

Modeling the Global Citation Network using the Scalable Agent-based Simulator for Citation Analysis with Recency-emphasized Sampling (SASCA-ReS)

Minhyuk Park¹, Haotian Yi¹, Tandy Warnow¹, George Chacko^{1*}

^{1*}Siebel School of Computing and Data Science, Grainger College of Engineering, University of Illinois Urbana-Champaign, 201 N. Goodwin Ave, Urbana, 21801, IL, USA.

*Corresponding author(s). E-mail(s): chackoge@illinois.edu;
Contributing authors: minhyuk2@illinois.edu; yi54@illinois.edu;
warnow@illinois.edu;

Abstract

The global science literature, represented as a network with articles as nodes and citations as edges, is a rich artifact for studies in scientometrics centered around historical, epistemological, sociological, and graph-theoretic perspectives. The structure of this network depends on its count of nodes and distribution of edges. We are interested in how the network has evolved from its origin to its present structure. Since extant theories of citation do not offer much in the way of quantitative explanation, we use a modeling approach that generates synthetic networks. Specifically, we have developed an idealized agent-based model of citations (SASCA-ReS) that can generate synthetic networks of size over 200 million nodes, which is comparable with the size of today's network. This model allows us to reason in an artificial world and to identify patterns of citation that may explain real-world scenarios. We report results from simulations under this model with different parameter settings and at various scales to explore counterfactual and hypothetical scenarios.

Keywords: agent based models, citation networks, scientometrics, Keyword4

1 Introduction

We conservatively estimate the body of scientific literature to be in the order of 100-300 million articles [1–3] today: a remarkable increase of around six orders of magnitude over 360 years since 1655 when the first scientific journal published less than a hundred articles in its inaugural year [4]. Representing this collection as a citation network with nodes (articles) connected by directed edges (citations) [5–8] enables scientometric studies from historical, epistemological, sociological, and graph-theoretic perspectives. Within this area of investigation, we are interested in understanding how the citations made by authors influence the structure of large citation networks. Given the incentives for authors to want to be well-cited, we are also interested in strategies that influence the accumulation of citations at the level of an individual article.

Preceding theories of citation offer some insight into the motivations of researchers that translate into citation behavior [9–13]. These motivations can be coarsely grouped into normative and rhetorical types, with the additional motivation of malpractice being worthy of mention today [14, 15]. However, motivations are not easily measured and are unlikely to be mutually exclusive. It seems more likely that each citation made by an author comprises multiple motivations that vary with time and context. Thus, quantitatively analyzing motivations is challenging. Nevertheless, citation behavior has been modeled on stylized normative and rhetorical behavior, although at a small scale and for the purpose of studying effects on “community health” [16].

Rather than attempt to quantify overlapping motivations, it is possible to encode the patterns of citation that have been previously observed and observe their effects on a growing network. In this respect, mathematical models of network growth [17–19] have been proposed that incorporate randomness, preferential attachment [20, 21], recency (immediacy) [5, 22], as well as fitness (epistemic quality) [23, 24].

In an approach complementary to mathematical modeling of citation dynamics, we have previously developed two idealized agent-based models (ABMs): PyABM [25], written in Python, followed by SASCA-s [26] (Scalable Agent-based Simulator for Citation Analysis with sampling), written in C++. In these models, publications are represented as nodes in citation networks and citations by directed edges.

During a simulation, batches of autonomous agents are created in time-steps of a year and make citations to existing nodes after which they serve as targets for other agents. The basis for selecting a target node is a composite formulation of the target’s features, its distance from the generator node, and the agent’s “citation phenotype”, which is a weighted combination of its bias towards preferential attachment, recency, and fitness. Thus, given an input of a “seed” network, the output is a larger citation network resulting from letting the random process operate over a period of years.

In the PyABM model, each new agent evaluates every other node in the network before selecting nodes to cite. This approach becomes expensive as the network grows and greatly limits scalability, essentially limiting PyABM simulations to 1-2 million nodes. SASCA-s addresses this limitation by only having each new agent scoring a random sample of the nodes within the 2-hop neighborhood of its generator node (i.e., the nodes that distance 1 or 2 from the generator); this approach, along with the benefits of a compiled language, allowed SASCA-s to scale to more than 100,000,000 nodes. A third generation of the initial model, SASCA-ReS (Scalable Agent-based

Simulator for Citation Analysis with Recency-emphasized Sampling) improved recency modeling by enforcing the distribution of citations made by each agent to produce a good fit to an empirically derived “recency table” that we computed from a citation network. This modification to SASCA-ReS provides the same scalability advantage of SASCA-s but with a better match to the several empirical properties of citation networks. The results presented in this manuscript are from the use of SASCA-ReS.

Of the various ways in which ABMs can be used [27], we use our model to support reasoning rather than to replicate real-world observations. We stress that SASCA-ReS is an idealized model in which realism is incorporated without attempting to copy it in detail. This offers advantages in terms of model parsimony and computational cost. Apart from the small number of variables, we have made additional simplifying assumptions: First, that every agent represents an article written by a single author and that no author has contributed more than one article to the network. Second, that we do not consider disciplinary habits or trends in citation behavior over time. The scenarios illuminated by the simulations reported below are *exploratory* and may or may not map to corresponding mechanisms of network growth in the real-world.

In the sections that follow, we report results from simulations under varied conditions resulting in networks ranging from just over a million nodes to greater than 100 million nodes. We focus both on network structure and the accumulation of citations by individual nodes. Throughout this manuscript, *in_degree* of a node is used to refer to citations *received* by a node while *references* and *out_degree* is synonymous with references, i.e., citations *made* to other nodes.

2 Materials and Methods

2.1 Data

Simulations shown in this manuscript used with one of three different seed networks as input: (i) the Stahl-Johnson (*sj*) network consisting of 491,532 nodes and 899,051 edges, which is derived from a real-world network from the biomedical literature and has been previously described in [25], (ii) the *sj_er* network, which is an Erdős-Rényi network modelled after the *sj* network, generated using the `sample_gnm` model in the R `igraph` package where $n = 491532$ and $m = 899,051$, and (iii) the *g10x10* network, which is a lattice network with 100 nodes arranged in a square grid with edges directed from one corner to the opposite corner in an acyclic pattern. After generation, nodes without edges were dropped. Data from the use of *sj_er* are shown in Supplemental Materials.

	# nodes	# edges	Reference
<i>sj</i>	491,532	899,051	[26]
<i>sj_er</i>	479,027	899,051	[26]
<i>g10x10</i>	100	180	

Table 1 Seed networks

The real-world networks used for comparison have been previously described. The Curated Exosome Network or CEN (*cen*) and Open Citations (*oc*) network with 13,989,436 and 75,025,194 nodes respectively [28] and the network extracted from Open Alex data (*oa*) with 111,453,719 nodes [29].

2.2 SASCA-ReS

Here we describe SASCA-ReS in more detail. Additional details are provided in Supplementary Materials and the code for SASCA-ReS is freely available from a Github site [30]. The design of SASCA-ReS incorporates a mid-level modeling approach where parameter settings for recency and out_degree quota are drawn from real-world data in order to focus on *more likely* rather than *formally possible* outcomes [31].

Each agent is created by vertex copying [32, 33] from a randomly selected node in the network, termed the “generator”. When created, each agent is assigned a quota of citations to make that is randomly drawn from a real-world distribution of reference counts (see out_degree distribution below) and randomly assigned a fitness value ranging from 1-1000 that is drawn from a power law distribution. Thus, every agent has a timestamp for when it was created (year), its fitness, and a quota of citations to make as attributes. Each agent within a time-step makes citations independently of other agents and the network is updated at the end of the time step. The duration of simulation and number of agents added each year are controlled by environment parameters set by the user for count of years and growth rate as a percentage of the number of nodes in the network.

Each agent has a parameter α that defines the fraction of its citations within the 2-hop neighborhood that go to the 1-hop neighborhood of its generator; thus, $\alpha = 1$ means that all its citations (other than the citation to the generator and the idiosyncratic citations that go outside the 2-hop neighborhood) are to nodes in the 1-hop neighborhood.

We computed a recency table (Supplementary Materials) based on references in publications in a real-world citation network [8]. Similarly we have constructed an out_degree distribution array from recent PubMed data, which is described in [26]. This recency table defines the expected fraction of its citations an agent will make to papers that were published n years ago, for each $n \geq 0$. Given the out_degree quota and recency table, we derive the expected distribution of citations to papers from a given year for each agent. Since achieving this distribution exactly is not always possible, we have implemented an approach that redistributes the out_degree when necessary to publications in adjacent years. This technique provides a good fit to the distribution defined by the recency table.

To account for citations of an idiosyncratic [34, p. 203] nature, a small fraction of citations made by agents is made to nodes selected by uniform random sampling from the entire network.

Since recency is now managed externally (as described above), the citation phenotype for an agent is a weighted composite of only two attributes—preferential attachment and fitness. In contrast, the citation phenotypes for SASCA-s and PyABM also included a weight for recency, which is not relevant for SASCA-ReS. Thus, in SASCA-ReS, the phenotype for an agent node is given by the weights it attaches to

preferential attachment and fitness, where the weights add up to 1.0. Agent phenotypes may be randomly assigned (*ra*) from a uniform distribution of possible phenotypes or fixed such that every agent has exactly the same phenotype.

Finally, like SASCA-s, SASCA-ReS only samples and scores at most 20,000 nodes in the 2-hop neighborhood of its generator node (that is, it samples and scores at most 10,000 node at distance 1 and 10,000 nodes at distance 2 from its generator). The small fraction of nodes that are cited from outside the 2-hop neighborhood are just randomly selected but not scored. In this way, SASCA-ReS only scores at most 20,000 nodes in the network, all of which are in the 2-hop neighborhood of its generator node, and thus achieves the same scalability as SASCA-s, while obtaining a better fit to the observed recency distribution of citation networks.

2.3 Simulations

SASCA-ReS simulations were conducted by passing four input files and a configuration file created by a user to the SASCA-ReS code. Additional details on how to perform simulations are available from the Github repository where the SASCA-ReS code is archived as open source [30]. The configuration file requires 23 entries of which the first 11 control simulation at the environment level, the next 7 specify the characteristics of individual agents, and the remaining manage parallelism, logging, and the location of output files (Supplementary Materials).

All simulations were performed on a compute cluster with up to 128 cores of parallelism and 512 GB of memory. Simulations were seeded with citation networks, Erdős-Rényi graphs, or lattice graphs with the fitness of seed nodes set to 1 in each case. A “standard” simulation was to simulate with the Stahl-Johnston seed set (below in Data) of around 500,000 nodes for 30 years at a growth rate of 3% using 16 cores of parallelism, which would result in an output network of around 1.2 million nodes in the order of 3-4 minutes. The default environment was to randomize agent phenotypes *ra* with respect to preferential attachment weight (*pa*) and α but results from partially static agent environments were also conducted where either α or preferential weight was set to fixed values for all agents. For a subset of the experiments reported here, simulations were scaled to (i) 10-20 million nodes, (ii) 70-80 million nodes, and (iii) 100-250 million nodes.

The largest network described in this manuscript has roughly 161 million nodes, and is termed the abm161m network (Table 2). Motivated by the two-dimensional epistemic grid used in [35, 36], we used a lattice graph of dimension 10×10 to approximate the number of publications in the inaugural year (1665) of [4]. The year of each node in the lattice was set to 1655, the fitness of each node was set to 1 (out of a possible range of 1 – 1000), and a constant growth of growth rate 4.0125% was applied for a simulation that lasted 360 years. For the first 150 years of this simulation, a normally distributed out_degree distribution ranging from 1 – 10 was used, after which it was exchanged for a PubMed derived out_degree distribution described in [26] that reflects more modern citation rates. For the two smaller networks, abm14 and abm76, the *sj* dataset was used as a starting point and the simulation allowed to proceed for 85 and 128 years respectively at a growth rate of 4.0125%.

2.4 Random Forest Analysis

A random forest regression analysis [37] was conducted to study model variables relative to time. This experiment was performed on the abm14 simulation where the final network size was 13,926,219 nodes and the simulation spanned 85 virtual years. A snapshot of node attributes for all agent nodes was recorded every 5 years, producing a total of 18 snapshots (including one from the final year of the simulation). For each snapshot, a random forest regressor was trained with Scikit Learn Python Library [38]. Each model consisted of 100 decision trees, with *max_features* = *n_features* and no *max_depth* constraint; other hyperparameters were left at default values. The input features were publication year, the preferential attachment weight, fitness, and *out_degree*. The target variable was *in_degree* (number of citations received). We computed two types of feature importance: impurity-based feature importance (reported in Supplementary Materials) and permutation feature importance using the R^2 score. Feature importances were evaluated by examining empirical trends observed in the agent-based simulations to analyze SASCA-ReS model behavior. All models were trained on the Illinois Campus Cluster with up to 128 cores and 512GB of memory. The training scripts and analysis procedures are available in our Github repository [39].

2.5 Clustering Analysis

The abm14 network as well as variants of it were clustered using the Leiden algorithm optimizing for the constant Potts model (CPM) [40] was used with the number of iterations set to 2 and resolution values of either 0.1 or 0.01. For abm14, abm76, and abm161, we attempted to cluster all three networks using a parallel Louvain implementation [41] on a machine with 64 cores and 950GB of RAM and a 4 hour time limit. Although the runs on abm14 and abm76 were successful, abm161 ran into a timeout.

3 Results

In this section, we present results from experiments using SASCA-ReS that are designed to explore recency modeling, scalability, the relative importance of model variables, community structure, and unconventional strategies to optimize citations to a document in a citation network.

3.1 Recency Modeling

A salient feature of citation behavior is recency or immediacy [5], the tendency of authors to cite more contemporary articles than older ones.

In earlier versions of our ABM (SASCA-s and PyABM), high *in_degree* nodes in a seed graph had an advantage in accumulating citations [42, pg5 and Fig. 1] on account of preferential attachment overriding recency. In SASCA-ReS, we improved recency modeling through binning approach (Supplementary Materials), which can be tuned to recency patterns of different fields. Our choice of bin size reflects patterns in the the biomedical literature and is derived from the *cen* network (Materials and Methods).

The effectiveness of this approach is shown in Fig. 1, where the proportion of citations made to preceding year nodes is more closely aligned with real-world data than in the case of the SASCA-s and PyABM models. The settings in Fig. 1 were used for all simulations reported in this manuscript.

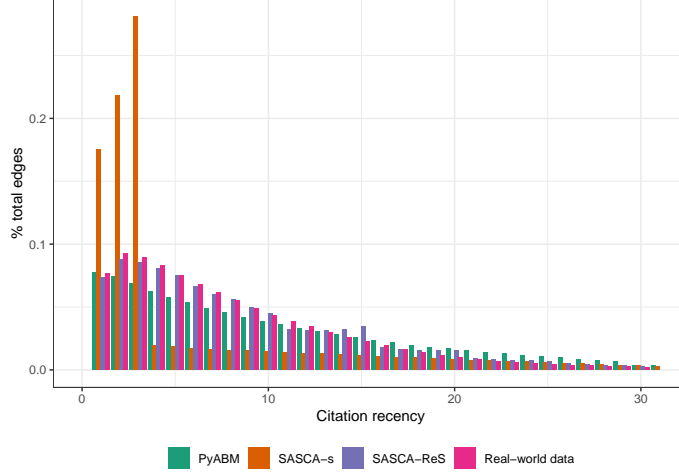


Fig. 1 SASCA-ReS improves recency modeling. Here we show the true recency distribution (real-world data) for comparison with observed recency distributions using SASCA-ReS, SASCA-s, and PyABM. For all models shown, the simulation was performed for 30 years at a growth rate of 3% using the *sj* seedset in a randomized agent environment. For SASCA-ReS, recency bins (Materials and Methods) were set to single-year bins with bin i containing all publications from i years ago for the first ten years, followed by 5-year bins until year 51. Those articles that were published at least 51 years ago were assigned to the last bin.

3.2 Scalability

To establish the scalability of SASCA-ReS, we simulated the growth of citation networks at three different scales: (i) 10-20 million nodes (ii) 50-100 million nodes, and (iii) greater than 100 million nodes. These simulations were conducted in the default environment, *ra*, where the citation phenotype of agents was randomized to represent a distribution of citation behaviors. These scales were chosen because of the availability of comparable real-world citation networks, although the *oa* real-world network is only about 70% as large as the abm161 network while the abm14 and abm76 networks are much more closely matched in size.

In the first and second cases, the simulated networks comprised 14,020,073 nodes (abm14) and 76,107,698 nodes (abm76) respectively (Table 2 and Figure 2). In the third case, also a metaphor for the growth of the scientific literature from 1665-2025 (see Introduction), the resultant graph (abm161) network consisted of 160,714,032 nodes with a runtime of just under 71 hrs. These simulated networks are less sparse than the real-world examples they were compared against, which we attribute to incompleteness in the data for our real-world networks and our generative model.

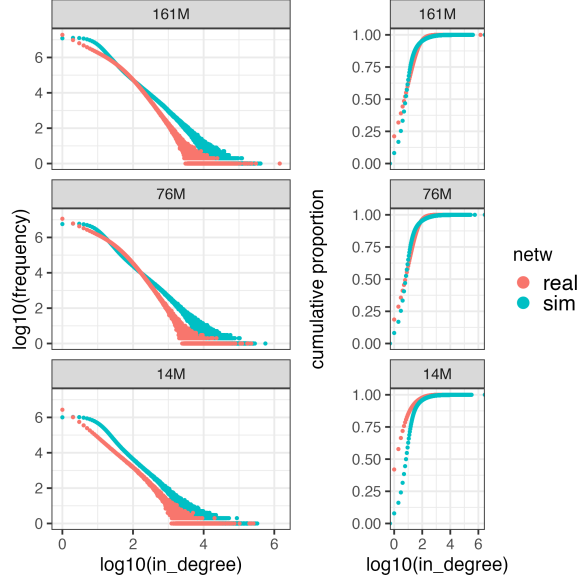


Fig. 2 Scalable simulations of citation networks. Synthetic citation networks were generated with SASCA-ReS at three different scales (the count of nodes for the synthetic networks is approximately 161M, 76M and 14M). Real-world networks are shown for comparison. *Left:* Frequency of each in_degree. *Right:* Cumulative distribution function of in_degree (both logarithmic).

Unsurprisingly, the software tools that we typically use to analyze networks do not scale well to the two larger networks (abm76 and abm161) (Supplementary Materials). While it is possible to compute node and edge counts and degree distributions, it is expensive to compute metrics such as diameter and clustering coefficients, and even more challenging to cluster these networks. While this finding underscores the opportunity to develop more scalable tools, it also restricted much of our analysis to networks at the 10-20 million node scale exemplified by the abm14 network. Consequently, we have not yet tried to cluster the largest network generated under SASCA-Res, which amounted to 217,839,850 nodes and 9,390,741,826 edges (Supplemental Materials) and for which we do not have a comparable real-world network.

3.3 Temporal Analysis

For the abm14 network, we examined how the in_degree accumulated by agents at the the end of a simulation varies according to when the agent was created (Figure 3). Agents that are created earlier in the simulation tend to have higher in_degree compared to those created later, likely since the agents created later have fewer years to accumulate citations in even with recency being enforced.

Examining the profiles of agents in the top 0.1% for in_degree (4,188–260,404 citations) in this simulation, the median values for out_degree and fitness amount to 103 and 18 respectively. Further, 92.04% of the agents in the top 0.1% have out_degree or fitness at least that of these medians. We evaluated the classifier for observations in the top 0.1% where in_degree of “out_degree > 103 *or* fitness > 18”, and noted

	# nodes	# edges	network	generation runtime
1	160,714,032	6,940,630,767	abm161	< 71hrs
2	75,598,213	3,245,034,596	abm76	< 4 hrs
3	13,926,219	581,472,876	abm14	< 30 min
4	111,453,719	2,148,788,148	oa	n.a.
5	75,025,194	1,363,303,678	oc	n.a.
6	13,989,436	92,051,051	cen	n.a.

Table 2 Summary statistics of simulated and real-world networks. Network sizes are shown for simulated networks (rows 1-3) and real-world networks (rows 4-6); we also show the time used to generate the simulated networks. The networks *oa* and *oc* are derived from the OpenAlex and Open Citations collections respectively [43], and *cen* [8] refers to the Curated Exosome Network, which is derived from the biomedical literature.

sensitivity and specificity of 0.926 and 0.948 respectively. These data strongly suggest that out_degree and fitness of an agent are influential, which is also consistent with our previous observations using the PyABM model [42]. In contrast, the median value of pa_weight was marginally different from the expected value of 0.5. However, the roughly 7.4% of agents in the false negative group suggest additional determinants in these stochastic simulations.

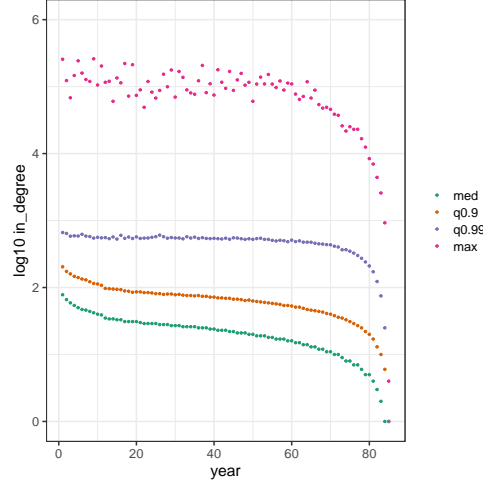


Fig. 3 Positional measures of in_degree for the abm14m network. The plot shows the minimum, first and third quartiles, median, and maximum in_degree for agents grouped by year in which they were created for the 85-year simulation that generated the abm14m network.

We also used a Random Forest regressor to infer the relative importance of fitness, out_degree, pa_weight, and time of creation (Figure 4). Across simulation years, the permutation feature importances (PFI) stabilize during the course of the simulation, suggesting that once the dataset becomes sufficiently large, random forest models capture the underlying relationships with some consistency. Fitness consistently emerges

	Min	Q1	Median	Q3	Max
abm161	—	—	—	—	—
abm76	678	8051	4,181,752	13,781,060	44,347,046
abm14	1,016,079	1,948,065	3,162,148	4,719,101	6,679,698

Table 3 Clustering with the Louvain algorithm. The abm14 and abm161 networks were clustered using the implementation of the Louvain algorithm in the Kuzu graph database. Summary statistics for the size of non-singleton clusters are shown.

as the dominant predictor while out.degree and publication year are also influential in making predictions at comparable magnitudes. In contrast, preferential attachment weight shows lower importance for predictive accuracy.

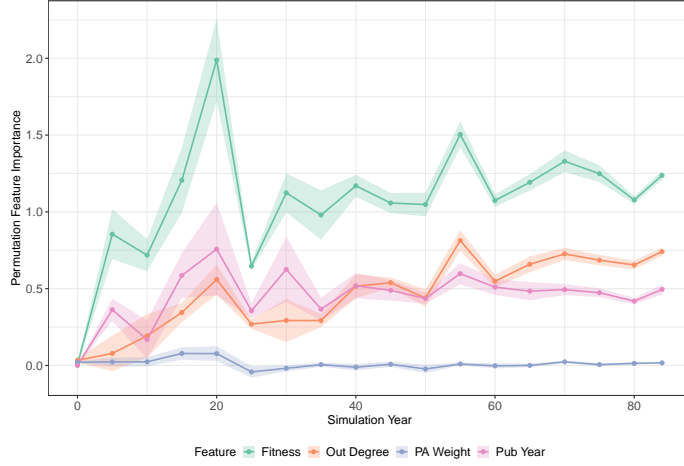


Fig. 4 Permutation feature importances for abm14. Here we show the permutation feature importances, measured with R^2 score, of random forest regressors trained on abm14 datasets at 5-year intervals during the simulation. Each random forest regressor consisted of 100 decision trees and predicts the final citation count received by an agent using four features: fitness, out.degree, preferential attachment weight, and publication year.

3.4 Community Structure

An expected feature of growing networks is the presence of clusters, suggesting specialization and research community formation, which has previously been reported in simulations [35]. To ask to what extent networks generated under SASCA-ReS exhibit community structure, we attempted to cluster the synthetic networks we generated.

While we were previously able to cluster the Open Citations network of 75,025,194 (Table 2) with the Leiden algorithm optimizing the Constant Potts model, the allocations of time and memory needed to analyze the abm76 and variants of it generated by varying parameter settings exceeded what was available to us. Further, clustering the abm76 and abm41 networks with the parallelized implementation of the Louvain

abm14 variants	alcc	25th	median	75th	90th	99th
ra	0.0839	5	12	24	49	360
ra.noalpha	0.0239	3	7	14	33	489
alpha0.05	0.0212	3	7	15	34	502
alpha0.95	0.1498	8	20	39	70	287
pa0.05	0.0824	5	12	26	52	408
pa0.95	0.0883	4	11	22	43	295
alpha0.00	0.0054	3	7	14	34	505
alpha0.20	0.0530	4	9	18	39	446
alpha0.40	0.0807	4	11	23	45	370
alpha0.60	0.1052	6	14	28	52	321
alpha0.80	0.1296	7	17	34	61	294
alpha1.00	0.1605	8	21	42	75	294

Table 4 In_degree statistics and average local clustering coefficient (alcc) for agents in different simulated networks with 14 million nodes. The networks are variants of the abm14 network produced by varying whether α and preferential attachment weight are randomized or static. These networks all have 13,926,219 nodes. For each network we display statistics for its in_degree distribution and alcc. Note that high α settings increase in_degree and alcc of agent nodes.

algorithm [44] available as an extension to the Kuzu graph database [41] resulted in a decidedly unsatisfying cluster size distribution (Table 3) with the largest cluster comprising 58% and 48% of the network for the abm76 and abm14 networks respectively. We do not consider these Louvain clusterings, very useful for evaluating community structure in these cases.

In an alternate strategy, we used the Leiden algorithm optimizing the Constant Potts Model [40] to generate a more intuitively sensible [45] cluster size distribution from clustered SASCA-ReS networks generated at the 10-20 million node scale under different conditions (Table 4). While altering the preferential weight did not result in much difference in the cumulative distribution function (CDF) for in_degree, a high α value of 0.95 caused an appreciable right shift of the CDF while a low α value caused a corresponding left shift relative to the default randomized conditions. Altering the pa_weight of agents did not result in appreciable differences relative to the control default simulations (Figure 5).

The results of this experiment are shown in Table 5 and indicate that a high α setting substantially increases the median, first and third quartile of cluster sizes relative to all other conditions tested while a high preferential weight setting results in the highest maximum cluster size. We also examined the “age” of each cluster, defining cluster age as the number of years since the earliest node in the cluster was created, and observed a pronounced decrease in the size of agent-only clusters in the high α simulation as cluster-age increases. Finally, we examined, under varying conditions, the relationship between the size of clusters that consisted only of agents and their age (Figure 6). To further examine the effect of α , we also examined node coverage (the

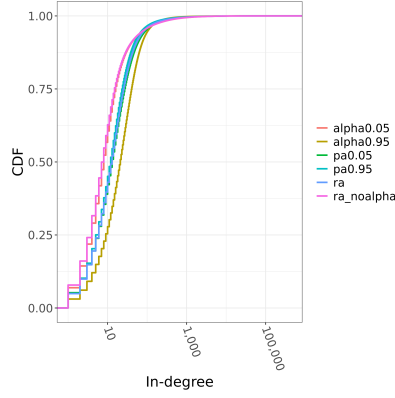


Fig. 5 In-degree CDF on 14m node networks. Here we show the CDF plots of in-degrees from six different simulations where different agent parameters were held static.

fraction of the network in clusters of size at least 2 or at least 10) across two different resolution values as α is varied.

agent phenotype	Min	Q1	Median	Q3	Max	num clusters	node cov.	sing.
alpha0.05	2	15	16	21	4714	569,362	1.00	20,647
alpha0.95	2	66	102	147	2936	118,428	1.00	6181
pa0.05	2	30	43	61	3532	260,943	1.00	927
pa0.95	2	31	44	62	<u>5576</u>	264,911	1.00	685
ra	2	30	43	61	4117	263,438	1.00	746
ra_noalpha	2	13	13	16	5419	659,344	1.00	25,866
alpha0.05	2	5	5	7	528	2,117,329	1.00	5156
alpha0.95	2	8	16	29	379	675,401	0.99	106,358
pa0.05	2	6	9	15	350	1,212,845	1.00	34,110
pa0.95	2	6	9	15	392	1,214,029	1.00	51,955
ra	2	6	9	15	351	1,213,577	1.00	38,788
ra_noalpha	2	4	5	5	<u>544</u>	2,323,650	1.00	7631

Table 5 Effect of α on cluster sizes SASCA-ReS simulations were performed in random and static agent environments at a growth rate of 4.0125% for 85 years to generate networks of size 13,926,219 nodes. Each of these networks was then clustered with the Leiden algorithm optimizing the Constant Potts Model (CPM) with top section at resolution value $r=0.01$ and bottom section at $r=0.1$.

3.5 Mavericks

Given the use of citation counts to evaluate productivity and impact, an incentive exists for authors to aspire to being well-cited. Results from PyABM suggest that, in addition to fitness, high-referencing and high values for α boost citation counts [25]. To explore unconventional strategies that might result in high in_degree for individual agents, we created customized agents that we refer to as mavericks, polysemically

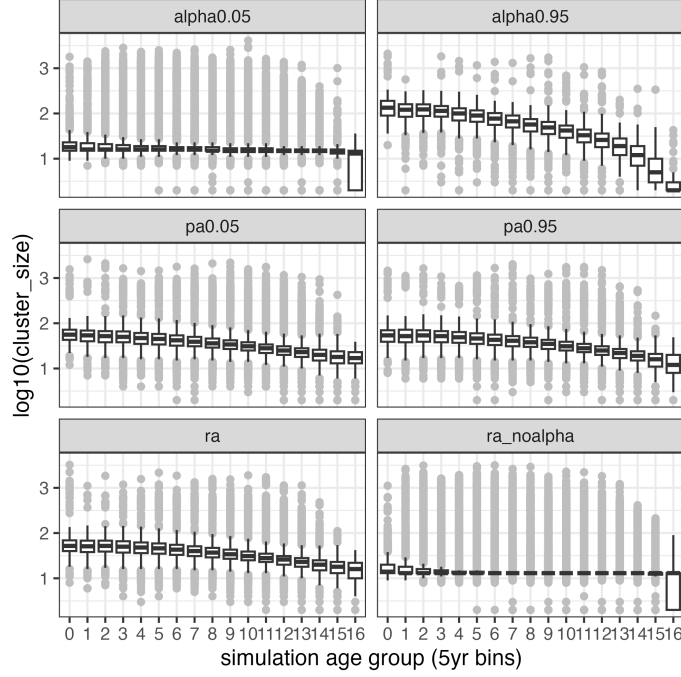


Fig. 6 Permutation feature importances for abm14. Here we show the permutation feature importances, measured with R^2 score, of random forest regressors trained on abm14 datasets at different years of the simulation. Each random forest regressor consisted of 100 decision trees and predicts the final citation received by an agent using four features: fitness, out_degree, preferential attachment weight, and publication year.

related to maverick in [36]. Mavericks follow a different rule set than the agents in SASCA-ReS in not being constrained by neighborhood, and may cite any other node in the network as long as they were published in a previous year. Each maverick is assigned a fitness value x and an out_degree quota y , and we allow these to vary.

Three different types of mavericks were designed: a *maximizer* that scans the entire network and randomly cites y random nodes from the top 0.1% of the network by total degree; a *randomnik* that scans the entire network and selects y random nodes; and a *minimizer* agent that scans the entire network and selects y random nodes (out_degree) from the bottom 0.1% of the network by total degree. Mavericks were created in the first year of the simulation and were assigned the same fitness level and out_degree quota. For the maverick parameters, we varied x from 1, 10, 100, and 1000 and y between 10 and 249. Given randomness in the SASCA-ReS model, we planted three mavericks of each type per simulation. Additionally, we set designated three non-maverick agents per simulation as controls.

Simulations were conducted for 30 years at a growth rate of 3% using the *sj* seed set where the nodes exhibit varying in_degree that varies from 0 – 88,435 produced a wide range of in_degree for mavericks to select from. The results (Figure 7) show that: (i) increasing fitness of mavericks increases their resultant in_degree and (ii)

increasing the out_degree of mavericks increases their in_degree (iii) the maximizer maverick outcompetes the randomnik, which in turn, outcompetes the minimizer.

We also computed disruption [46] for mavericks and control agents and observed that when $y = 249$, the maximizer maverick although very well-cited had disruption values considerably lower than the randomnik, minimizer, or control. We attribute this to the larger denominator in the case of the maverick on account of its references also being highly cited (Supplemental Materials). Thus, the original disruption formula of Wu and colleagues [46] may not be suitable for measurements in this context [47, 48] though a variant that is appropriately normalized to the context of this model may be more useful.

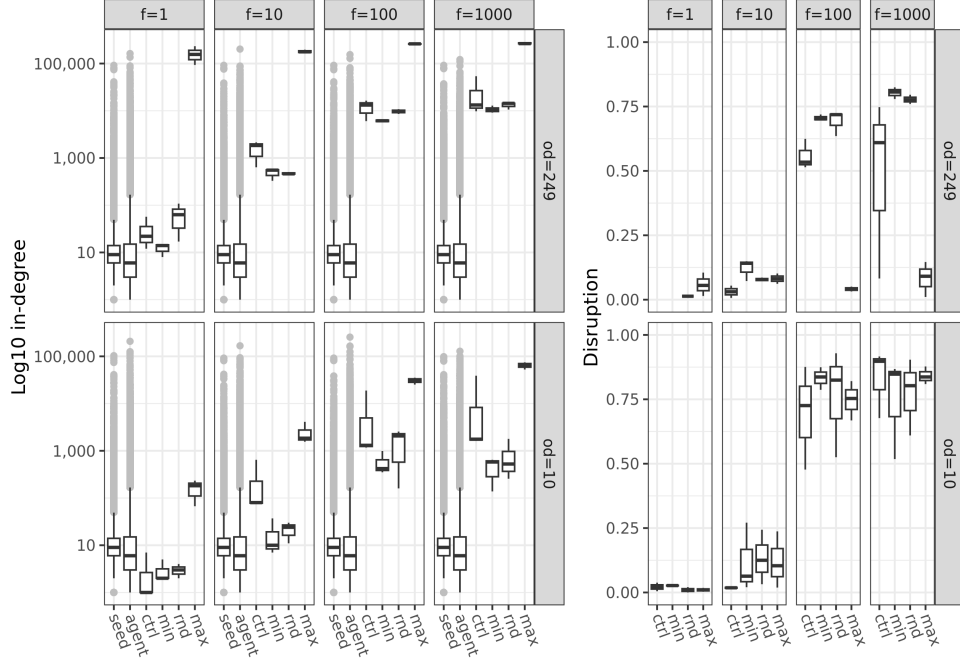


Fig. 7 Maverick citation strategies. Three each of *maximizer* (max), *randomnik* (rnd), and *minimizer* (min) mavericks were planted in the first year of 30-year 3% growth simulation on the *sj* seed network with randomized agent phenotypes. A *control* group of three randomized agent was also planted. Simulations in the upper row were performed with all planted agents (mavericks and controls) assigned out_degree quotas of 249 publications versus an out_degree of 10 in the bottom row. In independent experiments, mavericks and controls were assigned fitness values of 1, 10, 100, and 1000 (columns).

4 Discussion

Using the SASCA-ReS model, which features improved recency modeling, we modeled the growth of citation networks at scale. Simulations were conducted in randomized agent environments corresponding to our assumption of a broad variety of citation behaviors across individuals. Networks generated under the SASCA-ReS model are denser than the real-world examples we used for comparison. While we cannot exclude that this is an artifact of the model, we are inclined to believe that missing nodes and edges from real-world data also account for this difference. In the simulation protocol, all citations made by agents are to nodes in the growing network, whereas, publications in the comparable real-world network may cite or be cited by articles that are not included in the network. Thus, the simulated networks will have more edges and hence be denser than their corresponding real-world networks.

Models do not generate conclusive evidence but are useful for generating new hypotheses and revisiting existing ones. The SASCA-ReS simulations reported here suggest that fitness and out_degree are influential. While the model does not consider factors such as journal or author reputation, the effects of collaboration between authors, the limitations of peer review, and epistemic misconduct, the result suggests that publication quality is important even in a complex environment of individual behaviors. The out_degree effect is consistent with Small, Sweeney, and Greenlee’s interpretation in 1984 [49] of ‘high-referencing’ articles and the case for fractional citation counting. An incidental advantage of the idealization is that it is not possible to compute the h-index.

Simulations enable asking counterfactual questions. We asked what the structure of the global citation network might be today if researchers had historically exhibited extreme insularity, modeled by high α values, in their citation behavior. The results suggest that community structure would be enhanced perhaps at the cost of interdisciplinary work.

Using maverick agents, we examined the hypothetical question of whether discipline-independent citations would favor the accumulation of citations to individual documents. We observed that referencing only the most highly cited nodes in a network results in very high citations. In the real-world, this effect is likely counterbalanced by the insular behavior of researchers.

These initial studies were limited by the challenges of studying synthetic networks of over 75 million nodes in detail because of software and infrastructure constraints. That fitness is a driver, even in our idealized environment, makes a welcome case for quality in the world of citations. The result from the maximizer maverick suggests that more cynical strategies may also work. Using an abstract model for reasoning, however, generates ideas that should not be equated with evidence in the real-world. Our emphasis on model simplification in this manuscript also limited the extent to which we could explore the nonlinearity of citation dynamics proposed by Golosovsky [19]. Finally, we have not assessed statistical identifiability, but we expect that the SASCA-ReS model is not identifiable.

Supplementary information. This manuscript is accompanied by Supplementary Information in pdf format.

Acknowledgements. The authors thank the Illinois Computes program for allocations of computing time.

Declarations

- This study was partially supported by the Illinois:Inspire Partnership and NSF Award 2402559
- The authors declare that they have no conflicts of interest.
- Data availability. Data for the large networks generated in this study are available from the Illinois Data Bank under IDB-9265079. All other data are publicly available or can be regenerated through the protocols in the manuscript and the SASCA-ReS Code.
- Code availability The code for SASCA-ReS is available from Github
- Author contributions: Conceptualization (GC, MP, TW), Data Curation (MP), Formal Analysis (MP, HY, MP), Investigation (MP, HY, TW, GC), Methodology (MP, GC, TW), Supervision (GC, TW), Validation (MP, HY, GC), Visualization (HY, MP, GC), Writing- original draft (GC, TW, MP, HY), Writing- Review and Editing (GC, TW, MP, HY), Funding acquisition (GC, TW).

References

- [1] Digital Science. (2018-) Dimensions Software. <https://app.dimensions.ai>. Accessed: 2025-09-07 under licence agreement
- [2] Open Alex. <https://openalex.org>. Accessed: 2025-09-07
- [3] Peroni, S., Shotton, D.: OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies* **1**(1), 428–444 (2020) <https://doi.org/10.1162/qss.a.00023>
- [4] Philosophical Transactions of the Royal Society of London. <https://royalsocietypublishing.org/journal/rstl>. Accessed: 2025-09-07
- [5] Price, D.J.d.S.: Networks of Scientific Papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science* **149**(3683), 510–515 (1965) <https://doi.org/10.1126/science.149.3683.510>
- [6] Garfield, E., Cawkell, A.E.: Location of milestone papers through citation networks. *Journal of Library History* **5**, 184–188 (1970)
- [7] Boyack, K.W., Klavans, R.: Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology* **65**(4), 670–685 (2013) <https://doi.org/10.1002/asi.22990>
- [8] Wedell, E., Park, M., Korobskiy, D., Warnow, T., Chacko, G.: Center–periphery structure in research communities. *Quantitative Science Studies* **3**(1), 289–314 (2022) <https://doi.org/10.1162/qss.a.00184>

- [9] Gilbert, G.N.: Referencing as Persuasion. *Social Studies of Science* **7**(1), 113–122 (1977) <https://doi.org/10.1177/030631277700700112>
- [10] Cronin, B.: The Need for a Theory of Citing. *J. Documentation* **37**, 16–24 (1981)
- [11] Cozzens, S.E.: Taking the measure of science: A review of citation theories. *Newsletter of the International Society for the Sociology of Knowledge* **7**(1–2), 16–21 (1981)
- [12] Cozzens, S.E.: What do citations count? The rhetoric-first model. *Scientometrics* **15**(5–6), 437–447 (1989) <https://doi.org/10.1007/bf02017064>
- [13] Leydesdorff, L.: Theories of citation? *Scientometrics* **43**(1), 5–25 (1998) <https://doi.org/10.1007/bf02458391>
- [14] Biagioli, M., Lippman, A.: *Gaming the Metrics: Misconduct and Manipulation in Academic Research*. The MIT Press, Cambridge MA (2020). <https://doi.org/10.7551/mitpress/11087.001.0001>
- [15] Secchi, D.: A Simple Model of Citation Cartels: When Self-interest Strikes Science, pp. 23–32. Springer, ??? (2023). https://doi.org/10.1007/978-3-031-34920-1_3 . http://dx.doi.org/10.1007/978-3-031-34920-1_3
- [16] Bao, H., Teplitzkiy, M.: A simulation-based analysis of the impact of rhetorical citations in science. *Nature Communications* **15**(1) (2024) <https://doi.org/10.1038/s41467-023-44249-0>
- [17] Simkin, M.V., Roychowdhury, V.P.: A mathematical theory of citing. *Journal of the American Society for Information Science and Technology* **58**(11), 1661–1673 (2007) <https://doi.org/10.1002/asi.20653>
- [18] Goldberg, S.R., Anthony, H., Evans, T.S.: Modelling citation networks. *Scientometrics* **105**(3), 1577–1604 (2015) <https://doi.org/10.1007/s11192-015-1737-9>
- [19] Golosovsky, M., Solomon, S.: Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E* **95**(1) (2017) <https://doi.org/10.1103/physreve.95.012324>
- [20] Price, D.D.S.: A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27**(5), 292–306 (1976) <https://doi.org/10.1002/asi.4630270505>
- [21] Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)
- [22] Silva, F.N., Tandon, A., Amancio, D.R., Flammini, A., Menczer, F., Milojević, S., Fortunato, S.: Recency predicts bursts in the evolution of author citations. *Quantitative Science Studies* **1**(3), 1298–1308 (2020) https://doi.org/10.1162/qss_

- [23] Zhou, B., Holme, P., Gong, Z., Zhan, C., Huang, Y., Lu, X., Meng, X.: The nature and nurture of network evolution. *Nature Communications* **14**(1) (2023) <https://doi.org/10.1038/s41467-023-42856-5>
- [24] Touwen, L., Bucur, D., Hofstad, R., Garavaglia, A., Litvak, N.: Learning the mechanisms of network growth. *Scientific Reports* **14**(1) (2024) <https://doi.org/10.1038/s41598-024-61940-4>
- [25] Chacko, G., Park, M., Ramavarapu, V., Grama, A., Robles-Granda, P., Warnow, T.: An agent-based model of citation behavior. (Accepted) *Applied Network Science* (2025). A preprint version is available on arXiv at <https://doi.org/10.48550/arXiv.2503.06579>
- [26] Park, M., Lamy, J.A., Rodrigues, E.C., Ferreira, F.M., Vu-Le, T.-A., Warnow, T., Chacko, G.: Very Large Scale Simulations of Network Growth with the Scalable Agent-based Simulator for Citation Analysis with sampling (SASCA-s). in press, 14th International Conference on Complex Networks and their Applications (2025) (2025)
- [27] Macal, C.M.: Everything you need to know about agent-based modelling and simulation. *Journal of Simulation* **10**(2), 144–156 (2016) <https://doi.org/10.1057/jos.2016.7>
- [28] Park, M., Tabatabaee, Y., Ramavarapu, V., Liu, B., Pailodi, V.K., Ramachandran, R., Korobskiy, D., Ayres, F., Chacko, G., Warnow, T.: Well-connectedness and community detection. *PLOS Complex Systems* **1**(3), 0000009 (2024) <https://doi.org/10.1371/journal.pcsy.0000009>
- [29] Caetano Machado Lopes, L., Chacko, G.: A Citation Graph from OpenAlex (Works). University of Illinois Urbana-Champaign (2024). <https://doi.org/10.13012/B2IDB-7362697-V1> . <https://doi.org/10.13012/B2IDB-7362697-V1>
- [30] Park, M., Vu-Le, T.-A., Warnow, T., Chacko, G.: SASCA-ReS. <https://github.com/illinois-or-research-analytics/SASCA-ReS>. Github repository (2025)
- [31] Harnagel, A.: A mid-level approach to modeling scientific communities. *Studies in History and Philosophy of Science Part A* **76**, 49–59 (2019) <https://doi.org/10.1016/j.shpsa.2018.12.010>
- [32] Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: *The Web as a Graph: Measurements, Models, and Methods*, pp. 1–17. Springer, New York, NY (1999). https://doi.org/10.1007/3-540-48686-0_1
- [33] Krapivsky, P.L., Redner, S.: Network growth by copying. *Physical Review E* **71**(3) (2005) <https://doi.org/10.1103/physreve.71.036118>

- [34] Crane, D., Small, H.: American sociology since the seventies. In: Halliday, T.C., Janowitz, M. (eds.) *Sociology and Its Publics*. Heritage of Sociology Series. University of Chicago Press, Chicago, IL (1992)
- [35] Gilbert, N.: A Simulation of the Structure of Academic Science. *Sociological Research Online* **2**(2), 91–105 (1997) <https://doi.org/10.5153/sro.85>
- [36] Weisberg, M., Muldoon, R.: Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science* **76**(2), 225–252 (2009) <https://doi.org/10.1086/644786>
- [37] Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (2001) <https://doi.org/10.1023/A:1010933404324>
- [38] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay: Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**(85), 2825–2830 (2011)
- [39] Yi, H.: SASCA-RF. <https://github.com/AnoldGH/RandomForestTools>. Github repository (2025)
- [40] Traag, V.A., Waltman, L., Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**(1) (2019) <https://doi.org/10.1038/s41598-019-41695-z>
- [41] Feng, X., Jin, G., Chen, Z., Liu, C., Salihoğlu, S.: Kùzu graph database management system. In: CIDR (2023)
- [42] Chacko, G., Park, M., Ramavarapu, V., Grama, A., Robles-Granda, P., Warnow, T.: An Agent-based Model of Citation Behavior. Preprint: arXiv:2503.06579 (2025)
- [43] Park, M., Tabatabaee, Y., Warnow, T., Chacko, G.: Data for Well-Connectedness and Community Detection. University of Illinois Urbana-Champaign (2024). https://doi.org/10.13012/B2IDB-6271968_V1 . https://doi.org/10.13012/B2IDB-6271968_V1
- [44] Lu, H., Halappanavar, M., Kalyanaraman, A.: Parallel heuristics for scalable community detection. *Parallel Computing* **47**, 19–37 (2015) <https://doi.org/10.1016/j.parco.2015.03.003> . Graph analysis for scientific discovery
- [45] Šubelj, L., Eck, N.J., Waltman, L.: Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ONE* **11**(4), 0154404 (2016) <https://doi.org/10.1371/journal.pone.0154404>

- [46] Wu, L., Wang, D., Evans, J.A.: Large teams develop and small teams disrupt science and technology. *Nature* **566**(7744), 378–382 (2019) <https://doi.org/10.1038/s41586-019-0941-9>
- [47] Bornmann, L., Ettl, C., Leibel, C.: In search of innovative potential: The challenge of measuring disruptiveness of research using bibliometric data. *EMBO Reports* **25**(7), 2837–2841 (2024) <https://doi.org/10.1038/s44319-024-00177-8>
- [48] Bornmann, L., Devarakonda, S., Tekles, A., Chacko, G.: Are disruption index indicators convergently valid? the comparison of several indicator variants with assessments by peers. *Quantitative Science Studies* **1**(3), 1242–1259 (2020) https://doi.org/10.1162/qss_a_00068
- [49] Small, H., Sweeney, E., Greenlee, E.: Clustering the science citation index using co-citations. ii. mapping science. *Scientometrics* **8**(5-6), 321–340 (1985). DOI: <https://doi.org/10.1007/bf02018057>