Simulate the Evolution of Scholarly Network via Agent-based Modeling

Yingying Yu
College of Transportation Management
Dalian Maritime University
Dalian, China, 116026
uee870927@126.com

ABSTRACT

Agent-based modeling is a new modeling paradigm constituted by multiple autonomous agents and lots of basic rules to guide their behaviors and achieve their goals. In this paper, we propose a novel and flexible simulation approach using agent-based modeling for future scholarly network prediction. This model includes three components. 1) Agent: we represent author and publication as heterogeneous and self-contained agents that can make decisions for their future behaviors; 2) Relation: we treat the connections in publication repository as heterogeneous relations between agents; 3) Rule: the most important component that governs all the agents' behaviors and control the process of evolution. In accordance with the regulation of scholarly network evolution, we emphasize three generalized rules for future network simulation. Based on that, a series of specific rules combined with linear regression model and meta-path based random walk algorithm are elaborated on how to estimate new publications, how to compete for the coauthor and citation candidates and predict new links. Finally, the experiment results show that the proposed framework has a good performance on the simulation of future scholarly network.

Categories and Subject Descriptors

I.2.0 [**Computing Methodologies**]: Cognitive simulation

General Terms

Algorithms, Experimentation, Measurement

Keywords

Agent-based Modeling, Agent, Rules, Scholarly Network, Meta-path

1. INTRODUCTION

Recently there has been a lot of researches on the scientific network analysis, especially on the author collaboration network and paper citation network. Generally, scholars tend to focus on this problem from different perspectives. The very popular problem is link prediction problem such as infer the coauthor-ship or the citation possibilities in the future. Barabâsi [2] studied the evolution of scientific collaborations analytically and numerically. Han [8] predicted coauthor-ship combining the local and global network features. Sometimes, the researches only focused on the homogeneous network and adopted complex equation models to solve the problem. While different authors and papers have unique characteristics and behave differently, and thus is hard to find a closed-form equation model. Besides, the network is dynamic as a whole. It keeps expanding with the addition of new nodes and links. Also, for collaboration network and citation network, they are not independent. Börner [5] made a study and introduced a general process model to illustrate the simultaneous evolution of author and paper networks. In this paper, we view the whole scientific network as an evolving network containing numbers of different types of vertices and links. We take the interactions of the entities, which play an important role in the evolution of the scholarly network, into consideration. Our formulation differs from other models is that we employ the agent-based modeling to simulate the scientific publication repository from the overall point of view, which has been unprecedented in scholar network analysis up to our knowledge.

Agent-based modeling (ABM) is a relatively new approach to modeling complex systems with autonomous and interaction agents [13]. Applications of ABM are becoming widespread and presented in lots of fields, such as market analysis, manufacturing system [15] and so on. It turns out to be very effective when the system is complex and hard to analyze by equation modeling. In the simulation model, the agents with multiple attributes drive their behaviors to reach the goals. They adapt their behaviors to the environment and communicate with others according to the rules. This provides us an opportunity to simulate the scientific repository by ABM and discuss the interactions among the agents.

The object of this study is to simulate the construction of scholarly network and show the landscape of academic science using ABM. **The main contribution of our work is to propose a new idea on how to simulate the evolution of scientific publication repository**. There are three main components in our framework, including agent, relation and rule. The entities in the heterogeneous scholarly network are regarded as agents, which are self-contained and interactive. In this paper, we study a scholarly network with two kinds of agents and three types of relations. The author agent takes a leading role in the evolution who is able to make decisions and act according to a list of rules.

The prediction for possible links is designed as a series of competitions among agents, including publication competition and citation competition. For different kind of competitions, there are different detail rules for agents to follow. Our primary focus is on how to make up the rules rather than design complicated math-

ematical models, so that the agents will iteratively communicate with other agents and adapt themselves to the environment until they reach the goal. We introduce regression model and meta-path based agent competition rule is in our framework. The regression model will be adopted for future publication number prediction for each author. When competing for the candidate coauthors and citations, we will use random walk based on a list of meta-paths to search and calculate the ranking score of candidates.

In the remainder of this paper, we will : 1) review relevant works and methodologies for scientific prediction and agent-based modeling, 2) discuss the main problems and solutions, 3) describe the experiment setting and results, and 4) discuss the findings and limitations of the study and identify subsequent research steps.

2. LITERATURE REVIEW

In this section, we will review previous studies focusing on scientific prediction and agent-based model.

2.1 Scientific Publication Repository Prediciton

Prediction for future networks has gradually caught scholars' attention. The main focus is on the link prediction. Given a snapshot of a network, how to infer the edges that will be added to the network in the next time step is the link-prediction problem. A lot of methods were proposed to solve this problem. In [10], Liben-Nowell discussed different predictors performance about link prediction based on measures for analyzing the "proximity" of nodes in a standard network with homogeneous edges. The main focus is on the prediction of interaction between authors using coauthorship networks. However, it treats all the relationship equally or by separate the network into homogeneous network to study and ignore the dependency patterns across types. Such approaches lead to lossy representation of data. In order to deal with this problem, link prediction in heterogeneous networks is presented in [17] and [7]. The heterogeneous networks include multiple types of vertices or multiple types of links. Based on heterogeneous, [16, 17] proposed a meta-path based method to predict co-author relationship. [1,11] approached the link prediction as a supervised problem and explored many important features to reflect actual relationships and to build the learning model. [2,18] dug much into the evolution of co-authorship network. They argued that the network evolution is related with preferential attachment. However, in most occasions the nodes or links cannot stay forever. In [18], it took the aging of collaboration into consideration. Some authors especially students may stop writing papers when they leave school, so the collaboration of authors can be ceased after a given time window. In academic science, there are also a lot of researches on prediction of citation network. Statistical relational learning integrated with feature selection to building link prediction models was proposed in [14]. The model was used for citation prediction in the domain of scientific publications. Another study [6] predicted the future citation pattern of individual papers by stochastic model. [13] used undirected citation graph rather than directed graphs and illustrated three centrality measures. It found the correlation between future times cited and three centralities and proved that the papers that will take many citations were at a similar topological position in the past.

The proposed work differs from previous research in that we use agent-based modeling to simulate the evolution of scholarly network. The prediction is related to the time span. We intend to predict not only the nodes, but also the heterogeneous network with multiple relationships.

2.2 Agent-based Modeling

Agent-based modeling is a new approach to modeling systems comprised of autonomous, interacting agents. In [13], Macal gives detail introductions about agent's characteristics, the structure of agent-based modeling, several modeling applications and the implementation tools. In [9], a lot of agent-oriented methodologies are reviewed. It emphasized that there is no standard agent architecture, but a conceptual level for analyzing the agent-based system. The level should describe the characteristics of each agent and the relationships and interactions between agents. Agent-based simulation provides a new paradigm for simulating complex systems with many interactions among the entities in the system. Sophisticated ABM sometimes incorporates neural networks, evolutionary algorithm, or other learning techniques to allow realistic learning and adaptation [4]. In [4], it refers to three benefits of agent-based modeling that ABM captures emergent phenomena, provides a natural description of a system and is flexible. So that, many new agent-based modeling applications emerged in the last decades. Combination with other mining method, like using liner regression in [3]. In the micro-level, agents try to build local linear model by competition with each other, while in macro-level they build the global structure to simulate the cooperation with each other.

In our framework, we adopt linear regression model and meta-path based random walk method to help agents make decisions and compete with others.

3. PROBLEMS AND SOLUTIONS

As we know, in scientific publication repository, there are many kinds of important entities, including papers, authors, venues and so on. We can draw different heterogeneous networks containing different entities, like Figure 1 shows.

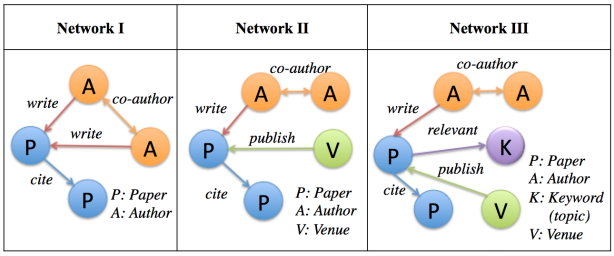

Figure 1: Heterogeneous networks of scientific publication repository

What we want to study is how to simulate the evolution of these networks. In this paper, we will mainly focus on Network I and verify the simulation via ABM. In the future, we will make further study on Network II and Network III.

3.1 How to Conceptualize Scientific Publication Repository?

During the process of agent-based modeling, the first step is to define the key elements, agents and relations, which can be extracted from the network we used .

Agent is considered as an independent individual. It is identifiable and autonomous [13]. According to Network I, we represent author and paper as heterogeneous and autonomous agents that interact with each other and make decisions such as publishing and citing papers based on prescribed rules. Thus, the two agents in discussion are author and paper. Author will initiate the interaction among different agents in the evolutionary process.

Besides, based on the Network I, we can extract three kinds of relationships among authors and papers as Table 1 shows, which can be defined as the relations in our agent-based models.

Table 1: Relations in scholarly network

Relation	Description
$A \xrightarrow{w} P$	author writes paper
$P \xrightarrow{c} P$	paper cites another paper
$A \overset{co}{\leftrightarrow} A$	author collaborates with another author

3.2 What kind of attributes do the agents have and how to measure them?

3.2.1 Author Agent

The main agents in our discussion include author and paper. First, we will move on to the author agent. We define several attributes and find ways to measure them as follows.

1) authority: the importance of each author. We suppose that the more important the author is, the more active he will be in the future.

This attribute could be obtained according to the PageRank algorithm. Based on Network I, we can generate a new network shows the citation relationship among authors. Then, with the PageRank approach, it could tell the ranking score of each author. This is the authority of each author.

2) publication number: denotes the number of publications of author A_i in every year.

According to the number of author A_i's publications in recent years, we will adopt simple linear regression model to estimate new amount $pNum(A_i)$for next year.

3) energy: the ability to write new papers. The author will gradually lose his energy when he keeps on writing papers. This attribute will control author's behavior.

The initial value $Energy(A_i)$ can be defined as the estimated $pNum(A_i)$ and decrease one if author writes one paper.

4) the distribution of coauthor number: given author A_i, $P(coNum|A_i)$ indicates that the probability of his coauthor number $coNum$.

5) the distribution of coauthor candidates: $P(A_x|A_i)$ denotes that for a given author A_i, the probability of finding candidate coauthor A_x.

In this paper, we will adopt meta-path based ranking algorithm for generating the candidate coauthors.

6) the distribution of citation number: $P(ciNum|A_i)$ represents that the number of his publications' citations $ciNum$ and its probability for a given author A_i.

3.2.2 Paper Agent

Paper's attributes include:

1) published year. This is the key attribute in our model since what we discuss is the annual evolution of scientific publication network.

2) importance: $PR(P_i)$ denotes the pageRank score of paper P_i.

Similar to author authority, the importance of paper can be inferred by PageRank algorithm on the paper citation graph.

3) the distribution of citation candidates: for a given paper P_i, $P(P_t|P_i)$ denotes the probability of its citation candidate P_t. We will infer the candidates via a list of meta-paths $P_i \rightarrow ... \rightarrow P_t$.

3.3 How to define the rules for guiding the agents' behaviors?

3.3.1 Generalized Rules for Scholarly Network Prediction

In this study, to simplify the model, we assume that the number of author remains stable. We are going to focus on a fixed set of authors and study their future behaviors and authority changes. Authors play a leading role in this process. They can write papers and decide their collaborators according to the basic rules. Generally, there are three basic rules for simulate this network, as Algorithm 1 shows. Give an existing network at year t denoted as $N_t(A, P, E_{A \overset{w}{\rightarrow} P}, E_{P \overset{c}{\rightarrow} P}, E_{A \overset{co}{\leftrightarrow} A})$, through three important rules to guide the interactions and communications of agents, we can infer the future network $N_{t+1}(A, P, E_{A \overset{w}{\rightarrow} P}, E_{P \overset{c}{\rightarrow} P}, E_{A \overset{co}{\leftrightarrow} A})$ at year $t+1$, where A denotes the set of author, P represents paper, $E_{A \overset{w}{\rightarrow} P}$ means all the "write" relations, $E_{P \overset{c}{\rightarrow} P}$ refers to all the "cite" relations and $E_{A \overset{co}{\leftrightarrow} A}$ stands for the "co-author" relations. Rule 1, Rule 2 and Rule 3 summarize the way of simulating the evolution of scholarly network. Note that both agents and rules are generalizable and extensible. We will separate the details and describe the specific rules in the following sections.

Algorithm 1 General Rules for Scholarly Network Simulation

Input : Network $N_t(A, P, E_{A \overset{w}{\rightarrow} P}, E_{P \overset{c}{\rightarrow} P}, E_{A \overset{co}{\leftrightarrow} A})$ at year t
Output: $N_{t+1}(A, P, E_{A \overset{w}{\rightarrow} P}, E_{P \overset{c}{\rightarrow} P}, E_{A \overset{co}{\leftrightarrow} A})$ at year $t+1$
Rule 1: Author makes decision to generate new papers $\{P_{t+1}\}$, and relations $\{E_{A_i \overset{w}{\rightarrow} P_{t+1}}\}$
Rule 2: Author and paper decide the coauthor candidates and generate new links $\{E_{A_j \overset{w}{\rightarrow} P_{t+1}, j \neq i}\}$ and update $\{E_{A_i \overset{co}{\leftrightarrow} A_j}\}$
Rule 3: Author and paper make decision to establish future citation links $\{E_{P_{t+1} \overset{c}{\rightarrow} P_t}\}$

As we mentioned above, all the rules are flexible and extensible. The generalized rules is a skeleton. We can define reasonable rules to better simulate the development of the scholarly network. Thus, in next section, we will discuss several detailed rules based on the aforementioned generalized rules, and elaborate how they guide the agents' behaviors.

3.3.2 Rule 1 for Generating New Publications

The simulation starts from the generation of new publications. Author has privilege of writing new papers. Rule 1.1 to Rule 1.4 are the specific rules about how to estimate new papers.

Rule 1.1: Author with higher rank owns higher priority in actions. So that, before creating new papers, rank authors by their authorities and take the top ranked author into consideration.

Rule 1.2: Suppose that the number of publication of each author follows the simple linear regression model. Based on the historical data, we could estimate the publication number of each author in a specific year.

Rule 1.3: The initial value of energy equals to the second attribute of author agent, which can be derived by Rule 1.2.

Rule 1.4: Only if the author has energy, he can write new papers. The behavior of creating new papers is determined by his " energy".

In this part, we introduce PageRank algorithm and regression model into the rules as **Algorithm 2** shows. At the beginning of each year, the authority of author will be refreshed. First, we build an updated graph $G(V, E)$, where V denotes all the author vertices, and E represents the relation of author cite author. The PageRank score on this graph decides the priority of each author. We suppose that the more important the author is, the more opportunities he will get. We take them into the first consideration during the prediction. Based on the analyzation of statistics data, the simple linear regression model is adopted to help each author decide the new number

of papers. The regression model is different from author to author.

Algorithm 2 Detail Rules for Scholarly Network Simulation

Calculate the authority of each author A_i at year t via PageRank algorithm;
Rank author by the authority in descending order;
Estimate $pNum(A_i)$ through linear regression model;
Set $Energy(A_i) = pNum(A_i)$;
loop For each author $A_i \in \{A\}$
  if $Energy(A_i)$>0 **then**
    Generate new papers $\{P_{i,t+1}\}$, the number of new papers equals $Energy(A_i)$;
    Set$Energy(A_i) = 0$;
    loop For Each Paper $P_k \in \{P_{i,t+1}\}$
      Generate new links $\{E_{A_i \xrightarrow{w} P_k}\}$;
      Coauthor-ship competition via **Algorithm 3**
      Citation competition via **Algorithm 4**
    end loop
  else
    continue;
  end if
end loop

3.3.3 Rule 2 for Coauthor-ship Competition

As we all know, the coauthor number of each paper is different. Author has rights to choose his collaborators. In this section, we will describe how to add other authors for new publications. The main rule is designed as the competition of coauthor candidates.

Rule 2.1: The main author A_i determines his coauthor number according to the attribute of "the distribution of coauthor number".

Rule 2.2: The coauthor candidate can be selected from A_i's direct or indirect collaborators. We could search the candidates A_x via path $A_i \overset{co}{\leftrightarrow} A_x$ and $A_i \overset{co}{\leftrightarrow} A \overset{co}{\leftrightarrow} A_x$. The one with higher ranking score has higher possibility to win.

Rule 2.3: A candidate can win only if his energy is positive. If the candidate wins, his energy will lose one unit at the same time. If he runs out of energy, that is $Energy(A_i)$ equals zero, he can not be chosen as a coauthor.

And so, arguably, the main author A_i has rights to decide the coauthor number and the candidate coauthors. In this study, we adopt roulette wheel selection to obtain the coauthor number. Through the historical data, it is easy to get the distribution of coauthor number of A_i, denote it as $P(coNum|A_i)$. Then add up the probabilities of different coauthor number successively. Generate a random number r between 0 and 1 and find the interval in the adding up distribution. Thus, we can infer the predicted number of co-authors for A_i.

As aforementioned, meta-path based random walk approach [12] is applied in the task of searching candidates.

The ranking score of each candidate is denoted as $P(A_x|A_i) = \sum_{s \in S} RW(s)$,

where S represents the meta-path from main author A_i to candidate author A_x, such as $A_i \overset{co}{\leftrightarrow} A_x$ or $A_i \overset{co}{\leftrightarrow} A \overset{co}{\leftrightarrow} A_x$. s is a path instance, such as $a_i \overset{co}{\leftrightarrow} ... \overset{co}{\leftrightarrow} a_x$.

$RW(s)$ denotes the random walk probabilities of path s, which can be calculated by

$RW(s) = \prod_{a_i, a_j \in s} w(E_{a_i \to a_j}) w(a_j)$.

$RW(s)$ is the product of all the transition probability in path s. $w(E_{v_i \to v_j})$ denotes the weight of link $E_{a_i \to a_j}$. $w(a_j)$ represents the importance of internal nodes, by which it means authority of author. The candidates will compete for the coauthor positions by their ranking score. The one with higher score has higher possibility to win. Note that, the energy of author is a decisive factor in this competition as **Algorithm 3** shows.

Algorithm 3 Coauthor-ship Competition

Input: P_k, A_i
Output: $\{E_{A \xrightarrow{w} P_k}\}$

Denote the number of authors for paper P_k as N_k
Infer N_k via $P(coNum|A_i)$;
According to $P(A_x|A_i)$, choose the candidate coauthors $\{Aco_i\}$;
loop For each author $A_m \in \{Aco_i\}$
  if $N_k > 0$ **then**
    if $Energy(A_m)$>0 **then**
      Generate new links $\{E_{A_m \xrightarrow{w} P_k}\}$;
      Strength link $E_{A_i \overset{co}{\leftrightarrow} A_m}$
      $Energy(A_m)$-1;
      N_k-1;
    end if
  else
    break;
  end if
end loop

3.3.4 Rule 3 for Citation Competition

In this section, we are going to define the rules for generating new citation links. This part can be viewed as a competition process among papers. We assume that the citing paper P_k cannot cite the papers written at the same year. In other words, new paper P_k at year $t+1$ only can cite the papers written before year t(inclusive).

Rule 3.1: The candidate cited paper was written before the predicted year.

Rule 3.2: The number of citation for paper P_k is determined by author's attribute "the distribution of citation number".

Rule 3.3: The citation candidate is determined by both social influence (random walk based on meta-path) and agent importance. Social influence includes three meanings.

Type I: author tends to cite his own papers;

Type II: author prefers to cite his collaborators' papers;

Type III: author is inclined to draw on experience from others and cite more papers that other cited.

Table 2: Path for Searching Citation Candidates

Type	Path	Description
I	$P_k \overset{w}{\leftarrow} A \overset{w}{\rightarrow} P_t$	the candidate paper has the same author with P_k
II	$P_k \overset{w}{\leftarrow} A \overset{co}{\leftrightarrow} A \overset{w}{\rightarrow} P_t$	the candidate paper's author cooperated with P_k's author
III	$P_k \overset{w}{\leftarrow} A \overset{w}{\rightarrow} P \overset{c}{\rightarrow} P_t$	the candidate paper was cited by other papers with the same author of P_k
	$P_k \overset{w}{\leftarrow} A \overset{w}{\rightarrow} P \overset{c}{\rightarrow} P \overset{c}{\rightarrow} P_t$	the candidate paper was cited by papers that were cited by the ones with the same author of P_k
	$P_k \overset{w}{\leftarrow} A \overset{co}{\leftrightarrow} A \overset{w}{\rightarrow} P \overset{c}{\rightarrow} P_t$	the candidate paper was cited by the papers that their authors cooperated with P_k's author
	$P_k \overset{w}{\leftarrow} A \overset{w}{\rightarrow} P \overset{c}{\rightarrow} P \overset{w}{\leftarrow} A \overset{w}{\rightarrow} P_t$	the candidate paper's author' other papers were cited by the ones with the same author of P_k

Based on the meta-paths shown in Table 2, P_k will be able to

find candidate cited papers set. It shows that all the path are start from the " written by" relations. The more important the paper is, the more frequent it will be cited. To make sure that new paper can seise the opportunity to be cited, we use meta-path based random walk to search those papers through their authors. Then all the candidate papers will compete for the cited positions by their ranking score. Note that, the score of each candidate cited paper is determined by both the random walk result and the importance of itself. We assume that λ equals 0.5. We adopt the roulette wheel selection method mentioned in 3.3.3 to get the citation number. The detailed algorithm is show as **Algorithm 4**.

Algorithm 4 Citation Competition

Input: P_k, A_i, $\{E_{A \xrightarrow{w} P_k}\}$
Output: $\{E_{P_k \xrightarrow{c} P_t}\}$

1: Denote the number of citations for paper P_k as C_k
2: Infer the citation number C_k based on the above distribution $P(ciNum|P_k)$;
3: Get $P(P_t|P_k)$ via a list of meta-path ($P_k \rightarrow ... \rightarrow P_t$);
4: Calculate the ranking score of each candidate cited papers: $\lambda P(P_t|P_k) + (1-\lambda)PR(P_t)$;
5: Get the top C_k cited papers from the candidate set;
6: **loop** For each cited paper P_c
7: Generate new links $\{E_{P_k \xrightarrow{c} P_c}\}$;
8: **end loop**

So far, we introduced the whole mechanism of simulating the evolution of future scholarly network. Figure 2 shows how the general rules guide this process.

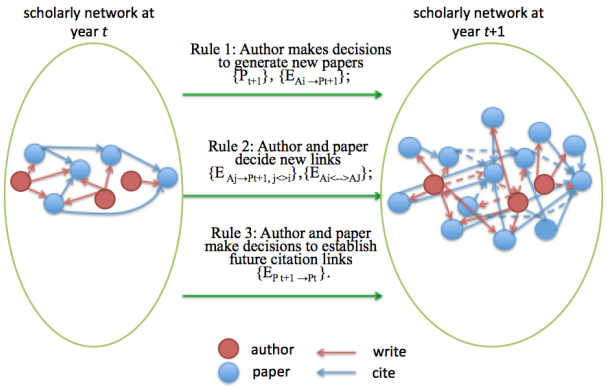

Figure 2: General Process for the Simulation of Network I

4. EXPERIMENT

In this section, we will describe the experimental setting and results. Our analysis and conclusions are presented in section 5.

4.1 Data and Network Construction

At present, we conduct the experiment based on a small dataset from ACM digital library. Larger dataset should be discussed in the future work. We extract 284 authors and published more than ten papers (inclusive) from 1990 to 2009. Besides, every author has continuous publications from 2002 to 2004. Thus, every author is relatively stable. There are 10,556 papers for these authors during this period. Meanwhile, we extract the citation relationships among these 10,556 papers. Then, we could use the papers published from 1990 to 2004 as training dataset, and the others published from 2005 to 2009 as testing dataset.

Then, we build the training heterogenous scholarly network using the training dataset. There are 284 author nodes, 5,943 paper nodes, 6,503 $A \xrightarrow{w} P$ relations, and 8,159 $P \xrightarrow{c} P$ relations. The weight of link from n_i to n_j is calculated by $w(n_i \rightarrow n_j) = \frac{1}{|E_{n_i \rightarrow N}|}$. It is decided by the number of outgoing links of starting node.

Through the training heterogeneous scholarly network, it is possible for us 1) to find the citation relationship between authors, and then via PageRank, we can get the authority of each author at 2004; 2) to get the number of publications of each author in every year, and then through linear regression model, we could estimate how many papers for each author in next year; 3) to set the initial energy of each author; 4) to get the distribution of coauthor's number for each author; 5) to get the distribution of coauthor candidates for each author A_i via two kinds of meta-paths : $A_i \overset{co}{\leftrightarrow} A_x$ and $A_i \overset{co}{\leftrightarrow} A \overset{co}{\leftrightarrow} A_x$ as we mentioned before; 6) to get the distribution of citation's number; (6) to get the distribution of citation's number; 7) according to PageRank, assign the score to each paper as importance; 8) to get the distribution of cited paper candidates via meta-paths mentioned in Table 2. So far, all the attributes for author and paper have been settled.

4.2 Experiment Result

Based on the training network, we implement the rules described before and simulate the evolution of publication repository. From the simulation, we can estimate how many publications and relations will appear in the next five years. Figure 3 shows the comparison of total number of publications in each year. Table 3 tells comparison of newborn relations between the actual data and simulation results(abm).

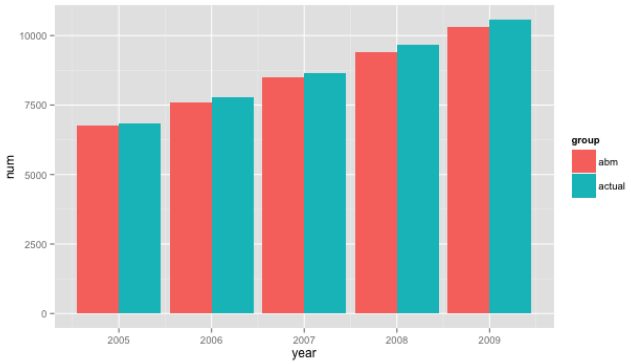

Figure 3: The number of all publications in each year

Table 3: The number of increased relations per year

year	$A \xrightarrow{w} P$		$P \xrightarrow{c} P$		$A \overset{co}{\leftrightarrow} A$	
	actual	abm	actual	abm	actual	abm
2005	1,005	873	1,829	1,948	143	62
2006	1,065	900	2,196	2,083	142	53
2007	1,000	924	2,138	2,159	130	56
2008	1,100	951	2,627	2,207	113	42
2009	988	980	2,271	2,341	84	52

Through Figure 4 and Figure 5, we can find that the data from ABM has the same tendency as actual data.

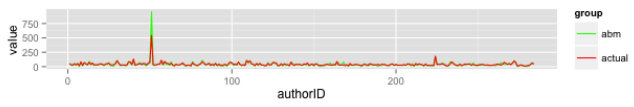

Figure 4: The number of publications of each author

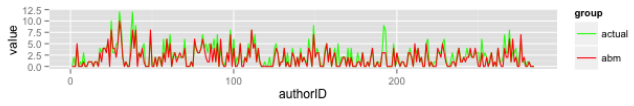

Figure 5: The number of coauthor-ship of each author

The comparison about number of new nodes and new relations is not enough for evaluation the network. Furthermore, we use an indirect way to make further evaluation about the evolutionary network. As we know, the citation relationship has a great effect on the authority of author. Based on $A \xrightarrow{w} P$ and $P \xrightarrow{c} P$, it is easy to infer the authority of each author via PageRank from 2005 to 2009. Then, via Spearman's rank correlation coefficient, we calculate the coefficient between actual author ranking and predicted author ranking. And the result is shown in Table 4.

Table 4: The Spearman's rank correlation coefficient

year	2005	2006	2007	2008	2009
coefficient	0.9683	0.8793	0.9043	0.8774	0.8602

5. ANALYSIS AND CONCLUSION

In this study, we propose a novel agent-based model for simulate the scholarly network. From the results, we can find that in 2005, the ranking result of author is highly relevant to the actual ranking result. With the time goes on, the coefficient declines. It is reasonable since the prediction for the next year is based on the previous result. The disparity between the actual and predicted ranking results will grow. Therefore, the performance will be better for the coming year. The new number of publications is not optimal. It is one-sided solution to estimate the accurate publications of each author based on liner regression model, so that the total number of papers sound not that good. While the estimated tendency of author's behavior, including writing papers and coauthor-ship, is consistent with the actual situation.

Through the result we can conclude that using ABM in scientific publication is effective and has a good performance. Without sophisticated equation models, the proposed rules derive from the mechanism of agent-based model.

6. FUTURE WORK

In this paper, to simplify the model, the number of paper is dynamic, but author is stable. Based on this hypothesis, we conduct a case study on the changes that will happen to these 284 authors. In the next step, we are supposed to study what would happen if new authors show up and other authors take off.

The method proposed in this paper is a prototype of agent-based modeling framework. We try to introduce the concept of agent into scholarly network and set up a list of rules. Nevertheless, there are more than author and paper in the network. As we mentioned at the beginning, in the future, we need to add venue and topic to enrich the network, and move forward to studies on Network II and Network III. Figure 6 illustrates the main idea of this approach and will guide our future work. We deem that every object and relation in the scientific publication repository could be extracted as the components (agent and relation) in ABM.

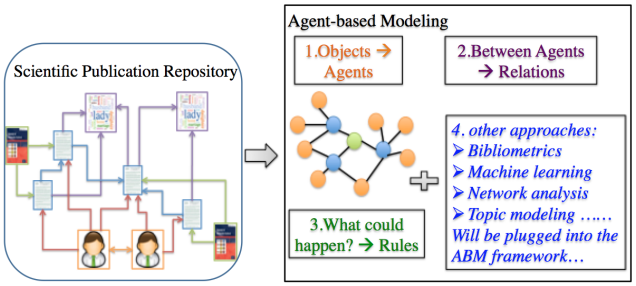

Figure 6: Sketch of Agent-based Simulation for Scientific Publication Repository

After we introduce new agents into the framework, especially keyword (topic), some new specific rules will show up and it will be helpful to make more credible predictions. For example, the agents based on Network III could be extracted as paper, author, venue and keyword. There are several kinds of relations among these agents, which will be studied in the future, such as 1) paper similarity, 2) paper publish at venue, 3) paper relevant to keyword, 4) keyword contribute by paper, 5) keyword contribute by author and so on. When authors search the coauthor candidates, they tend to find the one with the similar research area. When predict new citations, Paper P_i are more likely cite paper P_j because of their contribution to the topics, and so on. Not only the rules, lots of approaches could be applied in the framework, including network analysis, machine learning and so on. With the combination of other approaches, the rules could be more flexible and reasonable. Besides, it is necessary to use larger dataset, such as PubMed dataset, to test the proposed model and conduct further experiments.

7. REFERENCES

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[2] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.

[3] Pawel Bartoszczuk, Tieju Ma, and Yoshiteru Nakamori. Environmental kuznets curve for some countries-regression and agent-based approach. In *AIR POLLUTION-INTERNATIONAL CONFERENCE-*, volume 10, pages 93–102. WIT PRESS, 2002.

[4] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3):7280–7287, 2002.

[5] Katy Börner, Jeegar T Maru, and Robert L Goldstone. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5266–5273, 2004.

[6] Quentin L Burrell. Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, 54(5):372–378, 2003.

[7] Darcy Davis, Ryan Lichtenwalter, and Nitesh V Chawla. Multi-relational link prediction in heterogeneous information networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 281–288. IEEE, 2011.

[8] Shuguang Han, Daqing He, Peter Brusilovsky, and Zhen Yue. Coauthor prediction for junior researchers. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 274–283. Springer, 2013.

[9] Carlos A Iglesias, Mercedes Garijo, and José C González. A survey of agent-oriented methodologies. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings*, pages 317–330. Springer, 1999.

[10] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[11] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[12] Xiaozhong Liu, Yingying Yu, Chun Guo, and Yizhou Sun. Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 121–130. ACM, 2014.

[13] Charles M Macal and Michael J North. Tutorial on agent-based modeling and simulation. In *Proceedings of the 37th conference on Winter simulation*, pages 2–15. Winter Simulation Conference, 2005.

[14] Alexandrin Popescul and Lyle H Ungar. Statistical relational learning for link prediction. In *IJCAI workshop on learning statistical models from relational data*, volume 2003. Citeseer, 2003.

[15] Weiming Shen and Douglas H Norrie. Agent-based systems for intelligent manufacturing: a state-of-the-art survey. *Knowledge and information systems*, 1(2):129–156, 1999.

[16] Yizhou Sun, Rick Barber, Manish Gupta, Charu C Aggarwal, and Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 121–128. IEEE, 2011.

[17] Yizhou Sun, Jiawei Han, Charu C Aggarwal, and Nitesh V Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 663–672. ACM, 2012.

[18] Marco Tomassini and Leslie Luthi. Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and its Applications*, 385(2):750–764, 2007.