Does cumulative advantage affect collective learning in science? An agent-based simulation

Christopher Watts

Nigel Gilbert

Department of Sociology, Centre for Research in Social Simulation, University of Surrey, Guildford, Surrey, GU2 7XH, UK.

Tel: +44 (0)1483 682788

Fax: +44 (0)1483 689551

c.watts@surrey.ac.uk

www.simian.ac.uk

Abstract Agent-based simulation can model simple micro-level mechanisms capable of generating macro-level patterns, such as frequency distributions and network structures found in bibliometric data. Agent-based simulations of organisational learning have provided analogies for collective problem solving by boundedly rational agents employing heuristics. This paper brings these two areas together in one model of knowledge seeking through scientific publication. It describes a computer simulation in which academic papers are generated with authors, references, contents, and an extrinsic value, and must pass through peer review to become published. We demonstrate that the model can fit bibliometric data for a token journal, *Research Policy*. Different practices for generating authors and references produce different distributions of papers per author and citations per paper, including the scale-free distributions typical of cumulative advantage processes. We also demonstrate the model's ability to simulate collective learning or problem solving, for which we use Kauffman's *NK* fitness landscape. The model provides evidence that those practices leading to cumulative advantage in citations, that is, papers with many citations becoming even more cited, do not improve scientists' ability to find good solutions to scientific problems, compared to those practices that ignore past citations. By contrast, what does make a difference is referring only to publications that have successfully passed peer review. Citation practice is one of many issues that a simulation model of science can address when the data-rich literature on scientometrics is connected to the analogy-rich literature on organisations and heuristic search.

Keywords *Simulation; Cumulative advantage; Landscape search; Science models; Science policy*

Classification code:
MSC: 91D10 (primary); 91D30; 90B70
JEL: C63; D83; D85

Introduction

It has long been recognised that academic sciences show evidence of processes of cumulative advantage, or what Merton called 'the Matthew Effect' (Merton 1968; 1988): to those that have, more shall be given. That success breeds success Merton identified from Zuckerman's interviews with Nobel laureates (Zuckerman 1977). Within bibliometric data the telltale sign is a power-law, or scale-free, frequency distribution, as demonstrated for the numbers of papers per author (Lotka 1926; Simon 1955), and the numbers of citations per paper (Price 1976). Opportunities for publishing tend to go to those who already have papers to their name. References in a new paper tend to be made to publications already rich in citations. Most scientists will publish little and be cited little. A tiny minority of authors will enjoy a prolific publishing career, and a minority of publications will become citation classics. With this in mind, it might be asked what this concentration of resources and attention on so few academics and publications does for the advancement of scientific fields. Do collectives of academic scientists perform better for following the practices that generate cumulative advantage patterns? To answer such a question, it is not possible to rerun the course of science using an alternative set of publishing practices. Instead, we show in this paper that an agent-based computer simulation model of academic publication can provide insight into the role played by publication practices.

In what follows, we provide a brief introduction to the idea of modelling scientific publication. In the past this has been attempted through stochastic process models, and so we need to explain why agent-based simulation is called for. Both types of model employ simple micro-level mechanisms to generate macro-level patterns, but agent-based simulations can readily incorporate several such mechanisms in the one model. In particular, to address questions of scientists' performance we combine the mechanisms that lead to cumulative advantage with those of searching for better solutions to problems. These latter mechanisms, called heuristic search algorithms, have already been incorporated in simulations of organisational learning (March 1991; Lazer & Friedman 2007), and so our model connects this field to that of science modelling. We then describe the simulation model, flagging up some issues involved in its design and highlighting where alternative design decisions could be taken. With the help of bibliometric data from a real journal, *Research Policy*, we calibrate the model, and then perform an experiment with it, by varying the mechanism by which references are created for each new paper. This suggests that the cumulative-advantage process operating on citations has little or no effect on search performance, that is, on scientists' ability to find good solutions to scientific problems. This is in contrast to the effects on search performance of filtering papers for publication, achieved through peer review and through a preference for recency when selecting which papers to refer to. However, this paper is only intended to give an indication of the opportunities offered by computer simulation, and we end with some pointers for further research. In spelling out here the structure and issues behind the model we hope to inspire other attempts at designing and validating simulation models capable of providing insight into scientific organisation and performance.

Models of science

Empirical patterns to be explained with stochastic models

Activity by academics and scientists leaves a data trail of their publications that readily lends itself to modelling. In the age of electronic databases of journal publications this kind of data is widely available, but awareness of its patterns predates these. As noted already, Lotka (1926) showed that the numbers of papers per author followed a power-law or scale-free distribution, while Price (1976) found such a distribution for citations per paper. Price (1963) had earlier observed exponential growth rates in papers and authors in the field of physics and reflected on the implications of this.

To explain how these distributions come about, a model of a stochastic process, or *urn model*, is usually employed. Simon (1955) presented a simple stochastic-process model to generate a scale-free frequency distribution, and fitted it to Lotka's data. Items which could be new papers or references in new papers are allocated to selected 'urns' or categories, such as authors. In the case of the power-law distributions, the stochastic process involved growth over time in the number of urns, and some kind of cumulative advantage: those rich in papers or citations were expected to get richer.

Contributions to science modelling since Simon have explored the mathematical implications of such stochastic process models (Schubert & Glaenzel 1984; Glaenzel & Schubert 1990, 1995; Glaenzel & Schubert 1995; Burrell 2001). For example Burrell (2001) relates the citation process to the ageing and eventual obsolescence of papers. Redner (1998) argues for there being at least two mechanisms generating citations in the data he analyses, including one for more highly cited classics. Burrell (2007) employs a stochastic model to estimate the behaviour under different conditions of Hirsch's h-index for measuring research output and impact based on citations.

Beyond simple distributions of papers and citations, bibliometric data have also yielded networks of relations for analysis. Price (1965) relates papers by citations. Newman (2001a, 2001b, 2001c) relates co-authors by their having collaborated together on at least one paper. He then employs social network analysis metrics, including node degree or the number of co-authors, centrality and the shortest path between pairs of nodes, and clustering or the extent to which my neighbours are neighbours of each other. With increased interest in simple processes by which 'small-world' and 'scale-free' networks may be generated (Watts & Strogatz 1998; Barabasi & Albert 1999; Watts 2004), it seems desirable to extend science models to explaining network patterns as well as distributions (Boerner et al. 2004).

At this point it becomes desirable to employ computer simulations and so-called 'agent-based', 'multi-agent' or individual-level simulation models in particular (Axelrod 1997; Gilbert & Troitzsch 2005). Mathematical treatments of network formation, like those of distribution formation, are certainly possible (Newman 2003).

But for non-specialists, simulation has a number of advantages over mathematical models (Gilbert & Troitzsch 2005), not least that it can model agents as heterogeneous in behaviour rules, attributes and location in pre-existing network structures. When one wants to combine several interacting processes or factors in modelling the behaviour of scientific authors, mathematical analyses become too difficult compared to the programming and exploration of simulation models. In the case of modelling publications, mechanisms that generate the distributions of papers per author, the distributions of citations per paper, the growth over time in numbers of papers and authors, and ideally also the structures in the networks formed by co-authorship and co-citation all need to be combined.

A concept of knowledge seeking as organisational learning

Scientific publications offer the opportunity to inform others of one's findings, obtain validation for one's work through others' responses to it, and provide the starting points and stepping stones for future research. It is hard to represent this aspect of science through stochastic-process models. In the field of organisational studies, however, there is a tradition of constructing computer simulations of humans' collective problem solving, or 'organisational learning' (March 1991; Lazer & Friedman 2007). This offers two additions to the components of a science model: methods of problem solving, and representation of the problems to be solved or the knowledge to be found.

Beginning with Simon's conception of human actors as 'boundedly rational', it has been proposed that problem solving in the workplace involves heuristics (March & Simon 1958; Cyert & March 1963). Psychological studies in the 1970s and 1980s bore out this view (Kahneman et al. 1982). Heuristics are simple principles or rules of thumb for seeking solutions to problems that would require impractical amounts of time and other resources to solve by exhaustive search methods. While not guaranteed to find the optimal solution, heuristic search algorithms are used to obtain sufficiently good solutions within a reasonably short number of search steps. It has been proven that no heuristic algorithm will perform well in every situation (the 'no-free-lunch theorem', Worlpert & Macready 1997), but experience has shown several methods perform well on a variety of problems in which different combinations of values must be explored to find a solution.

The use of relatively simple rules to make decisions among combinations of fixed sets of values make heuristic search processes particularly easy to replicate in computer code. Many simulation models of learning in organisations employ combinations of heuristics, including trial-and-error exploration, learning from successful others (either through direct imitation or some more indirect channel for social influence) and recalling past successful ideas. In March's (1991) model of organisational learning, good search performance for a limited amount of search resources requires a balance between the *exploration* of new views, and the *exploitation* of those already evaluated. If the population of searching agents is too diverse, agents will spend much of their time exploring variations of poor solutions, not good ones. If the agents converge in their solutions too soon, however, search comes to an end, with a consensus solution that may not be very close to the optimum. Constraints on the

processes that lead to convergence thus become important for search performance (Lazer & Friedman 2007). These have included organising agents into social networks (which then restrict who can imitate whom) and restricting imitation so that imitator agents make only partial copies.

The behaviour of scientists publishing within academic fields has also been compared to heuristic search (Scharnhorst & Ebeling 2005; Chen et al. 2009). Various algorithms offer potential analogies for aspects of scientific publication. By insisting on new papers being original contributions to knowledge, authors are forced to explore more widely, in a similar manner to *tabu search* (Glover 1989; 1990). The combination of ideas from multiple co-authors and multiple references produces both exploration of new combinations, but also the exploitation of past experience, perhaps the main attraction of *genetic algorithms* (Mitchell 1996). The sharing between authors of information about past experiences resembles *particle swarm optimisation* (Clerc & Kennedy 2002). Search performance by swarms of agents can be improved through dividing the agents into *tribes*, whose members only communicate information within their own tribe, and *roles*, where 'managers' maintain a record of the best solution found so far while 'workers' concentrate on further exploration (Clerc 2006; Jin & Branke 2005). There may be scope for clustering and stratification among scientists to produce analogous effects.

Sandstrom (1999) compares information seekers to foraging ants, themselves the metaphor for another heuristic search algorithm, *ant-colony optimisation* (Corne et al. 1999). Recently successful foragers attract others to re-use their paths rather than the paths of the less successful foragers or those that are older and potentially out-of-date. Having attracted more foragers to them, the signals to good paths become renewed more often and with greater strength. Thus under a cumulative-advantage principle, relatively short paths to good sources become increasingly easy to identify from their relative popularity. In like manner when constructing reference lists for a new academic paper a preference for the already well-cited causes some papers to emerge as 'citation classics' that other researchers can be relied upon to be familiar with (Merton 1968). This suggests that practices among scientists that generate the Matthew Effect serve to simplify the task of new entrants to a field by selecting the most important texts, and the shortest path to the research frontier. But the organisational learning models also suggest that there may be a need for some balance between exploration and exploitation. Does cumulative advantage operate too fast among scientists?

It may not be possible to answer this question for real scientists. But one *can* begin to answer it for a simulated search of an artificial problem space and then draw an analogy with human systems (Steels 2001).

One analogy is based on the use of *similarity* or proximity. To get accepted by journals, scientists' publications must satisfy two sources of constraint on their contents: originality and similarity. With respect to similarity, they must be intelligible and relevant to readers, especially peer reviewers. With respect to originality, publications must differ from what has been published before, or at least from what a reviewer has read before, but their contents cannot be *too* unfamiliar. To be recognised as a contribution to the journal's field certain keywords, paradigm

problems or classic references must be mentioned because they are the symbols of membership to this field.

It might be asked, however, whether scientists also face some *extrinsic* source of value and constraint for their work, call it 'material reality'. In this conception, scientists' activities have costs in material resources and time, and their publications describe activities and equipment that may be prone to failure and breakages if the science justifying them is in some sense wrong, or out of tune with reality. To simulate problem solving activities addressing this external reality we can borrow the notion of a *fitness landscape*, often used when discussing problem solving using heuristic search methods (Kauffman 1993). The use of various tools and techniques is represented by a set of variables. In the simplest case, these are binary variables, representing presence or absence of some idea, material, technology or practice. The combinations of values of these variables describe the coordinates of a position in some multi-dimensional space. The fitness value of occupying that position, that is, the benefit or cost incurred by employing the particular combination of tools and techniques, can be thought of as the altitude at that point on a landscape. The most desirable combination becomes the tallest peak, but a rugged landscape may have multiple peaks of varying height. A heuristic search method, such as performing a random walk and rejecting any local step that goes downhill, may lead to a nearby peak, but is not guaranteed to find the tallest peak in the whole landscape. The more rugged the landscape, the harder it is to find good peaks. Heuristic search algorithms typically involve many agents each performing searches from various starting positions, sometimes with the sharing of information between search agents about the heights reached.

One fitness landscape suitable for simulating heuristic search is Kauffman's *NK* model. Initially presented as a theoretical model of biological evolution (Kauffman 1993), this has since been reapplied in models of technological evolution (Kauffman 2000), strategic management (Levinthal 1997; Levinthal & Warglien 1999; McKelvey 1999) and organisational learning (Lazer & Friedman 2007). Among its attractions are that it is relatively simple. A solution, or position on the landscape, consists of a string of N binary variables. It uses just one other parameter, K, the number of interdependencies between binary variables. Using this parameter, one is able to 'tune' the model to produce landscapes of varying ruggedness, and thereby varying degrees of difficulty for heuristic search. The use of this landscape in different models also means that it is possible to transfer code between and compare experiences of programs written for different audiences. Against the use of the *NK* model, however, is the fact that it lacks any empirical foundation. Whether one is interpreting it as a model of biological evolution or of organisational learning, the numbers that go into defining the *NK* fitness landscape are arbitrary and have no empirical referent.

That scientists seek better combinations of tools, techniques and other components is plausible enough. The literature from actor-network theorists contains many examples of scientists and other interested parties negotiating the satisfaction of their varied and often conflicting demands (Latour 1987), a pattern repeated in analyses of technological projects (Latour 1996; Law & Callon 1992). Kauffman (2000) also draws an analogy between technological evolution and constraint satisfaction problems. However, if the effectiveness of scientists' search practices is to be replicated in simulations it will be desirable to match as far as possible the structure of

the problems faced by real scientists. Realistic landscapes may be derivable from bibliometric data (Scharnhorst 1998), just as 'maps' of science have been drawn up based on co-citation and co-word relations. Scharnhorst (2002), for example, describes inferring the structure of a 'valuation landscape' from rates of change in the proportion of papers being published in particular areas. So although the *NK* landscape model has sufficed for models of organisation learning, more plausible looking landscapes for science models may yet emerge from future research.

Agent-based science models

Stochastic process models give insight into the generation of the patterns observed in bibliometric data. Models of organisational learning show how heuristic search algorithms applied to fitness landscapes can help with understanding how problem-solving performance in social groups depends upon communication practices, especially those that determine the rate at which past solutions are borrowed. A good science model should aim at combining these processes. It should generate patterns analogous to those seen in real journal publications and it should reflect the fact that scientists' activities serve a purpose, namely that of seeking knowledge or solving problems.

Agent-based simulation models already exist that capture some of these components. Whereas Simon's (1955) urn model simply generated a frequency distribution for papers per author, Gilbert (1997) represented individual academic papers with references to past papers and some contents. Using two continuous variables to represent paper topics, his model depicts an academic field as a two-dimensional plane. Subfields appear within this model as clusters of points. The *TARL* model ('Topics, Aging and Recursive Linking') of Boerner et al. represents both authors and papers, including references and 'topics' for papers, and generates network data. In both of these models, the contents of papers are constrained. For example, papers must be both original, that is, occupy distinct coordinates in the plane, and also sufficiently similar to the papers they refer to, that is, occupy a point within a radius of some given size from their reference papers. But in neither of these models do papers undergo any kind of selection for the extrinsic value of their contents. Drawing on models of organisational learning, we propose to remedy this. Weisberg & Muldoon (2009) also employ landscape search as a model for science. In the 2009 'Modelling Science' workshop at the Virtual Knowledge Studio, Amsterdam, there were several researchers working on simulation models of different aspects of science, including Muldoon on the division of labour among scientists, Payette on modelling 'science as process', and Wouters on the peer review system. (See the presentations available at http://modelling-science.simshelf.virtualknowledgestudio.nl/.) Agent-based models now promise to take the discipline of scientometrics far beyond the scope of stochastic-process models.

Outline of the model

Figure 1 summarises the simulation.[1] At initialisation, a number of 'foundational' publications are written. Being foundational these make no reference to other papers but may be referred to by later papers. Thereafter, at each time step a number of new papers are written. The number added grows geometrically over time at a given rate. For each new paper a number of authors, a number of references to past papers, some contents for the paper and their extrinsic fitness value, and a number of peer reviewers are generated. The paper then undergoes review by the reviewers. The prime determinant of the review's outcome is a paper's fitness as a solution to some extrinsic, complex problem, as defined by a fitness landscape. Papers that satisfy peer review become journal publications. Optionally, the mechanisms for generating authors, references and reviewers can be restricted to publications rather than all papers. Papers compete to become cited through the fitness value of their contents. The selection pressure placed on them is intended to produce ever-fitter solutions or knowledge as the academic field grows.

[Figure 1 goes about here.]

Generating authors

Every paper needs at least one author. As in Simon's urn model (Simon 1955), there is a given chance that this is a new agent with no previous papers in this field. A new agent has no past papers, but does have opinions concerning this scientific field, represented as a bit string of length N=20. If the author is not new, then one is selected from the stock of existing authors, using one of four methods (Table 1). There are two key distinctions involved. Firstly, we distinguish between using a past journal *publication* (options 2 and 4), and using a past, written *paper*, which may or may not have been published (options 1 and 3). Secondly, we distinguish between selecting authors from recent papers/publications (options 1 and 2), thus showing a preference for *recently prolific authors*, and selecting authors from papers contained in the reference lists of recent papers/publications (options 3 and 4), thus favouring *recently well-cited authors*. So depending on which option is chosen, prolific authors may become more prolific (options 1 and 2, a rich-get-richer principle), or writing opportunities may go to authors with *many publications* (option 2), or to those with *many citations* (options 3 and 4), the last being often suggested as a measure of the quality of a publication.

[Table 1 goes about here.]

When selecting authors, preference might be given to the most prolific and recent authors (emphasising recent quantity, not quality), but authors whose output is unread or unrated often command little respect and struggle to attract those resources (doctoral students, research funds, writing sabbaticals) that help in the generation of

[1]The simulation model, *CitationAgents1*, was developed initially in *VBA* within *Excel 2003*, and then, after a break of several months, reproduced using *NetLogo 4.1*. Replicating a simulation model in this way helps to verify that the program is working as intended. The extra work involved in replicating the model was worthwhile, as several minor errors in the original version were exposed. A version of it may be downloaded from *OpenABM*: http://www.openabm.org/model/2470 .

new papers. Selecting from publications rather than papers is one way to ensure that what is chosen has passed a quality assessment. If instead recent citations are preferred, those whose past works are currently in fashion or well read are rewarded. In addition an author may be spurred into action by a new paper critiquing one of his or her own, for conflict in intellectual social circles is particularly energising (Collins 1998). This points towards using options 3 and 4 in the experiments below.

Real authors also age and the author of a citation classic might not be active in the field anymore. Both Gilbert (1997) and Boerner et al. (2004) represent authors as having an 'age' or duration in the field. Authors' ageing may be added to future versions of the model, but for now it is assumed that even early arrivals last the whole of the simulated period (30 years).

To model the *recency* of papers stratified sampling is used. Past papers are weighted for sampling with a Weibull function of the age of the paper. There are several reasons for choosing the Weibull function for definitions of recency (as well as for numbers of authors and references per papers). It is parsimonious, taking only two parameters: alpha, controlling variability, and beta, controlling basic rate. It is faster to compute than certain other functions such as log normal, yet depending on its parameters it can approximate the bell curve of a normal distribution and the skew of a log normal, and produces the negative exponential when alpha is 1. Analysis of bibliometric data (see the next section) suggested it could be fitted via maximum likelihood estimation to the empirical distributions for authors per paper and references per paper. Boerner et al. (2004) employ it to represent *aging*, and we do likewise, but call it *recency*.

It is, however, based on a *continuous* random variable while time steps come in discrete values, as do the numbers of authors and references. To sample discrete values that are approximately Weibull distributed the continuous space is divided into discrete bands of equal width. So if $Weibull_{CDF}(x)$ is the cumulative distribution function, the probability of a discrete random variable taking the non-negative integer value x is given by:

$$P(X=x) = \text{Weibull}_{CDF}(x+1) - \text{Weibull}_{CDF}(x).$$

Papers can have more than one author in the model. The number of attempts to add co-authors varies according to a Weibull distribution. After the first, initiating author has been selected, selection of any co-authors employs the same chosen method described above. Although authors may vary in their beliefs or opinions concerning the field, in this version of the model there are no constraints on which authors may write together.

Generating references

The generation of the list of references is similar to generation of authors. A number of attempts are made to add items to a paper's reference list. A Weibull distribution determines this number. Several options are available for the method of selecting papers to become references (Table 2). As well as selecting any past paper or publication without preference (options 5 and 6), there are the options for selecting a

paper or publication with preference for recency alone (1 and 2), and selecting with preference for the recently cited (3, 4). Again a Weibull distribution's parameters control the definition of 'recent'.

[Table 2 goes about here.]

Two of the options (3, 4) equate to copying references from existing papers. As with Simon's (1955) model and the process for selecting authors, there should be the possibility of introducing new suggestions rather than always copying previous ones. Therefore, for these options, there is a fixed chance that the generated reference is directed at the (recent) paper selected, rather than directed at one of the selected paper's references. This parameter turns out to have some influence over the model's ability to approximate power law distributions of the numbers of citations per paper.

As when authors from past papers were sampled, there are again the options of rewarding the recent or the recently cited, and the publications that satisfied peer review. The organisational learning models (March 1991; Lazer & Friedman 2007) model only the generation of new solutions in one time step using the solution information held in the immediately preceding time step. This is in sharp contrast to science models that allow for copying references to cited papers that are potentially much older than the (recent or otherwise) citing papers. Academic fields vary in their use of older sources, from perhaps physics at one extreme, to footnotes-to-Plato philosophy at the other, suggesting that references can play different roles. For this paper, only their role in providing material for new solutions to problems is modelled, not any role that references to classics might play in signalling membership and evoking a sense of belonging to a tradition.

Generating contents and fitness

The authors and references for a paper are employed in the generation of that paper's contents. Each author has a vector of binary variables representing his or her opinions, beliefs or preferred practices within the field. The contents of papers are encoded as a vector of binary variables of the same length, N. When constructing a new paper, for each variable a value is sampled. With a fixed chance (0.01), this comes from a Bernoulli-distributed random process, in which case it represents the possibility of a new discovery or practice entering from outside the field, and hence not obtained from the literature. Otherwise, the value is sampled from the set of contents of all papers in the references and from the opinions of all authors of the new paper. Thus, like genetic algorithms (Mitchell 1996), the production of content involves both mutation and recombination processes.

When values have been sampled for every variable, a fitness value is calculated for the corresponding bit string. Like Lazer & Friedman (2007) in their model of organisational learning, the fitness value is taken from Kauffman's NK fitness landscapes using Lazer & Friedman's choice of parameters (N=20; K=5), which generates a moderately difficult landscape to explore. Descriptions of this fitness measure have been given in detail elsewhere (Kauffman 1993; 1995; 2000; Levinthal 1997) but we recap briefly here. For each of the N=20 bits or variables there exists a table of fitness values and dependency relations to other variables. A variable's fitness

table has one row for every combination of values (1 or 0) of that bit plus its K=5 dependency variables – i.e. 2^6 = 64 combinations or rows. The network of inter-variable dependency relations is randomly assigned at the start of the simulation. Given the current state of a variable and its K neighbours, the corresponding row of its table is examined. In each row there is a number, set at the start of the simulation by sampling from a uniform distribution [0, 1). This is the current contribution to total fitness for that variable. The actual fitness value for a paper or author is the mean fitness contribution from all N variables or bits.

Given fitness values, papers may be ordered as more or less fit in their contents or solutions, and authors in their beliefs. The fitness values have two consequences. Firstly, if an author has just co-constructed a paper with a better solution than that represented by the author's own beliefs, then the author updates its beliefs with the paper's contents. Secondly, fitness values are compared when peer reviewers evaluate a new paper.

Generating peer review and publication

A completed paper is evaluated to decide whether it will become a journal *publication*. Peer reviewers are selected for a new paper using a method chosen from the same list as that for selecting authors (Table 1). The choice might be between recently active authors, more likely to be junior researchers building their experience of the field, and well-cited, well-regarded senior academics, confident in their interpretation of what should or should not be accepted. Of course, juniors often collaborate with seniors on a paper so the distinction may not be as significant in the real world as it can be in the simulation. We shall focus on option 2 here: preferring as reviewers the authors of recent publications.

Nine attempts are made to find peer reviewers. Papers are rejected if the number of reviewers recommending the paper is below a threshold (set to 3). These numbers are chosen arbitrarily: although the journal *Research Policy* does claim to send submitted papers to three referees, how its editor chooses these we do not know.

A reviewer recommends a paper if:

- the reviewer is not an author of the paper,
- the paper's contents are not identical to the reviewer's own beliefs,
- the paper's contents are not identical to those of any of its referenced papers, and
- the paper's fitness value is not less than that of the reviewer.

Having non-inferior fitness to that of reviewers is a strong requirement. The simulated authors and reviewers have perfectly accurate estimations of the value of their papers and of their own beliefs. The costs and benefits of real academic papers may be much harder to judge and inferior papers do sometimes creep into print, their errors to be identified later. One solution to this modelling issue would be to introduce softer methods of fitness evaluation using stochastic elements, such as those used by the search algorithm of *simulated annealing* (Kirkpatrick et al. 1983), but this would add more parameters to the simulation and is omitted for the present.

Generating output data

The simulation outputs frequency distribution data: authors, references and citations per paper and per publication, and papers and publications per author. It also plots the growth of the field as the numbers of papers, publications and authors over time. Network data on collaboration (co-authorship) and citation relations can also be generated and analysed. In common with the organisational learning models, statistics concerning the fitness of the solutions currently contained in papers and agents' opinions are calculated. By plotting these over time search performance can be compared with the evolution in the field.

Calibrating the model: the case of *Research Policy*

Validation strategy

A good simulation model should provide knowledge and understanding that one would like to have had from a real-world system, but for practical reasons cannot obtain (Ahrweiler & Gilbert 2005). To address what-if questions about the science system, simulation models should occupy the middle ground between being, on the one hand, a detailed replica of some complex social system, and on the other hand some abstract mathematical construction that is difficult to derive real-world implications from. The former involve too much work in designing, programming, validating and computing to be of practical use. To obtain an answer, things need to be left out of the model. On the other hand, a science model needs to be a plausible representation of scientists' activities, and not just a mathematician's fiction. With this caveat in mind, parameter values can be sought that fit the simulation model to some real bibliometric data. This step is common to previous presenters of science models (Simon, 1955; Price 1976; Gilbert 1997; Boerner et al. 2004). The speed of the simulation is such that trial-and-error exploration of the simulation's parameter space suffices for obtaining the following fits. To achieve this, however, the model is simplified in one important respect: agents' problem-solving capabilities are omitted by setting all fitness values to 1, irrespective of paper contents or author beliefs. This means that peer review is doing nothing more than checking for originality. The *NK* fitness landscape is then reintroduced, but data fitting while using the fitness landscape is a much harder task. A summary of the parameters employed in the simulations is shown in Table 3.

[Table 3 goes about here.]

Growth over time

Desiring a small-scale simulation for faster runtimes during experiments, we took data from *ISI Web of Science* for a single journal, *Research Policy*, which sits at the heart of its particular field, innovation studies. Founded in 1974, this journal has shown fairly steady increases in its growth, in terms of both the number of papers per year

and the number of authors per year. Taking 1974 to be the model's year 1 (so omitting the foundational papers used for initialisation, which appear in year 0), Figure 2 shows the number of papers per year for each year thereafter, and the total number of papers for a typical run of the simulation, as well as the real data from *Research Policy* (hereafter *RP*). The results were obtained assuming 14 foundational papers, 16 papers in year 1, and each year thereafter the number of new papers was 1.067 times that of the year before. After 30 simulation iterations the number of papers generated was 1432, comparable with the 1389 papers published in *RP* by end of 2003, or model year 30. So far, this shows only that the field is growing exponentially.

[Figure 2 goes about here.]

Matching the number of authors per year is also straightforward. By 2003, the simulation model's Weibull function approximates the actual distribution of authors per paper for each year (Figure 3). The mean number of authors per paper rose slightly during the 30-year period for *RP*, but for simplicity constant parameter values were assumed in the simulation: *alpha* = 1.4; *beta* = 1.3. The number of authors available for writing the papers grows over time. Starting with an empty field, all the authors in the journal's volume for 1974 must be new to that journal. However Figure 4 shows a fall over time in the proportion of authors in a year's volume who are new, that is, publishing in the journal for the first time. Again for simplicity, the simulation assumes a constant chance of an author of a paper being new to the field: 0.6. This assumption of a constant chance is common to the models by Simon (1955), Gilbert (1997) and Bentley et al. (2009). The combination of number of authors per paper and proportion of authors who are new to the field gives the growth in authors seen in Figure 5, with the corresponding figures for *RP*.

[Figure 3 goes about here.]

[Figure 4 goes about here.]

[Figure 5 goes about here.]

Distributions: authors and papers

When the authors for a new paper are not new to the journal, they can be found from earlier papers. For sampling them, options 2 or 4 from Table 1 will generate a similar distribution. Figure 6 (first chart) shows the distribution from a typical run for option 4, plotted against the empirical distribution from *RP*. In addition, Figure 6 (second chart) shows a line with an exponent of -3.07, the average exponent from power laws fitted to the results of 20 simulation runs. The exponents for the 20 fits ranged from 2.99 to 3.18.

[Figure 6 goes about here.]

One other aspect of authorship remains to be defined, that of *recency* when selecting publications to obtain authors. To determine this, relations across time between papers having a common author are examined. Taking all pairs of papers that share at least

one author and restricting attention to those pairs where the latest paper of the pair was published in year 30, Figure 7 shows plots for the simulation and *RP* of the distribution of time gaps between these common-author papers. Like the papers-per-author distribution, and unlike the growth curves and authors-per-paper distribution, the simulation's distribution is a non-trivial outcome of its workings. Decisions concerning the method of selecting authors, including the definition of recency, will affect this distribution. The weighting of 'recent publications' used a Weibull function of the time since their publication with parameters *alpha* = 1.3 and *beta* = 1. The simulation output shown is the aggregate results of 20 simulation runs with 95% confidence intervals for each data point.

[Figure 7 goes about here.]

Distributions: references and citations

Turning now to the generation of references, there is a problem. The papers of any journal contain references to papers in other journals, including those written prior to the foundation of the target journal. To simulate the journals outside the target would be prohibitively complicated. The solution taken here is to restrict attention to references that point to other papers inside the target journal. Other modellers might try different solutions.

Having made this simplification, there are several data-fitting tasks analogous to those faced for authors-related distributions. The distribution of the numbers of references per paper can be taken straight from the bibliometric data. The Weibull distribution has again been used for this (*alpha* = 1, so equivalent to a negative exponential distribution; *beta* = 4.2). The distribution of time gaps between citing papers in year 30 and the papers cited by their references is used to guide the choice of parameters to define recency, when selecting past papers for their references. To generate the distribution in Figure 8 'recent publications' were defined using a Weibull function with parameters *alpha* = 1.3 and *beta* = 2. The mean figures from 20 simulation runs suggest papers published more than 20 time steps ago are receiving slightly too few citations, but the simulation has captured the general peak-and-decay pattern.

[Figure 8 goes about here.]

As in the model of Boerner et al. (2004), the initial or *foundational* papers in the simulation provide a means for representing papers outside the target dataset. An examination of the numbers of citations received by papers in each year (Figure 9) shows that the papers in the model's year 1 tend to receive slightly more citations than those in the next few years, despite the number of new papers in those years increasing gradually. (For these charts foundational papers have been included, and they receive far more citations.) Thereafter the curve rises and falls with the empirical data, though the simulation tends to generate too many citations compared with the empirical case.

[Figure 9 goes about here.]

The power law for papers per author was obtained with the help of a parameter for the chance of an author being new to the journal. Analogously for references, there is a chance of 0.3 of using a 'recent' publication as a reference in a new paper. If the recent paper is not used as the reference, then one of its own references is copied to the new paper. Thus papers that have recently been much cited are likely to be cited again, but there still exists a chance of a, potentially as yet uncited, recent publication gaining a citation. Twenty simulation repetitions were run and power laws were fitted in each case using maximum likelihood estimation (Figure 10). The exponents ranged from 1.75 to 1.81, with a mean of 1.80, whereas the data from *RP* call for an exponent in the range 1.7 to 1.8. Compared to *RP* the simulation tends to produce fewer papers with just one citation, and there is more spread in the single papers with high numbers of citations.

[Figure 10 goes about here.]

The distribution is sensitive to changes in the parameter that sets the chance of references pointing to recent papers rather than to papers referenced by recent papers. With extreme values, the results are far from a power law (Figure 11). If the recent paper rather than one of its references (equivalent to methods 1 and 2 in Table 2) is selected, then the resulting distribution is exponential, not power law. If the recent paper's referenced paper, never the recent paper itself, is chosen, then the distribution bends away from the kind of power law that would fit *RP*.

[Figure 11 goes about here.]

To summarise, using empirically grounded assumptions for growth in papers and in authors, and distributions for authors per paper and references per paper, so far the distributions of papers per author, citations per paper, the time gaps between authors' papers, the ages of papers when chosen for references, and the number of citations received per year found in *RP* has been matched closely by the output from the simulation. These fits were achieved through processes for selecting authors and references that included choosing parameters for the Weibull-distributed weighting of past papers in the definition of what constitutes a 'recent paper' (see Table 3), but also the choice of procedure for sampling papers to become new references. Papers were mostly selected with preference for those with recent citations. This is clearly better than selecting recent papers without regard to their citations, which fails to generate a plausible distribution of citations per paper, but over-concentration on citations would also generate the wrong distribution. So, although there is no guarantee that the former generation mechanism is the best, it is clear that some methods give distributions that clearly fail to fit.

Introducing fitness and peer review

So far, peer review has played little part in the model. Papers may be rejected by reviewers for being unoriginal, but this is a very weak constraint. With N=20 bits of information in the contents of each simulated paper, there are 2^{20}, or over one million, possible combinations of binary values for papers to explore during 30 time steps, and so a paper is rarely rejected for matching a reviewer's beliefs. Once *NK* fitness is

introduced, however, peer review places a significant selection pressure on papers. As Figure 12 shows for a typical simulation run, in most years only about 20% of the generated papers are accepted as publications. If methods for selecting authors or references are restricted to journal publications rather than including unpublished material (representing working papers, unreviewed conference proceedings, and drafts), reviewing for fitness is likely to have some impact on the model's ability to match the empirical distributions.

[Figure 12 goes about here.]

This turns out to be the case. The number of publications per year, the number of citations received and the time-gap distributions no longer match the *RP* distributions. To compensate for this, paper production can be increased, by raising year 1's papers from 16, in anticipation of the fact that only 20% will be published. Since each of these extra papers needs authors, the chance of an author being new to the journal and the impact of the recency functions may also have to be changed. However, raising the number of papers increases computation times for the simulation.[2] Instead of results being returned in seconds on a modern PC, they can take minutes, and the simulation no longer encourages user interaction. Further data fitting will have to wait for faster computers, or for the addition of automated methods for searching the parameter space.

We do not know what proportion of papers submitted to *Research Policy* is accepted. We would not expect to match it: the representation of contents and extrinsic fitness value was not intended to be that realistic. But data supplied by the journal *Management Science* suggest that an acceptance rate of roughly 10% is plausible for that journal. It has been found that many papers rejected from one journal eventually make it into print in another journal, typically one of lesser status (Bornmann 2010). Otherwise, bibliometric data do not generally reveal about what happens outside of publication success.

Experimenting with cumulative advantage

We turn next from generating plausible distributions to exploring the impact on search performance of the choice of method for generating reference lists in simulated papers.[3] Keeping the parameter values found during calibration, we experiment with varying the reference-selection method. The fitness of the best solution found during each simulation run is recorded. Figure 13 shows mean results for 200 simulation runs, with 95% confidence intervals for each mean, for each of the methods in Table 2. It is clear that methods involving publications (even-numbered methods) beat their

---

[2] The exact relation between computing time and model scale is difficult to state, and differs between the *VBA* and *NetLogo* versions, for reasons internal to those software environments.

[3] The options for generating author lists also include those that produce cumulative advantage (in papers per author), but preliminary testing showed that these mechanisms have smaller effects than those for generating references, perhaps because papers have more references than authors on average.

paper-based equivalents (odd-numbered). Secondly, the method associated with the scale-free distributions of citations per paper (method 4) is not the best, but the difference between it and method 2, which is known to generate a distinct distribution, is not statistically significant. So in this experiment the cumulative-advantage process for generating citations failed to make a noticeable difference to the search performance of the agents. The preference for recency in methods 2 and 4 helps when compared with method 6, which samples from any past publication with no preference for recency. However, the most important aspect of generating references is still the use of publications, that is, the use of filtering provided by peer review.

[Figure 13 goes about here.]

To test the sensitivity of these results, alternative methods of generating references were tried. For an alternative to methods 1 and 2, instead of sampling papers (or publications) with preference for recency, a recent year can be randomly chosen, then any paper (or publication) from that year selected. This would simulate a reference seeker who went to a particular journal volume simply because it was recent, and who ignored how many papers the volume had in it. This alternative did not differ statistically significantly from methods 1 and 2. As an alternative to methods 3 and 4, sampling stratified by recency and citations received was used, as opposed to stratification by citations received from recent papers. Methods 3 and 4 allow for generating references to 'citation classics' that are not themselves recent, but have recently been in fashion. The alternative method prefers well-cited items that are still themselves recent. Again, the results from the alternative method failed to differ statistically significantly. However, a field with a different growth rate might result in some differences. We leave this investigation for further work.

Discussion and future work

The model described in this paper has the ability to generate distributions and other patterns similar to those found in bibliometric data, including papers per author and citations per paper. Through its simulation of peer review it has the ability to perform searches on fitness landscapes, an analogy for problem solving. As demonstrated here, the pursuit of the second task turns out to disturb that of the first task. Obtaining distributions that fit a particular set of bibliometric data while searching the current, artificial fitness landscape would require further work, not least because of the extra computing time needed as the number of papers increases. Instead, we examined whether searching an artificial fitness landscape was affected by the choice of methods for generating references and citations. As it turned out, the impact on search performance was not *statistically* significant for cumulative-advantage processes. Whether it is significant in the sense of *important* is a question to be asked when we can replace *NK* fitness with landscapes that have some empirical grounding. By contrast, the use of publications rather than papers, that is, choosing to refer only to papers that have passed through peer review, did make a noticeable difference.

Of course, real academic authors and reviewers might not be employing the *optimal* publishing practices. They make choices about their own practices, and there is no

guarantee that their attempts to pursue their own individual interests will lead to the best *collective* performance. Nevertheless, simulation models of academic publication can suggest areas in which different practices might come with different costs. These suggestions can then be turned into theories to be investigated in further research, including through other approaches. When the practices that performed best in the simulation are not followed by real scholars, why is this so? What extra utility to those practices is not captured by the theory represent by the model?

Our simulation has a wealth of options and parameters, perhaps inevitably given the need to bring together several different processes from the science system. Even so we omit or simplify aspects of the generation of scientific papers, as well as other activities related to the dissemination of scientific knowledge such as conferences and teaching. But each choice of process or parameter value can be debated, with arguments drawing upon empirical sources, such as ethnographic studies of scientists at work (Latour 1987) and, of course, scientometric analyses. Clearly the exploration of this model has only just begun, but we pick out here a few suggested directions for future research, several of them demonstrating the interplay of endogenous and exogenous influences on the model's workings.

What are the effects on search performance of varying the parameters underlying the generation of recency and the numbers of co-authors, references and reviewers? For example, what would be the impact of a limit on the number of references per paper? At present, most authors can choose the number of references, and those that supply a higher number wield a larger amount of influence in determining which papers will become citation classics, and who among the field will receive the best citation count (Fuller 2000). Work in filtering the list of publications in a field and reducing it to a more manageable size for future researchers and students is not rewarded. Profligate reference creators are not penalised. Is there scope for gaming the system to favour one's own allies or research interests? Do some methods of selecting peer reviewers lead to elites controlling what gets published in a field?

There are several issues to address concerning how a science simulation represents connections to things outside its area of focus. For example, to keep the program fast and responsive, we only simulated the generation of as many papers as are found in a single journal. Generated references in these papers were to previous papers in the simulation. Contents for new papers were largely sampled from these references. The model's sensitivity to the number of foundational papers and the chance of innovation during contents sampling - the only extrinsic sources of information - should be examined, but better approaches to representing a journal's connections to other literature may exist.

The model's sensitivity to different types of landscape should be also investigated, both before and after learning more about the structure of real epistemic landscapes. For the *NK* landscape, increasing *K*, the number of interdependencies between components, is known to lead to more rugged landscapes, and more difficult searches. How does search difficulty affect distributions, and collaboration and citation networks? Will there be an analogue to Whitley's (2000) finding that the social organisation of scientists varied with the degrees of task uncertainty and mutual dependency faced by those scientists? For sufficiently long simulation runs, the difficulty of finding improvements and getting published will increase. Will this lead

to identifiable stages of growth in the research field, like those described by Mulkay et al. (1975): exploration, unification and decline / displacement?

As Scharnhorst's (2002) two-landscape conception of science highlights, the value of a particular position in science is dependent on the occupation by others of that and surrounding positions. Our model included one such constraint on where an agent can publish, that of originality. Future work will consider the impact of a requirement for a degree of similarity between publications and authors. Papers will be rejected if their contents are too unlike what has gone before in the referenced papers or in the experience of reviewers and potential co-authors. The evidence from Axelrod's cultural model (Axelrod 1997; Castellano et al. 2000), opinion dynamics models (Hegselmann & Krause 2002) and Gilbert's (1997) science model suggests that homophily (McPherson et al. 2001) or requirements for similarity or proximity, in our case defined as a required number of matching bits, will lead to the formation of clusters among publications and cultural groups among their authors. To be accepted by particular agents will require that many of their cultural practices, such as their terminology, techniques and assumptions, are matched by authors while still presenting them with something at least slightly novel. If originality and similarity are the only constraints on scientific publications, then the value of a paper's contents is endogenous to science and the meaning of a publication event is a social construction generated by the surrounding publication events. This is the situation represented by Gilbert's (1997) model and also by the *TARL* ('Topics, Aging and Recursive Linking') model of Boerner et al. (2004) which arbitrarily assigned 'topics' to each paper and to each author, then used these to restrict who could publish with whom and on what topic. Ideally a science model should combine these endogenous influences with exogenous influences, such as a fitness landscape.

A new, exogenous source of constraint on publishing activity would be social networks among author agents, be they informal, institutional, geographical, linguistic or cultural in origin. For example, authors who do not share a common language are unlikely to co-author a paper together. As Lazer & Friedman (2007) demonstrated, the network structure among agents can affect search performance. The presence of networks, whether endogenous or exogenous, will also raise questions of social capital. Do positions of brokerage and closure in the network (Burt 2005; Chen et al. 2009) lead to different roles in the generation and diffusion of novel ideas in the simulation?

As these proposals make clear, simulation models connecting, on the one hand, science models and bibliometric data, to, on the other hand, organisational learning models and heuristic search on fitness landscapes should provide a fruitful avenue for much research. By connecting these areas, understanding is gained into the extent to which authoring and publishing practices affect the ability to explore and exploit the range of positions in science.

Acknowledgements This research was supported by *SIMIAN* (Simulation Innovation: A Node), a part of the UK's National Centre for Research Methods, funded by the Economic and Social Research Council.

References

Ahrweiler, P., Gilbert, G. N. (2005) Caffè Nero: the evaluation of Social Simulation. *Journal of Artificial Societies and Social Simulation*, 8(4)14, http://jasss.soc.surrey.ac.uk/8/4/14.html.

Axelrod, R. (1997) The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton: Princeton University Press.

Barabási, A.-L., & Albert, R. (1999) Emergence of scaling in random networks. *Science*, 286, 509–512.

Bentley, R. A., Ormerod, P., Batty, M. (2009) An evolutionary model of long tailed distributions in the social sciences. *arXiv:0903.2533v1 [physics.soc-ph] 14 Mar 2009* http://arxiv.org/PS_cache/arxiv/pdf/0903/0903.2533v1.pdf . Accessed 1 May 2010.

Boerner, K., Maru, J. T., Goldstone, R. L. (2004) The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Science USA*, 101(suppl. 1), S266-S273.

Bornmann, L. (2010) Does the Journal Peer Review Select the 'Best' from the Work Submitted? The State of Empirical Research. *IETE Technical Review*, 27(2), 93-96.

Burrell, Q. L. (2001) Stochastic modelling of the first-citation distribution. *Scientometrics*, 52(1), 3-12.

Burrell, Q. L. (2007) Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1, 16-25.

Castellano, C., Marsili M., & Vespignani, A. (2000) Nonequilibrium Phase Transition in a Model for Social Influence. *Physical Review Letters*, 85(16), 3536-3539.

Burt, R. (2005) Brokerage and Closure: An Introduction to Social Capital. Oxford: Oxford Univerity Press.

Chen, C., Chen, Y., Horowitz, H., Hou, H., Liu, Z., & Pellegrino, D. (2009) Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3, 191-209.

Clerc, M. (2006) *Particle Swarm Optimisation*. London: ISTE.

Clerc M., Kennedy, J. (2002) The Particle Swarm—Explosion, Stability, and Convergence in a Multidimensional Complex Space. *IEEE Transactions on Evolutionary Computation,* 6(1), 58-73.

Collins, R. (1998) *The Sociology of Philosophies: a global theory of intellectual change*. London: Belknap Press, Harvard University Press.

Corne, D., Dorigo, M., & Glover, F. eds. (1999) *New Ideas in Optimisation*. London: McGraw-Hill.

Cyert, R. M., & March, J. G. (1963) *A behavioral theory of the firm*. Englewood Cliffs, New Jersey: Prentice-Hall.

Fuller, S. (2000) *The governance of science*. Buckingham: Open University Press.

Gilbert, N. (1997) A Simulation of the Structure of Academic Science. *Sociological Research Online*, 2(2)3, http://www.socresonline.org.uk/socresonline/2/2/3.html .

Gilbert, G. N., & Troitzsch, K. G. (2005) *Simulation for the Social Scientist*. Maidenhead, UK: Open University Press.

Glaenzel, W., Schubert, A. (1990) The cumulative advantage function. A mathematical formulation based on conditional expectations and its application to scientometric distributions.*Informetrics*, 89/90, 139-147.

Glaenzel, W., Schubert, A. (1995) Predictive aspects of a stochastic model for citation processes. *Information Processing & Management*, 31(1), 69-80.

Glover, F. (1989) Tabu Search — Part I. *ORSA Journal on Computing*, 1(3), 190-206.

Glover, F. (1990) Tabu Search — Part II. *ORSA Journal on Computing*, 2(1), 4-32.

Hegselmann, R., Krause, U. (2002) Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation. *Journal of Artificial Societies and Social Simulation*, 5(3)2, http://jasss.soc.surrey.ac.uk/5/3/2.html.

Jin, Y., & Branke, J. (2005) Evolutionary Optimization in Uncertain Environments – A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3), 303-317.

Kahneman, D., Slovic, P., & Tversky, A. (1982) *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.

Kauffman, S. (1993) The Origins of Order: Self-Organization and Selection in Evolution. New York: Oxford University Press.

Kauffman, S. (1995) At Home in the Universe: The Search for Laws of Complexity. London: Penguin.

Kauffman, S. (2000) *Investigations*. Oxford: Oxford University Press.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983) Optimization by Simulated Annealing. *Science*, 220, 671-680.

Latour, B. (1987) *Science in Action*. Cambridge MA: Harvard University Press.

Latour, B. (1996) *Aramis or the love of technology*. Cambridge MA: Harvard University Press.

Law, J. & Callon, M. (1992) The life and death of an aircraft: a network analysis of technical change. In Bijker, W. & Law, J. (Eds.), *Shaping Technology / Building Society: Studies in Sociotechnical Change*,(p.21-52). London, Cambridge MA: The MIT Press.

Lazer, D., & Friedman, A. (2007) The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly*, 52, 667-694.

Levinthal, D. A. (1997) Adaptation on Rugged Landscapes. *Management Science*, 43(7), 934-950.

Levinthal, D. A., & Warglien, M. (1999) Landscape Design: Designing for Local Action in Complex Worlds. *Organization Science*, 10(3), 342-357.

Lotka, A. J. (1926) The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences*, 16, 317-323.

March, J. G. (1991) Exploration and Exploitation in Organisational Learning. *Organization Science*, 2(1), 71-87.

March, J. G., & Simon, H. A. (1958) *Organizations*. New York, London: Wiley.

McKelvey, B. (1999) Avoiding Complexity Catastrophe in Coevolutionary Pockets: Strategies for Rugged Landscapes. *Organization Science*, 10(3), 294-321.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001) Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27, 415-444.

Merton, R. K. (1968) The Matthew Effect in Science. *Science*, 159(3810), 56-63.

Merton, R. K. (1988) The Matthew Effect in Science, II. Cumulative Advantage and the Symbolism of Intellectual Property. *ISIS*, 79, 606-623.

Mitchell, M. (1996) *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.

Mulkay, M. J., Gilbert, G. N., Woolgar, S. (1975) Problem Areas and Research Networks in Science. *Sociology*, 9, 187-203.

Newman, M. E. J. (2001a) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64, 016131.

Newman, M. E. J. (2001b) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64, 016132.

Newman, M. E. J. (2001c) The structure of scientific collaboration networks. *Proceedings of the National Academy of Science USA*, 98(2), 404-409.

Newman, M. E. J. (2003) The structure and function of complex networks. *SIAM Review* 45, 167–256.

Price, D. J. de Solla (1963) *Little Science, Big Science*. New York, London: Columbia University Press.

Price, D. J. de Solla (1965) Networks of Scientific Papers. *Science*, 149(3683), 510-515.

Price, D. J. de Solla (1976) A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, 27(Sep.-Oct.), 292-306.

Redner, S. (1998) How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*, 4, 131-134.

Sandstrom, P. E. (1999) Scholars as subsistence foragers. *Bulletin of the American Society for Information Science*, 25(3).

Scharnhorst, A. (1998) Citation–Networks, Science Landscapes and Evolutionary Strategies. *Scientometrics*, 43(1), 95-106.

Scharnhorst, A. (2002) Evolution in Adaptive Landscapes - Examples of Science and Technology Development. Discussion Paper FS II 00 - 302, Wissenschaftszentrum Berlin für Sozialforschung.

Scharnhorst, A., & Ebeling, W. (2005) Evolutionary Search Agents in Complex Landscapes. A New Model for the Role of Competence and Meta-competence (EVOLINO and other simulation tools). http://arxiv.org/abs/physics/0511232. Accessed 16 April 2010.

Schubert, A., & Glaenzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, 6 (3), 149-167.

Simon, H. A. (1955) On a Class of Skew Distribution Functions. *Biometrika*, 42(3/4), 425-440.

Steels, L. (2001) The Methodology of the Artificial. Commentary on Webb, B. (2001) Can robots make good models of biological behaviour? *Behavioral and Brain Sciences*, (2001), 24(6). http://www.csl.sony.fr/downloads/papers/2001/steels.html. Accessed 5 May 2008.

Watts, D. J. (2004) The 'New' Science of Networks. *Annual Review of Sociology*, 30, 243–70.

Watts, D. J., & Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, 393, 440–442.

Weisberg, M.& Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science*, 76 (2), 225-252.

Whitley, R. (2000) The Intellectual and Social Organization of the Sciences. Oxford: Oxford University Press.

Wolpert, D. H., & Macready, W. G. (1997) No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67-82.

Zuckerman, H. (1977) Scientific Elite: Nobel Laureates in the United States. New York: Free Press.

Tables and Figures

(Note to copyeditor: Many figures, such as Figure 2, consist of two plots. These may either be placed side by side, or one above the other. If one above the other, the captions need to be changed to refer to '(top') and '(bottom)', rather than '(left)' and '(right') as they are at the moment.)

```
Generate foundational publications
For each time step
        For each new paper
                Generate first author
                Generate list of co-authors
                Generate list of references
                Generate contents
                Calculate fitness of contents
                Generate list of peer reviewers
                Decide whether paper becomes journal publication
        Next new paper
        Output periodically statistics on simulation evolution
Next time step
Output final statistics and distributions
```

Figure 1 Stages in the simulation model

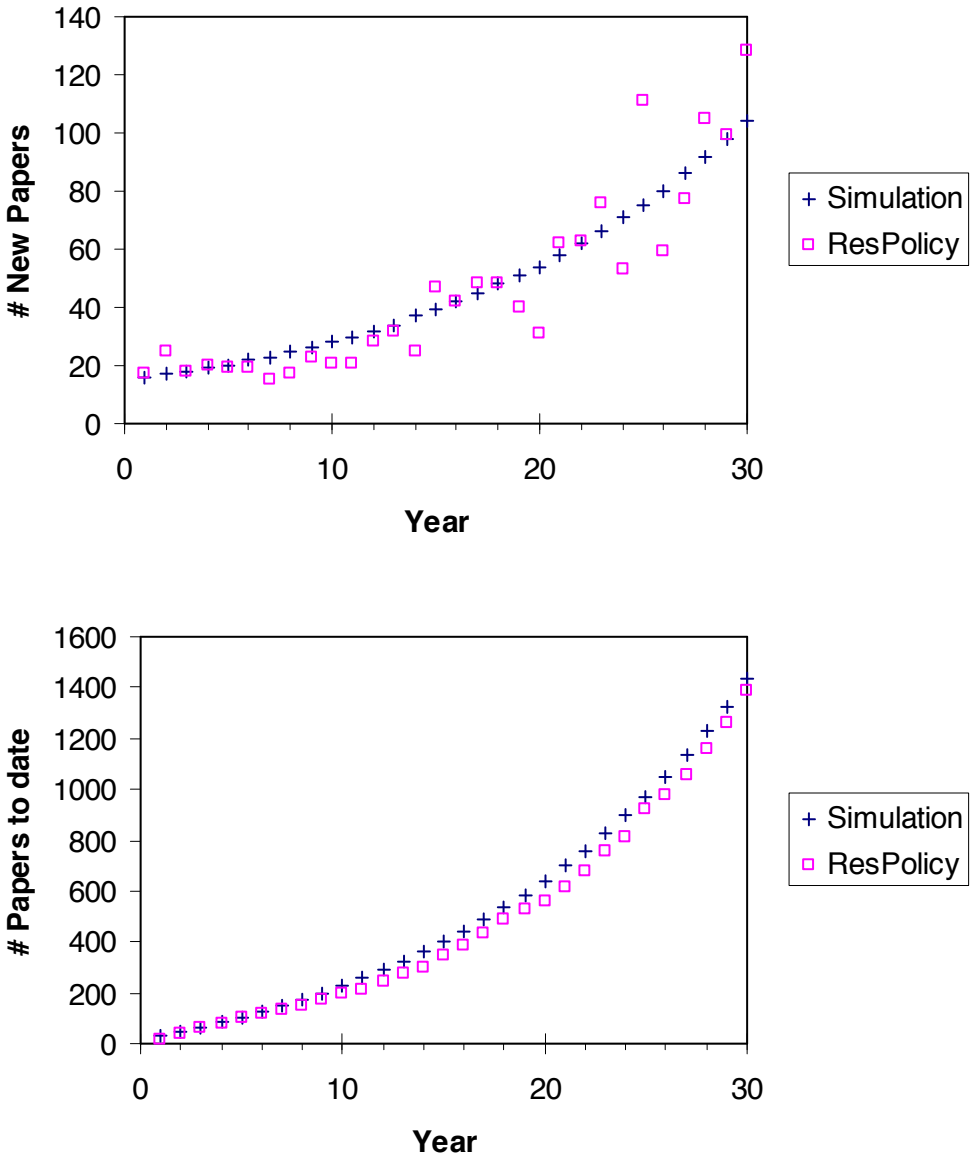

Figure 2 Field growth: (left) the number of papers added to the field each year and (right) the total number of papers to date. Output from a typical simulation run (crosses) is shown with actual data from the journal *Research Policy* (squares).

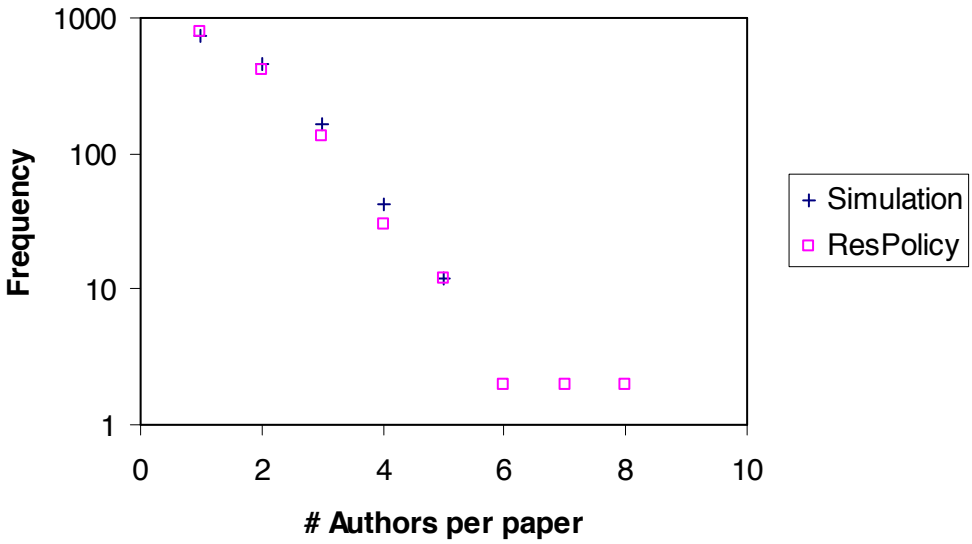

Figure 3 Number of authors per paper after 30 years: probability density functions fitted for a typical run of the simulation and for *Research Policy*

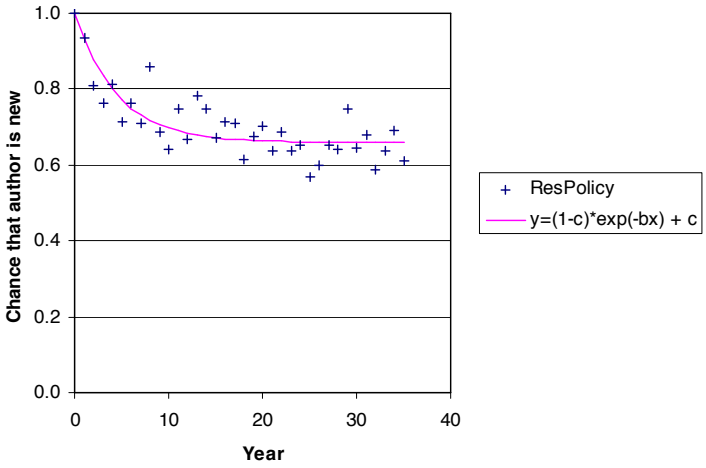

Figure 4 Proportion of the authors publishing in each year who are new to the journal *Research Policy*

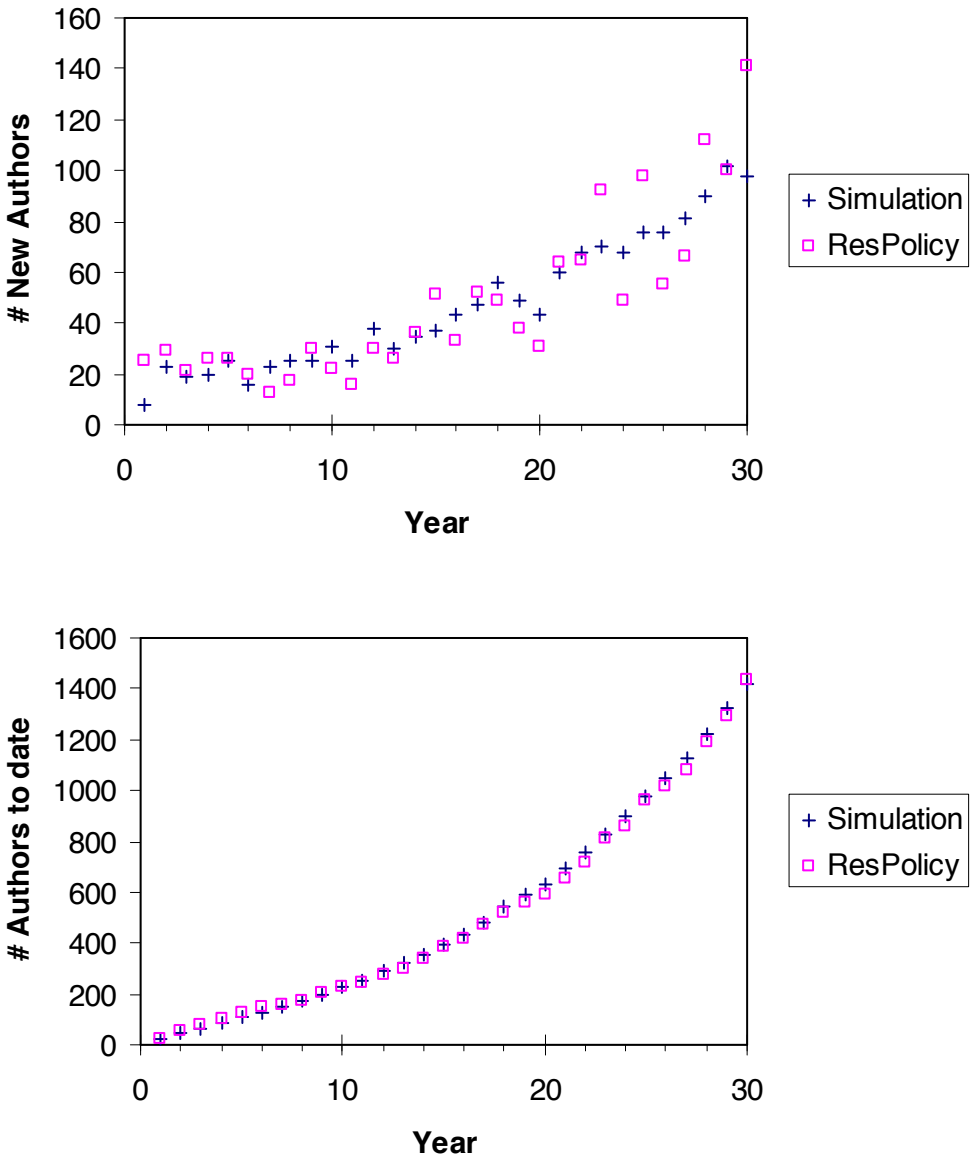

Figure 5 Growth in authors in a typical run of the simulation and in *Research Policy*: (left) new arrivals for each year; (right) total authors to date

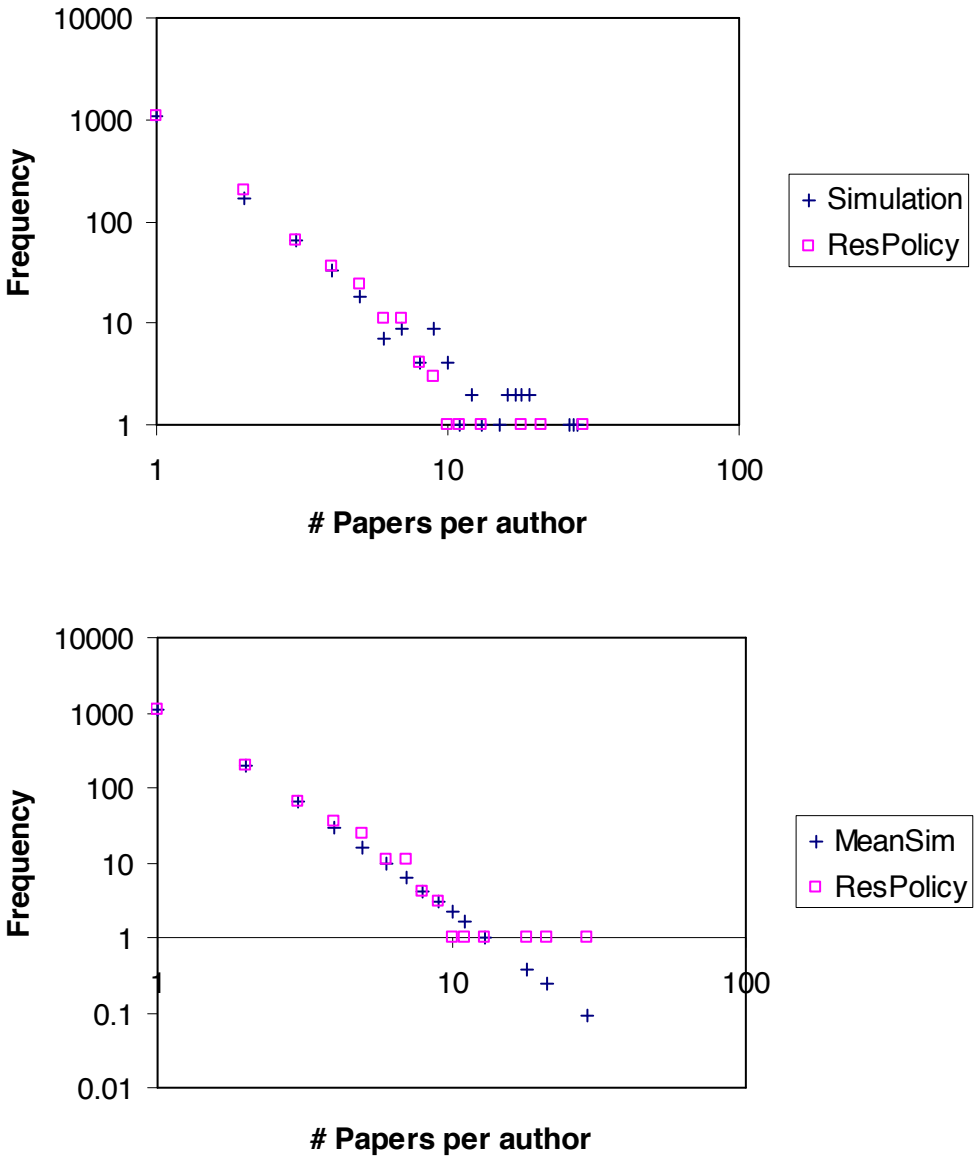

Figure 6 Frequency distributions of the numbers of papers per author: (left) a typical simulation run; (right) curve taking the mean exponent from power laws fitted to each of 20 simulation runs

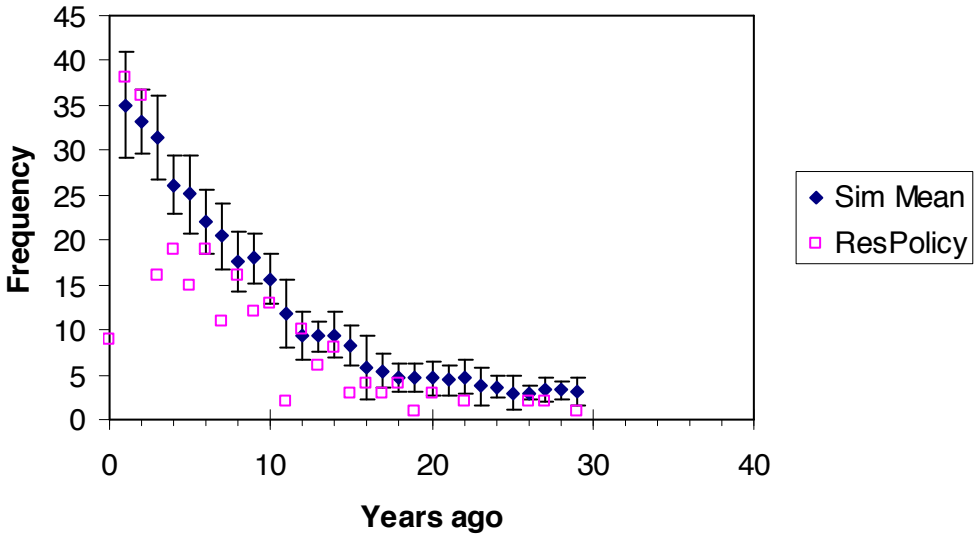

Figure 7 The time between authorship events. For all pairs of papers one published in year 30 and sharing at least one author, frequency distributions of the ages of the earlier paper: mean results from 20 simulation runs together with 95% confidence intervals for each point, and corresponding data from *RP*

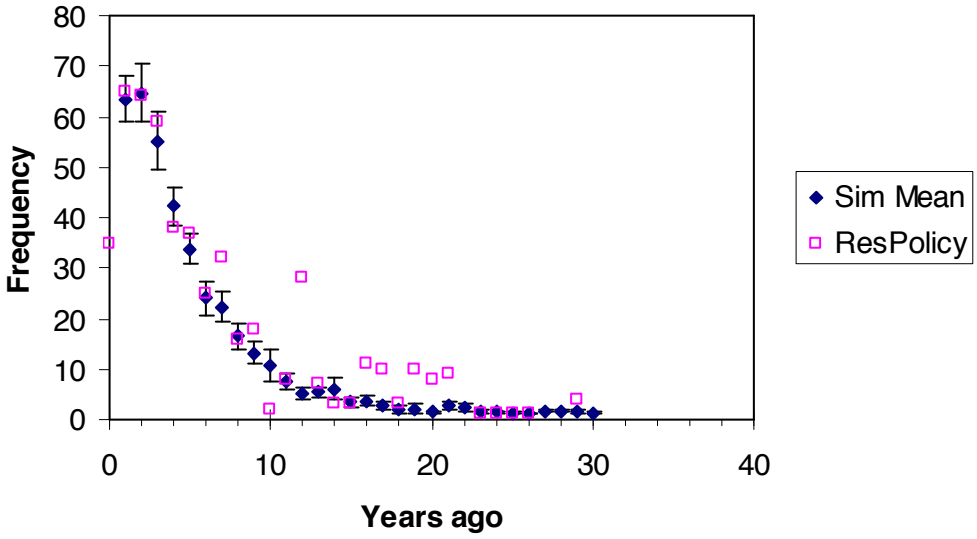

Figure 8 The time between citing and cited papers. Frequency distributions of the ages of all papers referenced by papers published in year 30 for mean results from 20 simulation runs, with 95% confidence intervals indicated, and corresponding data from *RP*

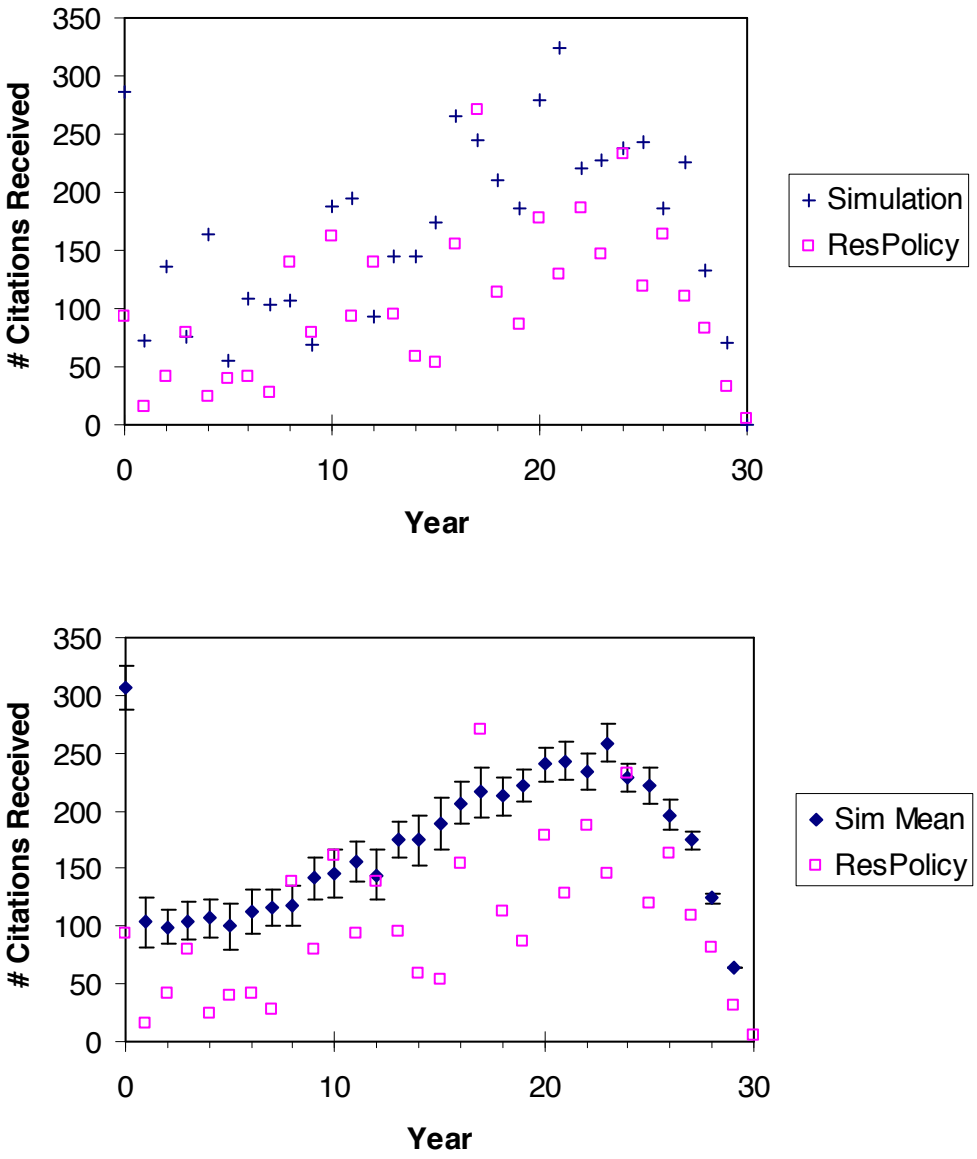

Figure 9 Citations received by papers in each year: (left) a typical simulation run; (right) mean results from 20 simulation runs, plus 95% confidence intervals

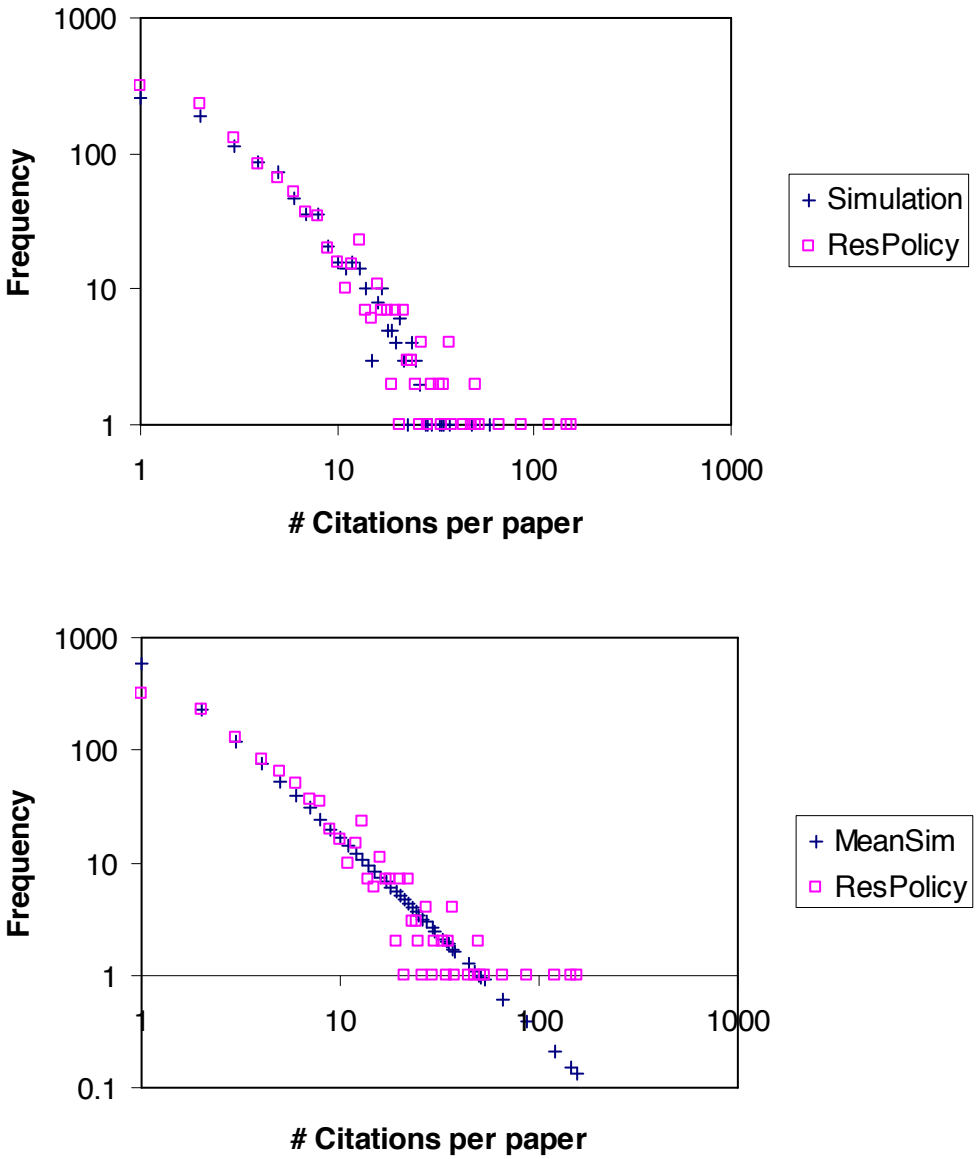

Figure 10 Frequency distributions for numbers of citations per paper: (left) a typical simulation run; (right) curve taking the mean exponent from power laws fitted to each of 20 simulation runs

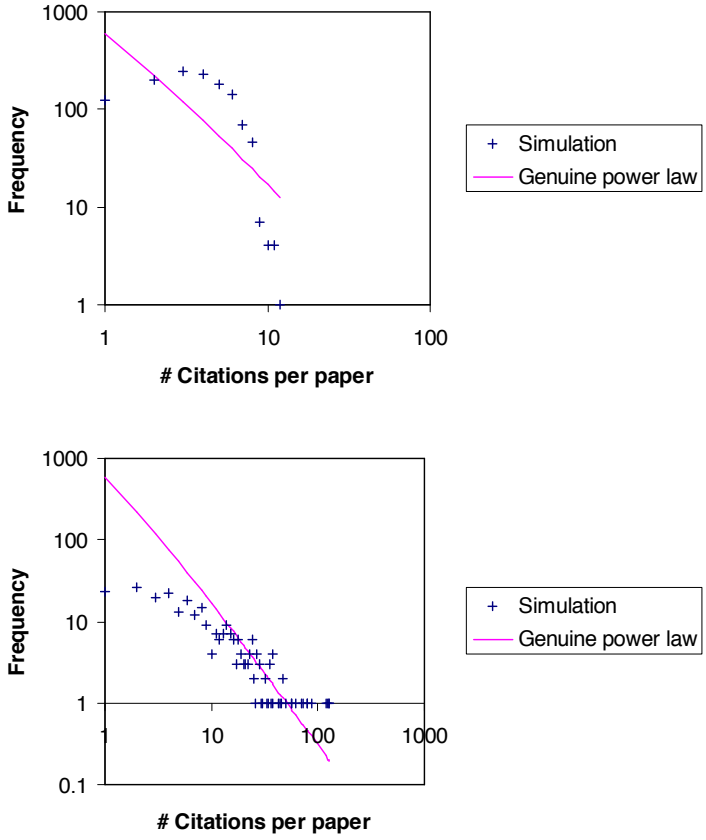

Figure 11 Citations per paper, but no power law. Results of (left) always choosing the recent paper instead of its reference, and (right) always copying the recent paper's reference, never the recent paper itself. The lines indicate what power laws would look like, and have gradients similar to those fitted to *RP*.

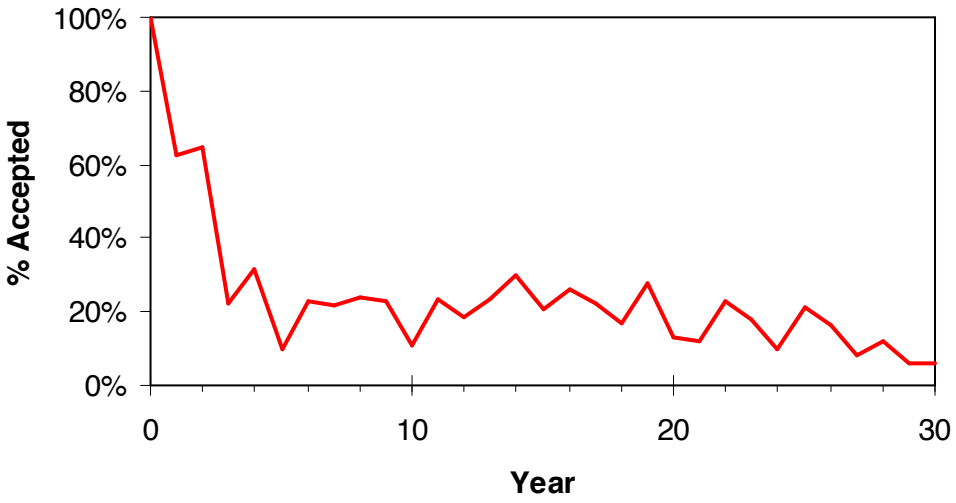

Figure 12 Percentage of papers being accepted for publication per year in a typical simulation run that includes a fitness function and peer review process

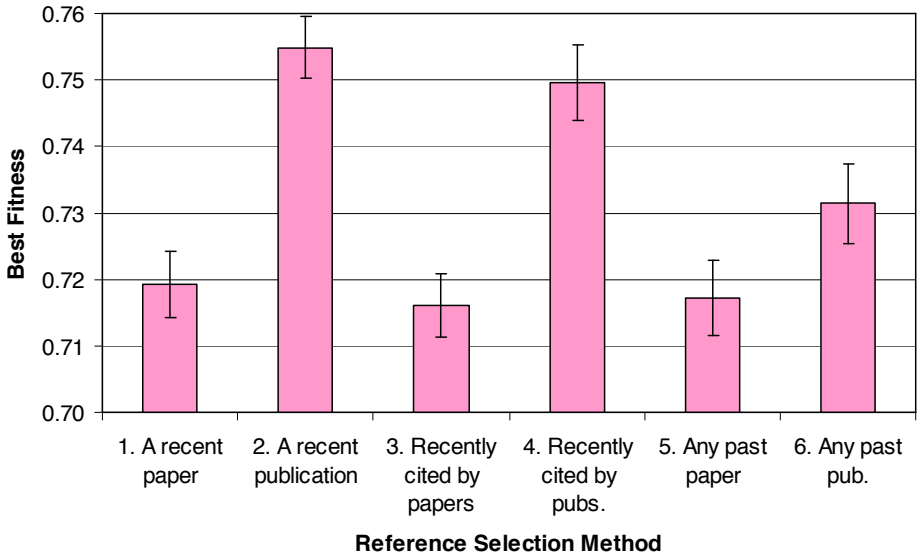

Figure 13 Best fitness found using various methods for selecting papers for references. Method numbers are those in Table 2. Results shown are the means of 200 simulation runs using each method, with 95% confidence intervals. For descriptions of the methods see the text.

Table 1 Author-selection options: for choosing which past authors become authors of a new paper

1. Select author from a *recent paper* (published or unpublished)
2. Select author from a *recent journal publication*
3. Select author from *a paper in the reference list* of a recent paper. (If the recent paper has nothing in its reference list, then select from its own authors instead.)
4. Select author from *a paper in the reference list* of a *recent journal publication*. (If the reference list is empty, select author from the authors of the recent publication instead.)

Table 2 Paper-selection options: for choosing which papers become references in a new paper

1. Select a *recent* paper (published or unpublished)
2. Select a *recent* journal publication
3. Select item from the reference list of a *recent* paper
4. Select item from the reference list of a *recent* journal publication
5. Select from all existing papers without preference
6. Select from all existing publications without preference

Table 3 Summary of model parameters, with example values

Parameter description:	**Example value:**
Field parameters:	
# time steps	30
# foundational publications	14
# papers added in first year after foundation, P_1	16
Field growth, G (# papers to be added at time $t = P_1 * G^{(t-1)}$)	1.067
Authors parameters:	
Method for selecting authors	2 or 4 (Table 1)
# authors per paper: 2 parameters (alpha, beta) for a Weibull distribution	1.4, 1.3
Chance of author being new to field	0.6
Author Recency: 2 parameters (alpha, beta) for a Weibull distribution	1.3, 1
References parameters:	
Method for selecting papers to cite	4 (see Table 2)
# references per paper: 2 parameters for a Weibull distribution	1, 4.2
Chance of using recent paper itself rather than copying its reference	0.3
Reference Recency: 2 parameters for a Weibull distribution	1.3, 2
Contents parameters:	
Chance of innovation in one bit during contents construction	0.01
# bits of information in paper (the *N* in *NK fitness landscape*)	20
# interdependencies between bits (the *K* in *NKfitness*)	5
Peer Review parameters:	
Method for selecting past authors to be peer reviewers	2 (see Table 1)
# attempts to find reviewers for paper	9
# recommendations required for publication	3
Reviewer Recency: 2 parameters for a Weibull distribution	1, 4