An Agent-Based Information Foraging Model of Scientific Knowledge Creation and Spillover in Open Science Communities

Özgür Özmen^(a), Levent Yilmaz^(b) M&SNet: Auburn Modeling and Simulation Lab Samuel Ginn College of Engineering Auburn University ^(a) 0200002@auburn.edu, ^(b) yilmaz@auburn.edu

ABSTRACT

Motivation and problem-domain preferences of scientists can affect aggregate level emergence and growth of problem domains in science. In this study, we introduce an agent-based model that is based on information foraging and expectancy theory to examine the impact of rationality and openness on the growth and evolution of scientific domains. We simulate a virtual socio-technical system, in which scientists with different preferences search for problem domains to contribute knowledge, while considering their motivational gains. Problem domains become mature and knowledge spills occur over time to facilitate creation of new problem domains. We conduct experiments to examine emergence and growth of clusters of domains based on local interactions and preferences of scientists and present preliminary qualitative observations.

Keywords: information foraging, agent based modeling, science of science, open innovation, knowledge spillover

1. INTRODUCTION

Knowledge spillovers are studied widely in literature and linked to innovation measures and outputs (Jaffe, 2008; Feldman and Audretsch, 1996). Spillovers are defined as the migration of knowledge beyond the domain borders (Fallah and Ibrahim, 2004). In our study we express spillovers not only in terms of knowledge transfer but also mobility of scientists. Hence, spillovers result in formation of new domains as a consequence of knowledge and skill transfer.

A research question of interest is to find out connection between individual rationality and aggregate efficiency. Axtell and Epstein (1999) discuss empirical data, which demonstrate that all individuals should not necessarily be rational to produce efficiency in macro level outcomes of a system. Given that individual rationality is bounded, they explore how much rationality should exist in a system to generate macro-level patterns. In this work, we do not propose to discern minimum level of rationality, but rather aim to address how rationality affects the spread of knowledge spillovers as well as growth and development of domains.

Scientists join or leave a problem domain on the basis of problems to be solved and tasks to be accomplished, and their position in the scientific landscape depends upon their knowledge, levels of interest, personal learning objectives, resources, and commitments (Hollingshead et al., 2002). Motivation of scientists as a personal interest is one of the main indicators for willingness of contribution. We take individual motivation as a main force driving the commitment of a scientist to a problem domain.

Metaphorically, scientists can be viewed as predators. Predators are expected to abandon their current patch (e.g., domain) when local capture rate (e.g., problem solving success) is lower than estimated capture rate in the overall environment (Bernstein et al., 1988). Information foraging theory, which is derived from this evolutionary phenomenon developed by Pirolli and Card (1999) assumes that people, if they have an opportunity, would adjust their strategies or the topology of their environment to maximize their rate of information gain. In our study, scientists join or abandon problem domains based on perceived cues about their performance in attaining the desired outcome. The cues are represented by the "instrumentality" component of an individual's motivation, which is described in detail in section 3. Also, in our model, motivated and successful scientists recruit new scientists just as genetically more adapted predators are more likely to have a offspring in their natural environment.

Interaction surface between scientists and information repositories in real life determines the time costs, resource costs, and opportunity costs of different information foraging strategies (Pirolli and Card, 1999). It is also suggested that people are selfish and lazy in applying their cost-benefit analyses (Nielsen, 2003). In accordance with these observations, we define three different characteristic of the scientists in relation to their cost- benefit decisions.

David (1998) defines the force of Open Science's universalist pattern as providing entry into scientific artifacts and open discussion by all participants, while promoting "openness" in regard to new findings. Carayol and Dalle (2007) explain open-science phenomenon as significant freedom of scientists to choose what they want to do and how they want to do.

In light of these observations, we present an agent based model, called "KnowledgeSpill" to create a virtual environment where scientists have limited omniscience. In the model, opportunities in a particular problem domain deplete over time. We visualized the impacts of individual rationality and openness on the growth of scientific domains. In section 2, we present the conceptual basis for our model. In sections 3 and 4, we describe the model structure and mechanisms in detail along with the initial parameter values and description of the visualizations. Section 5 discusses preliminary results and the qualitative observations derived from them. In section 6 we conclude by summarizing our findings in relation to reviewed literature and suggest potential avenues of future work.

2. BACKGROUND AND RELATED WORK

It is demonstrated that user innovation communities are self-organizing complex adaptive systems (Yilmaz, 2008). However, not all complex systems are self-organizing (Monge and Contractor, 2003). A system is self organizing when the network is self-generative (e.g., spawning agents), there is mutual causality between parameters, imports energy into system (e.g., creating new problem-domains and opportunities) and is not in an equilibrium state.

Agent based modeling is widely used to study such complex systems, and it is especially helpful to understand, explore, and parameterize those systems (Bonabeau, 2002). Monge and Contractor (2003) describe the main elements of complex systems in terms of the network of agents, their attributes or traits, the rules of interaction, and the structures that emerge from these micro-level interactions. Typical classes of agent traits are location, capabilities, and memory.

Simulation is used to study scientific domains by different scholars. For instance, Nigel (1997) introduces a model to determine whether it is possible to reproduce observed regularities in science using a small number of simple assumptions. His model generates knowledge structures consistent with observed Zipf distributions involving scientific articles and their authorship. Naveh and Sun (2006) explore the effects of cognitive preferences on the aggregate number of scientific articles produced and argue that simulations with credible cognition mechanisms may lead to creativity in academia.

In scientific knowledge generation, resources may include knowledge, people with skills and abilities, financial support and/or access to tools and raw materials (Mohrman and Wagner, 2008; Powell et al., 1996). The following assertion is stated by Carayol and Dalle (2007): "When the discipline grows, the relative rewarding of problems located in already developed fields increases: because their audience becomes larger, contributions to such domains are more likely to be cited." Another point expressed in Carayol and Dalle (2007) is that more specialized disciplines are more likely to get more specialized through time, and this phenomenon is more striking when the scientists are more focused on rewarding areas. We describe this phenomenon as imitation behavior. Furthermore, scientists make their decisions according to their perceptions and anticipations of their own performance. It is suggested that, intrinsic task motivation plays a critical role in creativity and innovations (Amabile, 1996; Kanter, 1988). Expectancy theory (Vroom, 1964) is a widely known popular model of motivation based on intrinsic motivation. The rationality perspective explored in this study can be defined as the behavior towards the maximum motivational outcome based on this theory.

3. CONCEPTUAL MODEL

The location of a scientist in our simulation context is a metaphor for disciplines. Scientists can not migrate between domains or become actively involved within the problemdomains in different disciplines without proper orientation and enculturation. So, the location can be perceived as a discipline and scientists can move in a limited environment at a given time during simulation. There are two basic agent types in the model:

- Scientists quest for knowledge in different problem domains. Scientists are members of different disciplines and are more likely to be aware of the problems within their area of expertise. Transferring to other disciplines or areas is difficult because of entrance threshold and the need for enculturation prior to making acceptable contributions. In our model, enculturation is interpreted as the process of searching environment to find a problemdomain to contribute. Our simplifying assumption is that scientists initially move randomly and have an omniscience of 1 cell around, known as Von Neumann neighborhood. A change in the area of awareness and shift to a more open environment with a lower entry threshold facilitates browsing a wider scope in the knowledge space.
- *Problem Domains* are different areas in disciplines (e.g., database management in computer science). Their maturity reflects the knowledge level. As maturity grows, the receptivity of the domain decreases due to higher levels of inertia in mature domains. Increased inertia results in knowledge spillover. Scientists migrate to new problem-domains formed as a result of these spillovers. (e.g., data-mining emerges via transfer of scientists from database management and machine learning fields)

3.1. Maturity and Receptivity

Maturity of a problem-domain is the knowledge level of that particular domain and increases with each contribution. This is represented in the model initially by assigning to a domain a randomly selected maturity value between (0, 0.5]. At every time tick, maturity level of domain *i* at time t + 1 is defined as follows:

$$M_{i,t+1} = Min(M_{i,t} + \sum_{j=0}^{C_i} \gamma_i, 1)$$
(1)

where C_i is the number of accepted contributions at time tin domain i and γ_i is the incremental increase on knowledge repository in the domain caused by each contribution which is drawn between 0 and maximum increment value. Receptivity R at time t + 1 of domain i is defined as:

$$R_{i,t+1} = 1 - M_{i,t} \tag{2}$$

3.2. Traits of Scientists

The area of perception (openness) and memory of past successes can be thought of as elements of scientist's traits. The

choice of the new domain in the area of awareness (scope) of a scientist is based on their preference. We assume that a scientist belongs to a character group with certain probability. There are three different groups:

- *Rationals* look for the least mature problem domain to contribute in their scope. Our interpretation is that less mature domains are more likely to accept contributions, so scientists are more likely to gather experience, early reputation, and motivation by working in those domains.
- *Imitators* have the characteristic of being influenced by the trend or crowd. They look around for a domain to contribute, but are more likely to choose crowded ones. Preferential attachment (Barabasi and Albert, 1999) is used as the guiding principle.
- *Random scientists* randomly choose their problem domains.

Scientists can be in two basic states. They can be "free" not working on a domain or can be "active" by contributing to a domain. Free scientists are searching their scope and when they find one or more domains, they make the decision based on the aforementioned traits above. Scientists, who are not free, contribute to the particular domain they reside on. At every time interval, they have fixed 10% chance of contribution, and once they contribute, the domain decides to accept it or not based on its receptivity level. Scientists also stop practicing when they reach to their maximum age.

As the simulation unfolds, experienced and highly motivated scientists spawn new scientists who are in an initial state. A highly motivated scientist inspires a new scientist with a fixed probability of 0.1. Also, each scientist has a susceptibility level. The larger the tenure of the scientist, the lower its susceptibility level. Susceptibility starts from a higher level and decreases exponentially over time but not below a certain threshold, which is different for each scientist. Susceptibility is defined as the following:

$$S_{i,t+1} = \beta_0 + \frac{1}{2}e^{-\beta_1 x A_i}$$
(3)

where $S_{i,t+1}$ is the susceptibility level of scientist *i* at time t + 1, β_0 is the lowest susceptibility level of a scientist, β_1 is the function parameter and A_i is the current age of the scientist *i*. Susceptibility is used to determine experience increment after every successful contribution. According to this definition experience of scientist *i* at time t + 1 is:

$$E_{i,t+1} = E_{i,t} + S_{i,t} \times \lambda_s \tag{4}$$

where λ_s is the success increment determined initially and fixed. Success increment can be described as the gain of experience after a successful contribution.

3.3. Motivation Theory and Memory

In expectancy theory, motivation is defined as a product of three factors: how much one desires a *reward* (*valence*), one' s estimate of the probability that effort will result in successful *performance (expectancy)*, and one's estimate that performance will result in receiving the *reward (instrumentality)*. It is given with the formula below:

$Motivation = Expectancy \times Instrumentality \times Valence$ (5)

We assume valence as a fixed value and is different for each scientist, while expectancy can be perceived as the experience of an individual. The experience is described as the attained level of skill through successful contributions. That is, the more experienced the scientists are, the more likely to be motivated they are. Instrumentality is a dynamic parameter which is updated over time, as it denotes the estimate of the award for successful performance. Scientists have a time window of θ which is used to calculate success rate below:

$$Sr_{i,t+1} = \frac{N_{i,t+1}}{\Theta} \tag{6}$$

while *N* is the number of successful contribution of scientist *i* during the time-window and θ is the pre-defined time-window length.

Also there is a memory factor α , which is different for each scientist. It denotes the impact of success rate on instrumentality and indicates the significance and weight of the current observation with respect to prior experience. A small value of α results in a conservative behavior by avoiding overriding of the past experience. The relative weight of past and present capture rate which is controlled by a parameter denominated the 'memory factor' is seen as a common approach in psychological models of simple learning (Bernstein et al., 1988). Following is the formula for instrumentality of scientist *i* at time *t* + 1

$$I_{i,t+1} = Min(1, (\alpha_i \times Sr_{i,t+1} + (1 - \alpha_i) \times I_{i,t}))$$
(7)

4. IMPLEMENTATION AND VISUALIZA-TION

Our model is coded in the RePast (Recursive Porous Agent Simulation Toolkit) environment, which is a free and opensource agent based modeling toolkit. Illustration of the aforementioned concepts are analyzed by simulating the spillovers over time. In the figures below, we illustrate that a mature domain spills over to the location on its north-east creating a new problem domain while also transferring 10% of its scientists. The brightness of the color represents maturity. The brighter the color of a domain is, the more is the maturity of that particular domain.



Figure 1. Creation of a new domain

Parameter Name	Initial Value
Rational rate	0.75
θ (Time window)	1
Imitator rate	0.2
Maximum Age	3000
Random rate	0.1
Minimum Age	1500
λ_s (Success)	0.01
Run time	3000
Contribution rate	0.1
Initial number of Scientists	500
Maturity threshold	0.8
Initial number of Domains	10
Spillover rate	0.1
World Height	50
World Width	50
Transfer rate	0.1
Motivation threshold(for spawning)	0.95
Initial Age	0
Spawning rate	0.1
Initial Experience	(0,0.5]
Motivation threshold(for leaving)	0.05
Initial Valence	(0.5,1]
Initial Maturity	(0,0.1]
β ₀	(0,0.3]
β_1	-0.001
Initial Instrumentality	(0,1]
α (Memory)	(0,1]
Increment of Maturity	(0,0.002]

 Table 1. Initial settings of the model

In order to understand the illustration better, snapshots of the model over time are shown in Figure 2. Also, Table 1 lists the initial values of the parameters of the model.





5. PRELIMINARY EXPERIMENTS AND OBSERVATIONS

We simulate the creation of knowledge (spillovers) under various scenarios observing the spread of domain structures, proportion of mature domains, proportion of the population with regard to motivation levels and the spillovers occurred over time. We examine the implications of the model under 9 basic scenarios, which are created by 3 dimensions of rationality and 3 dimensions of the openness as shown in Table 2.

Table 2.	Main	Scenarios
----------	------	-----------

Rationality	Low	Medium	High
Rational	25%	50%	75%
Imitator	65%	40%	15%
Random	10%	10%	10%
Scope	1 cell	5 cells	10 cells

Emergent patterns represented in figures 3, 4 and 5 are recorded at time tick 3000. 3000 time ticks can be interpreted as a scientist's maximum work-life, which we approximate as 60 years. Each time tick denotes 1 week. The first interesting observation is that the growth direction of the clusters changes with the ratio of the rationals. Another expected result is that new domains (less mature problem venues) are grouped at the edge of the main clusters.



(a) 25% Rationals (22% Occupation, Avg. Maturity - 0.67)

(b) 50% Rationals (40% Occupation, Avg. Maturity - 0.76)



(56% Occupation, Avg. Maturity - 0.78)

Figure 3. The visualization of the domain clusters for each rationality level after 3000 time ticks. Occupation rate indicates how many percent of the grid is occupied by domains and Avg. maturity is the average maturity level of all domains. This Scenario is with Scope of 1 cell.

When the proportion of rationals increases from 50% to 75%, the increase of the average maturity slows down. More interestingly, when agents can browse wider scope for opportunities (e.g., open-science cases) the occupation rate increases, but 5-cell scope results in higher rates of active do-



(a) 25% Rationals (24% Occupation,,Avg. Maturity -0.70) (b) 50% Rationals (74% Occupation, Avg. Maturity - 0.81)



(c) 75% Rationals (88% Occupation, Avg. Maturity - 0.84)

Figure 4. Scenario with Scope of 5 cells.





(a) 25% Rationals (20% Occupation, Avg. Maturity - 0.72)

(b) 50% Rationals (52% Occupation, Avg. Maturity - 0.78)



(c) 75% Rationals (85% Occupation, Avg. Maturity - 0.81)

Figure 5. Scenario with Scope of 10 cells.

main occupation in comparison to the 10-cell scope. This phenomenon needs further examination. The degree of openness matters and there seems to be diminishing return between openness and the develeopment of the domains. When the scope increases to 10-cells, increased levels of rational scientists does not yield higher levels of domain growth as compared to 5-cell scope.

In Figure 6, we examine the distribution of maturity levels across domains. The plots for different population characteristics suggests that there are more mature domains in the case of populations with higher rationals, while there are almost same amount of less mature domains in all scenarios. The graph shifts up with more omniscience scientists. The 5-cell



(c) Scope of 10 cells

Figure 6. Number of domains vs. the range of the maturity levels for each rationality level

(e.g., moderate openness) case outperforms all other cases regarding knowledge creation.



Figure 7. Average population per domain with different rational rates and different scopes

Although our expectation is to observe highly populated domains when the rationality is lower, increased levels of rational agents resulted in more spillovers, and spawning, and hence more number of scientists start practicing in the context. However, as shown in Figure 7, there is a level of diminishing returns. At moderate levels of openness, more scientists are practicing due to increasing spawning rates when the rationality is set at 50%. Beyond this point, average population per domain starts decreasing. To better understand this observation, we examine distribution of motivational levels among the members of the scientific community.

It is expected that increased rationality within a community yields higher levels of motivation throughout the population. In Figure 8, we can observe that the 5-cell case has the most motivated scientist population at 50% rational rate. Our interpretation is that when the scope gets larger, it diminishes the effect of rationality and at around level of 50% rationals, moderate level of openness results in higher levels of motivation. Furthermore, when we examine the distribution of scientists across domains, we observe that 20% of the domains host around 80% of the scientists under each scenario. Finally, as shown in Figure 9, to determine whether power law distribution over spillovers exists, we generate log-log plot of spillovers and their frequency. The results are indicative of existence of a power-law; that is, the frequency of spillovers



Figure 8. (a), (b) and (c) illustrate the fraction of the population in different range of motivation levels for each rationality level (d) Average motivation of a scientist for each rationality level and scope

is inversely proportional to its size.





scope

scope

Figure 9. Logarithm of the number of spillovers vs. logarithm of the frequency of that number

6. CONCLUSION

In this study, our main objective is to adopt and analyze the implications of computational mechanisms of well known theories such as information foraging and of behaviors such as motivation, susceptibility, and maturity on the growth and development of scientific domains. Preliminary observations indicate that with more rational populations, the allocation of efforts are distributed more efficiently, resulting in faster growth of domains and a community climate indicative of high motivation. These implications are consistent with our definition of rationality and expectations from information foraging theory. But when we increased the degree of openness, the growth was not significantly fostered with higher rational population size. Considering population dynamics, with less rational population (e.g., more imitators), our expectation is to observe dense domains; however, as a result of increasing motivation and spawning, increased rates of activity occurred rather in high rationality populations. The future work will explore if there is a diminishing return of openness in terms of motivation and number of mature domains. Potential extensions of the model include ecological aspects that relate to allocation of funds and resources across emergent domains. As feedback mechanisms, funding policies and their effects on the growth of clusters would be an interesting avenue of future research.

REFERENCES

- Amabile, T. (1996). Creativity in context: Update to the social psychology of creativity (boulder, co. Westview Press. Baer, J.(1998). The case for domain specificity of creativity. Creativity Research Journal 11, 173–178.
- Axtell, R. and J. Epstein (1999). Coordination in transient social networks: an agent-based computational model of the timing of retirement. *Behavioral dimensions of retirement* economics, 161–83.
- Barabasi, A. and R. Albert (1999). Emergence of scaling in random networks. *Science* 286(5439), 509.
- Bernstein, C., A. Kacelnik, and J. Krebs (1988). Individual decisions and the distribution of predators in a patchy environment. *The Journal of Animal Ecology* 57(3), 1007–1026.
- Bonabeau, E. (2002). 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.
- Carayol, N. and J. Dalle (2007). Sequential problem choice and the reward system in open science. *Structural Change and Economic Dynamics* 18(2), 167–191.
- David, P. (1998). Common agency contracting and the emergence of" open science" institutions. *The American Economic Review* 88(2), 15–21.
- Fallah, M. and S. Ibrahim (2004). Knowledge spillover and innovation in technological clusters. *Proceedings, IAMOT* 2004 Conference, Washington, DC.
- Feldman, M. and D. Audretsch (1996). Location, location, location: the geography of innovation and knowledge spillovers. *Discussion Papers FS IV 96 28*.
- Hollingshead, A., J. Fulk, and P. Monge (2002). Fostering intranet knowledge sharing: An integration of transactive memory and public goods approaches. *Distributed work*, 335–355.

- Jaffe, A. (2008). The "science of science policy": reflections on the important questions and the challenges they present. *The Journal of Technology Transfer 33*(2), 131–139.
- Kanter, R. M. (1988). When a thousand flowers bloom: Structural, collective and social conditions for innovation in organization. *Research in Organizational Behavior, eds. Staw BM and Cummings LL 10.*
- Mohrman, S. and C. Wagner (2008). The dynamics of knowledge creation: Phase one assessment of the role and contribution of the department of energy's nanoscale science research centers. Technical report.
- Monge, P. and N. Contractor (2003). *Theories of communication networks*.
- Naveh, I. and R. Sun (2006). A cognitively based simulation of academic science. *Computational and Mathematical Organization Theory* 12(4), 313–337.
- Nielsen, J. (2003). Information foraging: Why google makes people leave your site faster.
- Nigel, G. (1997, Jan). A simulation of the structure of academic science. *Sociological Research Online*.
- Pirolli, P. and S. K. Card (1999, Jan). Information foraging. *Psychological Review*.
- Powell, W., K. Koput, and L. Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly* 41(1), 116–145.
- Vroom, V. (1964). Work and motivation.
- Yilmaz, L. (2008). Innovation systems are self-organizing complex adaptive systems. Association for the Advancement of Artificial Intelligence.