Computational models of temporal co-authorship networks

Upasana Dutta, Department of Computer Science, University of Colorado Boulder, United States
Subhashis Majumdar, Department of Computer Science, Heritage Institute of Technology, India
Subhajit Datta, Department of Computer Science, Singapore Management University, Singapore

Collaboration represents a key aspect of the functioning of any scientific ecosystem. By sharing ideas and expertise through co-authorship of papers, researchers build the foundation of progress in scientific domains. In this paper, we propose a simple agent-based model to understand collaboration characteristics as they vary over time, and augment the model in light of established mechanisms of collaboration. Our models are validated using large-scale bibliometric data from multiple domains in the computing discipline. Simulations from the model on how authors are clustered in co-authorship networks closely match empirical observations. This indicates that our models are able to capture some of the essential aspects of citation dynamics. Our results offer a set of insights that might be of use to individual researchers, and research organizations.

Additional Key Words and Phrases: Coauthorship networks, NetLogo, preferential attachment

INTRODUCTION

Team work is one of the core characteristics of research as researchers collaborate with each other to communicate their findings to the scientific community. They publish their results by writing papers, which is co-authored by all the researchers who participated in the research project. Previous studies have shown that collaborative research produces better research output, both in terms of knowledge as well as impact, as compared to solo authors [29]. Hence, co-authorship is one of the building blocks of the research communities in many fields [15, 16, 21]. In this study, we build a simple computational model to explore how collaboration preferences of authors give rise to patterns in some global characteristics of the collaboration network formed by the authors as and when they work as co-authors together in papers. We propose two different models, and then validate them using real world co-authorship networks.

Every author holds a preference for the kind of authors they want to collaborate with and we consider two kinds of authors on the basis of their contribution. *Experienced authors* are authors who have been publishing papers in their research field for a considerable amount of time, are more known for their work, and have made more contributions to their domains through paper publications. *Newcomers* are authors who are comparatively new to the domain and are relatively less known in their research community. An author's inclination towards either of the two kinds of co-authors can influence the dynamics of a collaboration network. For example, in certain research domains, new authors may have more propensity towards collaborating with *experienced authors*, which would result in a co-authorship dynamics that is different from a case where new authors are more inclined towards collaborating with *newcomers*. This forms one of the key ideas that we use in our modelling approach. Each author acts as an agent and they decide which author they want to collaborate with on the basis of their individual preferences. These individual preferences shape the overall characteristics of the collaboration network. With this understanding, we build two computational models for co-authorship networks and use them in a number of scientific domains. Our model is useful because it offers some insights about the characteristics of co-authorship networks, which may contribute to the study of collaboration dynamics in various research domains. Specifically, the aim of this study is to answer the following research question -

RQ : By abstracting key features of developer collaboration dynamics and the collaboration characteristics of researchers, shall we be able to simulate their real world co-authorship patterns?

To answer this question, we first build a model using NetLogo agent based modeling environment [17, 28] with model parameters governing how the new vertices and edges are added to the co-authorship network. The parameters tuned in the models are individual level parameters, which decide the authors' propensity towards collaborating with either the *newcomers* in the field or the *experienced authors*, as well as the time span for which the authors continue publishing papers in their fields. We compare the clustering coefficient values of the networks produced by the models' simulations with the corresponding ones generating from empirical data [7, 27] across five different domains, namely Software Engineering, Artificial Intelligence, Networks, Operating Systems and Database. These empirical networks are cumulative in nature, meaning in each time-step new nodes and edges get added to the existing networks, and there are no deletion of nodes and edges in subsequent time-steps. We choose clustering coefficient as the metric for comparison because clustering coefficient has been observed to be one of the most fundamental properties of collaboration networks in scientific communities [20] and studies suggest a tendency for the occurrence of the small world network structure in many collaboration networks [8, 11, 12, 18, 21, 29]. We use mean absolute percentage error (MAPE) [26] to determine how close the model simulations are to the empirical data. We denote this model as *An Agent Based Model for Scientific Collaboration (ABM-SC)*. Next, we build another model that uses all the parameters used in the previous model, but is additionally enhanced following the phenomenon of the Barabási–Albert preferential attachment model [3], where the edges are added to the network following the notion of "rich get richer" phenomenon. We denote it as *An Enhanced Preferential Attachment Model for Scientific Collaboration (EPAM-SC)*. We then compare the clustering coefficient values of the models' simulations with those of the empirical data and we analyse which model did better and under what configurations of the parameters. Our findings show that by enhancing the initial model with the concept of preferential attachment, and then identifying the set of parameters that best fits the empirical data, we are able to simulate the clustering patterns that closely matches with the real world collaboration networks.

BACKGROUND AND RELATED WORK

There has been several studies in the past regarding the dynamics of collaboration networks involving different scientific fields. Parish et. al. [23] studied various factors that correlate with collaborative behaviour of authors of various fields. They showed that researchers in physics, medicine, infectious diseases and brain sciences show very high collaborative behaviour, whereas those in social science, computer science and engineering show relatively lower collaborative behaviour. They also found association between higher collaborations with higher citation impact, with its effect varying among different fields. Bales et. al. [2] studied the levels of internal and external collaborations across various academic departments and disciplines. They found that inclusion of at least one professor, research scientist, or a basic science researcher as an author, in particular, were all strongly associated with publication in high-impact journals. They also identified clear distinctions in authorship patterns between basic science and clinical departments, with regards to the ratio of outside-to-within-department collaborators as well as the network structure measured through connected components and centralisation.

Evans et. al. [9] studied homophily in collaboration networks, which is basically scientists collaborating with scientists who they are very similar to in some respect. Their study also indicates that "scientists use status positions for discriminating between potential partners by selecting collaborators from institutions with a rating similar to their own." In a similar direction, Gallivan et. al. [10] studied that across all over the world, the scientists pursuing research in the field of Information Science exhibit a propensity of collaborating with other researchers who are of the

same sex as they are, and also "the same PhD program than one would expect by chance".

These findings indicate that scientists, while collaborating with other authors in their fields, do have a preference regarding the kind of authors they would like to collaborate with, which can be based on factors like geography [1, 4, 22], gender [5, 6, 13], and so on. These factors in turn give rise to certain observable collaborative dynamics that varies from discipline to discipline.

MODEL DEVELOPMENT

Our aim is to build an Agent-based model that captures the dynamics of co-authorship networks across different research domains. In this paper, a co-authorship network of a given research domain is a network comprising of nodes and edges, where the nodes represent the authors who have published research paper(s) in the given field and two authors are connected with an undirected edge if they have co-authored in at least one paper germane to a given research field. Each edge has an associated edge-weight that signifies the number of papers the two adjacent authors have co-authored in. Each network spans over a particular time period, say T, which is divided into a number of time-steps $t_1, t_2, t_3, \ldots, t_n$. The networks are cumulative in nature, i.e., the network for any time-step t_{k+1} contains all the nodes and the edges of time-step t_k plus the nodes and the edges that got added to the network in the current time-step t_{k+1}. Hence, the network structure is extensible across the time-steps. Also, note that the interactions in the networks among the authors are generally at an individual level, and not aggregate(or population) level, since the extent of co-authorship between two authors does not depend on most of the other authors or their co-authorship. Since Agent Based Modelling (ABM) lets us model the phenomenon of individual authors' actions at the micro-level leading to macro-level characteristics, and at the same time is well suited for cases that are incremental in nature, we choose the ABM approach for building our computational model. Our model is governed by a set of parameters that simulate the generation of such co-authorship networks.

ODD Protocol

In 2006, Grimm et al. proposed a standard protocol, dubbed ODD (Overview, Design concepts and Details), for describing simulation models like individual based models (IBMs) and agent based models (ABMs). The protocol consists of three blocks, namely, **Overview, Design concepts, and Details**.

Two of these blocks are further subdivided into the following elements -

Overview - Purpose, State variables and scales, Process overview and scheduling
Details - Initialisation, Input, and Submodels.

We now describe our model with respect to the ODD Protocol.

Purpose : The purpose of the model is to explore and capture the dynamics of co-authorship networks. This model has been used to simulate the network attributes of co-authorship networks where authors and their papers are getting added to the network with time. Our main aim is to simulate how the clustering co-efficient value changes in these networks as and when the network grows. We also use this model to study the preferences of the authors in publishing papers with other authors who are either well established in the network, or are new to the network. The pattern observed in the network attributes are dependent on these preferences and on many

Table 1. Variables used to describe each author in the model

Authors	
Variable name	**Description**
Coordinates	X and Y coordinates of each author randomly assigned to them
Label	Determines if the author is "connected" to any other author in the network or not
TypeNode	Type of the author, determines whether the author was present in the previous time-steps, or it got introduced to the network in the current time-step
VerNo	Vertex number, a number that is unique to each vertex
Ticks_Lived	Number of time-steps for which the author has been present in the network
Lifetime	Total number of consecutive time-steps for which the author can be present in the network
MaxLinks	Maximum number of links an author can have in the first time-step of the simulation

other parameters of the simulation. This model helps in validating the patterns as observed in the empirical co-authorship networks.

State variables and scales : Our model consists of only one type of agents : Authors, whose primary behaviour is to establish an edge with the other authors of the model based on model parameters. Table 1 describes the variables used to define each agent of the model, which are the authors of the simulated networks.

Process and overview scheduling : In each time-step, the following processes take place -

```
For 1 to the defined number of new authors that get added to the network
     Create a new author
     Randomly assign the x and the y coordinates for the new author
     Mark its label as not connected
     Set its TypeNode as new
     Assign its Lifetime by choosing a normally distributed
          random number with mean Mean_V and standard deviation
          Stand_D
Calculate the number of edges to be added between old authors based on the
probability parameter 'oldold', between old and new authors based on the pa-
rameter 'oldnew' and between new authors based on the parameter 'newnew'.

As long as new edges are yet to be added [1]
     Choose the number of authors in each new paper with 90%
          probability of this number being between 2 to 5
          and 10% probability of this number being 6 or 7.
```

[1] if the number of new edges that remain to be added is less than the number of edges in the clique to be added to the network, we add the clique and then exit from the loop, assuming that the addition of a few number of extra edges to the network will have a negligible impact on the overall model. Note that the number of extra edges that can get added in each time-step is upper bounded by $\binom{7}{2} = 21$.

```
      Choose those many authors from the existing network given
           that all the authors of a given type (either old
           or new) have equal probability of getting chosen
           among other authors of the same type.
      Join those many authors in the network with edges to
           form a clique, signifying that the authors have
           co-authored with each other in a paper.

Increase the Ticks_Lived attribute of each author by 1
Update the TypeNode attribute of all the new authors to 'old'
Create Pajek file of the network and save it
```

Design Concepts :
Emergence : We build this model to explore the conditions under which a certain pattern is exhibited by the clustering coefficient attribute of the network. The pattern arises from activities of the authors based on their preferences. The parameters of our model control such preferences. With different set of parameters, emergence of different patterns in the network attributes can be observed.

Interaction : The preferences of an author is independent of the preferences of its neighbours, i.e. the preference of a given author will not depend on the authors they have already co-authored papers with.

Initialisation : The environment of our model is initialised by creating the initial number of authors in the network, based on the model parameter n which specifies the number of authors the model begins with. The maximum links that can be attached to each author during initialisation is assigned by choosing a Poisson-distributed random number with mean *Poisson_M*, since the degree distribution of random networks follow Poisson distribution [19, p. 393]. Edges are added between the authors as long as there is an author in the network that has no edge linked to it, respecting the limit on the maximum number of edges an author can be attached to. The publication span of each of these authors is assigned by choosing a normally distributed random number with mean *Mean_V* and standard deviation *Stand_D*. We conjectured that the time period for which authors actively publish in their research fields will follow a normal distribution, which is based on the previous research findings that human longevity conditioned on survival up to the modal age behaves like a normal distribution [14, 24]. No further initialisation is required for the authors.

Inputs : Once the model is initialised, no further inputs are required.

Submodels : The key process of the model involves addition of new authors to the network, and addition of edges between authors (both old and new) as and when the network grows with each time-step. The number of authors that get added to the network in a time-step is given by the f_v parameter. At each time-step t_k, $f\_v \times v$ vertices are added to the network, where v is the number of vertices in the network at time-step t_{k-1}. Therefore, f_v (<1) is the rate at which authors get added to the network. The number of edges that get added to the network in a time-step is given by the parameter *seed*. At each time-step t_k, $seed \times e$ edges are added to the network, where e is the number of edges in the network at time-step t_{k-1}. Therefore, seed is the rate at which edges get added to the network.

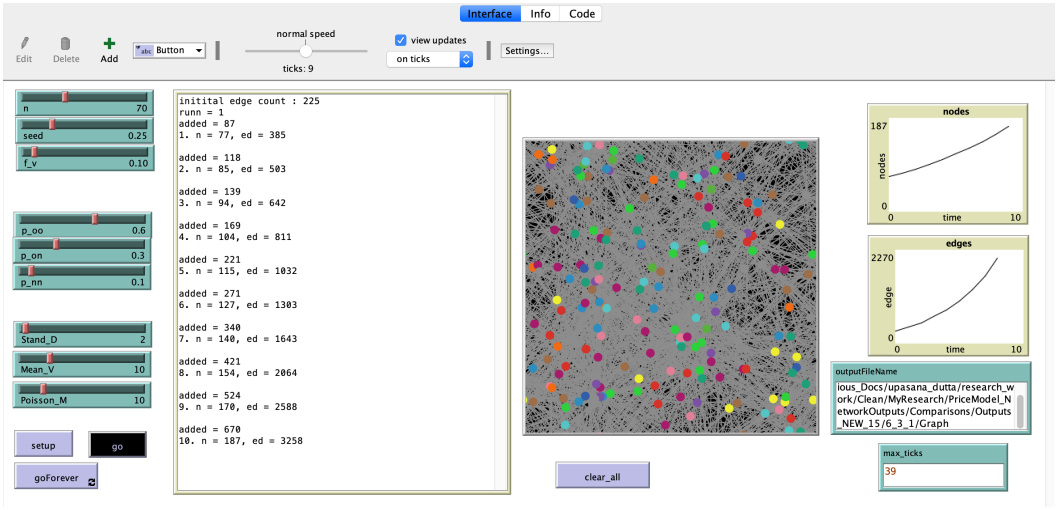

Fig. 1. A snapshot of the NetLogo Agent based Modelling environment

Agents

NetLogo is a programmable modelling environment used to build models that simulate natural and social phenomena. The model is run using various parameters which can be set to a given value using the sliders. Figure 1 shows a snapshot of the NetLogo Agent Based Modelling Environment. Now we describe the agents of the model.
1. The authors of the co-authorship networks are the turtles of the NetLogo model.
2. The weight of an edge between two authors represents the number of papers in which they have co-authored; these are represented as links in the model.

A co-authorship network consists of authors publishing papers with each other, where an author can be an established author in the network who has been publishing papers in the given research domain for several years, i.e., they already exist in the earlier time-step(s), or can be a new author who has entered the network in the current time-step. Since, the number of authors and the edges between them in the network is increasing, we keep certain parameters in our model that signify the rate at which edges and vertices are being added in the network. Also, in our research ecosystem, the co-author relationships between different authors vary. We repeat here that a co-authorship can exist between two authors who have been publishing papers in the domain for quite a few years, it can exist between two authors who are new to the network and it can also exist between an established author and a new author. It is crucial that we model these different co-author relationships through our model parameters because the type of these collaborations (co-authorship between different authors) plays an important part in deciding the dynamics of the network.

Model Parameters

Agent Based Model for Scientific Collaboration (ABM-SC):. We first describe the model parameters of the first model that we build using the NetLogo agent based modelling environment.

The model has two sets of parameters. At first, we define the parameters that set up the ecosystem of the model and are concerned with the introduction of new vertices and edges in the network.

1. n – The number of vertices at the first time-step t1.

2. f_v – At each time-step t_k, $f\_v \times v$ vertices are added to the network, where v is the number of vertices in the network at time-step t_{k-1}. Therefore, f_v is the rate at which authors get added to the network.

3. $seed$ – At each time-step t_k, $seed \times e$ edges are added to the network, where e is the number of edges in the network at time-step t_{k-1}. Therefore, seed is the rate at which edges get added to the network.

Now we define the parameters that are concerned with the type of co-authorship between the authors in the network. In our co-authorship network, an 'old' author in a time-step t_k is an author who exists in the network from time-step t_{k-1} or earlier. A 'new' author in a time-step t_k is an author who is added to the network in time-step t_k and was not present in the network in any of the previous time-steps.

4. p_oo – The probability of an old author to connect with another old author in the network at any time-step.

5. p_on – The probability of an old author to connect with a new author (i.e an author who got added in the current time-step) in the network.

6. p_nn – The probability of a new author to connect with another new author in the network.

In a co-authorship network, an author might not stay active in the network for the period of time T. They might publish research papers for a few years and then stop publishing, or they might shift to a different domain altogether. So, in our model, every author is allowed to publish new papers for only a specified number of time-steps. When the specified number of time-steps are over, no new edges (papers) are added to the vertex (author). We named the total number of time-steps an author can stay 'active' in a network to be the *Lifetime* of each author. Also, it is important to note that in a real world situation, the *Lifetime* of different authors may vary. Some authors may keep on publishing new papers for several years. On the other hand, some authors may stop publishing just after a couple of years. Hence the *Lifetime* of different authors in a network would be different. Therefore it will not be very meaningful to fix the *Lifetime* to a single constant value for all the authors in the network. So, we do not take a uniform distribution for the number of years an author remains active in the network. We instead take a normal distribution, such that majority of the authors have their *Lifetime*' close to one value (which is given by the mean of a normal distribution) but at the same time, a few authors also have their *Lifetime* much greater or much smaller than what is predominant in the network.

7. *Lifetime* of each author signify the total number of consecutive time-steps an author is allowed to get new edges attached to it, after the time-step in which the author got added to the network.

- $Mean\_V$: mean of the normal distribution of *Lifetime* of the authors.
- $Stand\_D$: standard deviation of the normal distribution of *Lifetime* of the authors.

8. To build the network for the first time-step, we use Poisson distribution for the network's degree distribution, since the degree distribution of random networks follow Poisson distribution [19, p. 393].

- *Poisson_M* : sets the mean of the Poisson distribution. The maximum allowable degree for each vertex present in the first time-step is also assigned using the Poisson distribution.

Enhanced Preferential Attachment Model for Scientific Collaboration (EPAM-SC):. This model has exactly the same set of parameters as the previous model; the only difference in this model is that all the authors belonging to a given category (viz. old author or new author) do not have equal probability of getting a new edge attached to it, instead, it depends on the number of edges the author already has, which is essentially indicative of the number of papers the author has already co-authored in. Hence, this model using the concept of preferential attachment [3] is an enhancement over our initial model, where nodes with higher degree have a greater tendency to get more edges attached to them, compared to the peers who have relatively lower degrees. The other facets in this model include defining the predispositions the authors have towards collaborating with the new authors and the well-established authors of the field, the Poisson degree distribution of the network the model starts its simulations with, and the normal distribution that gives the number of years the authors stay active in the network.

Clique Structure

We now discuss how the co-authorship networks show predominant clique structures in them [25] and how our model simulates this property. In co-authorship networks, since an edge exists between two authors if they have co-authored a paper, all the authors of a paper have existing edges between them because they all co-authored with each other. So, for each paper in the network there exists a clique formed by all the co-authors of that paper mutually connected to each other by edges. Hence in both the two models, the authors of any given research paper form a clique. The number of authors, *say* x, for each new paper added to the network is chosen between 2 to 7. However, it is not chosen uniformly at random. We keep a bias towards choosing x between 2 to 5 because most of the research papers today involve 2 to 5 co-authors in general. Papers with 6 or 7 co-authors are generally very uncommon. So, we set a bias so that there is a 90% chance of x being chosen between 2 to 5, and a 10% chance of x being chosen between 6 and 7. Once x is chosen, x nodes are randomly selected from the network (keeping the type of the nodes, i.e new node or old node, as a constraint) and then edges are added between those x nodes such that they form a clique amongst themselves.

MODEL VALIDATION

Once each of the models are developed, our next step is to validate how well the model simulates the real networks. Our aim is to identify the optimal set of parameters that best simulates the real co-authorship networks. With reasons discussed before, we choose the network's clustering coefficient as a metric to compare the simulated networks with the empirical ones. For a vertex v in a network, its clustering coefficient (CC) is calculated as follows : If the number of vertices directly linked to v is d_v, then the maximum number of links possible between the d_v neighbours of v is given by $\binom{d_v}{2} = \frac{d_v(d_v-1)}{2}$. Out of these, if there are l_v links actually existing between the neighbours of v, then CC for vertex v = $\frac{2l_v}{d_v(d_v-1)}$. The clustering coefficient of an entire network (GCC) is given by the average CC across all vertices. We do not set the model parameters to the values observed exactly in the empirical data, since that might result in the over-fitting of the model. We instead randomly assign n = 70, f_v = 0.1, seed = 0.25, and *Poisson_M*= 10, and then tune the other

parameters, p_oo, p_on, p_nn, $Mean\_V$ and $Stand\_D$ that decide the individual level characteristics of the authors in every time-step. How the tuning is done with different configurations of the parameters for both the models have been discussed in the following sub-section. Since both the models involve a number of parameters that use probabilities and probability distributions, the model outputs are stochastic in nature. Therefore, we average the model outputs over substantial number of iterations and then we check which configuration of parameters gives the least MAPE value in each of the 5 domains. We report the final MAPE value after averaging over 40 runs for each of the models, so that our results are statistically significant. The particular configuration that gives the least MAPE value in a given domain, averaged over 40 runs, is chosen as the best fit configuration of parameters for that domain.

Validation Strategy

Each of the models, the ABM-SC and the EPAM-SC, is run for 6 different configurations of the edge probability parameters - p_oo, p_on and p_nn. Also, note that the sum of these 3 probability parameters should sum up to 1 because each edge will either be added between two old authors, an old author and a new author, or two new authors. The 6 different configurations of the edge probability parameters are –

- p_oo = 0.1, p_on = 0.3 and p_nn = 0.6
- p_oo = 0.1, p_on = 0.6 and p_nn = 0.3
- p_oo = 0.3, p_on = 0.1 and p_nn = 0.6
- p_oo = 0.3, p_on = 0.6 and p_nn = 0.1
- p_oo = 0.6, p_on = 0.1 and p_nn = 0.3
- p_oo = 0.6, p_on = 0.3 and p_nn = 0.1

We observe that the total number of time-steps in the empirical data of Software Engineering domain is 39, that of the Artificial Intelligence domain is 39, that of the Databse domain is 39, that of the Networks domain is 38, that of the Operating Systems domain is 35. Hence, the maximum number of time-steps in the empirical data among all the five domains is 39. Therefore, in a single run of the model, the maximum number of time-steps the model would run for is chosen as 39.

We then experiment with the $Mean\_V$ and $Stand\_D$ values and we find out which values of $Mean\_V$ and $Stand\_D$ give the best simulations for each domain. We run our models with the following 3 values of $Mean\_V$ and $Stand\_D$ -

- $Mean\_V$ = 20, $Stand\_D$ = 5
- $Mean\_V$ = 15, $Stand\_D$ = 3
- $Mean\_V$ = 10, $Stand\_D$ = 2

When $Mean\_V$ = 20 and $Stand\_D$ = 5, about 68% of the authors stays active in the network for 15 to 25 time-steps. About 27% of the authors stays active in the network either for 10 to 15 time-steps or for 25 to 30 time-steps. About 4% of the authors stays active in the network either for 5 to 10 time-steps or for 30 to 35 time-steps. These values change accordingly when $Mean\_V$ and Stand_D are varied.

With each of the above 3 configurations, we run our models 6 times, each time with a different configuration of the probability parameters p_oo, p_on and p_nn as listed earlier, which finally gives us 18 sets of configurations of the model parameters in total. Our aim is then to average the model outputs over 40 runs (where each run outputs the clustering coefficient values of 39 networks simulated over 39 time-steps), and then identify the set of parameters, the simulations of which are closest to the empirical data for each of the five research domains.

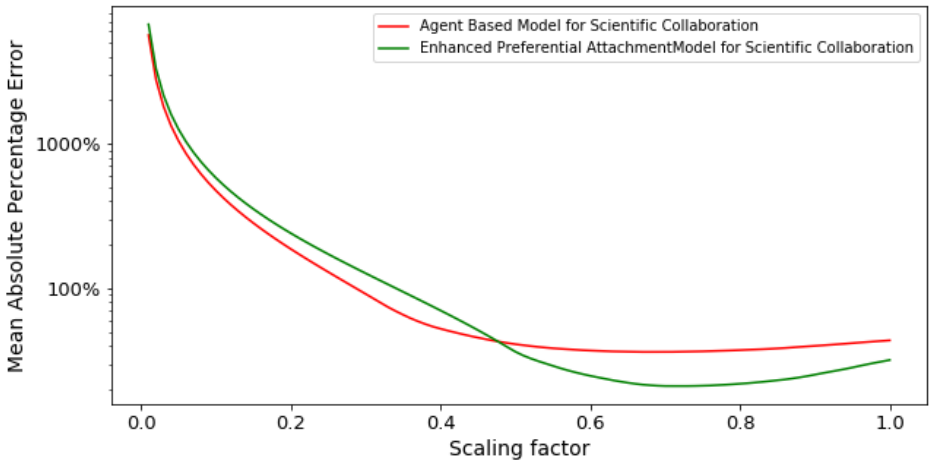

Fig. 2. Mean Absolute Percentage Error vs. Scaling Factor for both the models in the domain of SE for p_oo = 0.6, p_on = 0.3, p_nn = 0.1, $Mean\_V$ = 20 and $Stand\_D$ = 5.

Validation Results

Scaling factors : When comparing the clustering coefficient values of the simulations with the empirical data, it should be noted that the empirical networks are much bigger in size than the simulated networks. The empirical networks have tens of thousands of nodes (for example the empirical network of AI collaborations has approximately 39,000 nodes in its final time-step, whereas the simulated networks only have approximately 3,000 nodes in their final time-steps). The reason why the simulations were done in such low scales is the computational complexity in NetLogo, because of which it is very difficult to run models that simulate networks with sizes as huge as the empirical networks are. Therefore, before comparing the empirical data with the simulated results, we scale the empirical data using a scaling factor so that the empirical values of clustering coefficients can be scaled down to a smaller value, and hence the MAPE values can be minimised. Here scaling means multiplying each value of the empirical data with a scaling factor, which is a constant number. The optimal scaling factor, which is a fraction between 0 and 1, is found such that when each data-point in the empirical sample is multiplied by that fraction, and then the MAPE value is calculated, it gives the least MAPE value. Figure 2 shows a plot for MAPE values vs. Scaling factors for Software Engineering domain for both the models when the model parameters are set as p_oo = 0.6, p_on = 0.3, p_nn = 0.1, $Mean\_V$ = 20 and $Stand\_D$ = 5. It can be observed from the plot that as the scaling factor is increased from 0, the mean absolute percentage error decreases and then at a point it becomes minimum and then it increases again. The scaling factor for which the MAPE value is the minimum is taken as the best scaling factor for that model and for the given field. We further observe from Figure 2 that the MAPE value for both models fall precipitously for lower values of the scaling factor, but for ABM-SC, the MAPE values are near constant between 0.5 and 1.0 of the scaling factor. This indicates the utility of this model irrespective of how the outputs are scaled.

Findings : Table 2 shows the best configurations of parameters across the two models. Here SF denotes the best scaling factor corresponding to the domain and the model, Param denotes the

Table 2. The optimal set of parameters found after averaging over 40 iterations, in each domain, showing the model that had the lesser MAPE value in each domain.

Domain	Agent Based Model for Scientific Collaboration (ABM-SC)	Enhanced Preferential Attachment Model for Scientific Collaboration (EPAM-SC)	Better model for the domain
Software Engineering	MAPE = 18.24% SF = 0.48 Param = 1_3_6 Mean_V = 20 Stand_D = 5	**MAPE = 12.99%** SF = 0.62 Param = 3_1_6 Mean_V = 20 Stand_D = 5	EPAM-SC
Artificial Intelligence	MAPE = 7.87% SF = 0.48 Param = 1_3_6 Mean_V = 15 Stand_D = 3	**MAPE = 4.33%** SF = 0.69 Param = 6_1_3 Mean_V = 10 Stand_D = 2	EPAM-SC
Networks	MAPE = 15.43% SF = 0.37 Param = 1_3_6 Mean_V = 20 Stand_D = 5	**MAPE = 11.84%** SF = 0.48 Param = 1_3_6 Mean_V = 10 Stand_D = 2	EPAM-SC
Operating Systems	MAPE = 8.16% SF = 0.4 Param = 1_3_6 Mean_V = 20 Stand_D = 5	**MAPE = 5.58%** SF = 0.51 Param = 1_3_6 Mean_V = 20 Stand_D = 5	EPAM-SC
Database	MAPE = 13.74% SF = 0.39 Param = 3_1_6 Mean_V = 20 Stand_D = 5	**MAPE = 8.29%** SF = 0.46 Param = 6_1_3 Mean_V = 10 Stand_D = 2	EPAM-SC

edge parameters, such as Param = 1_3_6 means p_oo = 0.1, p_on = 0.3 and p_nn = 0.6, and so on. *Mean_V* and *Stand_D* are the mean and the standard deviation of the normal distribution used to assign the number of years each author stays active in the network. The findings clearly show that the Enhanced Preferential Attachment Model for Scientific Collaboration clearly outperformed the Agent Based Model for Scientific Collaboration since the MAPE values is lower for that model across all the five domains. Figure 3 shows 3D plots, one for each domain, which compares the MAPE value obtained with each of the 18 different configurations, after averaging over 40 iterations, across the SBM-SC Model and the EPAM-SM Model, shown using the legend. The plots show that the EPAM-SM Model consistently gives a lower MAPE value than the SBM-SC Model, for each of the 18 configurations of parameters, and across all the 5 domains. It is also observed that the MAPE values are between 4% to 13%, which shows that the model has been able to capture the clustering dynamics of the empirical collaboration networks to a reasonable extent, despite its simplicity.

Implications of Our Results

As it is done in any modeling activity, we have sought to isolate the essential elements of the complex dynamics of research collaboration in our models. Researchers operate in an ecosystem where peers engage with one another to fulfill collective goals. Accordingly, we have used the agent-based modeling paradigm as the foundation of our models. As established in the preceding discussion, by combining characteristics of ABM-SC with the preferential attachment mechanism, EPAM-SC is

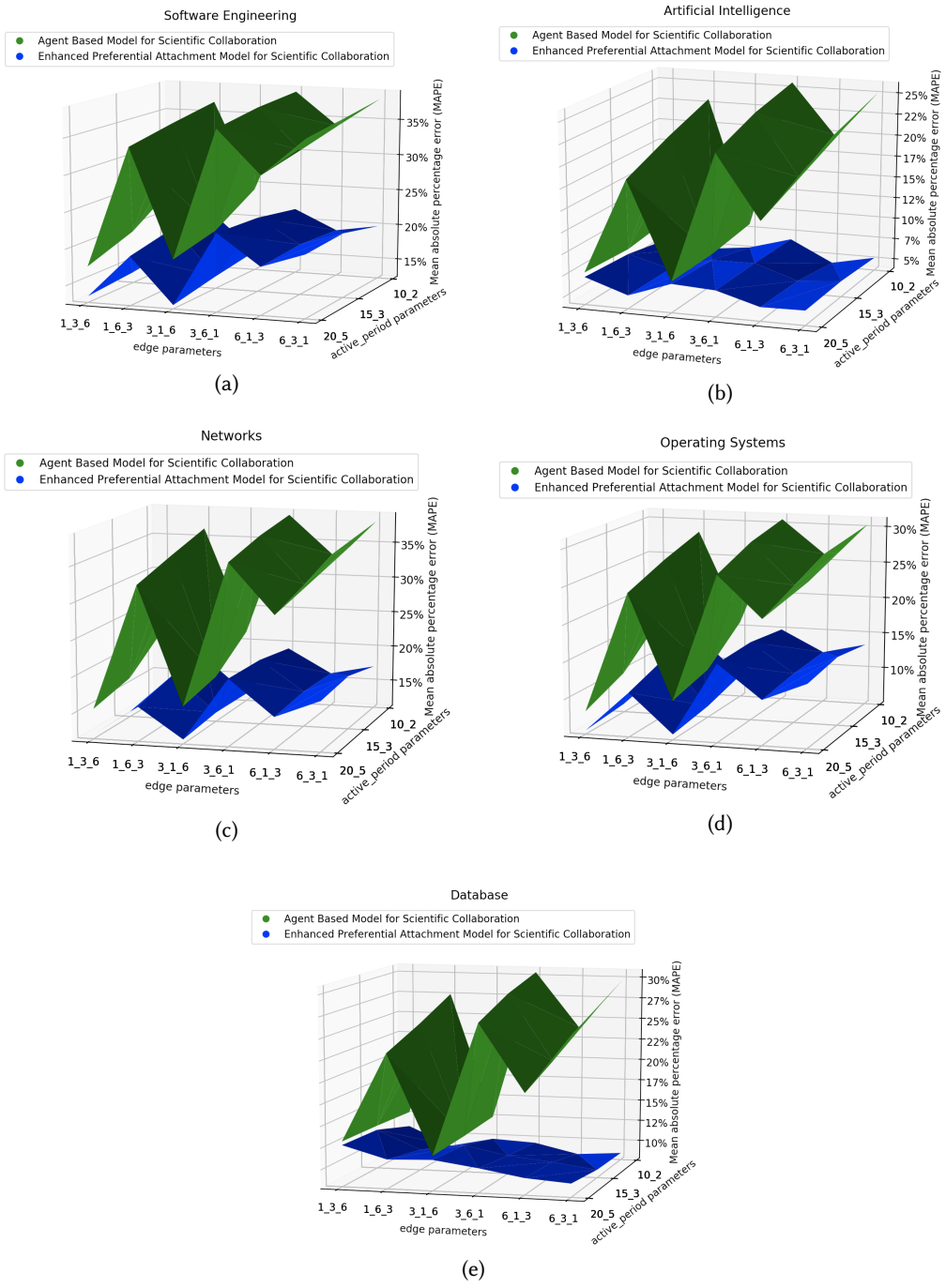

Fig. 3. MAPE value comparison with all 18 possible configurations of the model parameters across the two models, for all five domains. The x-axis has the edge parameters of the models, which are the configurations of the edge parameters such as '1_3_6'denotes p_oo = 0.1, p_on = 0.3 and p_nn = 0.6 configuration. The Y axis has the 'active period'parameters, which denote the values of $Mean\_V$ and $Stand\_D$ in the model, such as '15_3'denotes $Mean\_V$ = 15 and $Stand\_D$ = 3.

able to consistently offer a close fit to empirical data across all five computing domains we have studied. Preferential attachment finds wide application in diverse disciplines and is considered to be a fundamental driver of many interactive systems. However our results indicate the importance of augmenting this mechanism with factors that specifically relate to scientific collaborations, such as the probabilities of experienced researchers and newcomers collaborating within and among themselves, and the scientific lifespans of researchers. This has implications at multiple levels. For individual researchers, our results show the importance of choosing one's collaborators with a careful mix of youth and experience. Our results can inform research organizations in the setting-up and sustenance of research groups to successfully harness diverse researcher backgrounds. The burgeoning area of Science of Science (SciSci) aims to study the characteristics of scientific ecosystems using the same approaches that science brings to the study of natural and artificial phenomena. The progression from ABM-SC to EPAM-SC – along with the concomitant improvement in the accuracy of results – as demonstrated in this paper, offer a new SciSci perspective on how research domains in the computing discipline can be modeled and understood.

THREATS TO VALIDITY

To place our results in perspective, we discuss the following aspects of **threats to validity**: Threats to *construct validity* are concerned with the correct measurement of the variables. As explained in preceding sections, construction of the co-authorship network in this study follows a protocol that is common to studies in a similar context. Additionally, we have used some standard metrics from network science to represent the level of clustering in the networks. Thus, we do not see notable threats to construct validity. Threats to *internal validity* arise from the presence of systematic errors and biases. As we have validated our model using historical data from a publicly available bibliometric repository, common threats to internal validity from mortality and maturation of the subjects do not affect our results. Although we have included publications from major venues of the various computing disciplines considered in this study, we can not claim to have considered all research publications in each domain. This introduces a threat to internal validity; although the extent of this threat is minimal, given the overall sizes of our corpora. *External validity* is related to the generalizability of a study's results. As explained in preceding sections, our model has been validated using data from several domains within the computing discipline. Even as the model has been validated using data from various domains we do not claim our results to be generalizable as yet. Additionally, by introducing agent-based modelling as a tool in the study of collaboration dynamics, this study makes methodological contributions that can be generally useful in similar settings. A study's *reliability* comes from the reproducibility of its results. As discussed in preceding sections results from repeatedly running the model demonstrate a close match between the simulated and empirical values. The data used for validation is available in the public domain. Thus results from this study are reproducible.

SUMMARY AND CONCLUSIONS

In this study, we aimed at capturing the dynamics of co-authorship networks as they vary over time, by building agent-based models where the model parameters govern the individual-level characteristics and preferences of the authors. We then validate our models using empirical co-authorship network across five different research domains - Software Engineering, Artificial Intelligence, Networks, Operating Systems and Database. Our findings show that the Enhanced Preferential Attachment Model for Scientific Collaboration(EPAM-SC) consistently finds a closer fit to the empirical data across all the five domains mentioned above, and is enhanced with some additional parameters. Therefore this model, in spite of its simplicity, has been able to capture the dynamics of the collaboration networks to a reasonable extent.

REFERENCES

[1] Utku Ali Rıza Alpaydın. 2019. Exploring the spatial reach of co-publication partnerships of multinational enterprises: to what extent does geographical proximity matter? *Regional Studies, Regional Science* 6, 1 (2019), 281–298.

[2] Michael E Bales, Daniel C Dine, Jacqueline A Merrill, Stephen B Johnson, Suzanne Bakken, and Chunhua Weng. 2014. Associating co-authorship patterns with publications in high-impact journals. *Journal of biomedical informatics* 52 (2014), 311–318.

[3] ALBERT-LÁSZLÓ BARABÁSI. 2014. *NETWORK SCIENCE - THE BARABÁSI-ALBERT MODEL.* https://barabasi.com/f/622.pdf

[4] Laurent R Bergé. 2017. Network proximity in the geography of research collaboration. *Papers in Regional Science* 96, 4 (2017), 785–815.

[5] Gecia Bravo-Hermsdorff, Valkyrie Felso, Emily Ray, Lee M Gunderson, Mary E Helander, Joana Maria, and Yael Niv. 2019. Gender and collaboration patterns in a temporal scientific authorship network. *Applied Network Science* 4, 1 (2019), 1–17.

[6] Molly Callahan. 2019. *How Gender Bias Excludes Women From International Scientific Collaboration.* https://news.northeastern.edu/2019/08/07/how-gender-bias-excludes-women-from-international-scientific-collaboration/

[7] Subhajit Datta, Partha Basuchowdhuri, Surajit Acharya, and Subhashis Majumder. 2016. The habits of highly effective researchers: An empirical study. *IEEE Transactions on Big Data* 3, 1 (2016), 3–17.

[8] Ashkan Ebadi and Andrea Schiffauerova. 2015. On the relation between the small world structure and scientific activities. *PloS one* 10, 3 (2015).

[9] TS Evans, Renaud Lambiotte, and Pietro Panzarasa. 2011. Community structure and patterns of scientific collaboration in business and management. *Scientometrics* 89, 1 (2011), 381–396.

[10] Michael Gallivan and Manju Ahuja. 2015. Co-authorship, homophily, and scholarly influence in information systems research. *Journal of the Association for Information Systems* 16, 12 (2015), 2.

[11] Sanjeev Goyal, Marco J Van Der Leij, and José Luis Moraga-González. 2006. Economics: An emerging small world. *Journal of political economy* 114, 2 (2006), 403–412.

[12] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. 2005. Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308, 5722 (2005), 697–702.

[13] Luke Holman and Claire Morandin. [n.d.]. Researchers collaborate with same-gendered colleagues. ([n. d.]).

[14] Vaino Kannisto. 2001. Mode and dispersion of the length of life. *Population: An English Selection* (2001), 159–171.

[15] J Sylvan Katz and Ben R Martin. 1997. What is research collaboration? *Research policy* 26, 1 (1997), 1–18.

[16] Mehmet Ali Koseoglu. 2016. Growth and structure of authorship and co-authorship network in the strategic management realm: Evidence from the Strategic Management Journal. *BRQ Business Research Quarterly* 19, 3 (2016), 153–170.

[17] P. Parchman ML McDaniel RR Lanham HJ Agar M Leykum LK, Kumar. 1999-2016. *NetLogo User Community Models.* http://ccl.northwestern.edu/netlogo/models/community/Agent-Based%20Model

[18] Francesco Lissoni, Patrick Llerena, and Bulat Sanditov. 2013. Small worlds in networks of inventors and the role of academics: An analysis of France. *Industry and Innovation* 20, 3 (2013), 195–220.

[19] Mark Newman. 2018. *Networks.* Oxford university press. http://math.sjtu.edu.cn/faculty/xiaodong/course/Networks%20An%20introduction.pdf

[20] Mark EJ Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences* 98, 2 (2001), 404–409.

[21] Mark EJ Newman. 2004. Who is the best connected scientist? A study of scientific coauthorship networks. In *Complex networks.* Springer, 337–370.

[22] Raj Kumar Pan, Kimmo Kaski, and Santo Fortunato. 2012. World citation and collaboration networks: uncovering the role of geography in science. *Scientific reports* 2 (2012), 902.

[23] Austin J Parish, Kevin W Boyack, and John PA Ioannidis. 2018. Dynamics of co-authorship and productivity across different fields of scientific research. *PloS one* 13, 1 (2018).

[24] Henry T Robertson and David B Allison. 2012. A novel generalized normal distribution for human longevity and other negatively skewed data. *PloS one* 7, 5 (2012).

[25] Marcos Grilo Rosa, Inácio de Sousa Fadigas, Maria Teresinha Tamanini Andrade, and Hernane Borges de Barros Pereira. 2014. Clique approach for networks: applications for coauthorship networks. (2014).

[26] P. M. Swamidass (Ed.). 2000. *MAPE (mean absolute percentage error) MEAN ABSOLUTE PERCENTAGE ERROR (MAPE).* Springer US, Boston, MA, 462–462. https://doi.org/10.1007/1-4020-0612-8_580

[27] Jie Tang. 2005 - 2019. *AMiner: search and mining of academic social networks.* https://www.aminer.org/

[28] Uri Wilensky and William Rand. 2015. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo.* Mit Press.

[29] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science* 316, 5827 (2007), 1036–1039.