ and $h = 28.18$ (using formula 20) or $h = 27$ (as reported in their table 1). Another dataset of Redner, including citations to a cohort (1981) of papers (ISI data), gives from Tsallis and De Albuquerque's (2000)'s estimates $\gamma = 2.89$ and $r = 17.15$ (i.e. $b = 0.36$ and $h = 21.46$) and from Peterson et al.'s (2010) estimates $\gamma = 3.2$ and $r = 25$ (i.e. $b = 0.55$ and $h = 20.83$ or 17.3 from their paper). Peterson et al. (2010) analyze another dataset, made of publications from high h-index chemists, and find $\gamma = 2.935$ and $r = 37$ (i.e. $b = 0.48$ and $h = 39.57$ or 42 from their paper). Newman (2009) constructed a dataset with the 2407 papers citing five landmark publications of network science, and reported the estimated coefficients $r = 6.38$ and $\gamma = 2.28$, from which one finds $b = 0.22$ and $h = 23$. Golosovsky and Solomon (2012), with a dataset of 418,438 physics papers that were published during 1980-89 and their citations by July 2008, found $r = 10.2$ and $\gamma = 3.15$, from which one deduces $b = 0.53$ and $h = 11.73$. The estimates reported in table 2 for the last year of the High Energy Physics citation networks nicely fit between those results.[36]

The differentiated prediction of the two models regarding the effect of h can be tested both with time

[35] Not exactly the same data since the dataset has been corrected, see Sidney Redner's website.

[36] Atalay (2011) also fits a Waring distribution to the same data. His table 1 indicates $r = 14.82$ and $\gamma = 2.513$ for the last year of the Hep-Th network in excellent agreement with the last line, columns 5 and 9 of table 2 in this paper.

series and cross-section data, a task which is out of the scope of this paper. However, some heuristic arguments can be given for both cases, using recently published figures. Time-series and cross-sections are discussed in order.

It is known that h, understood here as the average bibliography size, has increased over time. For instance, Biglu (2008), using the whole ISI data, reports a value of h increasing from 8.40 to 34.63 between 1970 and 2005, This trend of h implies that, if we had a long dataset (say, 50 years), and with reliable values of h, we could try to discriminate between the two models.[37] I do not have access to such a dataset, but, with a number of necessary cautions, it is possible to comment on the one analyzed by Eom and Fortunato (2011). They use a dataset constructed as a growing network, as is needed here. The dataset is large and covers a very long period: 414,977 papers published between 1893 and 2008 in American Physical Society journals. They compare the fits given by the lognormal, the shifted power law, and the power law. They found that the shifted power law provides a better fit than a normal power law. However, because their purpose is to compare the fits given by different hypothesis, they do not fit the whole distribution, but only the part with values of k greater than k_{min}^{SPL}, where k_{min}^{SPL} is chosen so as to maximize the Kolmogorov-Smirnov statistics, as was proposed for power laws by Clauset et al. (2009). Eom and Fortunato (2011) reported the estimated shifted power law coefficient at in 1950 and at the end of the period, $\gamma_{1950} = 5.6$ and $\gamma_{2008} = 3.1$. Since they also report the average degree $\langle k \rangle_{1950} = 2.2$ and $\langle k \rangle_{1950} = 9$, we may use equation 20 to find $r_{1950} = 7.92$ and $r_{2008} = 9.9$. Hence, the increase of h leads both to a decrease of γ, as predicted by Price's model (assuming constant initial attractiveness), and to an increase of r, as predicted by the learning-based model (assuming constant b). In the *arXiv* datasets (table 2), except for early years of the Hep-Th network, the increasing in-degree is associated to a stable exponent, and an increasing shoulder parameter, which tends to support the learning-based over Price's model. It should be noted that Eom and Fortunato (2011) fit the shifted power law only after a certain value k_{min}^{SPL}, whereas the fits reported here are for the full distribution. Also, they fit a shifted power law whereas the fits here are done with a Waring distribution. Overall, time series evidence does not provide firm conclusions regarding the superiority of Price's or the learning-based model (see figure 5).

On the other hand, recently published figures allow us to look at the relationship between h and γ in cross-sections. This is also theoretically more attractive: disciplines with different citation practices are expected to have a different shape of the citation distribution. Albarrán et al. (2011) have studied 3.7 millions articles and their citations received after 5 years. They split the articles in different subcategories and reported the mean number of references made, $\langle k_{out} \rangle$, and the mean number of citations received, $\langle k_{in} \rangle$, both of which can be interpreted as h. They then fit for each subfield a power law and reported k_{min}, and γ_{PL}. We can use this evidence as follows.

[37] One issue in this exercise is that the models lead to steady-state distributions, and it is not totally clear that if h changes over time the exponent changes accordingly and fast enough for its effect being detectable in the data. In other words, it is not really appropriate to do comparative statics on non-equilibrium systems. In the complex network literature, the fact that growing networks feature an increasing average degree is known as *network acceleration* (Dorogovtsev and Mendes, 2001), *densification* (Leskovec et al., 2005), or *log-networks* (Krapivsky and Redner, 2005). Fitting the citation dynamic process would preferably be done by endogeneizing h in the theoretical model, solving for the non-equilibrium degree distribution (Cooper and Prałat, 2010) and fitting finite-time expressions. I refrain from going into such complications in this paper, where the main purpose is to add one layer (social embeddedness of learning) on the backbone model of Price to make a theoretical point.

First, as above, the learning-based model roughly predicts a positive relationship between h and k_{min} that Price's model does not. The bottom panels of figure 6 seems to give more credit to the learning-based model, especially if we consider h to be proxied by the *in*-degree (bottom right panel).

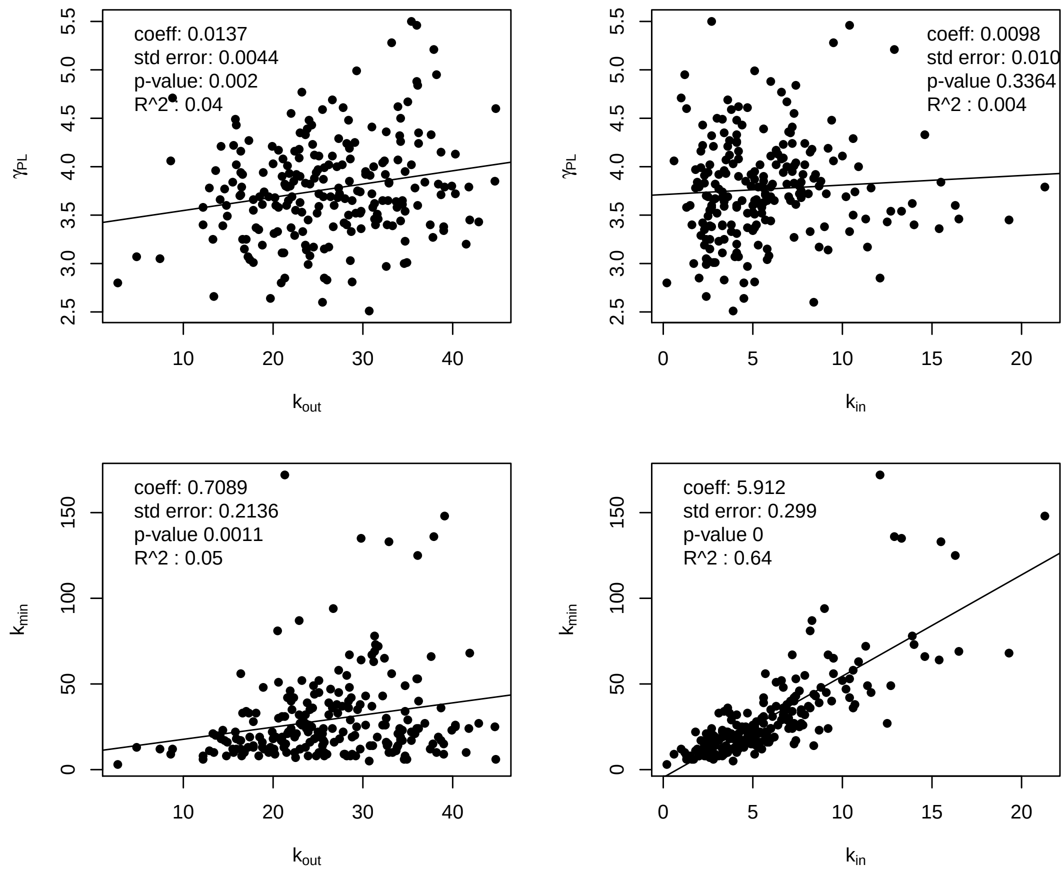

Figure 6: **Cross section correlations between the parameters** of a non pure power law estimation (γ_{PL} and k_{min}) and the mean number of citations received ($\langle k_{in} \rangle$) and mean number of references made ($\langle k_{out} \rangle$) (which are different because this dataset is not constructed as a network). All results are computed using table B p.71 and E1 p.99 in Albarrán et al. (2011). Each value corresponds to the estimation for a single Thomson Scientific sub-fields. See their paper for details. Based on intuitive heuristic arguments, one may regard the correlation between γ_{PL} and $\langle k \rangle$ (upper panels) as indicative of the correlation between γ_{SPL} and h, and the correlation between k_{min} and $\langle k \rangle$ (lower panels) as indicative of the correlation between r and h. This provides thus a direct way to discriminate between the Learning Based (LB) and Price's (P) models, because the LB predicts no correlation in the upper panels, and positive correlation on the lower panels, whereas the P model predicts a negative correlation in the upper panels, and no correlation on the lower panels. See the main text for a more detailed discussion.

Second, if we consider that $\gamma_{PL} \approx \gamma_{SPL}$,[38] we can test if the negative relationship between h and γ predicted by Price's model holds, or if there is no relation at all, as expected from the learning-based

[38]In fact only a milder assumption is needed, that is, for the interpretation to hold, it is needed that the derivatives of γ_{PL} and γ_{SPL} with respect to h have the same sign and a comparable order of magnitude. Even this is a strong assumption, as one can observe, for instance, that the Pearson correlation between the seventh and ninth column of table 2 is only 0.33 (Hep-Ph) and 0.08 (Hep-Th). Still, the strong theoretical link between γ_{PL} and γ_{SPL} legitimates the discussion of this subsection.

model. The top panels of figure 6 indicate again weak relationship, but, if any, it is positive. Surely, it does not validate the learning-based model, but it suggests that the negative dependence of the power law exponent on h in Price's model may not be a good feature.

It should be emphasized that in both models, the two parameters are assumed to be independent, but in the real world they could be correlated. For instance, considering the parameters of Price's model, one can think that the initial attractiveness of an article is linked to h, because an article is more likely to be effectively cited if many citations are to be distributed (a positive relation between a and h). As far as the parameters of the learning-based model are concerned, we can think that in a field where the learning rate is very high, agents will know more but produce fewer papers, so they might want to put more references in their papers (a negative relationship between b and h). Moreover, the fact that the models, by construction, have a mean in-degree equal to the mean out-degree may cause confusion for empirical work and its interpretation. Furthermore, both models take h as fixed, whereas it is well known that the distribution of the number of references per paper is far from being nicely peaked around its mean value. On the contrary, it is right skewed, although less than the distribution of citations (Albarrán and Ruiz-Castillo, 2011).

The discussion of this subsection is, to say the least, rather acrobatic. Power law and shifted power law exponents are clearly not the same thing, nor are k_{min} and r. However, it does not seem too unreasonable to conclude that the negative relationship between the shifted power law exponent and the mean in/out-degree predicted by Price's model should be questioned. If it is falsified, the learning based model appears as a natural candidate for explaining a shifted power law distribution but with an effect of the average in/out-degree on the shoulder, and not on the tail. That being said, the model is above all a theoretical one, and should not, in this form, be taken too seriously for empirical purposes. The following section discusses these and other limitations.

8 Limitations

Several aspects of the model, its analysis, and the empirical discussion deserve comments.

A classical critique to this type of model is that the monotonic relationship between age and expected (in)degree is not in line with reality. In fact, Adamic and Huberman (2000) disputed the original scale-free network model on the grounds that the world wide web does not feature such a relationship. A first answer is that the Adamic and Huberman's (2000) critique does not decrease the merit of the Barabási and Albert's (1999) model as a simple, canonical model capturing the essence of a phenomena while omitting perhaps important but not essential features, at least at a first step of the analysis. Second, Newman (2009) has argued that in the case of scientific publications, it does exist a "first-mover advantage". In spite of this fact, it would probably be exaggerative to say that no papers are better than others, and that only age and chance matter to garner citations. I conclude that the introduction of aging (as in Dorogovtsev and Mendes, 2000), burst dynamics (Ratkiewicz et al., 2010) and paper intrinsic value (as in Bianconi and Barabási, 2001) are important avenues towards a more realistic version of the model. In the same vein, one might refine the social side of the network dynamics, by assuming more realistic social network structure (see appendix G), or e.g., introducing specific processes for authorship and co-authorship dynamics.

Another extension would be to take into account the mesoscale, nearly-decomposable structure of science. In fact, growing two-mode networks were historically designed for a similar purpose. Although Yule was interested in the sizes of biological genera and did not think of his model as a network, applying Yule processes on a set of items and a set of categories produces a Yule-Simon distribution of knowledge categories' sizes.

Regarding the mathematical analysis, several shortcomings must be acknowledged. First, many results are derived using the mean-field continuous approximation, but this method does not always give very good results. The master equation method (sometimes called rate equation when written in continuous time) (Simon, 1955; Krapivsky et al., 2000; Dorogovtsev et al., 2000) provides a better approximation (as was shown for the distribution of popularity), but could not be used directly for the case of citation distributions. Moreover, the effect of population finiteness could not be analytically derived for the citation distribution. Finally, a rigorous analysis of the initial conditions has not been attempted. A more complete analysis, along the lines of Bollobás et al. (2001), should reveal the exact domain of validity of the results. I would like to stress here that there have been several controversies surrounding the mathematics of this type of system, thus, at best, I claim only reasonable use of generally admitted methods, certainly not exactness (some results may be exact under certain well defined limits, but it would require some work to prove it).

Besides the shortcomings of what has been done, one can comment on the lack of analysis of other properties of the model. Here, the focus was only on the degree distributions. There are many other topological quantities in networks that are informative and can be modeled and compared with empirical data, such as clustering, path length, assortativity, degree correlations, different measures of centrality, distribution of component sizes, etc. These quantities, well defined for one-mode undirected networks, can be more complicated to define and model on the two-mode and the directed networks that are studied here. As such, they have been omitted here but remain important avenues for future investigations.

Turning to the empirical validation, several issues are discussed in order. First, it is hard to test the theory because in practice we do not have data on b. How many papers are read by a scientist between two of her publications? Is it really constant over the career cycle? Does it differ from field to field? Is it constant over time? Jones (2009) has shown that, because of knowledge accumulation, it is harder and harder to reach the scientific frontier (the "burden of knowledge" effect). One might think that this implies a decrease of b (more papers need to be read before one can publish). Alternatively, one may argue that it is true that it is harder to reach the frontier, but conditional on being at the frontier, the number of papers to be read before one is able to publish does not depend on the amount of previously accumulated knowledge. Additional work is needed to clarify the empirical content of b. As far as the other parameter is concerned, the observed increase of h can be thought of as a natural consequence of the growth of the knowledge base (the more papers there are, the more a new paper needs to cite previous papers). In fact, it is possible to endogenize h, and similar power laws are obtained in finite time, but the system is significantly more complicated and has no steady state. Therefore I have chosen to study steady states, and discuss the effects of parameters as one generally does in comparative statics. I believe that this is a useful simplification to make the theoretical point that is the purpose of this paper. Nevertheless, a more realistic model would necessarily include network acceleration. In addition, the comparison between the learning-based and Price's models may be altered if acceleration is included in both.

Further effort in modelling this type of system should in any case help designing empirical investigations. For instance, many empirical papers study a cohort of papers (published in one single year) and their citations received during a certain time window (e.g. the ISI dataset in Redner, 1998), or during a time window which is expected to be long enough such there will be no more citations to those papers (e.g. Stringer et al., 2010). The study of Albarrán and Ruiz-Castillo (2011) take the papers published during a certain time span (1992-1997) and the citations they receive during a certain time window (1992-2007). A third type of studies take the whole dataset (e.g. Eom and Fortunato (2011) and this paper). In fact, there is no reason why the distribution within a cohort should be the same as the distribution for the whole dataset. As a matter of fact, cohort distributions are much less skewed for cohort data than in whole datasets. For instance Stringer et al. (2010) found that a discrete log-normal fits well their cohort data. while Newman (2009), Eom and Fortunato (2011), and this paper argue that the shifted power law works well for whole datasets. Ideally, one would model and test both distributions.

The comparison of the learning-based and Price's models with data should not, in my view, be taken too seriously. Many other very important factors shape the structure of citation network, which also implies that we do not know exactly what is captured under the label of "initial attractiveness" or "relative share of learning/producing" when comparing the models with data. Both models privileged parsimony and analytical tractability over detailed empirical content. The point of Price was that papers benefit from a cumulative advantage: the more they are cited, the more they are likely to be cited again, a prediction confirmed empirically to a good extent by Newman (2009). My point is that *socially embedded learning is one of the plausible mechanisms behind the cumulative advantage driving the self-organization of science* and its materialization into a citation network. I intended to shed light on the humans in the shadow of citation networks, advocating a view of knowledge items networks (scientific publications or others, linked by citation or another relational feature) more systematically based on the human agency underpinning them, being it related to rational individual decision making, psychological aspects of attentional behavior, or social embeddedness. For instance, in a stand-alone citation network model such as Price's, the decision-making process is about "what to cite", whereas in the learning-based it is firstly about "what to learn", creating space for modelling search explicitly. A model encompassing social and bibliographic search would thus be a natural follow-up of the research presented here.

9 Conclusion

In this paper, it was argued that the cumulative advantage enjoyed by scientific publication is a consequence of the social embeddedness of learning. It was shown that growth (creation of new papers) and preferential attachment (learning of popular papers) generates power law-like distributions for papers' popularity and papers' number of citations received. The trade-off between the production and the consumption of papers is the key parameter which was estimated using data on citation networks.

The model presented here has been discussed in the context of scientific evolution, because the topic is of high importance for economic development, is interesting in itself, and leads to a nice translation of the theoretical concepts for empirical work. I believe nonetheless that it is more general and can be useful for thinking about a number of other situations, in particular all contexts in which actors and their ideas/artefacts co-evolve.

The core of the mathematical system, a growing two-mode network with polynomial attachment kernel, was analyzed using hypergeometric functions solutions of the master equation and mean-field continuous approximations. This system may apply in all cases where only one set of nodes of a two-mode network is growing.

The model in its complete form may be useful in applications where one needs to study the evolution of a two-mode network and its relationships with associated one-mode networks. It demonstrates that the classical self-reinforcing dynamics of a given network ("preferential attachment"), traditionally micro-founded by a "copying" mechanism (i.e. linking to friends-of-friends), can also be explained by a similar mechanism taking place on the associated two-mode network: papers are cited preferentially because they are learned preferentially, because agents learn random ideas of random friends.

This feature of the model also draws attention to the time scales involved in the growth of author-paper networks. From the viewpoint of the citation networks, the time scale is defined by the arrival of new papers, neglecting all the "learning events" happening in between. I tried to convince the reader that these two-mode network events do, however, matter a lot in shaping the (one-mode) citation network. I would tentatively suggest that this is also the case in many other one-mode networks, and expect further research along this line to be fruitful for understanding interrelated complex systems.

References

Adamic, L. and Huberman, B. (2000), 'Power-law distribution of the world wide web', *Science* **287**(5461), 2115–2115.

Albarrán, P., Crespo, J., Ortuño, I. and Ruiz-Castillo, J. (2011), 'The skewness of science in 219 sub-fields and a number of aggregates', *Working Paper, Universidad Carlos III de Madrid* (11-09).

Albarrán, P. and Ruiz-Castillo, J. (2011), 'References made and citations received by scientific articles', *Journal of the American Society for Information Science and Technology* **62**(1), 40–49.

Atalay, E. (2011), 'Sources of variation in social networks', *Mimeo, University of Chicago, Department of Economics* .

Barabási, A. and Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509.

Beguerisse-Diaz, M., Porter, M. A. and Onnela, J.-P. (2010), 'Competition for popularity in bipartite networks', *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20**(4).

Bianconi, G. and Barabási, A. (2001), 'Competition and multiscaling in evolving networks', *EPL (Europhysics Letters)* **54**, 436.

Biglu, M. (2008), 'The influence of references per paper in the SCI to Impact Factors and the Matthew Effect', *Scientometrics* **74**(3), 453–470.

Bollobás, B., Riordan, O., Spencer, J., Tusnády, G. et al. (2001), 'The degree sequence of a scale-free random graph process', *Random Structures & Algorithms* **18**(3), 279–290.

Börner, K., Maru, J. and Goldstone, R. (2004), 'The simultaneous evolution of author and paper networks', *Proceedings of the National Academy of Sciences of the United States of America* **101**(Suppl 1), 5266.

Bramoullé, Y., Currarini, S., Jackson, M., Pin, P. and Rogers, B. (2012), 'Homophily and long-run integration in social networks', *forthcoming, Journal of Economic Theory* .

Breiger, R. (1974), 'The duality of persons and groups', *Social Forces* **53**, 181.

Clauset, A., Shalizi, C. and Newman, M. (2009), 'Power-law distributions in empirical data', *SIAM Review* **51**(4), 661–703.

Cooper, C. and Prałat, P. (2010), 'Scale-free graphs of increasing degree', *Random Structures & Algorithms* **38**(4), 396–421.

Cowan, R. (2005), Network models of innovation and knowledge diffusion, *in* S. Breschi and F. Malerba, eds, 'Clusters, networks and innovation', Oxford University Press New York, pp. 29–53.

Csárdi, G., Strandburg, K., Zalányi, L., Tobochnik, J. and Érdi, P. (2007), 'Modeling innovation by a kinetic description of the patent citation system', *Physica A: Statistical Mechanics and its Applications* **374**(2), 783–793.

Dorogovtsev, S. and Mendes, J. (2000), 'Evolution of networks with aging of sites', *Physical Review E* **62**(2), 1842.

Dorogovtsev, S. and Mendes, J. (2001), 'Effect of the accelerating growth of communications networks on their structure', *Physical review E* **63**(2), 025101.

Dorogovtsev, S., Mendes, J. and Samukhin, A. (2000), 'Structure of growing networks with preferential linking', *Physical Review Letters* **85**(21), 4633–4636.

Eom, Y. and Fortunato, S. (2011), 'Characterizing and modeling citation dynamics', *PloS one* **6**(9), e24926.

Evans, T. and Plato, A. (2008), 'Network rewiring models', *Networks and Heterogenous Media* **3**(2), 221–238.

Faust, K. (1997), 'Centrality in affiliation networks', *Social Networks* **19**(2), 157–191.

Freeman, L. (2003), Finding social groups: A meta-analysis of the southern women data, *in* 'Dynamic Social Network Modeling and Analysis: workshop summary and papers', pp. 39–77.

Gabaix, X. (2009), 'Power laws in economics and finance', *Annual Review of Economics* **1**(1), 255–294.

Gehrke, J., Ginsparg, P. and Kleinberg, J. (2003), 'Overview of the 2003 kdd cup', *ACM SIGKDD Explorations Newsletter* **5**(2), 149–151.

Ghiglino, C. and Kuschy, N. (2011), 'Are patent citations driven by quality?', *Mimeo, University of Essex* .

Golosovsky, M. and Solomon, S. (2012), 'Runaway events dominate the heavy tail of citation distributions', *The European Physical Journal-Special Topics* **205**(1), 303–311.

Goyal, S. (2007), *Connections: an introduction to the economics of networks*, Princeton University Press.

Irwin, J. (1963), 'The place of mathematics in medical and biological statistics', *Journal of the Royal Statistical Society. Series A (General)* pp. 1–45.

Jackson, M. O. (2008), *Social and Economic Networks*, Princeton University Press.

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

Jones, B. (2009), 'The burden of knowledge and the death of the renaissance man: Is innovation getting harder?', *Review of Economic Studies* **76**(1), 283–317.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. S. (1999), The Web as a Graph: Measurements, Models, and Methods, *in* T. Asano, H. Imai, D. T. Lee, S.-i. Nakano and T. Tokuyama, eds, 'Computing and Combinatorics', Vol. 1627 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, Berlin, Heidelberg, chapter 1, pp. 1–17.

Krapivsky, P. and Redner, S. (2005), 'Network growth by copying', *Physical Review E* **71**(3), 036118.

Krapivsky, P., Redner, S. and Leyvraz, F. (2000), 'Connectivity of growing random networks', *Physical Review Letters* **85**(21), 4629–4632.

Laherrere, J. and Sornette, D. (1998), 'Stretched exponential distributions in nature and economy:fat tails with characteristic scales', *The European Physical Journal B-Condensed Matter and Complex Systems* **2**(4), 525–539.

Latapy, M., Magnien, C. and Vecchio, N. (2008), 'Basic notions for the analysis of large two-mode networks', *Social Networks* **30**(1), 31–48.

Latour, B. (2005), *Reassembling the social: An introduction to actor-network-theory*, Oxford University Press, USA.

Latour, B. and Woolgar, S. (1979), *Laboratory life: The construction of scientific facts*, Princeton University Press.

Leskovec, J., Kleinberg, J. and Faloutsos, C. (2005), Graphs over time: densification laws, shrinking diameters and possible explanations, *in* 'Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining', ACM, pp. 177–187.

Liu, C., Yeung, C. and Zhang, Z. (2011), 'Self-organization in social tagging systems', *Physical Review E* **83**(6), 066104.

Mandelbrot, B. (1953), 'An informational theory of the statistical structure of language', *Communication theory* pp. 486–502.

Merton, R. (1968), 'The Matthew effect in science: the reward and communication systems of science are considered', *Science* **159**(3810), 56.

Newman, M. (2006), 'Finding community structure in networks using the eigenvectors of matrices', *Physical review E* **74**(3), 036104.

Newman, M. (2009), 'The first-mover advantage in scientific publication', *EPL (Europhysics Letters)* **86**, 68001.

Newman, M. (2010), *Networks: an introduction*, Oxford University Press.

Perline, R. (2005), 'Strong, weak and false inverse power laws', *Statistical Science* **20**(1), 68–88.

Peruani, F., Choudhury, M., Mukherjee, A. and Ganguly, N. (2007), 'Emergence of a non-scaling degree distribution in bipartite networks: a numerical and analytical study', *EPL (Europhysics Letters)* **79**, 28001.

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

Petkovšek, M., Wilf, H. and Zeilberger, D. (1996), *A=B*, Wellesley, MA, http://www.math.rutgers.edu/ zeilberg/AeqB.pdf.

Price, D. (1965), 'Networks of scientific papers', *Science* **149**(3683), 510.

Price, D. (1976), 'A general theory of bibliometric and other cumulative advantage processes', *Journal of the American Society for Information Science* **27**(5), 292–306.

Ramasco, J., Dorogovtsev, S. and Pastor-Satorras, R. (2004), 'Self-organization of collaboration networks', *Physical Review E* **70**(3), 036106.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F. and Vespignani, A. (2010), 'Characterizing and modeling the dynamics of online popularity', *Physical Review Letters* **105**(15), 158701.

Redner, S. (1998), 'How popular is your paper? an empirical study of the citation distribution', *The European Physical Journal B-Condensed Matter and Complex Systems* **4**(2), 131–134.

Roth, C. and Cointet, J. (2010), 'Social and semantic coevolution in knowledge networks', *Social Networks* **32**(1), 16–29.

Sanditov, B. (2005), 'Patent citations, the value of innovations and path-dependency', *KITeS Working Papers* (177).

Seglen, P. (1992), 'The skewness of science', *Journal of the American Society for Information Science and Technology* **43**(9), 628–638.

Simkin, M. and Roychowdhury, V. (2011), 'Theory of citing', *ArXiv preprint* (1109.2272).

Simon, H. (1955), 'On a class of skew distribution functions', *Biometrika* **42**(3/4), 425–440.

Star, S. and Griesemer, J. (1989), 'Institutional ecology, translations and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39', *Social Studies of Science* **19**(3), 387–420.

Stringer, M., Sales-Pardo, M. and Amaral, L. (2010), 'Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal', *Journal of the American Society for Information Science and Technology* **61**(7), 1377–1385.

Stumpf, M. and Porter, M. (2012), 'Critical truths about power laws', *Science* **335**(6069), 665–666.

Styles, E. (2006), *The psychology of attention*, Psychology Press.

Tsallis, C. and De Albuquerque, M. (2000), 'Are citations of scientific papers a case of nonextensivity?', *The European Physical Journal B-Condensed Matter and Complex Systems* **13**(4), 777–780.

Valverde, S., Solé, R., Bedau, M. and Packard, N. (2007), 'Topology and evolution of technology innovation networks', *Physical Review E* **76**(5), 056118.

Vázquez, A. (2003), 'Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations', *Physical Review E* **67**(5), 056104.

Vega-Redondo, F. (2007), *Complex social networks*, Cambridge University Press.

Yule, G. (1925), 'A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S.', *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* **213**, 21–87.

Zeng, A., Yeung, C., Shang, M. and Zhang, Y. (2012), 'The reinforcing influence of recommendations on global diversification', *EPL (Europhysics Letters)* **97**(1), 18005.

Zipf, G. (1949), *Human behavior and the principle of least effort.*, Addison-Wesley Press.

A Evolution of individual paper's popularity when n is infinite

Let us write a difference equation for the evolution of the degree of paper j (recall that here we assume $n \to \infty$).

$$k_j(t+1) - k_j(t) = (1-b)\frac{k_j(t)}{t} \tag{A.1}$$

which becomes

$$\begin{aligned} k_j(t+1) &= \left(1 + \frac{1}{\hat{b}t}\right) k_j(t) \\ &= \left(1 + \frac{1}{\hat{b}t}\right)\left(1 + \frac{1}{\hat{b}(t-1)}\right)\dots\left(1 + \frac{1}{\hat{b}t_j}\right) k_j(t_j) \\ &= \prod_{i=t_j}^{t}\left(1 + \frac{1}{\hat{b}i}\right) \end{aligned} \tag{A.2}$$

One can then express the product in terms of Gamma functions or Pochhammer symbols. Evaluating the solution at t instead of $t+1$ and choosing a simplification in terms of the Beta function one obtains 6 □

Equation 6 is the exact solution for the deterministic approximation of the evolution of paper j assuming $n \to \infty$.

B Popularity distribution when n is infinite

Assuming an infinite population, the model is similar to Simon's (1955). To be precise, Simon's model assumes that, in our context, there is one and exactly one new actor per period, but since we are counting the edges attached to papers, it does not matter what is on the other side of the edges. What matters is that exactly one edge is added at each period. With probability b, it comes with a new paper, and with complementary probability it attaches to an existing paper.

Simon used a technique known now as the master equation method. It consists in specifying the probability that, during any time step, a given bin of the histogram increases or decreases.

The master equation is as follows. The number of papers with degree k in $(t+1)$ is equal to the number of papers with degree k in t, plus the number of papers with degree $k-1$ in t that have been learned, minus the number of papers with degree k in t that have been learned. Putting this together

$$P_{t+1}(k) - P_t(k) = (1-b)P_t(k-1)A_t(k-1) - (1-b)P_t(k)A_t(k) \tag{B.1}$$

From equation 2, the attachment kernel for $n \to \infty$ is k/t. Putting this and $P_t(k) = w_t p_t(k)$ in equation B.1, and simplifying, one gets

$$(t+1)p_{t+1}(k) - tp_t(k) = (1-b)p_t(k-1)(k-1) - (1-b)p_t(k)k$$

Now, assume that a stationary distribution exists for $t \to \infty$, so $p_{t+1}(k) = p_t(k) = p(k)$ to obtain

$$p(k) = (1-b)\left[(k-1)p(k-1) - kp(k)\right] \tag{B.2}$$

For $k = 0$, we know that $p(0) = 0$ since each paper enters with a degree equals to 1. For $k = 1$, equation B.1 is modified thus

$$P_{t+1}(1) = P_t(1) + b - (1-b)P_t(1)\frac{1}{t}$$

Using the same simplification as before leads to

$$p(1) = \frac{1}{2-b} \tag{B.3}$$

The system (B.2-B.3) was solved recursively by Simon (1955) to yield the Yule-Simon distribution, equation 9, which is a power law because of the following relation (derived using Stirling approximation).

$$B(a,b) \propto \Gamma(a)b^{-a} \text{ if a is fixed and b is large} \tag{B.4}$$

C Popularity distribution when n is finite

The master equation B.1 does not change, except that now it has the attachment kernel given by equation 2, without assuming infinite n. Using the usual simplifications and assuming a steady state gives the recurrence

$$p(k)(k(n-k)+\hat{b}n\mu) = p(k-1)(k-1)(n-(k-1)) \tag{C.1}$$

This equation can be solved using different methods. Here I choose to present the derivation using the definition of generalized hypergeometric functions. The reason is that the final result contains a quantity, $\frac{n-4\pm\sqrt{n(n+4\hat{b}\mu)}}{2}$, the origin of which is not straightforward to see unless we use the method below and observe that these are simply the roots of the denominator of the term used to calculate the ratio of two consecutive terms of the degree distribution. Another reason to use generalized hypergeometric functions is that they often naturally represent the generating function of probability mass functions involving special functions, as is the case for the Yule-Simon, the Waring, and the distribution below. It is common in growing network models to solve the master equation by solving the Ordinary Differential Equation (ODE) involving the generating function of the degree distribution. The solution of the ODE is most often a generalized hypergeometric function (see e.g. Dorogovtsev et al.'s (2000) equation 8). I suggest that this result can be obtained directly by writing the series associated to the degree distribution, and avoid using generating functions[39]. The method below and how and why this relates to symbolic computation and computer science (a simonian theme again), is nicely described in the book A=B (Petkovšek et al., 1996).

Let us start by writing the recurrence for the degree distribution in terms of $k+1$

$$\frac{p(k+1)}{p(k)} = \frac{k(n-k)}{\hat{b}n\mu + (k+1)(n-(k+1))}$$

Now, consider the series for the degree distribution

$$\begin{aligned}
\sum_{k=1}^{\infty} p_k &= p(1) + p(1)\frac{1(n-1)}{\hat{b}n\mu + 2(n-2)} + p(1)\frac{1(n-1)}{\hat{b}n\mu + 2(n-2)}\frac{2(n-2)}{\hat{b}n\mu + 3(n-3)} + \dots \\
&= p(1)\Big(1 + \frac{1(n-1)}{\hat{b}n\mu + 2(n-2)} + \frac{1(n-1)}{\hat{b}n\mu + 2(n-2)}\frac{2(n-2)}{\hat{b}n\mu + 3(n-3)} + \dots\Big) \\
&= p(1)\Big(1 + \sum_{K=1}^{\infty}\prod_{i=1}^{K}\frac{i(n-i)}{\hat{b}n\mu + (i+1)(n-(i+1))}\Big)
\end{aligned} \tag{C.2}$$

Now, consider the following series

$$\frac{t_{m+1}}{t_m} = \frac{(m+1)(n-(m+1))}{\hat{b}n\mu + (m+2)(n-(m+2))} = \frac{(m+1)(m+1-n)}{(m-r_1)(m-r_2)} \tag{C.3}$$

where r_1 and r_2 are the two roots of $\hat{b}n\mu + (m+2)(n-(m+2)) = 0$, namely

$$\{r_1, r_2\} = \frac{n-4\pm\sqrt{n(n+4\hat{b}\mu)}}{2} \tag{C.4}$$

[39] The method works nicely for Simon's and Price's models. In these cases, the function involved is a Gauss Hypergeometric function and one can use Gauss's Hypergeometric Theorem to check that the result is a proper probability mass function (i.e. $\sum_k p_k = 1$).

Also assume that $t_0 = 1$. If we write the series t_m explicitly, we have

$$\begin{aligned}\sum_{m=0}^{\infty} t_m &= 1 + \frac{1(n-1)}{\hat{b}n\mu + 2(n-2)} + \frac{1(n-1)}{\hat{b}n\mu + 2(n-2)}\frac{2(n-2)}{\hat{b}n\mu + 3(n-3)} + \dots \\ &= 1 + \sum_{K=1}^{\infty}\prod_{i=1}^{K}\frac{i(n-i)}{\hat{b}n\mu + (i+1)(n-(i+1))}\end{aligned} \tag{C.5}$$

so that substituting C.5 into C.2, we have a simple relationship between the series t_m and p_k

$$\sum_{k=1}^{\infty} p_k = p(1)\sum_{m=0}^{\infty} t_m \tag{C.6}$$

On the other hand, by definition of hypergeometric series (see e.g. Petkovšek et al., 1996), series C.3 implies

$$\sum_{m=0}^{\infty} t_m = {}_3F_2\begin{bmatrix}1, & 1-n, & 1 \\ & -r_1, & -r_2\end{bmatrix}; 1 \tag{C.7}$$

where the generalized hypergeometric series is defined as

$${}_pF_q\begin{bmatrix}a_1, & a_2 & \dots & a_p \\ b_1, & b_2 & \dots & b_q\end{bmatrix}; z = \sum_{s=0}^{\infty}\frac{(a_1)_s\dots(a_p)_s}{(b_1)_s\dots(b_q)_s}\frac{z^s}{s!} \tag{C.8}$$

where the Pochhammer symbol is defined in equation 18. Substituting equation C.7 into C.6 gives

$$\sum_{k=1}^{\infty} p(k) = p(1)\ {}_3F_2\begin{bmatrix}1, & 1-n, & 1 \\ & -r_1, & -r_2\end{bmatrix}; 1$$

Using equation C.8, this becomes

$$\sum_{k=1}^{\infty} p(k) = p(1)\sum_{s=0}^{\infty}\frac{(1)_s(1-n)_s}{(-r_1)_s(-r_2)_s}$$

Choosing $s = k-1$, one gets

$$\sum_{k=1}^{\infty} p(k) = p(1)\sum_{k=1}^{\infty}\frac{(1)_{k-1}(1-n)_{k-1}}{(-r_1)_{k-1}(-r_2)_{k-1}}$$

By comparison of the two sides we can get rid of the summation term, leaving us with equation 16 □

Finally, the initial condition is found by writing the master equation for $k = 1$

$$P_{t+1}(1) = P_t(1) + b - (1-b)P_t(1)\frac{n-1}{\mu n t}$$

Assuming steady state and using the same simplifications as before, one obtains 17 □

D Discrete citations distribution when n is infinite

To derive the exact distribution of citations, one would need to model popularity and citations together by writing the master equation for the joint probability mass functions of k and k^C. This task is arguably out of the scope of this paper. Nevertheless, I can propose here a turn around which does lack mathematical rigor but indeed provides a formula which is better than that from the mean-field continuous approximations, because it is a probability *mass* function, i.e. it *sums* up to one.

The idea is to derive the mean-field continuous approximation of a very related model for which the exact discrete degree distribution is known, namely Price's model (Dorogovtsev et al., 2000; Newman, 2010). By comparing the two mean-field continuous approximations (Price's and learning-based models) and looking at the discrete degree distribution of the known model (Price's), one can guess the discrete degree distribution of the learning-based model.

Price's model is described as follows. At each period, a new paper arrives, and it cites h existing papers with probability proportional to their citations count plus a constant. The constant, denoted a, is the so-called "initial attractiveness" of a paper. It is necessary because since papers have in-degree equal to zero at birth, they would never have a chance to be cited if that constant was not there. Denote q_j the degree of paper j. The time index is w. We have

$$\frac{dq_j(w)}{dw} = h\frac{q_j + a}{\sum_{j=1}^{w}(q_j + a)} = h\frac{q_j + a}{w(h + a)}$$

With the initial condition that $q_j(w_j) = 0$, one obtains

$$q_j(w) = a\left(\left(\frac{w}{w_j}\right)^{\frac{h}{a}+1} - 1\right)$$

and by the usual method

$$p(q) = \left(\frac{a}{h} + 1\right) a^{\frac{a}{h}+1}(q + a)^{-(2+\frac{a}{h})} \tag{D.1}$$

Clearly, this is a shifted power law with $r = a$ and $\gamma = 2 + \frac{a}{h}$. Now, let us look at the exact solution found by Dorogovtsev et al. (2000) (see also Newman, 2010) using a master equation technique[40]:

$$p(q) = \frac{B(q + a, 2 + \frac{a}{h})}{B(a, 1 + \frac{a}{h})} \tag{D.2}$$

We wish to compare D.1 and D.2. We need not care about the constants, since they come from the normalization constraint. We are left with the simple observation that $B(q + a, 2 + \frac{a}{h})$ compares with $(q + a)^{-(2+\frac{a}{h})}$. Indeed, the second is the approximation of the first assuming that $(q + a) \to \infty$ (using Stirling approximation). Note that this type of relationship between the mean-field continuous approximation and the master equation solution can also be observed for Simon's model (see equations 8 and 9)

Now, if we compare the mean-field continuous approximation of the learning based model (equation 12) with the mean-field continuous approximation of Price's model, we see that the role played by $q+a$ in the latter is played by $k + \hat{b}bh$ in the former, and that the role played by $2 + a/h$ in the latter is played by $2 + \hat{b}b$ in the former. Substituting these values in equation D.2, we get 13 □

[40] Dorogovtsev et al. use the parameter $\bar{a} = a/h$ to solve the model, so equation (9) in their paper looks identical to the distribution of citations in the learning-based model (or in a mixed attachment model such as Jackson and Rogers's), but it is not. After converting back to the original parameters, we observe that the power law exponent depends on h and the shoulder parameter does not: this is Price's model.

E On Lotka's law

In the model above, actors are treated symmetrically. As a result, they have a similar number of papers learned and created. This is clearly against the empirical evidence known as Lotka's law, which states that the number of publications of individual scientists are power law distributed. In the model, this distribution is binomial. For simplicity, consider the number of papers known instead of the number of papers produced. To "know" k_a papers at time t, a r.c. agent needs to have been chosen exactly k_a times, and not chosen exactly $(t - k_a)$ times. Thus it follows that:

Proposition E.1. *The distribution of actors' number of papers known is the binomial distribution*

$$p_t(k_a) = \binom{t}{k_a} \left(\frac{1}{n}\right)^{k_a} \left(1 - \frac{1}{n}\right)^{t-k_a}$$

Hence, one can say that the model shows that it is possible to have skewed citations network *without* heterogeneity of author's productivity or author's knowledge amount.[41] If one wants to obtain Lotka's law, then it is a simple matter of adding the Ramasco et al.'s (2004) model of co-authorship "on top" of the learning-based model. This opens interesting avenues because the way one specifies how learning in teams takes place can alter some results. Therefore an extended model can highlight the interplay of team-based work and collective learning in scientific evolution, without loosing analytical tractability.

F The model with productivity parameters

A notably unrealistic aspect of the model is that during one period, the r.c. agent learns or produce exactly one paper. Let us assume here that during one time period, an agent either learns λ_L papers or produce λ_P papers. The evolution of paper j's popularity is given by

$$\frac{dk_j(t)}{dt} = (1-b)\lambda_L \frac{k_j(t)}{t(b\lambda_P + (1-b)\lambda_L)}$$

which, using $k_j(t_j) = 1$, has solution

$$k_j(t) = \left(\frac{t}{t_j}\right)^{\frac{1}{1+\hat{b}b\lambda_P/\lambda_L}} \tag{F.1}$$

The usual method then gives a power law similar to (8) but with exponent $2 + \hat{b}b\frac{\lambda_P}{\lambda_L}$ instead of $1 + \hat{b} = 2 + \hat{b}b$, and appropriate normalizing constant. The interpretation is straightforward: learning many papers per period implies a flatter distribution because learning is biased towards popular papers. Producing many papers per period steepen the distribution because produced papers arrive with degree one (they are known solely by their author) which increases the first bin of the histogram. In other words, the parameters λ_P and λ_L plays the same role as, respectively, b and $1 - b$.

Again using the standard method followed in the main text, the dynamics of the number of citations received by paper j is

$$\frac{dk_j^C(t)}{dt} = bh\lambda_P \frac{1}{t(b\lambda_P + (1-b)\lambda_L)/n} \frac{k_j(t)}{n} = bh\lambda_P \frac{k_j(t)}{t(b\lambda_P + (1-b)\lambda_L)}$$

[41] The variance of this distribution increases with t, so agents heterogeneity grows unboundedly, but the coefficient of variation (mean/variance) is stable and equal $1 - 1/n$. The variance is always much lower than what it would be under Lotka's law.

where $k_j(t)$ is given above by (F.1). The distribution of citations received is found to be a shifted power law with exponent equal to that of the popularity distribution $\gamma = 2 + \hat{b}b\frac{\lambda_P}{\lambda_L}$ and the shoulder parameter $r = \hat{b}bh\frac{\lambda_P}{\lambda_L}$. Hence to make the point that socially embedded learning generates a shifted power law citation distribution, it is not necessary to introduce complications linked to the relative productivity of learning/producing papers.

G The model with a social network

In the text, a full social network was assumed. This is a very strong assumption, and in view of the huge amount of research on research on the effects of network structure on different processes and outcomes, it should be expected that assuming more realistic network structure would change the results. A detailed study is out of the scope of this paper, and we shall contend ourselves here with some simulations to be seen as a set of robustness checks of the results obtained under the assumption of a full graph. It is basically shown that random graphs delivers a behavior very similar to full graphs, whereas circles (one dimensional boundaryless lattice) lead to noticeably different results. The conclusion that imitating random ideas of random friends and citing randomly within one's knowledge set leads to skewed distribution of popularity and citations is preserved in any case.

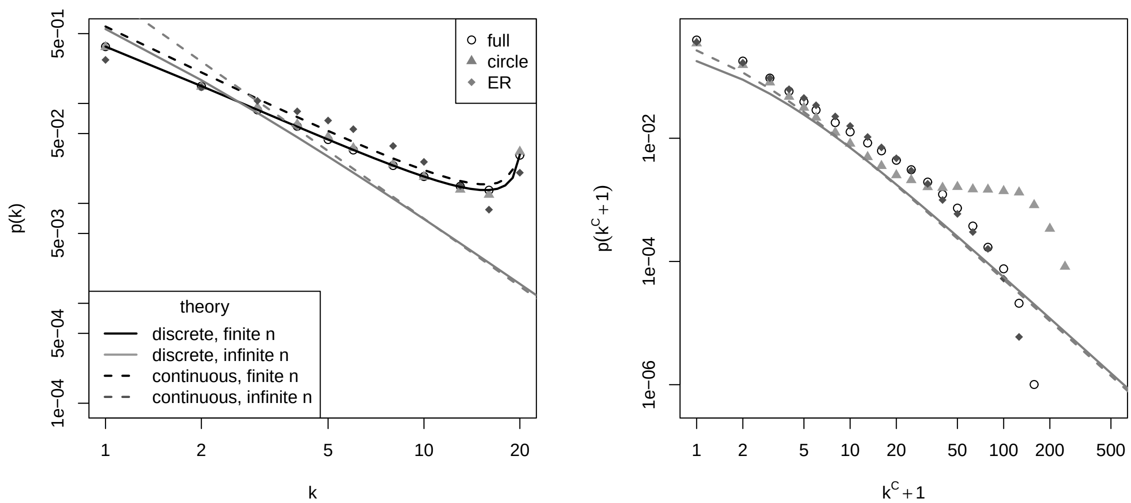

Figure 7: **Effect of social network- One dimensional lattice and Erdős-Renyi**. The simulation setup is $b = 0.2$, $n = 20$, $h = 5$, 50000 time steps, and 100 simulations are averaged. The circle network is a one dimensional boundaryless lattice where each agent is connected to two agents on her left and two on her right. The Erdős-Renyi graph is a (connected) random graph with 40 edges. The plain lines are the same than in figure 2

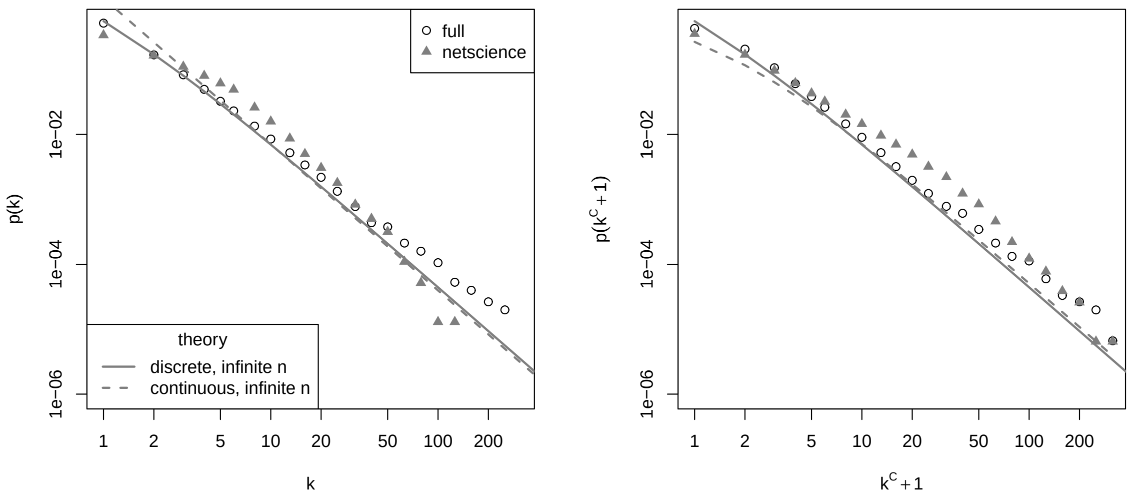

Figure 8: **Effect of social network- Real-world network**. The simulation setup is $b = 0.2$, $h = 5$, 150000 time steps, and 5 simulations are averaged. The social network is either a full one, either a real world network, namely the collaboration network from papers about network science as collected and made available by Newman (2006) (379 nodes and 914 edges).

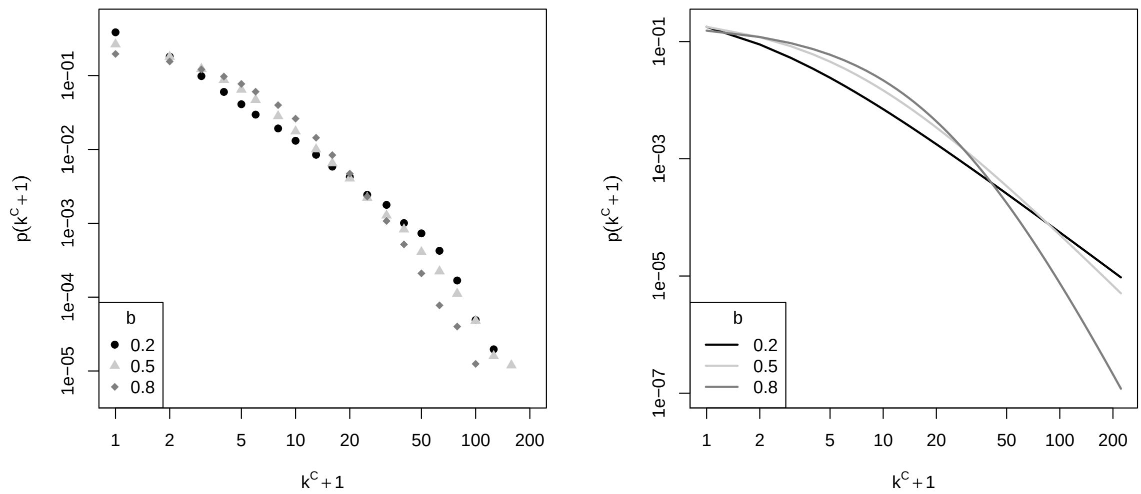

Figure 9: **Effect of sparse social network- effect of** b. The simulation setup is three different values of b, $n = 100$, $h = 5$, 50000 time steps,and 10 simulations are averaged. The social network is a circle where agents are connected to four agents on each side. The left panel shows the simulations and the right panel shows theoretical results (using equation 13)

The first simulation exercise compares stark theoretical structures, namely a one dimensional boundaryless lattice (a circle) and an Erdős-Renyi graph (a random graph) with results known with great detail, i.e. for $n = 20$. The results, depicted in figure 7, show that the distribution of popularity stays the same to a large extent. The distribution of citations, while similar when a random graph is used, exhibits a "break" in the middle of the curve when the social network is a circle.

The second test uses the largest component (379 nodes) of a real-world collaboration network (Newman, 2006), which is known to have a skewed degree distribution. In this case, the difference observed between the full and the sparse network seems similar for popularity and for citations: the curve for sparse graph is more "round" and cuts the curve for full graph at two different points, i.e., it has a lower constant and a lower tail, as one observes in figure 8.

The last set of simulations is designed to check that the effect of the parameters b and h is the same when we consider a sparse graph. Figure 9 shows that the effect of b on the distribution of citations is as expected, even in the worst case scenario which is a circle. The effect of h deserves more attention since it is used to discriminate between Price's and the learning-based model empirically. Figure 10 shows that the effect of h on the distribution of citations is as expected in the case of the random graph, but not in the case of the circle network. In the latter, the change in h does not lead to a parallel shift of the curve, and there is also a more marked cut-off. This somehow limits the span of the results presented in the main text under the full graph assumption, and calls for further research. That being said, once we compare the results obtained under the same parameter conditions (same h) and changing only the underlying social network (see figure 11), we see that the overlap seems really good when the probability mass distribution is plotted (in particular, the break in the citation distribution found for $n = 20$ in figure 7 disappears in this set of simulations where $n = 100$). However, a significant difference between the full and random on the one hand and the circle network on the other hand reappears when plotting the counter-cumulative version of the same raw data.

Overall, I conclude that sparse social networks lead to degree distributions which are significantly different, but not completely at odd with those obtained under the full network assumption. To the extent that the effects of b and h are preserved, and this is roughly the case, the full network assumption is not only a mathematical convenience but a requirement following from the exigence of parsimony.

Obviously, a lot of interesting research is ahead if one wishes to understand in detail the consequences of the social network structure on popularity and citation distributions.

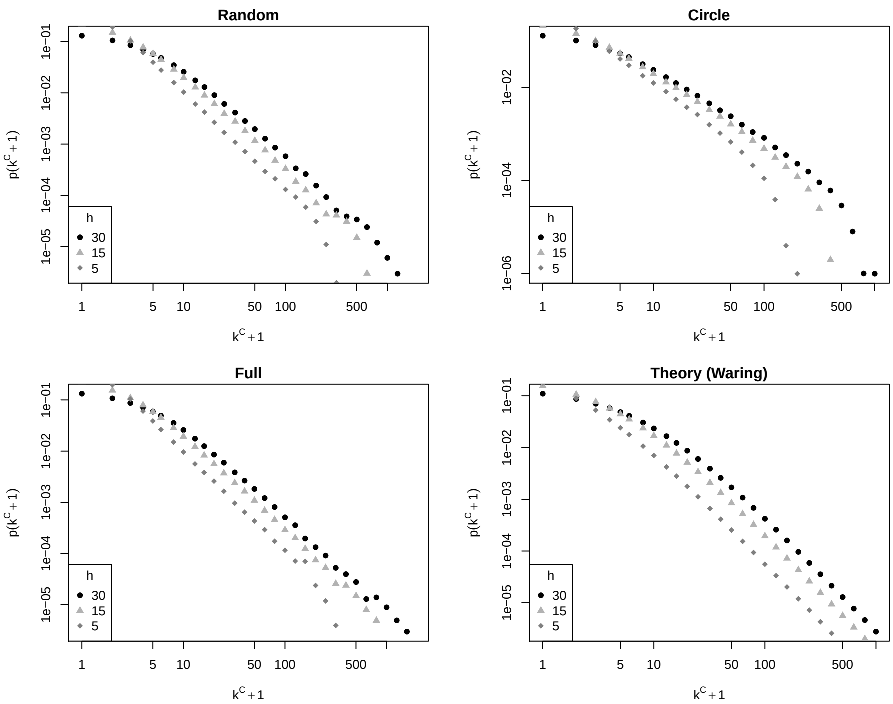

Figure 10: **Effect of social network- effect of** h. The simulation setup is $b = 0.2$, $n = 100$, three different values of h, 50000 time steps, and 100 simulations are averaged. Three social networks are compared: a circle where agents are connected to four agents on each side, an Erdős-renyi graph with mean degree equal to 10, and a full graph. The bottom-right shows the simulations and the right panel shows theoretical results (using equation 13)

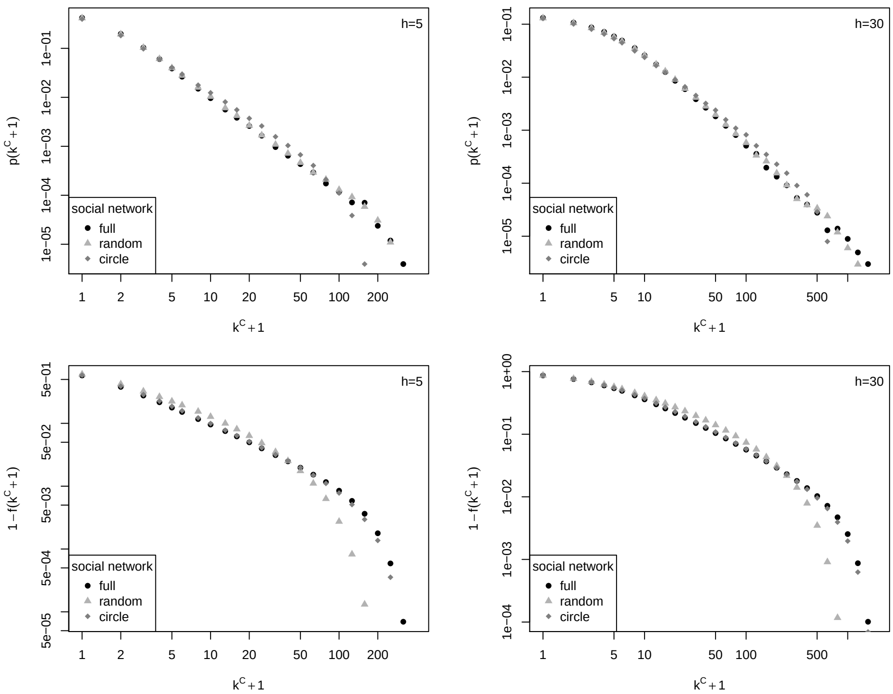

Figure 11: **Effect of social network- effect of** h. The raw data in the same than in figure 10. The bottom panels display the counter cumulative distribution. The same pattern is obtained for $h = 15$ and is omitted.

The UNU-MERIT WORKING Paper Series

2012-01 *Maastricht reflections on innovation* by Luc Soete
2012-02 *A methodological survey of dynamic microsimulation models* by Jinjing Li and Cathal O'Donoghue
2012-03 *Evaluating binary alignment methods in microsimulation models* by Jinjing Li and Cathal O'Donoghue
2012-04 *Estimates of the value of patent rights in China* by Can Huang
2012-05 *The impact of malnutrition and post traumatic stress disorder on the performance of working memory in children* by Elise de Neubourg and Chris de Neubourg
2012-06 *Cross-national trends in permanent earnings inequality and earnings instability in Europe 1994-*2001 by Denisa Maria Sologon and Cathal O'Donoghue
2012-07 *Foreign aid transaction costs* by Frieda Vandeninden
2012-08 *A simulation of social pensions in Europe* by Frieda Vandeninden
2012-09 *The informal ICT sector and innovation processes in Senegal* by Almamy Konté and Mariama Ndong
2012-10 *The monkey on your back?! Hierarchical positions and their influence on participants' behaviour within communities of learning* by Martin Rehm, Wim Gijselaers and Mien Segers
2012-11 *Do Ak models really lack transitional dynamics?* by Yoseph Yilma Getachew
2012-12 *The co-evolution of organizational performance and emotional contagion* by R. Cowan, N. Jonard, and R.Weehuizen
2012-13 *"Surfeiting, the appetite may sicken": Entrepreneurship and the happiness of nations* by Wim Naudé, José Ernesto Amorós and Oscar Cristi
2012-14 *Social interactions and complex networks* by Daniel C. Opolot
2012-15 *New firm creation and failure: A matching approach* by Thomas Gries, Stefan Jungblut and Wim Naudé
2012-16 *Gains from child-centred Early Childhood Education: Evidence from a Dutch pilot programme* by Robert Bauchmüller
2012-17 *Highly skilled temporary return, technological change and Innovation: The Case of the TRQN Project in Afghanistan* by Melissa Siegel and Katie Kuschminder
2012-18 *New Technologies in remittances sending: Opportunities for mobile remittances in Africa* Melissa Siegel and Sonja Fransen
2012-19 *Implementation of cross-country migration surveys in conflict-affected settings: Lessons from the IS Academy survey in Burundi and Ethiopia* by Sonja Fransen, Katie Kuschminder and Melissa Siegel
2012-20 *International entrepreneurship and technological capabilities in the Middle East and North Africa* by Juliane Brach and Wim Naudé
2012-21 *Entrepreneurship, stages of development, and industrialization* by Zoltan J. Ács and Wim Naudé
2012-22 *Innovation strategies and employment in Latin American firms* by Gustavo Crespi and Pluvia Zuniga
2012-23 *An exploration of agricultural grassroots innovation in South Africa and implications for innovation indicator development* by Brigid Letty, Zanele Shezi and Maxwell Mudhara
2012-24 *Employment effect of innovation: microdata evidence from Bangladesh and Pakistan* by Abdul Waheed

2012-25 *Open innovation, contracts, and intellectual property rights: an exploratory empirical study* by John Hagedoorn and Ann-Kristin Ridder

2012-26 *Remittances provide resilience against disasters in Africa* by Wim Naudé and Henri Bezuidenhout

2012-27 *Entrepreneurship and economic development: Theory, evidence and policy* by Wim Naudé

2012-28 *Whom to target - an obvious choice?* by Esther Schüring and Franziska Gassmann

2012-29 *Sunk costs, extensive R&D subsidies and permanent inducement effects* by Pere Arqué-Castells and Pierre Mohnen

2012-30 *Assessing contingent liabilities in public-private partnerships (PPPs)* by Emmanouil Sfakianakis and Mindel van de Laar

2012-31 *Informal knowledge exchanges under complex social relations: A network study of handloom clusters in Kerala, India* by Robin Cowan and Anant Kamath

2012-32 *Proximate, intermediate and ultimate causality: Theories and experiences of growth and development* by Adam Szirmai

2012-33 *Institutions and long-run growth performance: An analytic literature review of the institutional determinants of economic growth* by Richard Bluhm and Adam Szirmai

2012-34 *Techniques for dealing with reverse causality between institutions and economic performance* by Luciana Cingolani and Denis de Crombrugghe

2012-35 *Preliminary conclusions on institutions and economic performance* by Denis de Crombrugghe and Kristine Farla

2012-36 *Stylized facts of governance, institutions and economic development. Exploring the institutional profiles database* by Bart Verspagen

2012-37 *Exploring the Panel Components of the Institutional Profiles Database (IPD)* by Luciana Cingolani and Denis de Crombrugghe

2012-38 *Institutions and credit* by Kristine Farla

2012-39 *Industrial policy for growth* by Kristine Farla

2012-40 *Explaining the dynamics of stagnation: An empirical examination of the North, Wallis and Weingast approach* by Richard Bluhm, Denis de Crombrugghe and Adam Szirmai

2012-41 *The importance of manufacturing in economic development: Past, present and future perspectives* by Wim Naudé and Adam Szirmai

2012-42 *Lords of Uhuru: the political economy of elite competition and institutional change in post-independence Kenya* by Biniam Bedasso

2012-43 *Employment and wages of people living with HIV/AIDS* by Pilar García-Gómez, José M. Labeaga and Juan Oliva

2012-44 *Prescriptions for network strategy: Does evidence of network effects in cross-section support them?* by Joel A.C. Baum, Robin Cowan and Nicolas Jonard

2012-45 *Perspectives on human development theory in democracy promotion: A comparison of democracy promotion programmes in Egypt through the lenses of classical and revised modernisation theory* by Inger Karin Moen Dyrnes

2012-46 *Nonlinearities in productivity growth: A semi-parametric panel analysis* by Théophile T. Azomahou, Bity Diene and Mbaye Diene

2012-47 *Optimal health investment with separable and non-separable preferences* by Théophile T. Azomahou, Bity Diene, Mbaye Diene and Luc Soete

2012-48 *Income polarization and innovation: Evidence from African economies* by Théophile T. Azomahou and Mbaye Dien

2012-49 *Technological capabilities and cost efficiency as antecedents of foreign market entry* by Fabrizio Cesaroni, Marco S. Giarratana and Ester Martínez-Ros
2012-50 *Does the internet generate economic growth, international trade, or both?* by Huub Meijers
2012-51 *Process innovation objectives and management complementarities: patterns, drivers, co-adoption and performance effects* by Jose-Luis Hervas-Oliver, Francisca Sempere-Ripoll and Carles Boronat-Moll
2012-52 *A systemic perspective in understanding the successful emergence of non-traditional exports: two cases from Africa and Latin America* by Michiko Iizuka and Mulu Gebreeyesus
2012-53 *Determinants of quadic patenting: Market access, imitative threat, competition and strength of intellectual property rights* Can Huang and Jojo Jacob
2012-54 *Envy and habits: Panel data estimates of interdependent preferences* by Jose Maria Casado, Francisco Alvarez-Cuadrado, Jose Maria Labeaga and Dhanoos Sutthiphisal
2012-55 *The impact of Medium-Skilled immigration: A general equilibrium approach* by Joan Muysken, Ehsan Vallizadeh and Thomas Ziesemer
2012-56 *Tax incentives or subsidies for R&D?* by Isabel Busom, Beatriz Corchuelo and Ester Martinez Ros
2012-57 *The impact of development aid on education and health: Survey and new evidence from dynamic models* by Thomas Ziesemer
2012-58 *Do R&D tax incentives lead to higher wages for R&D workers? Evidence from the Netherlands* by Boris Lokshin and Pierre Mohnen
2012-59 *Determinants of the prevalence of diarrhoea in adolescents attending school: A case study of an Indian village school* by Shyama V. Ramani, Timothée Frühauf, Arijita Dutta and Huub Meijers
2012-60 *Communication costs and trade in Sub-Saharan Africa* by Evans Mupela and Adam Szirmai
2012-61 *Differential welfare state impacts for frontier working age families* by Irina S. Burlacu and Cathal O'Donoghue
2012-62 *Microeconometric evidence of financing frictions and innovative activity* by Amaresh K Tiwari, Pierre Mohnen, Franz C. Palm, Sybrand Schim van der Loeff
2012-63 *Globalization and the changing institution for sustainability: The case of the Salmon farming industry in Chile* by Michiko Iizuka and Jorge Katz
2012-64 *Chronic and transitory poverty in the Kyrgyz Republic: What can synthetic panels tell us?* by Mira Bierbaum and Franziska Gassmann
2012-65 *Worker remittances and government behaviour in the receiving countries* by Thomas H.W. Ziesemer
2012-66 *Switching the lights off: The impact of energy tariff increases on households in the Kyrgyz Republic* by Franziska Gassmann
2012-67 *The dynamics of renewable energy transition in developing countries - The case of South Africa and India* by Radhika Perrot
2012-68 *Government R&D impact on the South African macro-economy* by Radhika Perrot, David Mosaka, Lefentse Nokaneng and Rita Sikhondze
2012-69 *The determinants of home based long-term care utilisation in Western European countries* by Sonila M Tomini, Wim Groot and Milena Pavlova
2012-70 *Paying informally for public health care in Albania: scarce resources or governance failure?* by Sonila M Tomini and Wim Groot

2012-71 *Learning and the structure of citation networks* by Francois Lafond