Chapter 1: Cultures of Collaboration: Simulating Scientific Co-Authorship Networks

WORKING PAPER. DO NOT REPRODUCE OR DISTRIBUTE WITHOUT DIRECT PERMISSION FROM AUTHOR

Authors: Peter Revay, Claudio Cioffi-Revilla

1.1 Introduction and Background

Scientific collaboration is a complex phenomenon that is crucial to the effective diffusion, communication and exchange of scientific knowledge. Scientific collaboration has arguably only increased in complexity and effect since the advent of modern information technology. It is perhaps for this reason, that research into networks of scientific collaboration has surged in frequency in recent decades. One of the most direct manifestations of the pathways of scientific collaboration is the co-authorship of scientific literature. The study of scientific co-authorship networks has originated in the social sciences (e.g. Endersby et al. 1996; Moody, 2004) but has since been joined by efforts in the natural sciences as well (e.g. Barabasi et al., 2002; Newman, 2004). The focus in most studies has always been the analysis of the structure of the co-authorship networks, whether quantitative or qualitative. Barabasi et al. (2004) have compared networks generated from different datasets and noted the scale-free structure of such co-authorship networks. They then applied the preferential attachment model to successfully generate similar structures. Newman (2004) has further corroborated the scale-free nature of co-authorship networks in his study. Many other efforts to probe the underlying mechanisms of scientific collaboration through network analysis have appeared

since: from analyzing the choice of data (De Stefano et al., 2013) and network definitions (De Stefano et al., 2011), through community detection approaches (Perianes-Rodrigues et al., 2010), qualitative approaches (Velden et al., 2010), to investigating the effects of network position on scientific performance (Abbasi et al., 2011, Abbasi et al., 2012).

These research efforts have two important elements in common: First, the focus is almost exclusively on network concepts as causal explananda, and second, the studied networks are either historical snapshots or cumulative graphs. We depart from this approach in both aspects. We study the issue of scientific collaboration as a multi-faceted one, and perceive co-authorship networks as a result of a combination of social mechanisms, cultural mechanisms, network structure effects and temporal (evolutionary) effects. Furthermore, we are interested in the analysis of co-authorship networks as constantly evolving structures, such, where pathways of collaboration can be forged, but also abandoned over time.

We thus propose and test a model of scientific co-authorship. In this model we assume that actors (scientists) make choices regarding collaboration with others based on both their *cultural* and *social* preferences, which are in turn affected by their current position in the network and past experiences of their own, as well as those of their previous collaborators. To test the model we develop an agent-based simulation.

The *cultural* aspect of the model rests on the logic of dual-inheritance theory, specifically on the mechanisms of indirectly biased transmission and guided variation; both being suggested drivers of cultural evolution (Boyd & Richerson, 1985). Here, indirectly biased transmission refers to the evaluation, adoption and subsequent diffusion of specific cultural behaviors and attitudes on the basis of possession of initially unrelated external markers. Meanwhile, guided variation refers to the evaluation and adoption of cultural behaviors based on self-generated and self-explored alternatives. We introduce these concepts into our model, because we posit (a) that cultural considerations, such as the actor's approach to research organization and management, communication, or writing, not to mention the choice of scientific paradigm (from potentially many within any given discipline) are all important determinants of successful collaborations, as measured by the actors' satisfaction

with the process and the end result, and (b) that these cultural factors are often unknown to others beforehand. Therefore, we assume, that the actors are often forced to rely on externally observable markers, or cultural signals, such as institutional affiliation, rank, past publication record, etc., as proxies for actual cultural behavior.

The *social* aspect of the model then rests on the observation that actors in social networks tend to cluster together (Watts & Strogatz, 1998), i.e. the probability of two actors collaborating increases with the degree to which their ego-based networks overlap. In fact, today we see many scientific collaborations where the individual actors do not necessarily know each other, and participate in the effort together solely on the basis of a shared acquaintance that also happens to be part of the project.

Finally, we model temporal and structural effects via an *evolutionary algorithm*. Evolutionary algorithms are used both as optimization tools, but also as a means to simulate the evolutionary dynamics of social systems (De Jong, 2005, p. 28). In our case, we assume that current actors may leave the network entirely (e.g. retirement), and that new actors may enter it at any point. Furthermore, we assume that new actors joining the network adopt cultural behaviors and preferences of current actors to a certain extent (e.g. graduate students from their advisors, post-docs from PIs, or junior faculty from senior faculty), and that successful, high-performing actors gain relatively more "disciples" on average than others.

Finally, we assume that the maintenance of active scientific collaborations is costly in terms of time, effort and resources.

To test the model we compare its output to empirical data on scientific publishing. Specifically, we use the Microsoft Academic Graph[1] (Sinha et al., 2015) which is a large database that includes information on over 126 million scientific publications, written by over 114 million different authors, from over 50,000 fields, over the course of more than 150 years. Despite certain shortcomings of the dataset, such as the extent of missing, incorrect, or duplicate data, it correlates well with other major publishing databases (e.g.

[1] we will further refer to the Microsoft Academic Graph by the abbreviation MAG.

CORE, Scimago, or Mendeley), and is currently considered the most comprehensive publicly available dataset of its kind (Herrmannova & Knoth, 2016).

In the following sections we first describe the agent-based implementation of our model in full detail. We then proceed to outline the experiment design and the data manipulation process. Finally, we report the observed results and discuss the most important findings.

1.2 Methods

We devise a model of the evolution of academic co-authorship networks which rests on *cultural* (or *institutional*) forces, as well as *social* forces. To test the model we carry out multi-agent simulations that are based on the model introduced in Chapter 4 and extended further to control for the specifics of the academic publishing context. As in the original cultural evolution model, the agents possess one of many possible variants of a trait that is unknowable to other agents *a priori*; similarly, the agents possess one of several possible *tags*, i.e. directly observable external markers. The agents seek to collaborate with others, but are only successful if their trait variants match those of their partners. The agents are therefore forced to rely on the tags to select adequate partners, and as a result they form distinct preferences for different tags over time.

There are two major departures from the original model in this version. The first change is related to population size. Unlike in the original model, here the population sizes are not necessarily constant. In fact, the populations keep growing as time progresses. This is true not only of the agent population, but also of the populations of possible trait variants and tags. As time progresses we assume that new variants and tags will be "discovered" or "invented". The second major change comes with the addition of a *social attraction* mechanism, which seconds the cultural mechanism of indirect bias as a driving force for the agent network evolution. We will discuss these additions in greater detail later in this section.

1.2.1 Entities and Variables

The model consists of a number of agents who are interested in collaborating to produce value. The agents are defined by a list of state variables. These are presented in table 1.1. Most of these are identical to the agent state variables described in Chapter 4. We have already discussed the *CulturalTrait* and *Tag*. As before, the *PositiveThreshold* is defined as the minimum base-level activation that a tag has to clear, for it to be considered by the agent when creating new ties. Similarly the *NegativeThreshold* is the minimum base-level acitvation a tag has to clear, for it to *not* be included when deleting ties. The *Neighborhood* is simply the set of all of the other agents that connect to the ego via an immediate link and *Fitness* is the sum of all successful interactions during an agent's lifetime minus the number of unsuccessful interactions. The *LastOutcome* variable keeps track of whether the most recent interaction was successful or not. The sole addition to the list of state variables is the variable *YearsLeft*. Unlike before, this version of the model is non-generational. Instead, new agents are introduced into the simulation and old agents leave on a step-by-step basis. The agent's lifetimes are initiated at their "birth" into the *YearsLeft* variable, and its value is decremented every step (or "year") to keep track of the agent's remaining lifetime, before it is retired from the population. Finally, as in the original model, the remaining variables figure in calculating the preference levels for different tags (their meaning is described in subsection 1.2.2).

Table 1.2 lists the global model parameters. We experiment with two configurations of the model. One in which both the *social* and the *cultural* mechanism are in play, which we will refer to as the *biased* configuration, and a baseline configuration, in which these mechanisms are omitted. We refer to the baseline as the *unbiased* configuration. We explain the differences between the two configurations in subsection 1.2.2. The *InitPopsize, InitTraits,* and *InitTags* give the initial number of agents, available trait variants and tags at the beginning of the simulation. The *PopGrowth, TraitGrowth,* and *TagGrowth* variables control the rates at which new agents, trait variants or tags are introduced into the simulation at each step. As before, the *AdjacencyMatrix* represents the initial configuration of the

Name	Domain	Scale	Type
CulturalTrait	Integer	Categorical	Static
Tag	Integer	Categorical	Static
PositiveThreshold	Integer	Cardinal (ratio)	Static
NegativeThreshold	Integer	Cardinal (ratio)	Static
Neighborhood	List of agents	Categorical	Dynamic
Fitness	Integer	Cardinal (ratio)	Dynamic
LastOutcome	Ordered pair of integers	Boolean/categorical	Dynamic
YearsLeft	Integers	Cardinal ratio	Dynamic
NumGood	List of integers	Cardinal (ratio)	Dynamic
NumBad	List of integers	Cardinal (ratio)	Dynamic
FirstGood	List of integers	Cardinal (ratio)	Dynamic
FirstBad	List of integers	Cardinal (ratio)	Dynamic
LastGood	List of integers	Cardinal (ratio)	Dynamic
LastBad	List of integers	Cardinal (ratio)	Dynamic

Table 1.1: Agent variables

agent social network and the *SuccessPayoff* and *FailurePayoff* give the fitness increments (decrements) for each successful (unsuccessful) interaction between two agents. Similarly, the *MaintenanceCost* refers to the cost the agent bears every step for maintaining a single link to another agent and the *MutationRate* defines the probability with which agent states are stochastically modified after each step. The value of *SocialWeight* indicates the relative strength of the social attraction mechanism in the model. Finally, the *TimeMatrix* holds information on the next scheduled activation of links between agents: the value stored in position (i, j) refers to the next activation of the link between agents i and j. The *TagInnovation* and *TraitInnovation* variables give the probabilities of adopting newly discovered trait varians and tags by surviving agents. The *Lifespans* distribution is used for sampling lifetimes when agents are created. Similarly, the *ActivationIntervals* distribution is used to sample times remaining until the next activations of links in the agent networks.

1.2.2 Process Overview and Scheduling

The model processes can be broken into three distinct parts. This includes *neighborhood maintenance* along with *interaction* done by the agents, and the *evolutionary algorithm* which controls the nature of the agent, trait and tag populations over time. In each step

Name	**Domain**	**Scale**
Configuration	Integer	Categorical
InitPopsize	Integer	Cardinal (ratio)
InitTraits	Integer	Cardinal (ratio)
InitTags	Integer	Cardinal (ratio)
PopGrowth	Floating-point number	Cardinal (ratio)
TraitGrowth	Floating-point number	Cardinal (ratio)
TagGrowth	Floating-point number	Cardinal (ratio)
SuccessPayoff	Integer	Cardinal (ratio)
FailurePayoff	Integer	Cardinal (ratio)
AdjacencyMatrix	Matrix of Booleans	Categorical (Boolean)
TimeMatrix	Matrix of integers	Cardinalratio
MaintenanceCost	Floating-point number	Cardinal (ratio)
SocialWeight	Floating-point number	Cardinal ratio
TraitInnovation	Floating-point number	Cardinal ratio
TagInnovation	Floating-point number	Cardinal ratio
Lifespans	Histogram of integers	Cardinal (ratio)
ActivationIntervals	Histogram of integers	Cardinal (ratio)
MutationRate	Floating-point number	Cardinal (ratio)

Table 1.2: Model Parameters

every agent is activated and gets a chance to maintain its neighborhood, and potentially to interact with one other agent. The order in which agents are activated is randomized at the beginning of each step.

The *neighborhood maintenance* phase is carried out similarly as in Chapter 4. First, each agent determines its sets of preferred and disliked tags. Additionally, the agent then computes a *social* score for each of its current neighbors. The social score largely depends on the degree to which the neighborhoods of the two agents overlap. If a current neighbor fails to reach a certain threshold degree of overlap *and* it possesses a disliked tag, the agent will cut ties to it. In a similar fashion, the agent will propose to create a new link to anyone with whom it is not yet connected, if their neighborhoods overlap sufficiently *and* the other agent possesses a preferred tag. Once more, link creation is a mutual act, and thus, the link is created only if the other agents accepts by following the same protocol. At the conclusion of this phase we adjust the agents' fitness values by the maintenance costs, which is determined by the number of their connections.

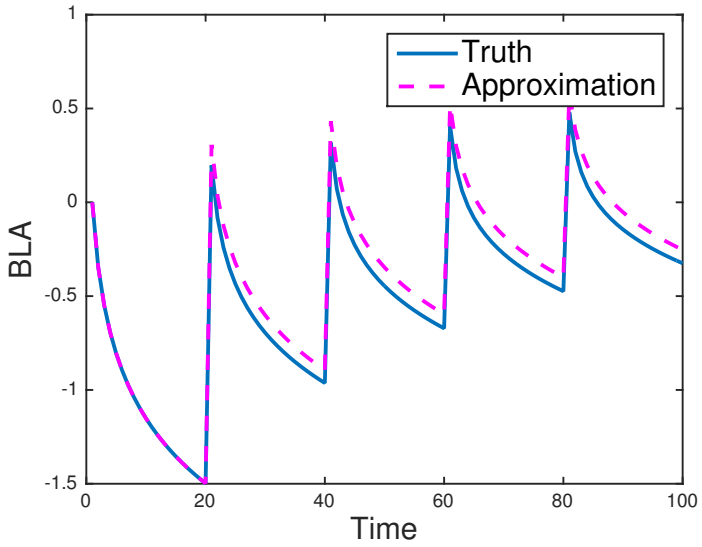


Figure 1.1: Example of a base-level activation for some input over time. The spikes in the chart coincide with instances of processing the input, followed by gradual decay.

The preference for a given tag is determined by calculating its base-level activation value. This quantity is taken from the ACT-R memory model (Anderson & Lebiere, 1998). The base-level activation τ_T^+ for successful interactions with tag T is calculated as follows:

$$\tau_T^+ = \ln\left[\sum_i^n t_i^{-d}\right] \approx \ln\left[t_n^{-0.5} + \frac{2(n-1)}{\sqrt{t_1}+\sqrt{t_n}}\right] \quad \text{(when } d = 1/2\text{)} \tag{1.1}$$

Here t_i is the time elapsed since the i-th successful interaction with an agent bearing tag T, while $n = NumGood_T$ is the total number of such experiences, $t_n = LastGood_T$ is the time since the most recent experience, and $t_1 = FirstGood_T$ is the time since the first experience. Finally, d is the rate of decay. Due to the computational complexity of the above relationship for large n, we use the above approximation (Petrov, 2006). In line with convention, we use $d = \frac{1}{2}$. Dislike for a given tag is calculated in the same way, using unsuccessful interactions as input. Figure **??** shows an example of the change in base-level activation for some input over time.

The *social* score of an ordered pair of agents (i, j) is computed in multiple steps. First, the overlap $\mathcal{O}$ in their neighborhoods is defined as:

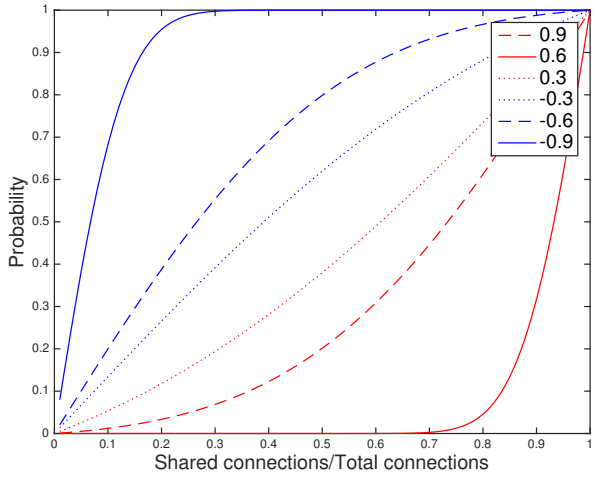


Figure 1.2: Examples of the the social force function F for different values of x.

$$\mathcal{O} = \frac{|N_i \cap N_j|}{|N_i|}.$$

Here, $|N_i|$ is the size of i's neighborhood. Next, we take the truncated normal distribution on the interval $[0, 1]$ with $\mu = 0, \sigma = 1 + w$, if $w < 0$, and $\mu = 1, \sigma = 1 - w$, if $w > 0$. Finally, we take the value s of the CDF of the chosen distribution at $\mathcal{O}$ (Figure **??** shows examples of the CDFs for several values of w). This gives us the social score.

When deleting links, we take $w = -SocialWeight$ and the social condition for deletion is met when $s < r$, where r is a random draw from the uniform distribution on the interval $[0, 1)$. When creating new links, we set $w = SocialWeight$, and the condition is met when $s \geq r$.

The *interaction* phase of the simulation is executed exactly as in Chapter 4: The two agents simply compare their trait variants, and if they match they both increment their fitness by the *SuccessPayoff*. If the trait variants do not match, the fitness of both agents is decremented by the *FailurePayoff*.

Once every agent has taken its turn in the current step, the evolutionary algorithm is invoked. First, the number of new trait variants, tags and agents is calculated. Next, any

agents scheduled for removal are removed from the simulation. Finally, new agents are created and added to the network.

Reproduction is performed locally. For each new agent that is set to enter the population a random node location in the network is chosen. A set of candidates for that location is assembled by taking the agent occupying the chosen node itself along with all of its immediate neighbors. Once the candidate set is defined, we calculate the mean and standard deviation of the fitness distribution within this set. Next, we select those candidates whose fitness is at least one standard deviation above the set mean to become the *parents* to the new agent. The agent is created by performing uniform attribute-wise crossover on the parent set, i.e. for each attribute the new agent copies its value from one of the parents, chosen with uniform probability. Finally, mutation is introduced by modifying each agent attribute by a small amount with probability equal to *MutationRate*. For categorical variables mutation amounts to uniformly random switching. In the case of cardinal variables mutation is carried out by adding small perturbations sampled from Gaussian distributions. The *Fitness* of a new agent is set to zero. Once the agent is created it is automatically linked to its parents.

The links between surviving agents remain intact during the evolutionary procedure. However, the trait variants and tags of surviving agents may be modified with a small probability equal to *TraitInnovation* and *TagInnovation* respectively, to reflect occasional in-life adoption of new tags or variants.

1.2.3 Initialization and Inputs

A complete specification of the model is given by providing the values of the model parameters in table 1.2. The agents are placed on nodes of the network specified by the initial *AdjacencyMatrix*. The trait variants and tags are assigned to agents uniformly from the initial distributions, and independently of each other. Each agent is assigned a lifetime by sampling the *Lifespans* distribution. The *TimeMatrix* tracks the times to the next scheduled activation of specific links in the network (i.e. interactions between specific pairs of

agents). Once a link is activated and the interaction between agents i and j takes place, the value of $TimeMatrix_{ij}$ is re-seeded with a random variate from the *ActivationIntervals* distribution.

1.2.4 Experiment Design and Data Processing

The tested model parameter values are listed in table 1.3. We tested the sensitivity of the model to four variables: the model configuration (biased, unbiased), the cost of maintaining social relationships, the weight of the social mechanism and, crucially, the scientific *field*.

To control for differences between scientific disciplines we randomly chose the fields of Economics and Artificial Intelligence. Our only requirements for the choices of fields were that they are not too obscure (i.e. that the number of authors and publications in the field is substantial) and that they have a significant history. The MAG database holds records of over 50,000 scholarly publications by thousands of different authors in both Economics and Artificial Intelligence. However, certain differences remain: Artificial Intelligence (first appearing in the database in 1946) is younger than Economics (first appearing in 1931), and AI is currently just reaching its peak volume and popularity, while Economics has been long considered an established discipline.

Working with the MAG, we first interrogated the *Fields of Study* table and filtered out all of the fields whose name included the (case-insensitive) strings "economics" or "artificial intelligence" respectively. We then filtered the *Papers* table to include only those publications whose *Field* value matched one of those extracted in the previous step. Finally, we then joined the *Papers* and the *Authors* tables on the *Title, Author* and *Year* fields, to create a table of unique Paper–Author–Year triplets for each of the two field groups.

The population growth rates were established by analyzing the totality of authors and their appearances in the MAG for the two respective fields. The counts of new author appearances in given years were then fitted with exponential models for each field separately, using non-linear least squares. These models were then used to generate the number of new agents introduced into the simulation at each time step. We took a similar approach when

modelling the lifespans and activation intervals for the two fields. We measured the career span of each author by noting the time elapsed between their first and last publications in the given field. We also measured the interval lengths between successive co-authored publications for all pairs of co-authors in the respective fields. We then created histograms from the observed values (binned by year). We ran the simulations for 70 steps, which is equal to the number of years for which Artificial Intelligence (the "younger" of the two fields) appears in the MAG prior to 2015 (the last complete year in the database at the time of experimentation).

The initial population size was set, somewhat arbitrarily to 8 agents. In both fields, there were only a few active authors in the beginning years. However, we chose a slightly higher number, because for lower initial population sizes the number of runs that went "extinct" prematurely in the first few steps (due to stochasticity in the agent lifespans coupled with the chosen growth models) was inconveniently high. However, preliminary observations concluded that final results were not significantly affected by slight changes in the initial population size. The values of the *InitTraits, InitTags, TraitGrowth* and *TagGrowth* variables were also chosen with a degree of arbitrariness. We initially intended to deduce these values from analyzing the MAG as well (by observing variables such as the numbers of unique journals, institutional affiliations, etc.), however we soon learned that the records for the relevant fields were insufficiently reliable, with an excessive number of duplicities, missing values and errors. We were thus forced to provide our own values, which we believe represent reasonable estimates (we posit that growth in cultural trait variants is faster than in relevant external markers, such as institutional affiliation, rank, possible publication venues, etc.). We carried out 100 simulations for each tested parameter combination.

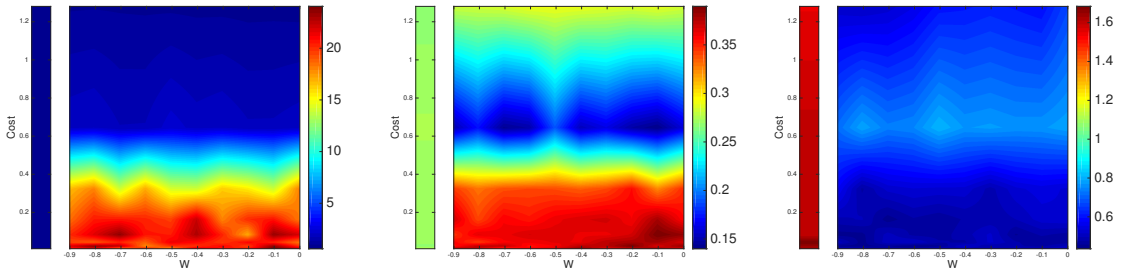
Name	Value
Configuration	biased, unbiased
NumSteps	70
InitPopsize	8
InitTraits	3
InitTags	3
PopGrowth	Economics, AI
TraitGrowth	0.04
TagGrowth	0.02
TraitInnovation	0
TagInnovation	0
Lifespans	Economics, AI
ActivationIntervals	Economics, AI
SuccessPayoff	1
FailurePayoff	-1
AdjacencyMatrix	random
MutationRate	0.01
MaintenanceCost	$(0, 1]$
SocialWeight	$(-1, 0]$

Table 1.3: Model parameter values

1.3 Results

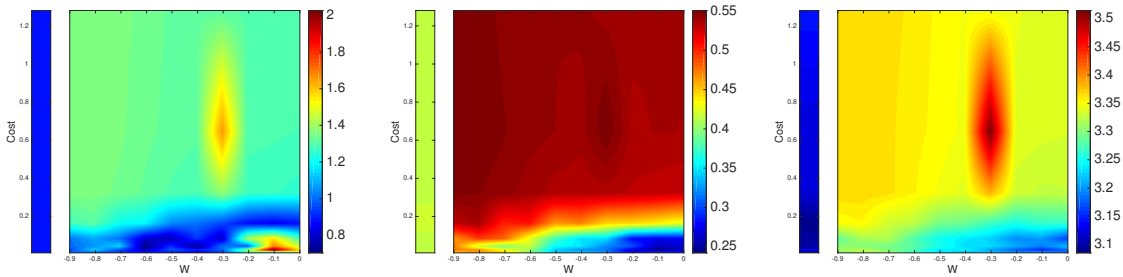
We measure three important attributes of the agent networks as they evolve over time: the average local clustering coefficients, average node degree, and the average shortest path lengths. We then compare the simulated values with the values observed in the data extracted from the MAG. Figures 1.3 and 1.4 show the root-mean-squared errors between the simulated and empirical time series for each of the three attributes as a function of *MaintenanceCost, SocialWeight* and the model configuration. The root-mean-squared error is measured only after the first 20 steps (years), once the empirical populations are large enough and the trends settle sufficiently (see figures 1.5 and 1.6 for comparison of the actual time series). We first notice that there are significant differences in both the performance and the sensitivity of the models between the two fields. One may observe, for example, that while the Economics simulations seem to be sensitive only to the value of *MaintenanceCost*, the AI simulations show sensitivity with respect to both the *MaintenanceCost* and the *SocialWeight*. Furthermore, the scale of the errors differs as well: where the average degree

RMSE ranges from negligible levels to over 20 in the case of Economics, it remains between 0.7 and 2.1 for the AI runs. Similar discrepancies can be seen in the clustering coefficients and the average path lengths of the networks.



(a) (b) (c)

Figure 1.3: Heatmaps of the RMSE in terms of (a) average degree, (b) clustering coefficient and (c) average path length relative to the Economics data for the biased (right) and the unbiased (left bar) model configurations as a function of social weight and maintenance cost.



(a) (b) (c)

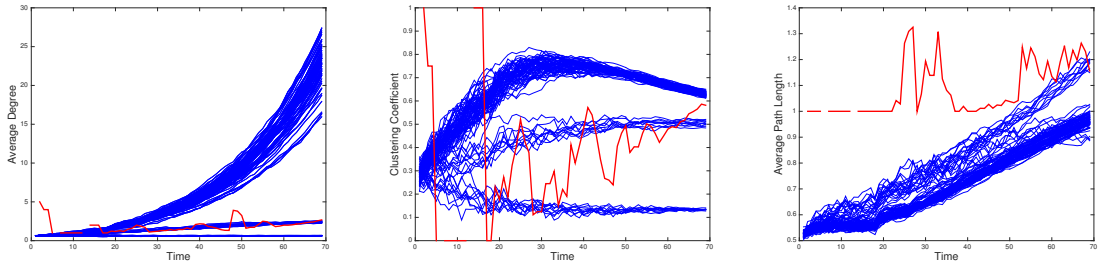
Figure 1.4: Heatmaps of the RMSE in terms of (a) average degree, (b) clustering coefficient and (c) average path length relative to the AI data for the biased (right) and the unbiased (left bar) model configurations as a function of social weight and maintenance cost.

As we focus on the AI simulations, we note that in the biased model the error in terms of the clustering coefficient as well as the average path length decreases with both the *SocialWeight* and the *MaintenanceCost*. A similar trend appears at first in the case of average degree, however it becomes noticeably reversed at the furthest extremes of the tested ranges. Nonetheless, we observe that the unbiased model is able to follow the empirical trends in network evolution just as well, if not better, in terms of average path length and

average degree. It also performs well in terms of clustering coefficient, relative to the biased model.

The situation is different when considering the Economics simulations. First, we notice that in certain regions of the search space the biased model produces a better bit fit than the unbiased configuration in terms of clustering coefficient and average path length. It also performs similarly well in specific regions in terms of average degree.

Moreover, we point out that the biased model seems to achieve the best goodness of fit in the critical region of $MaintenanceCost \approx 0.6$. In fact, one may observe distinct trifurcations in the resulting network characteristics (see figure 1.5), with one of the branches following the empirical trends significantly better than the others. As we move away from this region into lower values of *MaintenanceCost*, the error increase both in terms of clustering coefficient as well as average path length. In the opposite direction, as *MaintenanceCost* increases, the clustering coefficient error increases once again. Meanwhile, the unbiased model does not seem to be significantly sensitive to changes in the *MaintenanceCost* parameter; this observation holds in the AI simulations as well.

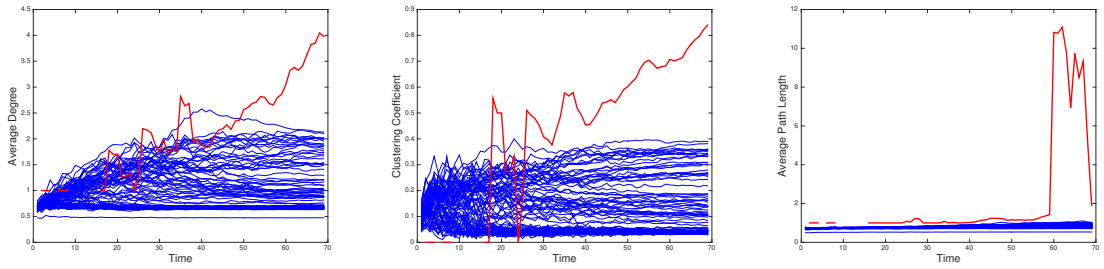


(a) (b) (c)

Figure 1.5: Time series of the (a) average degree, (b) clustering coefficient and (c) average path length for the Economics biased model runs (blue) and the Economics data (red).

1.4 Discussion

The model results are illustrative in multiple aspects. First, they suggest that different scientific fields are marked by different sets of mechanisms driving collaboration. Our model



(a) (b) (c)

Figure 1.6: Time series of the (a) average degree, (b) clustering coefficient and (c) average path length for the AI biased model runs (blue) and the AI data (red).

performed very differently when applied to the Economics networks rather than the Artificial Intelligence networks. In the case of the Artificial Intelligence populations, neither the cultural nor the social mechanisms showed any significant effect on the resulting co-authorship networks. The parameters that seemed to matter were the cost of maintaining relationships and the weight of social forces. Conversely, the Economics populations showed good fit with the biased model, but demonstrated sensitivity only to costs of collaboration and not with respect to the social weight. This result further carries the implication that in certain scientific populations the hypothesized social and cultural mechanisms potentially do guide the dynamics of the system.

The fact that the AI and Economics cultures are different, are reflected in the empirical measures of their network structures. This is not as surprising, considering that the field of Artificial Intelligence is rooted mainly in computer science, while Economics is mostly categorized among the social sciences. The cultural differences between the social and the natural, or exact sciences have been noted before. For example, the rate of co-authorship (as opposed to sole authorship) is significantly higher in the natural sciences (Moody, 2004). This is perhaps reflected in the higher observed clustering coefficient and the higher average observed degree in the AI networks. Moreover, computer scientists are specific in that they give disproportionately more weight to conference articles rather than journal articles, relative to other disciplines. Other cultural differences might be at play, such as thresholds for co-authorship, or the number of appropriate publication venues.

We also suspect that another reason for the seemingly different modes of collaboration is the time of maturation of both fields. Because Economics has matured earlier, any culture that has been established in the field in the past is potentially deeply rooted in the ways collaborations are forged. Meanwhile, the younger field of AI could be potentially more influenced by the recent proliferation of modern channels of communication and knowledge exchange.

Although the model that we have developed and analyzed has shown some promise, we are wary to draw any hard conclusions, as we were forced to leave a large region of the parameter space unexplored. This was mostly due to the lacking availability of data. Although the MAG dataset is currently the most extensive and complete database of its kind, it still has its shortcomings. This is reflected in our limited ability to operationalize concepts such as the growth in cultural and external markers, or the rates of innovation. We believe that, as the Information Age matures, availability of data of this nature will only increase, in turn extending the possibilities for testing and validating models such as the one presented here.

1.5 Conclusion

We have developed and tested an agent-based model of scientific collaboration founded on social, cultural, as well as evolutionary principles and the complex interactions among them. We have validated the performance of our model against extensive empirical co-authorship data in two scientific disciplines. The results show that there are large differences in the underlying mechanisms of collaboration between scientific fields. They also demonstrate that under certain assumptions the proposed model fits well with the empirical data. However, further refinement of the model with respect to its parameters is needed, resulting in an antecedent need for higher quality and higher resolution data.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Infometrics*, *5*(4), 594-607.

Abbasi, A., Chung, K. S. K., & Hossain, L. (2012). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, *48*(4), 671-679.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: LEA Publishers.

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schebert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3-4), 590-614.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago, Illinois: University of Chicago Press.

De Jong, K. A. (2005). *Evolutionary Computation: A Unified Approach*. Cambridge, Massachussetts: MIT Press.

De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, *35*(3), 370-381.

De Stefano, D., Giordano, G., & Vitale, M. P. (2011). Issues in the analysis of co-authorship networks. *Quality and Quantity*, *45*, 1091-1107.

Endersby, J. W. (1996). Collaborative research in the social sciences: multiple authorship

and paper credit. *Social Sciences Quarterly*, *77*, 375-392.

Herrmannova, D., & Knoth, P. (2016). An Analysis of the Microsoft Academic Graph. *The Magazine of Digital Library Research*, *22*(9-10).

Moody, J. (2004). The structure of a social science: disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*, 213-238.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(1), 5200-5205.

Perianes-Rodriguez, A., Olmeda-Gomez, C., & Moya-Anegon, F. (2010). Detecting, identifying and visualizing research groups in co-authorship networks. *Scientometrics*, *82*(2), 307-319.

Petrov, A. A. (2006). Computationally Efficient Approximation of the Base-Level Learning Equation in ACT-R. In D. Fun, F. Del Missier, & A. Stocco (Eds.), *Proceedings of the seventh international cconference on cognitive modeling.* Trieste,Italy: Edizioni Goliardiche.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hu, B. J., & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th international conference on world wide web (www 15 companion).* New York, NY: ACM.

Velden, T., Haque, A., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics*, *85*, 219-242.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440-442.

Curriculum Vitae

Peter Revay has earned bachelor's degrees in Mathematics and Sociology from the Masaryk University in Brno, Czech Republic in 2012. He has then attended University of Vermont where he received his Master of Science in Mathematics along with a Certificate in Complex Systems in 2014. In 2015 he has also received a Master of Science in Sociology from the Masaryk University. Peter started his PhD in Computational Social Science at the George Mason University in 2014. He has served as Graduate Research Assistant on multiple projects for Dr. Claudio Cioffi-Revilla.