ColorScape: A Creative Artificial Ecosystem Model of Communication and Collective Creativity in Global Participatory Science

by

Guangyu Zou

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 7, 2012

Keywords: Global Participatory Science, Innovation, Agent-based Simulation, ColorScape

Approved by:

Levent Yilmaz, Chair, Associate Professor of Computer Science and Software Engineering
Drew Hamilton, Professor of Computer Science and Software Engineering
Wei Shinn Ku, Assistant Professor of Computer Science and Software Engineering

Abstract

With the increasing use of cyberinfrastructure and popularity of e-Science initiatives, science is becoming truly globalized, reducing barriers to entry and enabling formation of open and global networked innovation communities. Yet, relatively little is known about the mechanisms that govern such globalized communities. Meanwhile, creative artificial ecosystem metaphors and interaction processes among communities have potential to shed light on the effects of communication styles in the emergence of global knowledge communities. So, this study explores how networks of scientific communities and epistemic cultures form and evolve, what network patterns emerge from different socio-technical communication theories, and the relationship between environmental constraints, community traits, and innovation performance and potential. Understanding scientific communities and their associated communication networks is key to understanding the dynamics of knowledge creation, as well as formation and growth of scientific communities to facilitate informed science and innovation policy-making. A benefit of this research is to offer federal agencies a computer-aided decision-making tool so as to evaluate investment decision and policies. To this end, an agent-based simulation model combining boundary processes and theories of communication is developed. The model is verified and validated with respect to empirical network data. Simulation results suggest that communities with highly connected clusters are likely to thrive if resource availability is low. So far as the resource allocation strategy is concerned, key area investment with technology transferring results in the highest variety. Exploration of the impact of socio-technical communication theories suggest that under low communication frequency, openness and receptivity lead to higher variety. On the contrary, variety decreases with increasing receptivity under high communication frequency.

Acknowledgments

I would like to express special gratitude to my advisor, Dr. Levent Yilmaz, associate professor, Department of Computer Science and Software Engineering at Auburn University, for his instruction, guidance, encouragement and patience in completion of the research and dissertation. In particular, his suggestions, criticisms and materials greatly contributed to this thesis.

Thanks also to my advisory committee members, Dr. Drew Hamilton, Dr. Wei Shinn Ku and the professors and staff members in the Department of Computer Science and Software Engineering at Auburn University for their kindness and help through these three years. Especially thanks to the university reader, Prof. Shu-Wen Tzeng, who provides valuable comments to this dissertation.

Finally, sincere thanks to my wife Ying Zhao. She gave her greatest support and encouragement to help me succeed in finishing all the research work. Also, I thank my parents, who poured enormous effort into supporting my study during these years.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

1.1 Problem

Creativity is the production of novel and useful ideas by an individual or group of individuals working together [4]. Innovation is extension of creativity, as it is the successful implementation, adoption, and transfer of creative ideas, products, processes, or services [98]. Collective creativity emphasizes the collaboration and coordination of all members in a community rather than individual works. Scientific communities provide a concrete basis to facilitate scientific discovery and collective creativity. So, the study of scientific communities is beneficial to understand the dynamics of knowledge creation, as well as their formation and growth to facilitate informed science and innovation policy-making.

A scientific community consists of scientists, domain knowledge as well as their relationships, and interactions. It is normally divided into "sub-communities" each of which works on a particular field within science, and objectivity is expected to be achieved by the scientific method [105]. As the access to and production of knowledge are increasingly becoming transparent, the practice of science is now more open and global [113], where communication is carried by networks, and shared knowledge is documented in electronic medias such as software and electronic documents. The cyber-infrastructure transcends the geographical boundaries so that members around the world can collectively make contributions in the virtual scientific community. Such virtual collaboratories include Open Source Science (OSS) communities such as OBO Foundry (Open Biomedical Ontologies) [86], NanoHUB (Simulation, Education, Technology for Nano Technology) [62], and NEES Grid (Network for Earthquake Engineering Cyberinfrastructure) [63]. It leads to an evolving collective knowledge-base that is governed and maintained by community members without

central authority. We call such communities of practice and epistemic cultures as Global Participatory Science (GPS) communities [108].

GPS is based on a self-organizing network in which scientists work together not because they are asked to but because they desire it [93]. Social scientists have proposed several theories of communication to interpret the underlying mechanisms of forming such self-organized networks [61]. Therefore, it is important to explore the effects of different theories of communication on patterns of emerging networks and innovation performances.

Based on these observations, we focus on the following problems:

1. How do scientific communities' networks form and evolve, and what network patterns emerge from different socio-technical communication theories such as Cognitive Theories, Self-interest Theories, Exchange and Dependency, Homophiliy & Proximity, and Preferential Attachment [61]?

2. How do scientific communities respond to environmental changes such as funding and resource allocation across research areas? Since communities sustain themselves by adapting to changing environmental conditions, while shaping their cognitive niches, how can we design innovation environments that influence overall innovation potential and performance of the landscape of scientific communities?

3. What is the impact of scientific community traits (i.e., receptivity, flexibility, reorganization tendency) and environmental constraints (i.e., interaction topologies, maximum community number, level of external funding) on the innovation performance (e.g., diversity and resilience) of GPS?

4. Which metrics measure innovation performance and potential based on science of networks and complex adaptive systems perspective? What are the underlying interrelationships between communities' configuration parameters, network metrics, and diversity, resilience, as well as innovation?

1.2 Significance of the Problem

The globalization driven by advances in computer and communication technology, as well as the collective economic and political processes brings dramatic changes in organizational forms and communication networks [61]. The key for such dramatic changes of organizational landscape is the emergence of social communication networks among organizations. Furthermore, the underlying mechanisms for such social communication networks can be abstracted into several theories of communication. Therefore, it is important to study these theories that shape the communication networks.

Understanding scientific communities and their associated social communication networks is key to understanding the dynamics of knowledge creation, as well as formation and growth of scientific communities to facilitate informed science and innovation policy-making. Some Federal agencies, such as NIH and DOE, have begun to use social network analysis techniques to understand the process of innovation [91]. Lack of knowledge of science and innovation dynamics can lead to serious and unintended consequences [91]. For example, Federal encouragement of universities to transfer technologies to industry has resulted in universities putting more attention on near-term research rather than long-term basic research. In addition, Shane [83] examines the effects of Bayh-Dole Act in the United States on one aspect of technology commercialization i.e. university patenting, and suggests that the Bayh-Dole Act provided incentives for universities to increase patenting.

Studying the formation and behavior of scientific communities could avoid unnecessary duplication by predicating what final forms the community could evolve into. For instance, in [47], Kaiser presents a computational model to predict the emergence and development of scientific fields.

"Although the importance of investment in science, technology is understood, the rationale for specific scientific investment decisions lacks a strong theoretical and empirical basis" [91]. So, an interdisciplinary research theme, called "the science of science policy" has recently emerged. This is a theme that aims to provide a scientifically rigorous quantitative

basis from which policy makers and researchers could assess the impacts and likely outputs, while improving the understanding of its dynamics [28]. It is critically important to develop science of science policy because the U.S. Federal government's total R&D budget reached $139 billion in 2007, and it is essential to make use of such a significant amount of funding effectively so as to maximize social and economic benefits.

Research funding could be structured to encourage the formation of new communities, as is currently occurring through the large Federal investment in the nanoscience [82] and synthetic biology [10] communities. Investment in innovation capacity is the key to higher productivity, higher wages, and higher economic growth [91]. Although more emphasis have been put on investment analysis, there is little understanding of how scientific communities respond to changes in funding within research areas. The understanding of how communities of science evolve would have clear implication for investment decisions.

1.3 Strategy

Under globalization driven by advances in computer and communication technology, the flow of information that transmits through communication networks is independent of space and time, because people can share knowledge and make contributions simultaneously anywhere in the world [61]. Furthermore, the mechanisms for the emergence and evolution of communication networks can be abstracted into several communication theories. So, the first perspective for GPS is a global communication system.

In the communication network, communities affect and are influenced by peer communities through boundary processes, during which cooperation and competition occur. GPS communities operate in ways similar to ecosystems in that communities act to ensure their survival and success by accessing resources, creating knowledge, and keeping attractiveness within the social communication network in which they want to thrive [60]. So, the second perspective for GPS is a creative ecosystem.

As a communication system, the interconnections between communities are emphasized. As an ecosystem, it focuses on cooperation and competition driven by which communities form, develop, fade, and coevolve. These two properties can be combined under the framework of a complex adaptive system, because the complex system is a system composed of interconnected parts that as a whole exhibit one or more properties not obvious from the properties of the individual parts [99]. Meanwhile, agent based modeling (ABM) is an ideal way to study complex systems because even a simple ABM can exhibit complex behavior patterns and provide valuable information about the dynamics of the real-world system that it emulates [12]. ABMs provide theoretical leverage where global patterns of interest are more than the aggregation of individual attributes, but at the same time, the emergent pattern cannot be understood without a bottom up dynamic model of the micro foundations at the relational level [55].

The strategy adopted here is to explore how communities' innovation networks form and evolve under a specific communication theory using the complex adaptive systems perspective. Furthermore, the environment is designed to maximize innovation outputs based on the understanding of communities' responses to varying investment strategies. So the study is guided by network theory, boundary processes, and the theory of complex adaptive systems.

1.4 Contributions

1.4.1 Contributions to Theory of Agent-Based Modeling

Agent-based Modeling provides theoretical leverage to explore complex systems where global patterns result from interactions of multiple agents. In a large-scale ABM, it is essential to standardize the communication among interacted agents, since agents may be developed by different programmers. Agent Communication Language (ACL) [31] is proposed by the Foundation for Intelligent Physical Agents (FIPA) as a standard language for agent communications. ACL mainly focuses on the structure of message sent and received

by agents, which includes four mandatory parameters: performative, sender, receiver, and content [31]. Because FIPA-ACL is a basic speech-act theory based communication primitive, it does not provide any specific rules to guide how the communication is carried out. My research using social communication theories as behavioral rules of agents can advance existing ACL by providing a new layer above it. The new layer is named as communication protocol that defines how to choose communication targets, when the communication happens, and when the communication dissolves.

Agent-Based Models (ABMs) are often criticized for relying on informal, subjective, and qualitative validation procedures [27]. Because most ABMs are highly abstract and are built from bottom up, their emergent behavior is often unpredictable. Furthermore, ABMs are often developed for studying complex adaptive phenomena, which involve uncertainty and ambiguity in terms of their underlying behavioral mechanisms. Models that focus on human and social dynamics are especially prone to ambiguity and uncertainty. To gain empirical insight into such problems and to be able to generate behavior that mimics expected or theoretical scenarios, model development and refinement should be coupled with evaluation and validation. The validation strategy used here is a Robust Generative Validation (RGV) [111] method that refers to the strategies used by scientists in generating and validating knowledge. The main steps of RGV consist of generating ensembles, initiating schema, evaluating schema, and transforming schema, where each ensemble refers to a single hypotheses space and each schema refers to the set of configurations corresponding to the ensemble. The model introduced in this dissertation is validated using a simplified RGV by replacing a network of ensembles with two independent experiments. These two experiments aim to compare simulation outputs against overlay map and OBO Foundry respectively.

1.4.2 Contributions to Science and Innovation Policy Development

In order to address the questions raised in section 1.1, a multi-agent model is built, in which behavioral rules of agents can be varied based on various social communication theories including homophily, preferential attachment, structural hole, exchange, and balance theory.

Homophily theory has been identified by social scientists as an important mechanism that explains communication networks are created, maintained, and reconstituted [61]. Homophiliy means a community would like to communicate with similar others and is highly influenced by similar peer communities. Similarity is thought to ease communication, foster trust, and reciprocity, and improve diffusion of knowledge [15]. *Structural holes* are those knowledge spaces where communities are not connected so that other communities may exploit them by investing their social capital to indirectly link two or more unconnected communities [61]. The community that fills the structural hole becomes a broker in relationship to others. A *preferential attachment* is a process where resource is distributed among individuals according to how much they already have, i.e., rich gets richer. Under suitable circumstance, preferential attachment can generate power law [25] that exists in many social systems, for instance, the number of papers published by authors, the citation index of papers etc. *Exchange and dependency* theories seek to address how communication emerges based on the distribution of information and resources across the members. Heider's *balance* theory [38] states: "my friend's friend is my friend; my friend's enemy is my enemy; my enemy's friend is my enemy; my enemy's enemy is my friend", which means friends have similar attitudes, while enemies have different opinions on the third object. Using homophily, preferential attachment, structural hole, exchange, and balance theory, we analyze the interaction between communities and compare them in terms of emergent network patterns and innovation metrics.

In addition, experiments using agent-based simulation have been conducted. Among these experiments, we examine six types of topologies (i.e., 1D grid, 2D grid, random network, random group network, scale-free network, and dynamic network) and observe the

emergent patterns of communities and their interrelationship with innovation performance. Simulation results show that scale-free network has the highest resilience compared with random and random group network. In addition, the relationship between variety and density is a concave-like function, to which the relationship between variety and centrality is similar. Furthermore, policy-makers may encourage communities to build highly connected clusters if resource availability is low. As far as the resource allocation strategy is concerned, key area investment with technology transferring results in the highest variety. Considering the situation where communities communicate with one another guided by structural hole, preferential attachment, or homophily theory, decision-makers may encourage communities to be open to accept influences from peers in order to foster innovation. In addition, under low communication frequency, openness and receptivity lead to higher variety. On the contrary, variety decreases with increasing receptivity under high communication frequency.

The rest of the dissertation is organized as follows. In Chapter 2, we present background on GPS from the perspective of communication system, ecosystem, and complex adaptive system. Chapter 3 introduces the design and formalization of the model, which embodies the mechanisms of boundary processes, Homophily theory, and HSB (Hue, Saturation and Brightness) color model that is used to visualize emergent community landscapes. Chapter 4 describes the implementation of the model. The verification and validation are conducted in Chapter 5, where a novel generative validation method is devised to instill confidence in the operational behavior of the model. Chapter 6 delineates metrics and indicators used to measure network structure and innovation output, as well as evaluation using these metrics. Chapter 7 examines the impacts of socio-technical communication theories on innovation performance. Finally, in Chapter 8, we conclude by summarizing our findings and point out potential avenues of future research.

Chapter 2

Literature Review

The research conducted in the dissertation views global participatory science as a global communication system and a creative ecosystem. These two perspectives can be combined under the framework of a complex adaptive system.

2.1 Characteristics of GPS Communities

Recently, a number of virtual scientific collaboratories emerged and continue to successfully bring together scientists over the globe to collaborate to not only share and aggregate data, but also create new knowledge [93]. Such virtual collaboratories include Open Source Science (OSS) communities such as OBO Foundry (Open Biomedical Ontologies) [86], which is a form of GPS. So, we choose OSS communities as a research object to study, develop, and explore models of innovation in collective knowledge creation communities. OSS communities are immune to process loss through production blocking because all team members can contribute ideas simultaneously. In addition, OSS communities reduce cognitive failures and enhance the synergistic effects of group brainstorming using electronic media to communicate, because access to the data is unrestricted by individual recall [108]. Furthermore, compared with traditional scientific teams, OSS is located with distributed structure of network, as well as more open and transparent due to decentralized decision-making style. Besides OBO, the following are among such open science communities: NanoHUB (Simulation Education Technology for Nano Technology) [62], and NEES Grid (Network for Earthquake Engineering Cyberinfrastructure) [63]. Table 2.1 describes the comparison between traditional scientific teams and open source communities.

Table 2.1: Traditional Scientific Teams vs. Global Participatory Science

Criteria	**Additional Criteria**		*Traditional Scientific Teams*	*Open Source Communities*
Distribution	Space		Co-located	Distributed
	Time		Synchronous	Asynchronous + synchronous
Communication			Face to face	Virtual meeting
Organization	Structure		Hierarchical	Networked
	Style		Team/Formal Group	Community/Market
Openness	Product	Access	Push-driven	Pull-driven
		Transparency	Complete product	Incomplete product
		Integration of contributions	Pre-production decisions	Pre and post-production review
	Process	Decision-making	Closed/Centralized	Open/Decentralized
Mobility	Entry threshold		High	Low
	Turnover rate		Low	High

In the invisible college [93], researchers complement each other by sharing equipments, ideas, knowledge, techniques, and tools. In other words, scientific curiosity and ambition are the driving forces for researchers to work together in an invisible college. As far as these networks are concerned, they are neither pre-designed, nor random. Rather, these networks organize and operate based on self-organizing processes, which are also the main focus of this research. With better understanding of such rules, policymakers could make better policy decisions in terms of improving innovation performance and investment efficiency.

2.1.1 Understanding GPS as a Communication System

Social network theory is often used to describe the structure of scientific communities, but little research is conducted on the formation of network of communities [64].

Communication networks and the organizational forms of the 21st century are undergoing rapid and dramatic changes [32]. There are many theories that focus on the role of

social communication mechanisms in explaining the emergence and evolution of community networks. Table 2.2 summarizes selected social theories.

Table 2.2: Selected Social Theories [61]

Theories	Sub-Theories
Theories of Self-interest	Social Capital Structural Holes Transaction Costs
Mutual Self-Interest & Collective Action	Public Good Theory Critical Mass Theory
Cognitive Theories	Semantic/Knowledge Networks Cognitive Social Structures Cognitive Consistency Balance Theory
Contagion Theories	Social Information Processing Social Learning Theory Institutional Theory Structural Theory of Action
Exchange and Dependency	Social Exchange Theory Resource Dependency Network Exchange
Homophily & Proximity	Social Comparison Theory Social Identity Physical Proximity Electronic Proximity
Theories of Network Evolution	Organizational Ecology NK(C)

Theories of self-interest explain how people make decisions based on their personal favorites and desires [20]. Theories of mutual interest and collective action focus on why outcomes produced by coordinated activity are unattainable by individual action [21]. Contagion theories examine how ideas, messages, and beliefs spread through some forms of direct connection [18]. Cognitive theories address the role that knowledge and perception play in socio-technical communication networks [88]. Exchange and dependency theories seek to address how communication emerges based on the distribution of information and resources

across the members [43]. Homophily and proximity theories explore the emergence of communication networks based on the similarities of network members' traits [15]. Theories of network evolution study the mechanisms of variation, selection, retention, and competition [95].

2.1.2 Understanding GPS as a Creative Ecosystem

Scientific communities behave in similar ways to an ecosystem in that there exist both competition and cooperation over the use of resource; that is, interacting species (i.e., communities) compete to gain resources from their environment to survive and grow, while also cooperating to develop symbiosis and to improve their chance for survival. Artificial ecosystems have grown as a generalized evolutionary approach for creative discovery, since their applications across different domains have been developed such as economics [6], ecology [59], and social science [29]. Arthur [6] extends the frameworks of economics from viewing economic activities within an equilibrium steady state, to viewing economic activities continually changing, and constantly adapting and co-evolving within a dynamically changing environment. Mitchell [59] abstracts the natural evolution at a high level into two phases: evolution using genetic operators (e.g., combination, mutation etc.), and selection of descendents based on fitness. Epstein [29] studies the underlying rules and develops models about how the decentralized local interactions of heterogeneous autonomous social individuals could generate regularities observed in the real world.

We can summarize characteristics of artificial ecosystems under eight basic concepts and processes listed as follows [56]:

- The phenotype used in artificial ecosystem forms the basis of an individual.
- A collection of individuals represent a species.
- Individuals are distributed and move within the environment.
- Individuals inhabit and interact within environment.

- Individuals have abilities to modify and change the environment.
- Individuals have a scalar measure to represent success, i.e., health.
- Individuals undergo stages of development, i.e., life-circle.
- An energy-metabolism cycle describes the resources cycle.

In relation to dynamics of ecosystems, scientific communities exhibit the following characteristics:

- The domain of a scientific community is its phenotype, which is composed of norms, practices, and skills.
- Clusters of communities are comprised of epistemic cultures that correspond to species.
- Communities are distributed globally.
- An explicit model of environment (e.g., funding agencies) influences decisions of communities by altering the availability and distribution of resources.
- Scientific communities have the ability to change and modify their environment as a result of research and technology transfer.
- Communities have a scalar measure to represent success, i.e., fitness.
- Scientific communities undergo stages of coalescing, growth, stability, and renewal.
- Scientific communities adopt external funding and transfer human capital and knowledge into technology and products, which is similar to metabolism.

Based on the characteristics above, scientific communities can be viewed as a creative ecosystem.

2.1.3 Boundary Processes in Knowledge Creation and Diffusion

Boundary refers to something that indicates a limit or extent, which often has negative meanings because it leads to limitation and a lack of access. Boundaries exist between communities; for example, there are technical communication challenges when communities of psychology and computer science jointly hold a meeting. Unlike in traditional scientific teams, where boundaries are usually well defined due to officially sanctioned affiliation, the boundaries of the open source communities are rather fluid because they engage in interdisciplinary research. On the other hand, boundaries are also important for a learning system, because boundaries offer learning opportunities, and the learning opportunities are different from those within a community [96]. In a community, the competence and experience converge since it is the basis for a community to be stable. However, the competence and experience tend to be diverse so as to expose communities to a foreign competence [97]. Therefore, both strong core activities within a community and active boundary processes determine the learning and innovation potential of a social learning system [57].

The influence among communities is bidirectional, which means that each scientific community influences other communities by publishing papers and holding conferences. At the same time, they are also affected by peer communities. Such processes are called boundary processes, which "arise from different enterprises; different ways of engaging with one another; different histories, repertoires, ways of communicating, and capabilities" [97]. Through boundary processes, communities with common interests that promote each other become closer and closer so that clusters emerge. Communities in a cluster share similarities in terms of discipline, norms, skills, and expertise, and strongly connect with each other. In addition, there are still interconnections among clusters, which are important for different ideas to diffuse, although such interconnects are not as strong as those inside clusters. At last, the environment communities inhabit is another noteworthy item. It is expected that environments influence the behavior of communities by investments policies [91].

In order to build bridges across boundaries, some communities may act as brokers between communities that are originally disconnected. Brokering makes boundary processes occur, through which knowledge diffuses. Concomitantly, communities that act as brokers are likely to thrive because they can benefit from different experiences and views [97]. This view is similar to the theory of structural hole in that communities invest social capital in a structural hole to gain profits, as a broker builds links between those communities that originally disconnected.

Effects of boundary processes on GPS are different from those on traditional scientific teams in terms of strength, because they have different decision-making styles. The decision-making style in traditional teams is centralized, which results in traditional teams being less likely to be affected by boundary processes. On the other hand, GPS communities are more likely to be influenced by boundary processes since its decision-making style is decentralized. In addition, GPS communities have lower entry threshold than traditional ones, which causes higher mobility that in turn leads to higher influence of boundary processes on GPS communities. Therefore, to deal with different effects on GPS or scientific teams, the simulation model presented in the dissertation adjusts the receptivity (ratio of weights of neighbors to weights of itself). In other words, the receptivity for GPS is larger than that for traditional scientific teams.

2.1.4 Understanding GPS as a Complex Adaptive System

A complex system is composed of interconnected parts that as a whole exhibit one or more properties (behavior among the possible properties) not obvious from the properties of the individual parts. In essence, complexity is concerned with emergency, that is, the process where the global behavior of systems results from the actions and interactions of agents [99].

The behaviors of scientists and scientific communities have the characteristics of complex adaptive systems. While scientists and scientific groups adapt their behavior to fit their changing environment, they also actively shape it to create cognitive niches to improve their

resilience and success. The forming, dissolution, and maintenance of emergent collaboration structures in reaction to opportunities, resources, and environmental (e.g., science policy) interventions can be viewed as a dynamic ecosystem [108]. Therefore the formation of scientific domains, problem areas, and disciplines occurs in the context of a complex adaptive system [41].

Knowledge creation in GPS is comprised of a large population of decentralized networked individuals and groups of scientists who interact with and influence each other to form aggregate emergent communities of interest around focal problem domains. As a complex adaptive system, a GPS has the following characteristics: [108]

1. Problem solving behavior, as well as emergence and co-evolution of communities are results of activities and interactions of scientists.

2. Communities compete and cooperate to form and sustain cognitive niches and interact through boundary processes.

3. Scientists and knowledge have mobility across boundary of scientific communities. Mobility fosters innovation [25].

4. Consists of many complex subsystems (e.g., scientific communities, academic institutions, R&D institutes).

Unlike in a traditional research project, where scientists are guided by a central authority, scientists in a GPS aim at not only advancing science but also choosing a problem based on their self-interests. During the scientific process, scientists generate new problems by solving existing problems, which in turn attracts more scientists with similar interests to participate. Thus a circle of positive feedback forms. In [70], Pirolli illustrates how the dynamics of information foraging play a negative feedback when solutions become routine and become less novel due to diminishing rate of returns. Under such positive and negative feedback, scientists adapt their behavior to improve their fitness and success [108].

2.2 Environmental Constraints

The environmental constraints in our research refer to investments and policies that policy-makers can leverage to influence the activity of scientific communities and further foster the innovation performance. One of the policy instruments available to policy-makers is the investment strategy in science and technology. The other is to foster the role of competition and cooperation in the promotion of discovery [91]. However, the impacts of these various policies on innovation are largely unknown. McCormack [56] states that the design of environments based on which creative behavior is expected to emerge is at least as important as the human capital which is expected to evolve within the environment. The lack of knowledge about impacts of policies can lead to unintended consequence [91]. For example, the goal of College Opportunity and Affordability Act of 2007 [68] is to stabilize the state higher education spending and decrease the cost of colleges. But there is an undesired consequence that the rate of growth of state higher education spending in the future is also reduced.

Although the importance of (public and private) investment on science is widely accepted, there is still a lack of theories and methodologies to evaluate the nature and distributions of investments. Reed et al. [76] propose a seven-step framework to help program managers develop a well-structured impact evaluation: 1. Identify scope, objective, and priorities; 2. Select the types of evaluation to be completed; 3. Select the aspects of deployment-induced changes to be evaluated; 4. Identify research questions and metrics; 5. Design the evaluation; 6. Conduct the evaluation; 7. Report and use the results and data. Knezo [49] concludes that Federal agencies allocate funding based on topical or field-specific priorities that have shifted over time, by studying the trends of Federal R&D budget in the last half century.

The other policy instrument policy-makers can make use of is to foster competition and cooperation. Competition for resources leads to various kinds of alliances/mutual relationships and to the establishment of various symbiotic relationships [3]. Axelrod [8] undertakes

a variety of simulations related to iterated Prisoner's Dilemma, drawing conclusions based on these experiments about the relationship between selfishness in the short run versus cooperation in the long run. Axelrod also figures out ways in which groups form, adhere, oppose or join other groups.

Based on the understanding of how scientific communities respond to changes in funding across research areas, we can design an environment by making policies to guide communities to act in the ways agencies expect them. It is to analyze the configuration of an environment given an expected behavior or output of communities. Since the relationship between the environment and communities' behavior may not be one to one, it is feasible that multiple environmental configurations correspond to the same innovation outputs. The basic allocation decision is the choice of which items to fund in the plan, and what level of funding it should receive [104]. There are two contingency mechanisms dealing with unexpected situations. One is to determine which community will be funded if more resources are available. The other is to determine which community will be sacrificed if total resources have to be shrunk.

For designing the environment, there are at least two potential aspects that need to be explored further.

- Find out available investment strategies that can be used to compare effects on innovation performance or potential. For example, broad investment in all domains, key support for some specific domains, random allocation, and dynamically changing allocation based on contributions.

- Find out the investment strategy to maximize the innovation performance or innovation potential. For example, what is the investment strategy so as to maximize the diversity of communities?

One difference between traditional teams and GPS is that GPS may require fewer resources to create and maintain because resources can be shared by collaboration in GPS.

In addition, GPS is driven by the desire of scientists to do original and creative research [93], which in turn can reduce costs to hire scientists. Therefore, the difference between traditional teams and GPS is reflected by the maintenance resource model with adjusting the maximum/minimum resources needed. In other words, traditional teams have higher maximum/minimum resources needed than GPS. So the model presented in this study tries to deal with both traditional teams and GPS by adjusting the parameters of the model.

2.3 Relation to Earlier Work

Earlier studies pertaining to the application of computational models to scientific discovery processes focus on simulating cognitive processes and re-enacting discoveries [48]. Specifically, computational modeling of concept formation is viewed as central to discovery and has a long history [44]. More recent and complex applications of computational models include artificial intelligence and machine learning techniques that view science from the perspective of problem solving [84]. Most of these techniques focus on mimicking the discovery processes employed by individual scientists. Yilmaz [109] develops an agent simulation model conducted to examine the impact of culture and conflict management styles on collective creativity in open source innovation systems. How collectives govern and coordinate the actions of individuals to maximize innovation output is examined in [110] to better understand the emergence of collective creativity.

Besides computational models, there are also other methods to study scientific activities. The overlay map presented in [75] is a visualization technique that intends to catch the reforms that most science and technology institutions are undergoing to transcend the traditional boundary of disciplines. Visual analytics is a new field of research that is focusing on how people interact with information to make decisions [24]. Such visualization techniques are also applied in other domains. Chemists use visualization to present a visual comparison of properties or states in two or more systems so as to present visual prediction of properties or states in the future [33]. Gloor [36] introduces an alternative method of measuring the

success of knowledge workers. In [24], the visualization technique is used to address how the public investment in science affects the lives of U.S. citizens. As a new emerging tool, visualization techniques also have some limitations. The data for visualization has so far been limited to publication and patent data. In addition, in many cases it is not possible to pool many cross-country data-sets because the data is gathered in different ways. Furthermore, if there is no understanding of the underlying dynamics, the use of visualization does not advance metrics [24]. Therefore, it is reasonable to combine visualization techniques with computational models to better understand the underlying mechanism and to better present results.

Although significant research has been conducted on scientific communities from the network perspective, simulation modeling of such communities is rare. In [34], Gilbert develops a model where citation patterns and growth of knowledge are simulated to exhibit empirical regularities observed in scientific communities. However, this study does not aim to consider social processes pertaining to enculturation and innovation. On the other hand, the simulation study presented in [112] views scientific discovery as a social process. However, it focuses on the interactions between single scientists so that it does not analyze the pattern of network of communities formed by single scientists. In the context of innovation, the use of simulation of collective invention and innovation diffusion [22] revealed the significance of social network structure in knowledge creation and diffusion. Furthermore, in [56], McCormack uses the HSB model to represent artificial species so as to demonstrate similar species with similar color, which ignites the idea of using the HSB color model to vividly depict the states of scientific communities. In order to analyze the inner dynamics, an organization is often divided into several interconnected components such as organizational structure, agent, and environment [26]. Recently, the Simulating Knowledge Dynamics in Innovation Networks (SKIN) [85] emerges as a tool to simulate knowledge dynamics in innovation networks, which has been applied in learning competence [35], the university-industry links [2], and technology spillover [73].

Building on these earlier studies, the model introduced herein is:

1. a computational model, that can provide not only qualitative but also quantitative analysis.

2. a model based on complex system theories, boundary processes, and communication network theories.

3. an adaptive learning system of communities that can change their behavior based on their fitness.

4. using communities as the unit of analysis to track the scientific impact of investments.

Chapter 3

Design Concepts and Details

3.1 Overview

Three major components are selected to represent the status of scientific communities: domain, maturity, and resources. Domain refers to discipline or task characteristics. Maturity is an attribute that indicates the scientific sophistication and degree of advancement in a specific domain. Resources that a community holds are vital to undertake scientific activities. In order to visually depict the states of communities, the HSB color model is used. Hue indicates the domain of a community, as it determines the basic color such as red (0^o), blue (120^o), green (240^o) etc. Saturation represents maturity as it serves as an indicator for the level of growth. Brightness corresponds to the level of resource. Figure 3.1 shows the visual snapshots of our model with grid and network topology, respectively. Each cell represents a community whose color indicates its internal states. As the figure depicts, the community network pattern looks like a color landscape. Hence, the simulation model of GPS communities is named the ***Colorscape*** model. Additionally, Colorscape is generic, as it can be used to represent both traditional scientific teams and GPS communities by adjusting model's parameters.

Figure 3.2 represents the components of the simulation model and their relationships. In Figure 3.2, there are three kinds of relationships: *interaction within a community*, *interaction between communities*, and *interaction between community and environment*. These three interactions are the fundamental driving forces for the dynamics of scientific systems composed of interconnected scientific communities.

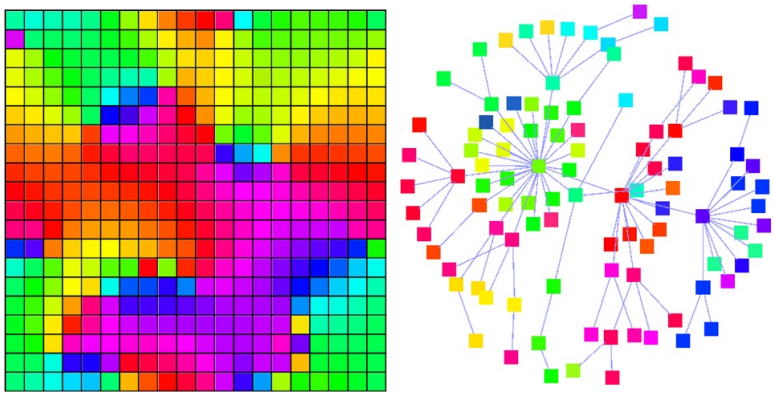


Figure 3.1: Snapshots of Colorscape Model

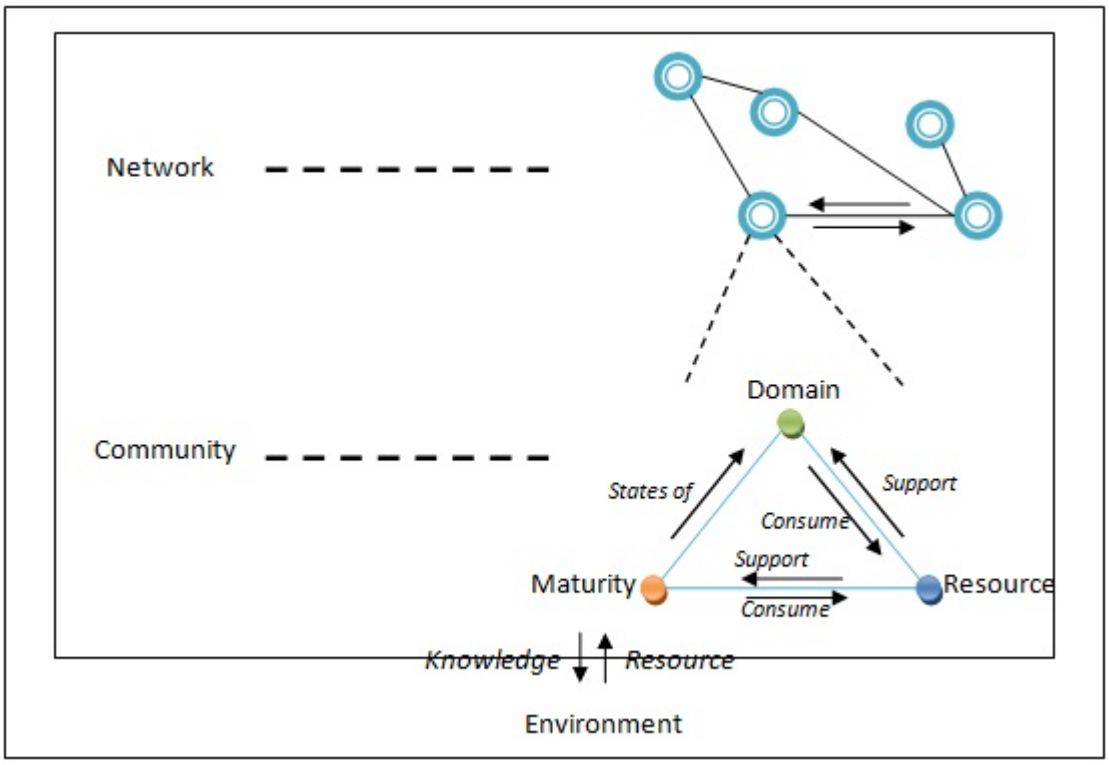


Figure 3.2: Network of Scientific Communities

3.2 Process

As shown in Figure 3.3, the process of our simulation model mainly consists of activities specified by six sub-processes: resource allocation, interaction within community, learning, innovation, growth, and fade. Resource allocation refers to strategies used to distribute resources to communities. Interaction within community refers to scientific activities at

the macro level i.e., a community is driven by funding to improve its maturity. Learning and innovation between communities mimic the boundary processes of communities i.e., communities affect and are influenced by peer communities. Growth is defined as the process through which communities improve their extent so as to increase their influences. Fade refers to disappearance of a community due to loss of resources and attractiveness. These six sub-processes are discussed in detail in the following sections.

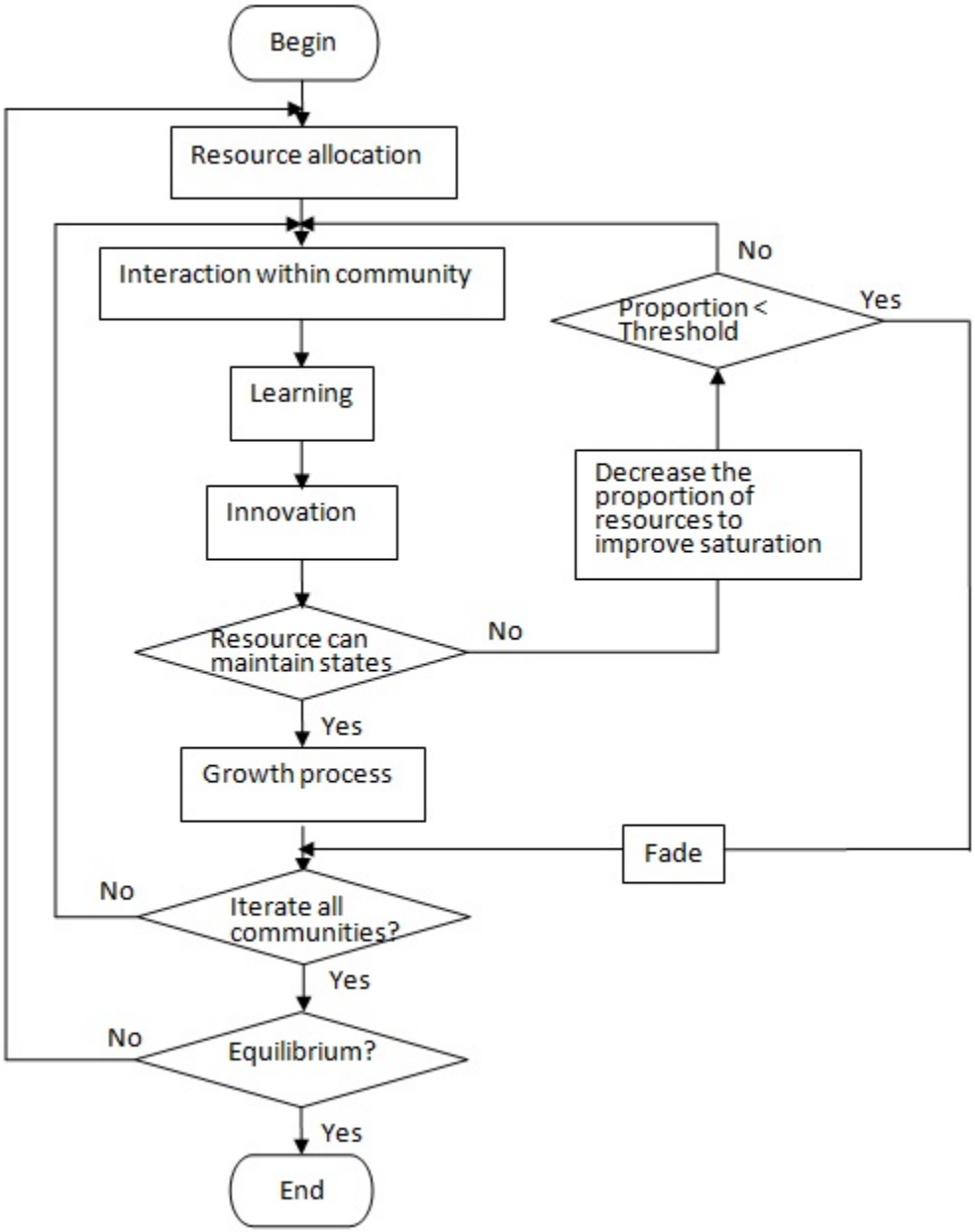


Figure 3.3: The Activity Flow of the ColorScape Model

3.3 Resource Allocation

As shown in Figure 3.4, the triple helix of University-Industry-Government [30] is a spiral model of innovation that captures multiple relationships in the process of knowledge capitalization. Governments provide subsidies and grants as resource. Academia generates knowledge, licenses, and graduates as input to industry; industry generates products as input to innovation, as well as returns on capital and investment capital as input to financial system.

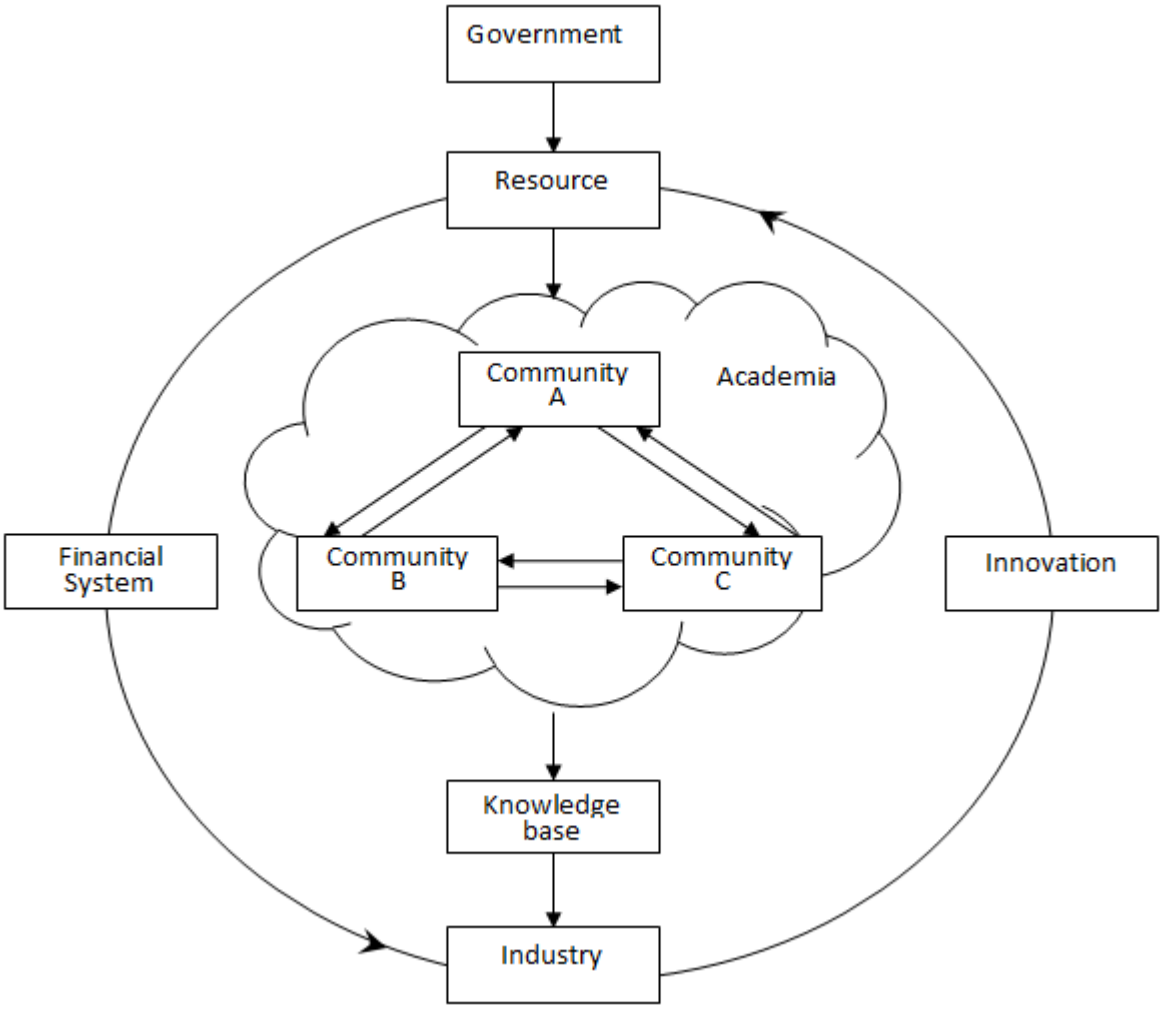


Figure 3.4: Triple Helix of University-Industry-Government Relations

My research focuses on the relationship between government and academia, so Figure 3.4 is reduced to Figure 3.5. In the baseline model, the strategy for resource allocation is

uniform allocation; that is, the total resources are distributed among all communities equally. The total amount of resources available for allocation is equal to sum of the contributions of communities coupled with external funding. Contributions by communities are based on the premise that produced knowledge can be transferred to technology, which in turn results in economic growth.

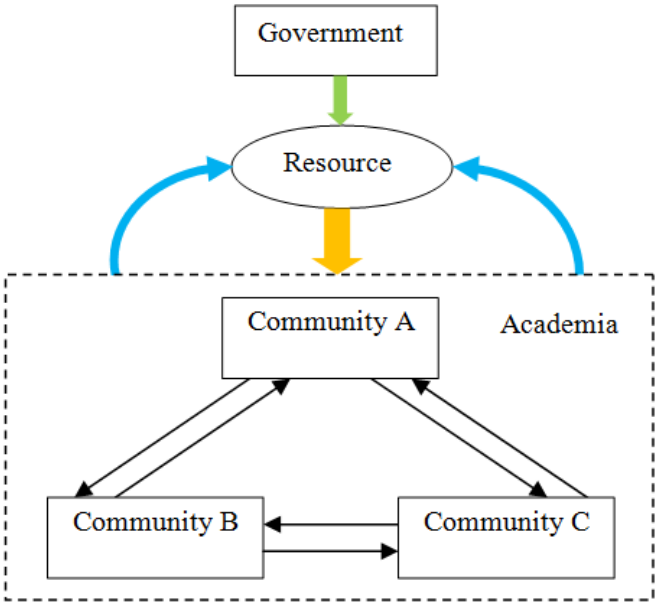


Figure 3.5: Resource Allocation

Contributions provided by a community is moderated by the product of its maturity and resource. This is based on the hypothesis that communities with higher maturity and resources are expected to be more productive. This is expressed in Equation 3.1:

$$R_t = \sum_{i=1}^{\#communities} (F_{i,t} + S_{i,t} \times B_{i,t}), \tag{3.1}$$

where R_t indicates the total resource available at time t. $F_{i,t}$ denotes the external funding allocated to community i at the time t. $S_{i,t}$ and $B_{i,t}$ indicate maturity and resources of community i respectively.

3.4 Interaction within Community

Interaction within the community refers to the scientific activities at the macro level, i.e., the community is driven by funding to improve its maturity. The interaction process is depicted in Figure 3.6.

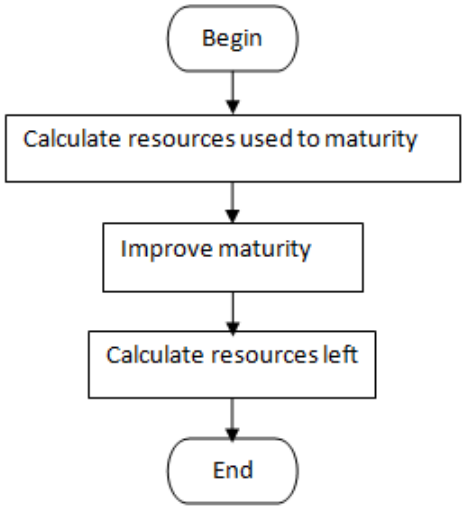


Figure 3.6: Interaction within Communities

3.4.1 Relationship between Maturity and Resources

Riss et al. [78] develop a model of knowledge maturation that includes three phases: coalescing, maturing, and transformation. At the phase of coalescing, the efforts to improve maturity of knowledge are high, because it is a period of exploration toward a solution for a new problem. As the problem and methodology become clear, the maturity improves faster as a result of aggregation of knowledge of individual scientists during the phase of maturing. In the last stage, significantly high efforts are required to standardize knowledge artifacts to make them reusable [78] and resolve conflicts among stakeholders because tension builds up as maturity passes the threshold for the problem and method to be settled. Thus, a U-shaped trend between maturity and efforts arises.

The relationship between resource and saturation is defined as follows:

$$\begin{aligned} R_{m,t} &= \frac{R_{min} - R_{max}}{c} S_t + R_{max} && if\ S_t \leq c \\ R_{m,t} &= \frac{R_{max} - R_{min}}{1 - c} S_t + \frac{R_{min} - cR_{max}}{1 - c} && if\ S_t > c \\ 0 < & \quad c \quad < 1 \end{aligned} \tag{3.2}$$

where $R_{m,t}$ is the resources needed to maintain the current saturation. S_t is the current level of saturation, and c is the critical value that divides the trend into two categories. R_{max} and R_{min} are the maximum and minimum resources needed respectively, which are adjustable and based on types of communities.

At each time step, a community receives resources via external funding. But not all available resources can be used to push forward the maturity of community, i.e., only part of the resources helps advance maturity, because the following learning and innovation processes also require resources. How much saturation the community can gain by these resources is determined by the following equation:

$$S_{t+1} = S_t + \alpha_t \times (1 - S_t) \times R_{s,t}, \tag{3.3}$$

where S_{t+1} is the maturity of the community at the time $t + 1$. $R_{s,t}$ is the resources that could be used to increase maturity, which is a proportion of $(TotalResource - R_{m,t})$. The increase in saturation is adjusted by α_t, which is an exponential decay function with gradually decreasing slope over time, because more efforts are needed to sustain a community with increasing maturity.

$$\alpha_t = e^{-t'/\tau} \times (S_{max} - S_{min}) + S_{min}, \tag{3.4}$$

where α_t is an adjusting parameter to control the increment of saturation. t' is the time period during which saturation increases continuously. τ is a constant coefficient to

control the slope of the function curve. S_{max} and S_{min} refer to the maximum and minimum increment for saturation respectively. At the time when saturation just begins to increase, α_t is equal to S_{max}. As saturation increases, α_t gradually decreases toward S_{min}.

3.4.2 Resources Consumed

As maturity increases, it is necessary to consume resources because technology develops based on research and development costs. Resource level is updated as Equation 3.5,

$$B_{t+1} = B_t + R_t - R_{m,t} - R_{s,t}, \tag{3.5}$$

where B_{t+1} is the resource the community has at the end of this interaction process. B_t denotes the resource the community already holds at the current time. R_t is the new resource allocated to the community at the current time. $R_{m,t}$ is the resource to maintain the current state, which is based on equation 3.2. $R_{s,t}$ is the resource needed to change maturity.

3.5 Learning between Communities

Learning between communities mimics the boundary processes between communities, i.e., communities affect and are influenced by peer communities. In the Colorscape model, based on the assumptions predicated on the Homophily theory [61], a community is more likely to communicates with similar others and is highly influenced by similar peer communities. As depicted in Figure 3.7, when a scientist or domain knowledge in community 2 transfers to community 1, the scientists of community 1 who interact with and are influenced by the new knowledge are pulled toward to the new direction.

Figure 3.8 depicts the process of learning, which mainly includes update neighbors, update weights of neighbors, update discipline, maturity and resources of the current community through boundary processes based on the homophily theory.

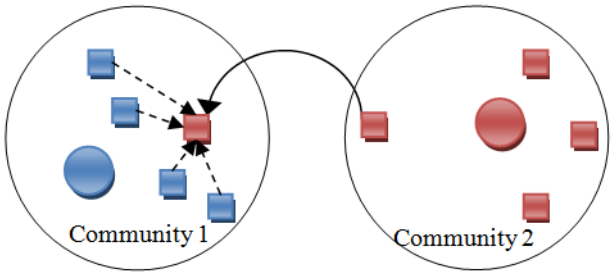


Figure 3.7: Learning Process

3.5.1 Updating the Intensity of Communities' Influences

The influences that communities exert or receive are reflected by their interaction frequency. Interaction frequency between communities is depicted by the weights associated with links in the evolving communication graph. According to the Homophily theory, the more similar communities are, the stronger the influences. So, the intensity of community j's influence on community i is defined as follows:

$$\begin{cases} W_{ji,t} & = W_{ji,t-1} + C_W \times I_{ji,t} \times (1 - W_{ji,t-1}) \qquad if\ I_{ji,t} \geq 0 \\ W_{ji,t} & = W_{ji,t-1} + C_W \times I_{ji,t} \times W_{ji,t-1} \qquad otherwise \end{cases} \tag{3.6}$$

where $W_{ji,t}$ is the influence of neighbor j at the current time. C_W is a number between 0 and 1 and is inversely proportional to inertia (resistance to change in a community). $I_{ji,t}$ is the intensity of change in the influence, which is defined as:

$$I_{ji,t} = (1 - D_{ji,t})^4 - (1 - \overline{D_{i,t}})^4, \tag{3.7}$$

where $D_{ji,t}$ is the dissimilarity which is equal to the distance between community i and community j in terms of current hue at the time t whose equation is 3.8. $\overline{D_{i,t}}$ is the average distance between community i and all of the neighbors at the time t. This function grows much faster when dissimilarity between i and j becomes smaller in comparison to average dissimilarity, resulting in higher intensity $I_{ji,t}$.

The equation for the dissimilarity between community i and j is defined as follows:

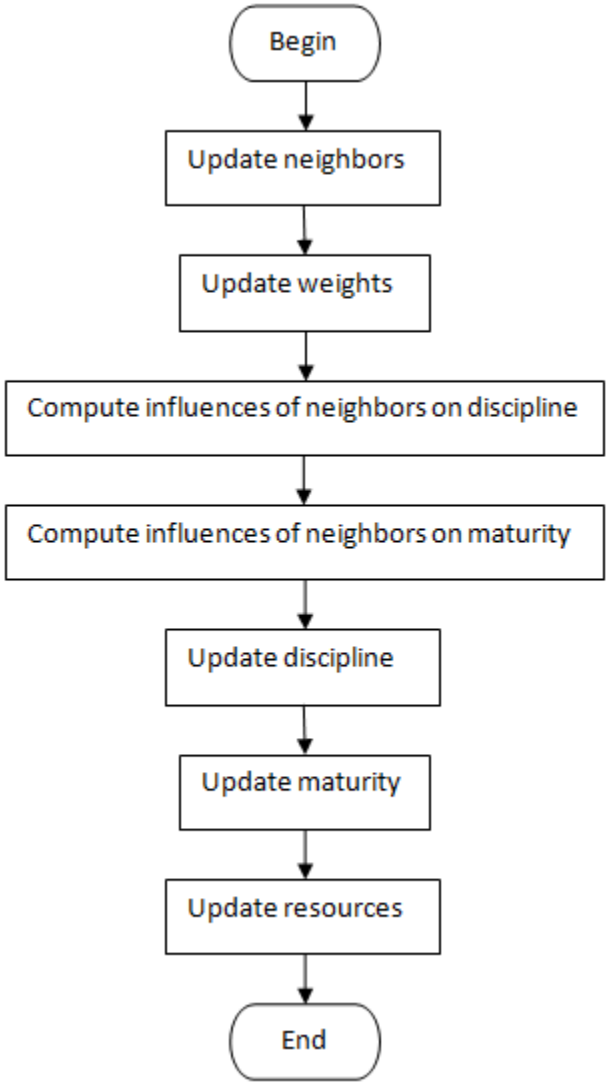


Figure 3.8: Flow Chart of the Community Learning Process

$$D_{ji,t} = Dissimilarity(H_{i,t}, H_{j,t}), \tag{3.8}$$

where $H_{i,t}$ is the hue of community i at the time tick t. $H_{j,t}$ is the hue of community j at the time j.

$$Dissimilarity(x, y) = \begin{cases} \frac{|x-y|}{180} & if\ |x-y| \leq 180 \\ \frac{360-|x-y|}{180} & otherwise \end{cases} \tag{3.9}$$

3.5.2 Updating the Maturity of a Community

Learning among communities affects both saturation and discipline. Saturation refers to maturity of the domain that is the state or quality of being fully grown or developed. The reason for change in saturation by learning is that scientists can borrow theories and methodologies from other domains to improve the skills and knowledge necessary to solve problems in their own domain.

As shown in Figure 3.9, the circle refers to hue and the vector refers to saturation. Length of the vector indicates the strength of saturation. The longer the vector is, the larger is the saturation. Angles represent differences between communites in terms of their domains. The larger the angle is, the more different these domains are. S_1, S_2 and S_3 are saturation of community 1, community 2, and community 3, respectively. S_2 and S_3 are in different domains from S_1. But both S_2 and S_3 have effects on S_1 moderated by the angles α and β, respectively. So, the influence from S_2 is equal to $W_{2,1} \times S_2 \times cos(\alpha)$. On the other hand, the influence from S_3 is $W_{3,1} \times S_3 \times cos(\beta)$, where $cos(\beta)$ is negative since β is obtuse.

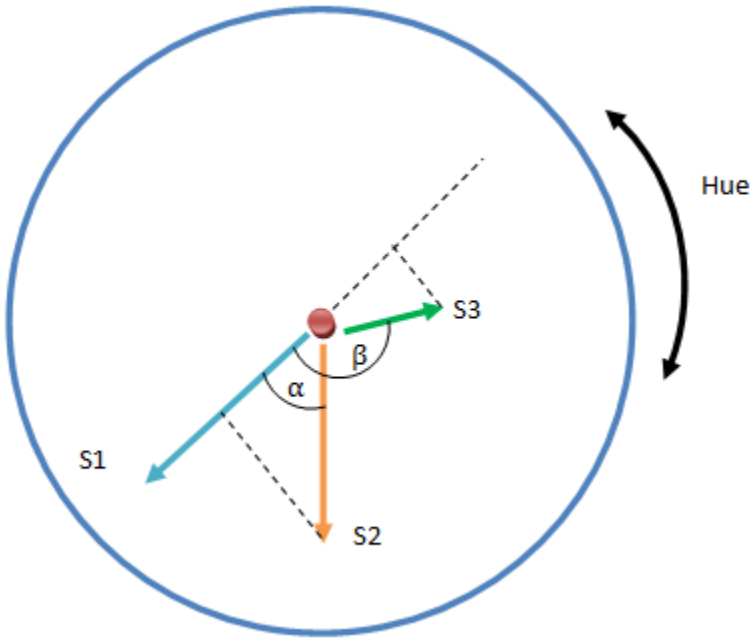


Figure 3.9: Updating Maturity during the Learning Process

The change of saturation during the learning process is the sum of influences from peers, which is defined by equation 3.10,

$$S_{i,t+1} = S_{i,t} + A_S \times \sum_{j=0}^{\#neighbors} W_{ji,t} \times S_{j,t} \times cos(\alpha_{ji,t}), \tag{3.10}$$

where $S_{i,t+1}$ refers to the saturation of community i at the time $t+1$. $\alpha_{ji,t}$ refers to the angle between the hues of communities i and j. A_S is a function of susceptibility defined in Equation 3.11:

$$A_S = e^{-\mu \times R}, \tag{3.11}$$

where R refers to the resources the community currently holds. μ is a constant coefficient to control and calibrate the rate of change in susceptibility.

3.5.3 Updating the Discipline of a Community

Learning can lead the current community to change its hue i.e., discipline (specific norms, practices, and relevant skills) due to influences from neighbor communities. Concomitantly, the community itself is inclined to realize its own target norms as shown in Figure 3.10, where the circle refers to hue and the vector refers to saturation. Angles between vectors represent differences of communities in terms of their domains. The larger the angle is, the more different these domains are. $H_1^{current}$, $H_2^{current}$ and $H_3^{current}$ are the current hue of community 1, community 2, and community 3, respectively. $H_2^{current}$ and $H_3^{current}$ are in different domains from $H_1^{current}$. But both $H_2^{current}$ and $H_3^{current}$ have effects on $H_1^{current}$ with angle of α and β, respectively. Additionally, $H_1^{current}$ is pulled by its target hue H_1^{target} due to its intention to realize its target. So, the change of hue during the learning process is the sum of influences from peers and its own inclination, which is defined by equation 3.12.

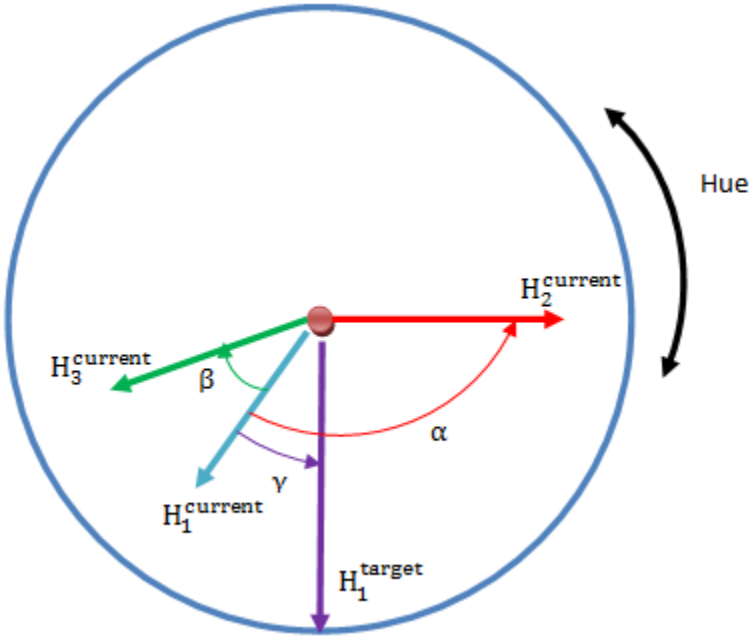


Figure 3.10: Domain Update during the Learning Process

$$H_{i,t+1}^{current} = H_{i,t}^{current} + A_H \times \left(\left(\sum_{j=1}^{\#neighbors} W_{ji,t} \times (H_{j,t}^{current} - H_{i,t}^{current}) \right) + W_{i,t} \times (H_{i,t}^{target} - H_{i,t}^{current}) \right) \tag{3.12}$$

where $H_{i,t+1}^{current}$ refers to the new hue after the learning process. $H_{i,t}^{current}$ refers to the current hue of the community i. $H_{j,t}^{current}$ refers to the current hue of the community j. $W_{ji,t}$ refers to the influences of community j on community i at the current time. $W_{i,t}$ refers to the resistance of community i to reach its own target hue. $H_{i,t}^{target}$ refers to the current target hue of community i. A_H denotes susceptibility and is defined in Equation 3.13 as:

$$A_H = e^{-\frac{\mu \times R}{S}}, \tag{3.13}$$

where S refers to saturation. Other parameters are the same as Equation 3.11.

A_S and A_H are community's susceptibility to influence on saturation and hue respectively. Susceptibility to influence on saturation (A_S) and hue (A_H) decreases with increasing

resources, because a community obtains greater success if the community acquires more resources, which in turn inhibits the strength of influences exerted by other communities. In addition, as saturation increases, a discipline becomes more susceptible to change. When the resource level is high and the discipline is saturated, members are more likely to experiment with new ideas.

3.5.4 Updating the Resource of a Community

During the learning process, communities purchase instruments, organize meetings and forums, or spend time on new materials etc. The amount of resources consumed is proportional to the degree of change in saturation and hue. Resource consumption for learning is defined as follows:

$$B_{i,t+1} = B_{i,t} - (C_H \times |\Delta H| + C_S \times |\Delta S|), \tag{3.14}$$

where $\Delta H = Dissimilarity(H_{i,t+1}^{current}, H_{i,t}^{current})$

$\Delta S = S_{i,t+1} - S_{i,t}$

$B_{i,t+1}$ refers to the resource of domain i at the next time. $B_{i,t}$ refers to the resource of domain i at the current time. C_H and C_S are constant numbers to convert changes of hue and saturation to resource respectively.

3.6 Innovation Process

Innovation changes the norms of the community i.e., target hue in the Colorscape model, because changing target hue is a strategy for a community to adapt to its environment. Moving target hue of a community toward its current hue can decrease resource consumption during the learning process, which in turn improves its sustainability. The distance between current and target hue is defined as flexibility that is an important requirement for innovation [90]. Therefore, a requirement for innovation is flexibility greater than a threshold as:

$$D_{ii,t} \geq T_{Innovation}, \tag{3.15}$$

where $D_{ii,t}$ is the distance between current hue and target hue. $T_{Innovation}$ is the tolerance.

In addition, there are two kinds of innovation patterns, one of which is reorganization, the other is specialization. Reorganization means that the community starts transforming itself by moving its accepted target toward the current state. On the other hand, specialization means branching out new communities. Whether reorganization or specialization happens is determined by a parameter called reorganization tendency. The innovation process is depicted in Figure 3.11.

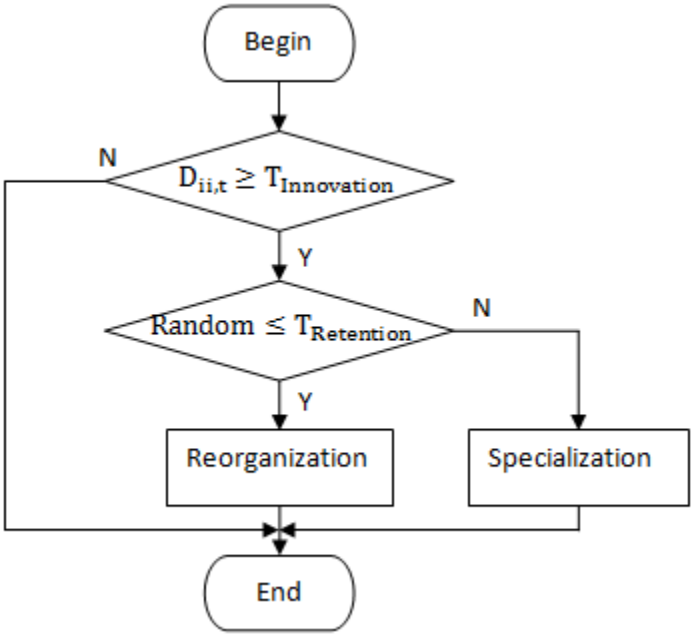


Figure 3.11: Flow Chart of Innovation

3.6.1 Reorganization

Reorganization process affects the hue of the target color, which is the weighted sum of target colors of the influential neighbors and resistance to change. Communities are

influenced by target colors of neighbors because the target of a community determines its future direction and can be seen as its vision. As shown in Figure 3.12, H_1^{target} is influenced by H_2^{target}, H_3^{target}, and $H_1^{current}$.

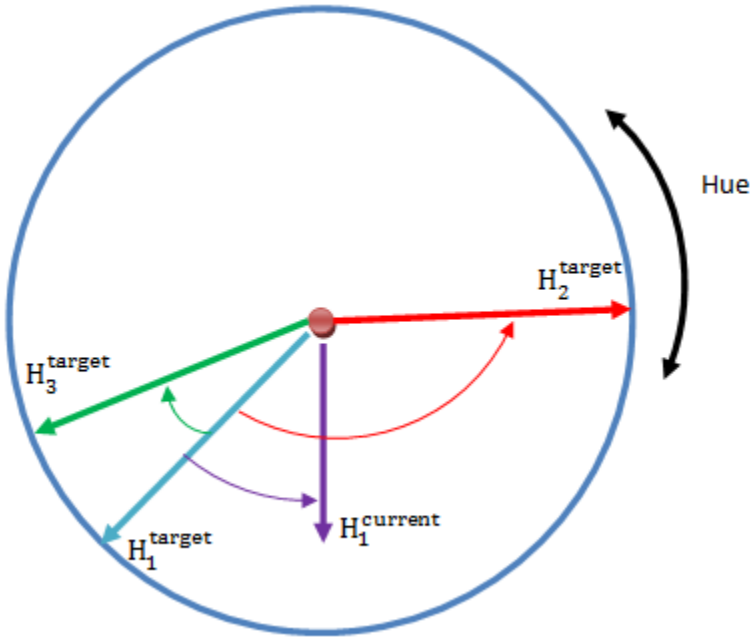


Figure 3.12: Updating the Domain during the Innovation Process

$$H_{i,t+1}^{target} = H_{i,t}^{target} + A_H \times \left(\left(\sum_{j=1}^{\#neighbors} W_{ji,t} \times (H_{j,t}^{target} - H_{i,t}^{target}) \right) + W_{i,t} \times (H_{i,t}^{current} - H_{i,t}^{target}) \right) \tag{3.16}$$

where $H_{i,t+1}^{target}$ refers to the new target hue after the reorganization process. $H_{i,t}^{target}$ refers to the current target hue of the community i. $H_{j,t}^{target}$ refers to the current target hue of the community j. $W_{ji,t}$ refers to the influence of community j on community i at the current time. $W_{i,t}$ refers to the resistance of community i to retain its own current hue. $H_{i,t}^{current}$ refers to the current hue of community i. A_H is susceptibility of community i and is defined in Equation 3.13.

Innovation may refer to incremental and emergent or radical and revolutionary changes in thinking, products, processes, or organizations [101]. Therefore, innovation requires additional resources. Resource consumption during innovation is defined as follows:

$$B_{i,t+1} = B_{i,t} - C_H \times |\Delta H|\,, \tag{3.17}$$

where $B_{i,t+1}$ refers to the brightness of domain i at the next time. $B_{i,t}$ refers to the brightness of domain i at the current time. ΔH refers to the changes of target hue of domain i. C_H is a constant value used to convert hue to resources needed for the innovation process, which is the same as C_H in equation 3.14.

3.6.2 Specialization

Specialization corresponds to the fact in the real world that new communities are split from the original community if the current community cannot match the expectations of all members. When specialization occurs, a new community is created. The new community occupies the nearest empty cell to the current community. If there are no empty cells, then specialization cannot happen. The underlying reason is the carrying capacity that is defined as the maximum number of communities that the current environment can sustain [72].

After the new community is created, the current color of the new community is the same as the original community. On the other hand, the target color of the new community is generated randomly within a range, as shown in Figure 3.13.

3.7 Grow and Fade

Following the innovation process, if the resource of a community cannot maintain its current state, then $R_{s,t}$ is decreased, and the processes of interaction, learning and innovation starts over. The iteration process continues until the remaining resources can maintain the current state or $R_{s,t}$ is equal to 0. When $R_{s,t}$ is equal to 0, the community fades and

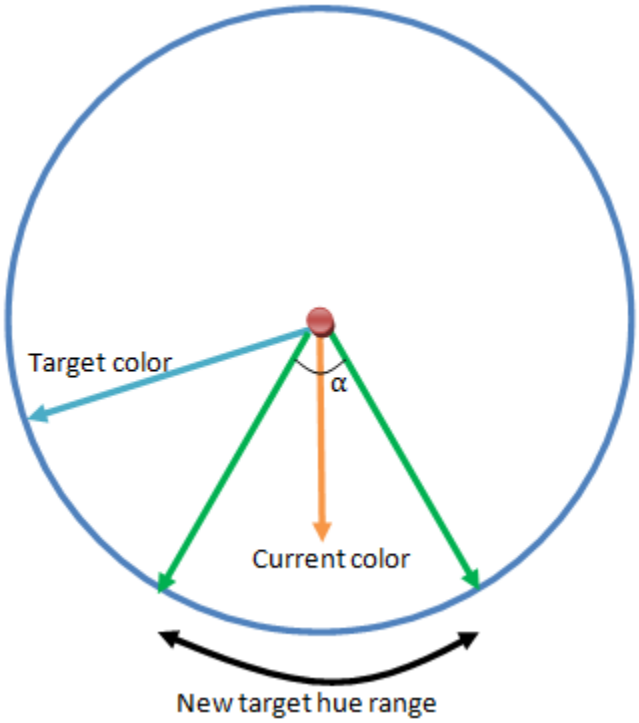


Figure 3.13: Specialization

is removed from the current context. On the other hand, if the community has enough resources to maintain and the neighbor cell is empty, then the community is likely to extend to occupy neighbor cells with a small probability. This captures evolutionary dynamics by retaining those communities that are fit to survive in the current environment.

3.8 Heterogeneous Adaption

Individual communities can adaptively change the weights of interconnections with other communities based on the environmental feedback so as to maximize their fitness [58]. The fitness refers to the resource the community gains. The more resources the community gains, the higher its fitness becomes. On the other hand, the fitness decreases if fewer resources are acquired. Furthermore, the weights of interconnection evolve along with the fitness. The equations for weights to change are as follows: two groups that correspond to weights of neighbors and weights of self respectively.

$$\begin{cases} W_{ji,t+1} = W_{ji,t} + \frac{W_{ji,t}}{W_{i,t}+\sum_{k=1}^{\#neighbors} W_{ki,t}} \times (1 - W_{ji,t}) & if\ fitness\ inceases \\ W_{ji,t+1} = W_{ji,t} - \frac{W_{ji,t}}{W_{i,t}+\sum_{k=1}^{\#neighbors} W_{ki,t}} \times W_{ji,t} & if\ fitness\ decreases \\ W_{ji,t+1} = W_{ji,t} & otherwise \end{cases} \tag{3.18}$$

$$\begin{cases} W_{i,t+1} = W_{i,t} + \frac{W_{i,t}}{W_{i,t}+\sum_{k=1}^{\#neighbors} W_{ki,t}} \times (1 - W_{i,t}) & if\ fitness\ inceases \\ W_{i,t+1} = W_{i,t} - \frac{W_{i,t}}{W_{i,t}+\sum_{k=1}^{\#neighbors} W_{ki,t}} \times W_{i,t} & if\ fitness\ decreases \\ W_{i,t+1} = W_{i,t} & otherwise \end{cases} \tag{3.19}$$

where $W_{ji,t}$ refers to the original interconnection weights of community j to community i. $W_{ij,t+1}$ refers to the new interconnection weights after feedback based on fitness. $\sum_{k=1}^{\#neighbors} W_{ki,t}$ refers to the sum of weights of all neighbors. $W_{i,t}$ refers to the tendency of community i to reach its own target hue. The range of weight is between 0 and 1. If fitness rises, the weight should increase toward 1 in proportion to the contributions of community j i.e., $W_{ji,t}/(W_{i,t} + \sum_{k=1}^{\#neighbors} W_{ki,t})$. On the contrary, if fitness falls, the weight should also decrease toward 0 in proportion to the contributions.

It is worth noting that the link between j and i is removed, if W_{ji} is smaller than a threshold.

3.8.1 Initialization

Table 3.1 describes all the state variables and their initial values in the simulation model. The significances of each variable are discussed in the following.

1. *Carrying capacity* (initial community number) refers to the size of the whole scientific society that is composed of single communities interconnected with each other. In biology, the term of minimum viable population [103] is the lower bound of population

Table 3.1: Initial Values of State Variables

Parameters Name	Range	Initial Value
Carrying Capacity	[10, 200]	100
Stop Time	$[1, \infty)$	1000
Startup Funding	[1, 2]	2
Parameter $F_{i,t}$ in equation 3.1	[0.1, 1]	0.5
Tolerance	[0, 1]	0.2
Reorganization Tendency	[0, 1]	0.5
Parameter c in equation 3.2	[0, 1]	0.5
Parameter R_{max} in equation 3.2	[0, 1]	0.9
Proportion of resources to advance maturity	[0, 1]	Random
Max Increment of Saturation Per Step (S_{max} in equation 3.4)	[0.1, 1]	0.5
Min Increment of Saturation Per Step (S_{min} in equation 3.4)	[0, 1]	0.1
τ in equation 3.4	$(0, \infty)$	100
C_W in equation 3.6	[0, 1]	0.5
μ in equation 3.13	$(0, \infty)$	3
Resources Cost to Push Hue (C_H in equation 3.17)	[0, 1]	1
Resources Cost to Push Saturation (C_S in equation 3.17)	[0, 1]	1
Current color	HSB range	Random
Target color	HSB range	Random
Initial weight of Self	[0, 1]	Random
Initial weight of neighbor	[0, 1]	Random
Weight to grow	[0, 1]	Random

of species so that it can survive. So, it is expected that the initial community number is related to diversity and resilience.

2. *Startup Funding* (parameter $F_{i,t}$ in equation 3.1) indicates the external funding allocated to the community. Research funding could be structured to encourage the formation of new communities [91]. Also research funding has effects on the developments of existing communities.

3. *Tolerance* (threshold for innovation to happen) and *reorganization tendency* determine innovation occurrence frequency and which type of innovation occurs. Since innovation changes the norms of the community, it is of interest to investigate the relationship between the type of innovation and diversity.

4. *Parameter c and R_{max} in equation 3.2* determine the form of the maintenance function that in turn decides whether or not the community could fade out.

5. *S_{max}, S_{min} and τ in equation 3.4* determine how much maturity can be gained per time tick. On the other hand, the more maturity a community gains, the more resources the community consumes, which increases the likelihood of fading of the community. So S_{max}, S_{min} and τ are important parameters when total available resources are limited.

6. *C_W in equation 3.6* determines changes of weights of links that in turn determine influences of peer communities during the process of learning and innovation.

7. *μ in equation 3.13* determines the slope of the curve of susceptibility of a community. It determines the extent to which the community can be changed further.

8. *C_H and C_S in equation 3.17* determine the relationship between domain change intensity, maturity, and corresponding resources consumption. So, these two parameters are expected to affect the rate of fading.

Chapter 4

Implementation of Simulation Model

4.1 Introduction to Repast

Repast is an acronym for the Recursive Porous Agent Simulation Toolkit [77] that is a free and open source agent-based modeling toolkit that simplifies model creation and use. Repast Simphony provides a rich variety of features including the following:

- The model development can use pure Java, Groovy, flowcharts, and any mixture of them.
- A pure Java model execution environment includes built-in results logging and graphing tools that make it easy to change the appearances of agents.
- The context is based on a flexible hierarchy that can realize the modeling and visualization of 2D environments and 3D environments.
- The discrete event scheduler is fully concurrent multithreaded.
- All the models developed by Repast are object-oriented.

In general, the standard model using Repast is based on contexts and projections. There are some frequently used projections including grid, continuous space, network, and geography. Figure 4.1 shows how context, sub context, and projection interact.

4.2 Implementation of Agents

Figure 4.2 represents part of the class diagram of this simulation model of open science communities, where there are four main classes: Community, SubCommunity, Neighbor and

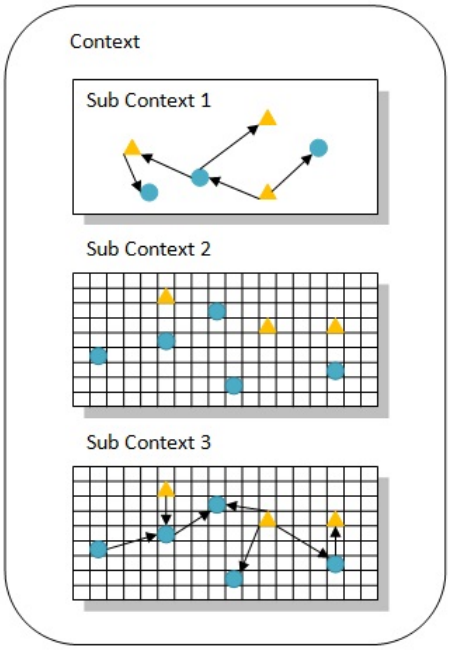


Figure 4.1: Contexts and Projections

CommunityStyle. Community is the major research object in the simulation model, which has two types of state variables and three functions. SubCommunity is used when a community occupies multiple cells. A community is comprised of one or more SubCommunities. The CommunityStyle class is used to render Community and SubCommunity, so as to show the correct color according to the states of domain. The Neighbor class represents communities connected with the current community.

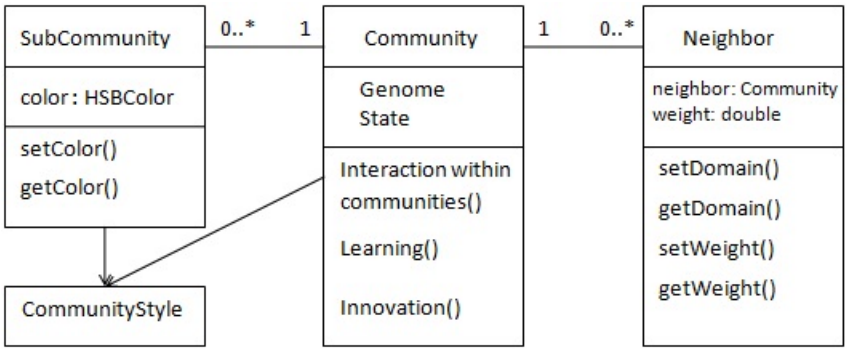
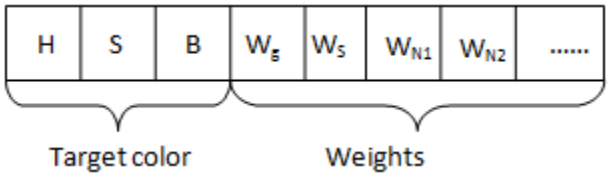


Figure 4.2: Class Diagram of Model

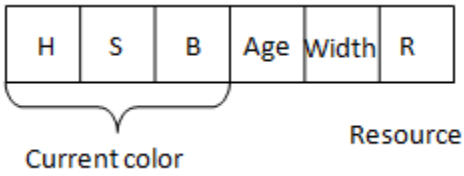
A community is represented by its genome and state whose details are described as follows:

- Genome



  Target color is (H, S, B) = (H, 1, 1), where H refers to discipline of community whose range is [0, 360). W_g is the probability for a community to grow i.e., occupy neighbor places. W_S is propensity of community to move toward the target. W_{Nk} denotes the influence exhibited on the community by the k_{th} neighbor

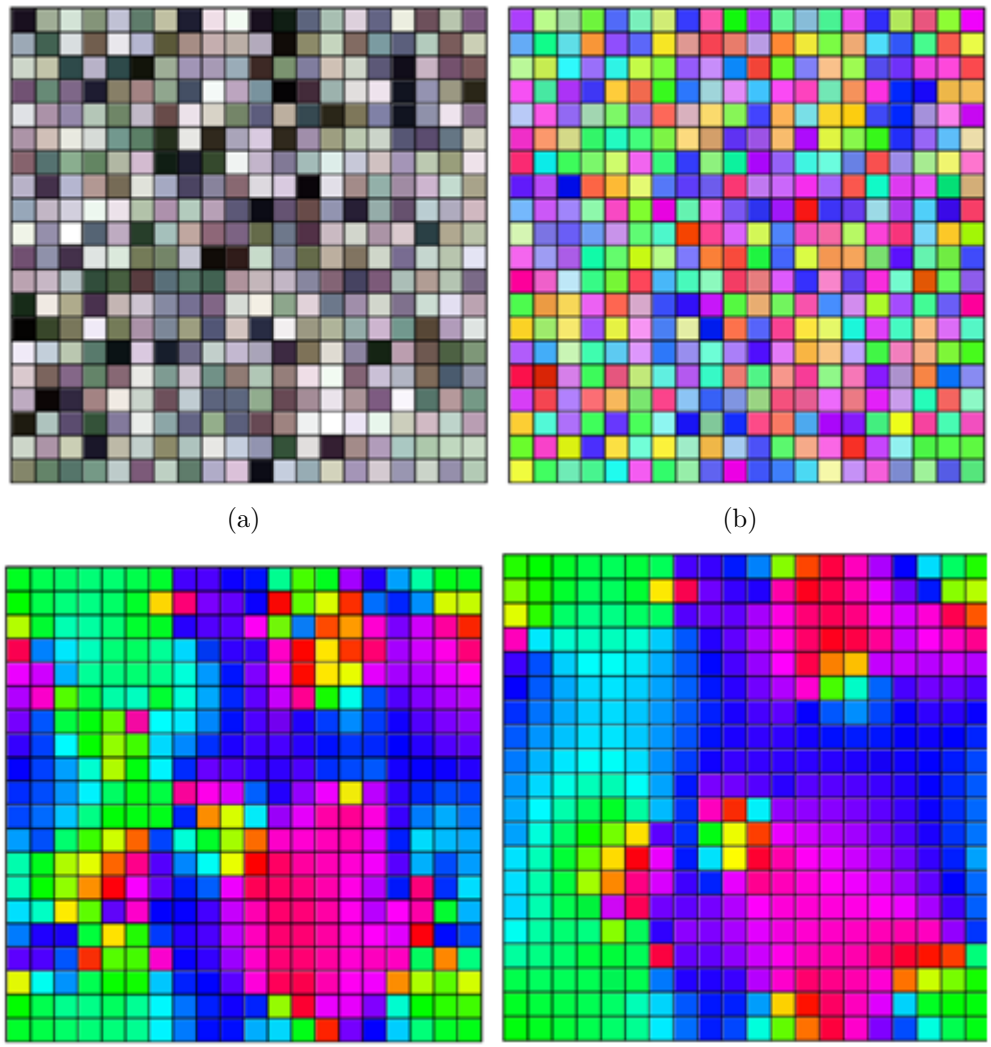
- State



  H, S, and B of current color represent domain, maturity, and resource respectively. Age refers to the time period when the community exists in the context. Width is the number of cells the community occupies. The resource level R allocated at the current time is different from B, which represents the overall resources held by the community.

4.3 Visual Snapshots of the Simulation View

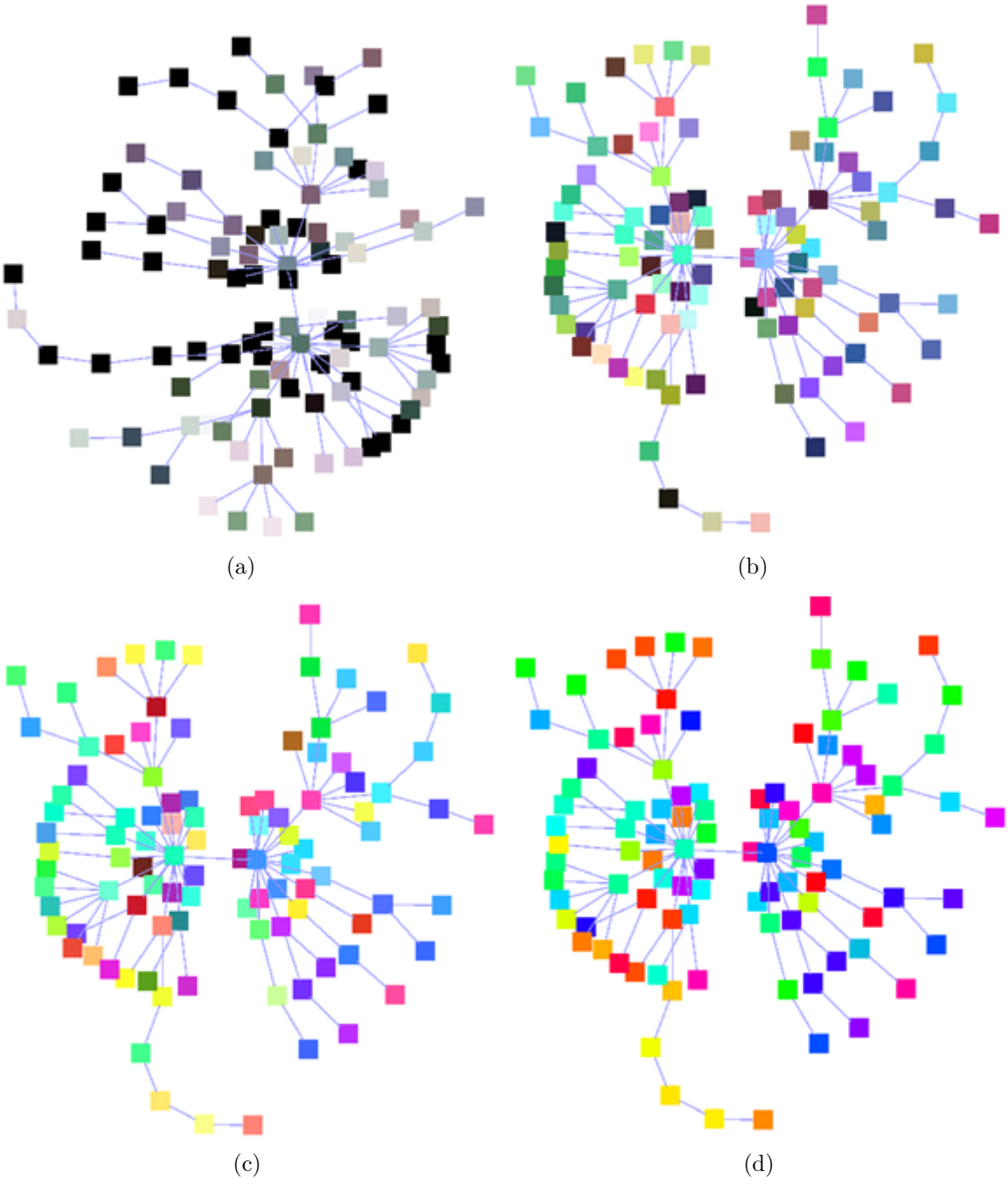
The following three groups of figures depict snapshots of the Colorscape model over time with 2D, scale-free, and dynamic communication context respectively. Among these figures, there are two points in common, one of which is that communities become more colorful as the result of increasing maturity. The other is that clusters of similar communities emerge as the result of boundary processes among communities.



(a) (b)

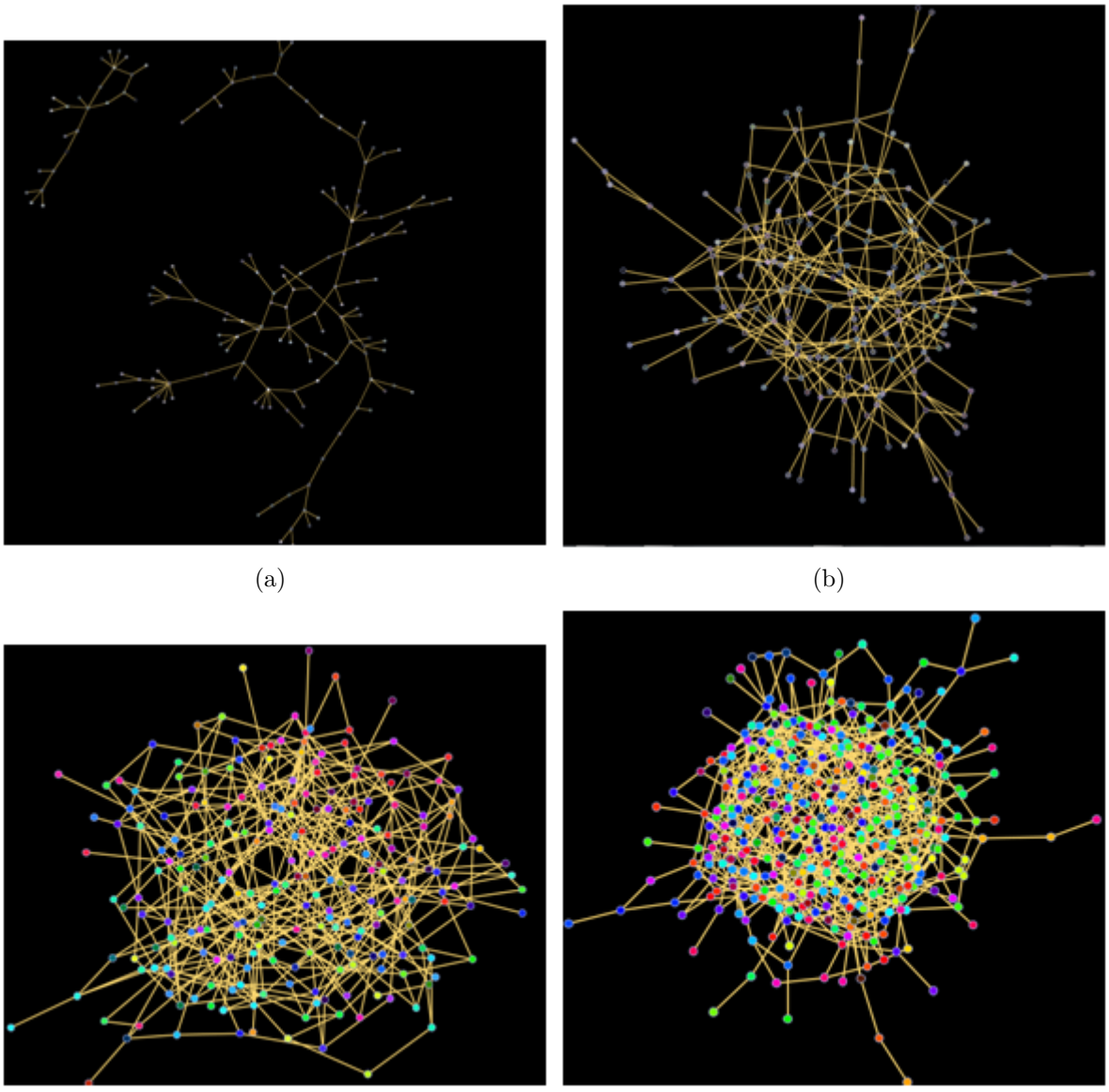
(c) (d)

Figure 4.3: Snapshots of 2D Communication Context



(a) (b) (c) (d)

Figure 4.4: Snapshots of Scale-free Communication Context



(a) (b)

(c) (d)

Figure 4.5: Snapshots of Dynamic Communication Context

Chapter 5

Verification, Validation and Evaluation

The evaluation of a simulation model involves two major activities, one is verification, and the other is validation that includes conceptual and operational validation [9]. Conceptual validation aims to assure that the conceptual model is consistent with the system under investigation [79]. Operational validation substantiates the accuracy of model's behavior against the system behavior for its intended purpose and domain of applicability. [79].

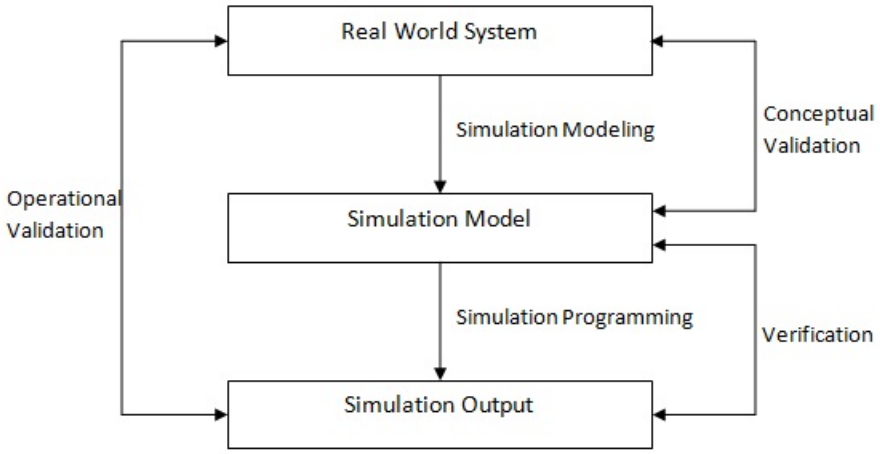


Figure 5.1: Overview of Verification and Validation [92]

Also, verification and validation can be conducted at the micro and macro level respectively. Thus, the strategies for different levels of verification and validation are summarized as shown in Table 5.1.

5.1 Verification

Verification is the process of determining that a computer model, simulation, or federation of models and simulation code and their associated data accurately represent the

Table 5.1: Verification and Validation at Micro and Macro Level

	Verification	**Conceptual Validation**	**Operational Validation**
Micro	Unit test at the level of single function	Ontological Adequacy: ground each equation on theory	Activity of single agent against theory
			Sensitivity analysis with respect to single agent
Macro	Integration test at the level of components of agent	Conceptual validity against theory	Activities of set of agents against theory
			Activities of set of agents against empirical evidence
		Conceptual validity determined by experts	Cross-model validation
			Sensitivity analysis with respect to set of agents

developer's conceptual description and specifications [1]. To achieve this goal, unit and integration tests are used.

5.1.1 Micro Verification

Micro verification is carried out by unit test at the level of single function. Unit testing involves determining the correctness of the simulation program at the function level. All functions are tested using boundary and error conditions, and the outputs are observed for consistency against expected regularities. For demonstration purposes, the following example illustrates testing of the resource allocation module. The resource allocation strategy in the baseline model is used to allocate resources evenly across the communities. The total amount of resources each community receives is equal to resources allocated to each cell times the number of cells the community occupies plus the contributions to transfer technology.

1. When communities make no contributions and one cell per community:

Resources allocated per cell	Resources received by each community
0	0
0.5	0.5
1	1

2. When communities make no contributions and two cells per community:

Resources allocated per cell	Resources received by each community
0	0
0.5	1
1	2

3. When communities make contributions 1 and one cell per community:

Resources allocated per cell	Resources received by each community
0	1
0.5	1.5
1	2

4. When communities make contributions 1 and two cells per community:

Resources allocated per cell	Resources received by each community
0	1
0.5	2
1	3

5.1.2 Macro Verification

Macro verification is carried out by integration tests at the level of collectives of agents such as interaction, learning, reorganization, specialization, fade, and growth etc. Integration testing is the activity of software testing in which individual software modules are combined and tested as a group [102]. For our simulation model, we focus on the behavior of the *Community* class since community behavior is the focal aspect of our study.

The learning process is composed of several functions such as updating influences of neighbors, calculating changes of hue, saturation, and brightness etc. Table 5.2 records the precondition and expected values (the column Next Color) of the integration test for learning process.

Table 5.2: Summary of the Integration Test for the Learning Process

Community 1			Neighbor	Community 1
Current Color	Target Color	Receptivity	Current Color	Next Color
(0,0,0)	(0,1,1)	1	(180,1,1)	(180,0,0)
(0,1,0)	(0,1,1)	1	(180,1,1)	(180,0,0)
(0,1,0)	(90,1,1)	1	(180,1,1)	(180,0,0)
(0,1,1)	(0,1,1)	1	(180,1,1)	(8.96,0.95,0.9)
(0,0,0)	(0,1,1)	0	(180,1,1)	(0,0,0)
(0,1,1)	(0,1,1)	0	(180,1,1)	(0,1,1)
(0,0,0)	(0,1,1)	0	(90,1,1)	(0,0,0)
(0,0,0)	(180,1,1)	0	(90,1,1)	(180,0,0)

In the above table, the first three rows show that the current color of a community with receptivity of 1 will be changed to the same as the current color of interacted neighbors after the learning process, no matter what the community's target color is. The difference of the fourth row from the first three rows is that the quotient of the community's resource divided by saturation is equal to 1, which in turn makes its susceptibility (Equation 3.13) greater than 0. So, the current color under the case of the fourth row is the result computed by Equation 3.12. The last four rows illustrate that the current color of a community with receptivity of 0 is always pulled toward its target color, which is independent of influences from neighbors.

5.2 Validation

There are two major types of validation: conceptual validation and operational validation.

5.2.1 Conceptual Validation

The conceptual validation refers to ontological adequacy by grounding the underlying generative mechanisms of the model on theories and/or empirical evidence. Table 5.3 lists the evidence used to validate each subprocess of the Colorscape model.

Table 5.3: Summary of Conceptual Validation of Each Subprocess

Subprocess	**Evidence**
Interaction between environment and community	Prey-predator models[46]
	Observed trends in NSF investments [60]
Relationship between maturity and resources	U-shaped model of the knowledge maturing process [78]
	Kuhn's paradigm change theory [53]
Updating intensity of communities' influences	Homophily theory [15]
Domain dynamics of a community	Boundary processes [96], Social learning theory [88].
Maturity dynamics of a community	The formation process of DNA computing [5].
Community reorganization	Panarchy theory[37]
Community specialization	Panarchy theory[37]
Fading process dynamics	Panarchy theory[37]
Community growth dynamics	Panarchy theory[37]

Panarchy is the structure in which systems of nature and systems of humans, as well as combined human-natural systems are interlinked in continual adaptive cycles of growth, accumulation, restructuring, and renewal [37]. Therefore, it is used in this study to validate the subprocesses of reorganization, specialization, grow, and fade.

5.2.2 Micro Operational Validation

As far as the micro operational validation is considered, we compare the behavior of a single agent to the expected regularities. The followings illustrate three micro operation validation strategies used for the Colorscape model.

1. According to the Homophily theory, the intensity of influences from similar peers is greater than that from different peers. So, one way to conduct micro operational validation is to investigate the impact of differences between weights of influences of neighbors associated with an agent.

2. The other strategy is to undertake sensitivity analysis with respect to single agents. For example, an agent with fewer resources is more likely to fade than agents with more resources. In addition, an agent with higher receptivity has stronger intention to change its domain toward its neighbors, compared with agents having lower receptivity.

3. When an unexpected phenomenon occurs, we need trace it back to the internal mechanisms of the model by viewing it as a white box. If the rationale behind the unexpected phenomenon is found and it matches either existing theories or empirical rules, then the model is validated with respect to the case.

5.2.3 Macro Operational Validation

For macro operational validation, we focus on the global emergent behavior based on agent interactions i.e., external validation against real world. There are various strategies for macro operational validation, such as comparison of simulation outputs to target systems, empirical rules, cross-model validation, and sensitivity analysis [80]. Firstly, validation can be conducted via comparison of simulation model outputs and the actual data collected from the system under investigation. If data are not available, empirical rules can be used to determine the validity of the model, e.g., presence of power law in cities' population, financial market, and internet sites [11]. In addition, we can evaluate impacts of a specific parameter by changing its value but keeping others unchanged. If the result is consistent with the expected regularities, then we increase our confidence about the correctness of the simulation model under this specific case. Finally, unexpected phenomena could emerge since the Colorscape model aims to study complex systems. When such unexpected phenomena

occur, it is necessary to trace back to the model and check the control and data flows step by step. If the unexpected phenomenon can be interpreted reasonably, then the model is validated under this phenomenon.

5.2.3.1 Emergence of Communities

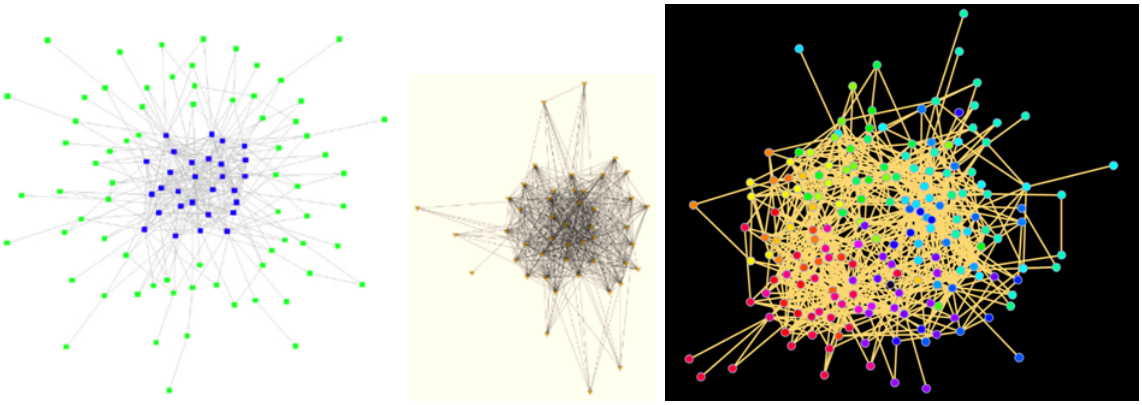
Figure 5.2 presents evolving states of communities over time during a single run of the Colorscape model. Initially, the colors of communities are grey due to their low maturity. As the simulation unfolds, states of communities become increasingly colorful due to increasing maturity through community sustainment, interaction, learning, and innovation processes. After a long run, clusters with similar color patterns emerge, which suggests formation of related disciplines as a result of communication and boundary processes.

Figure 5.2: Growth and Formation of Community Clusters

Figure 5.3(a) presents the ideal core/periphery network pattern. Figure 5.3(b) depicts the domain-domain network pattern of OBO community. Figure 5.3(c) is a snapshot of the network of the Colorscape model. From the visual comparison of these three figures, one can observe similar network structures such as the presence of core communities with a large number of links surrounded by periphery communities.

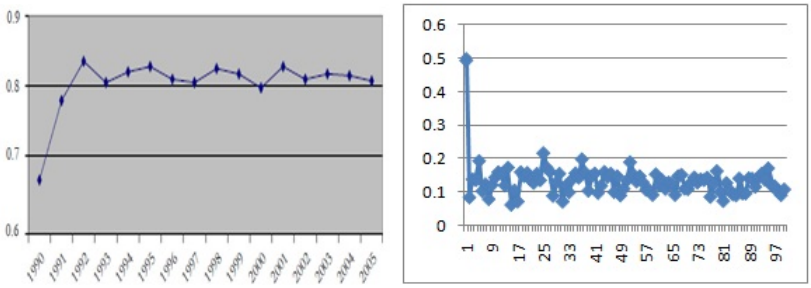
5.2.3.2 Comparison with Institutions around Department of Energy

Figure 5.4(a) depicts how the clustering coefficient of DOE in nanoscale science changes from 1990 to 2005 [60]. Figure 5.4(b) depicts the clustering coefficient gathered by running the Colorscape model from time step 1 to 100. From these two figures, we observe very similar trends i.e., clustering coefficient oscillates within a limited range.



(a) Core/Periphery Network (b) OBO Domain-Domain Network (c) Snapshot of the Colorscape Model

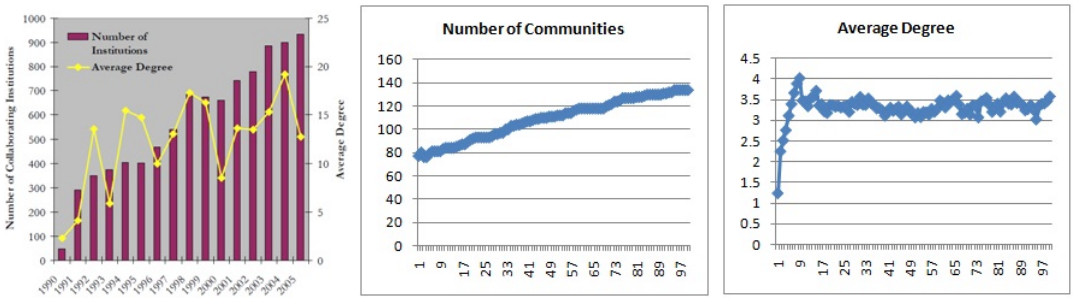
Figure 5.3: Emergent Network Patterns



(a) (b)

Figure 5.4: Comparison of Clustering Coefficient

Figure 5.5(a) shows the number of institutions and average degree of institutions of collaborations in nanoscale science with DOE from 1990 to 2005 [60]. Figure 5.5(b) and 5.5(c) present the number of communities and average degree respectively gathered via running the Colorscape model from time step 1 to 100. From these three figures, we observe very similar trends, i.e., the number of communities increases gradually and the average degree fluctuates within a limited range. This increases our confidence in the Colorscape model introduced in this dissertation, because of its capability of generating similar network patterns and metric outputs to corresponding indicators such as institutional structure involved in nanoscale science in Department of Energy (DOE).



(a) (b) (c)

Figure 5.5: Comparison of Communities Number and Average Degree

5.3 A Robust Evolutionary Framework for Validation

In this section, a robust evolutionary framework for validation based on genetic algorithm is introduced to find the appropriate configuration parameters of the Colorscape model to produce results similar to the overlay map and OBO data in terms of the number of nodes, density, centrality, clustering coefficient, average path, and core/periphery ratio.

5.3.1 Design of the Validation Framework

The strategy used in the operational validation of the ColorScape model is shown in Figure 5.6. Each step will be discussed in detail in the following sections.

5.3.2 Gene Encoding

Gene Design has two subprocesses, one of which is gene encoding. The other is gene decoding.

Gene encoding refers to the process of converting the configuration parameters to genes that evolve toward parameter space that exhibits accurate results with respect to system data. In general, genes are presented in binary strings, where each element is 0 or 1. It is essential to determine how many bits are needed to represent a configuration parameter, which in turn is determined by the value range and the degree of precision needed. If the configuration parameter is integral, then the binary presentation of the integer is used as

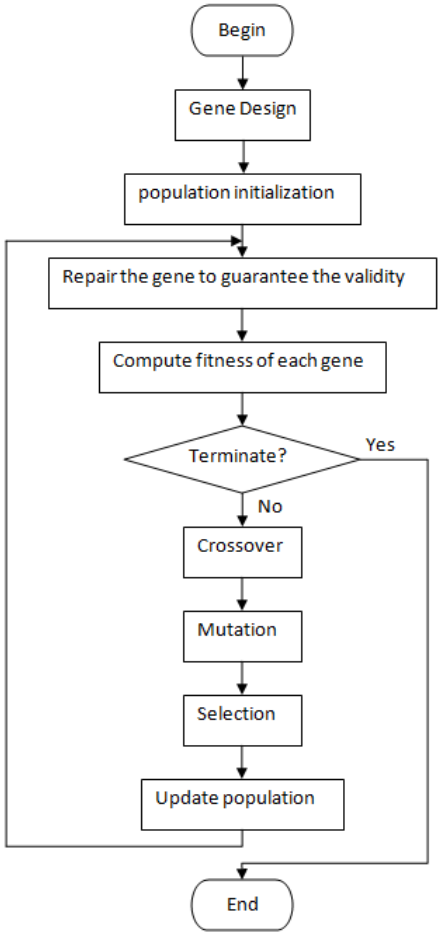


Figure 5.6: Validation Framework

the gene. If the configuration parameter is float, the degree of precision must be set up in advance so that the binary presentation of the float number satisfies the requirements. If the configuration parameter has a fixed amount of feasible values, then the number of bits of the corresponding gene is determined by the total number of feasible values. The final gene is a string that consists of all the configuration parameters. As shown in Figure 5.7, we select two parameters from a collection as an example, one of which is integer and ranges between 0 and 6. The other is float and the range is from 0 to 1. For the integral parameter m, three bits are required since the maximum value is 6. For the float parameter with value equal to 0.75, there are four possible values since the precision is set to 0.25. So two bits are required to represent parameter n.

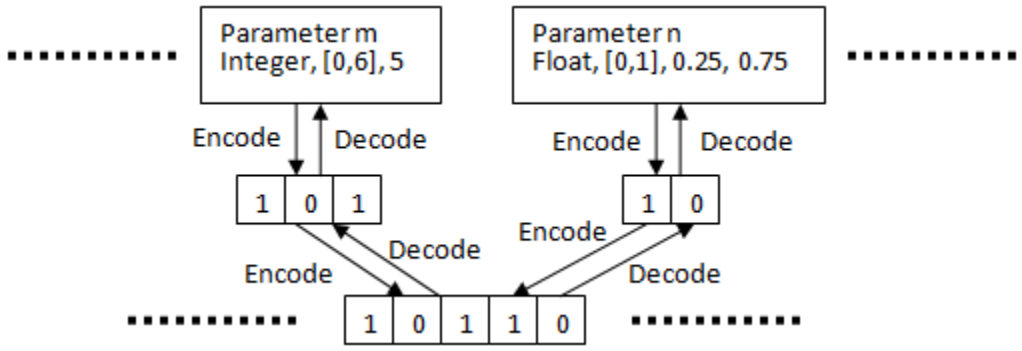


Figure 5.7: Gene Encoding

Since there are twelve allocation strategies, four bits are needed to represent allocation strategies. Based on the same rationale, the number of bits of other configuration parameters is determined. Figure 5.8 shows a sample of gene after encoding the configuration parameters of the Colorscape model, where the total number of bits is 21.

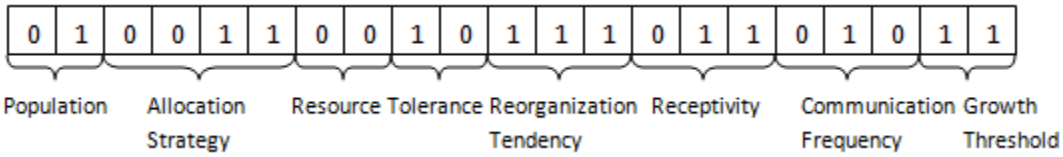


Figure 5.8: Gene Example

5.3.3 Gene Decoding

As an inverse process of gene encoding, gene decoding aims dividing the gene into parts corresponding configuration parameters.

To calculate the fitness of each gene, the gene has to be translated (decode) into configuration parameters of the Colorscape. Then the Colorscape model is batch run given the parameters and return the outputs.

The total number of bits of the gene is 21, among which different bits have different meaning. The following table interprets the relationship between bits and the corresponding meanings.

Table 5.4: Gene Decoding

Bits	Variable Name	Code	Value
0-1	Population	00	10
		01	50
		10	100
		11	200
2-5	Allocation Strategy	0000	Uniform allocation with fixed external resource
		0001	Uniform allocation with technology transferring
		0010	Allocation proportional to contribution with fixed external resource
		0011	Allocation proportional to contribution with technology transferring
		0100	Allocation proportional to cluster size with fixed external resource
		0101	Allocation proportional to cluster size with tech transferring
		0110	Allocation proportional to importance of domains with fixed external resource
		0111	Allocation proportional to importance of domains with technology transferring
		1000	Competition allocation
		1001	P2PAllocation
		1010	Random allocation with fixed external resource
		1011	Random allocation with technology transferring
6-7	Resource	00	0.1

Continued on next page

Table 5.4 – continued from previous page

Bits	Variable Name	Code	Value
		01	0.4
		10	0.7
		11	1.0
8-9	Tolerance	00	0
		01	0.3
		10	0.6
		11	1.0
10-12	Reorganization Tendency	000	0
		001	0.1
		010	0.3
		011	0.5
		100	0.7
		101	0.9
		110	1.0
13-15	Receptivity	000	0
		001	0.1
		010	0.3
		011	0.5
		100	0.7
		101	0.9
		110	1.0
16-18	Communication Frequency	000	0.1
		001	0.2
		010	0.3

Continued on next page

Table 5.4 – continued from previous page

Bits	Variable Name	Code	Value
		011	0.4
		100	0.5
		101	0.6
		110	0.8
		111	1.0
19-20	Growth Threshold	00	0.5
		01	0.7
		10	0.8
		11	1.0

There is one noteworthy aspect, that is, specific values of bits may not have real meanings, for instance, receptivity of 111. If it happens, a recovery strategy is required. The recovery strategy used here is to mod the old value by the maximum practical value. For the receptivity of 111, the new receptivity is equal to 111 % 7 = 000.

5.3.4 Population Initialization

Population initialization involves generating a collection of genes with the predefined total number. Each bit of a gene is assigned randomly as 0 or 1. Once the total number of genes is assigned, the population is generated automatically. The size of the population used in the validation framework is 100.

5.3.5 Repair to the Genes

During the evolution, the generated new genes may be out of the feasible range. In this case, a repair is needed to guarantee the validity of the gene. Considering the parameter m

shown in Figure 5.7, if the generated value is 111, it is beyond the feasible range since its maximum value is 6. Two strategies can be used to make the repair. One is to mod the new value by the maximum value, i.e., 7%6 = 1. Then the repaired gene is 001. The other strategy is to randomly map it into the feasible domain.

5.3.6 The Fitness Function

In biology, natural selection is the process of eliminating members of a species that do not adapt to the environment well. In genetic algorithm, the selection of genes is based on their fitness that is the indicator showing how close the gene's outputs are against the target. To quantitatively validate a simulation model, metrics have to be chosen and the corresponding values of these metrics with respect to the real system are computed. The values of metrics of the real system are the target. The fitness of each gene is inversely proportional to the distance between its outputs and the target metrics, which is defined by Equation 5.1:

$$f(g) = \frac{1}{\sqrt{\sum_{i=1}^{6}(x_i - t_i)^2}}, \tag{5.1}$$

where g is the gene. x_i is the i_{th} element of output metric vector given the gene. t_i is the i_{th} element of target metric vector. The metric vector includes six elements, each of which corresponds to a metric: the number of nodes, density, centrality, clustering coefficient, average path, and core/periphery ratio.

As a general measure of the degree of socio-technical interaction, we use and interpret density, centrality, clustering coefficient, average path length and core/periphery ratio so as to identify the target networks including OBO and overlay map. Except the core/periphery ratio, the definition of other metrics and their relation to creativity and innovation potential are presented in [113]. The core/periphery network pattern is considered as a stable, sustainable, and innovative structure [52]. Given the same number of core members, increasing level of periphery members is beneficial for bringing new external ideas. The core/periphery

ratio is used to measure the percentage of the members in the core to the members in the periphery. The algorithm shown in Figure 5.9 describes the strategy used to compute the core/periphery ratio.

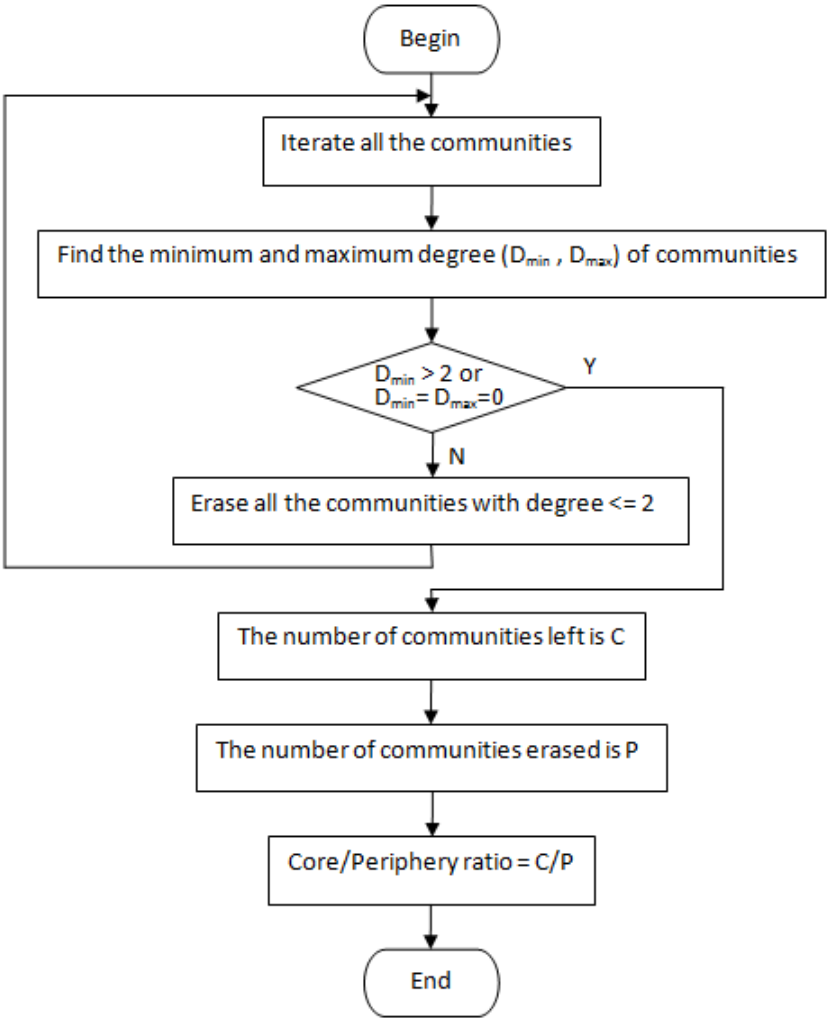


Figure 5.9: Core/Periphery Ratio

5.3.7 Termination Condition

There are two conditions to terminate the validation process:

1. The maximum iteration times are reached.

2. A gene with fitness greater than a predefined threshold emerges.

5.3.8 Crossover

Reproduction is the process of generating the next generation of genes, where two genetic operators are used, i.e., crossover and mutation. For crossover, there are two options: one-point crossover and two-point crossover. One-point crossover means that two genes exchange the parts beginning at the randomly selected cross point. Two-point crossover is defined as that two genes exchange the part between the first and the second cross point.

5.3.9 Mutation

Crossover is a binary operator. On the other hand, mutation is a unary operator. There is a very small probability of mutation, i.e., 1%. Iterate all genes, if a randomly generated number is less than the mutation probability, then mutation happens. When mutation happens, our strategy randomly chooses a mutation point and then flips the bit.

5.3.10 Selection

Selection is the process of updating population in terms of a fitness-based function. The higher the fitness of a gene is, the more likely the gene is selected. In the population including both parents and children, a fixed number of genes are selected as the next generation. For each gene, the probability for it to be selected is based on its fitness:

$$pp_i = \sum_{j=1}^{i} p_j, \tag{5.2}$$

$$p_j = \frac{f_j}{\sum_{k=1}^{N} f_k}, \tag{5.3}$$

where p_j is the probability for gene i to be selected. pp_i is the accumulated probability. f_j is the fitness of gene j. Only if $pp_i \geq rand(0,1) \geq pp_{i-1}$, the gene i is selected.

5.3.11 Equilibrium

During the evolution process, if the population does not change, then this is an indication that equilibrium is reached. The potential reason is that the population converges to a local optimal point. To break the equilibrium and continue the evolution, a mechanism named kick the ball is used. Vividly imagining the whole searching range as a mountain, kicking the ball aims to transfer a solution from one valley to another, where it may evolve to be a better solution. In the genetic algorithm, when the population of genes does not change, a part of population are randomly selected and their bits are randomly mutated. Compared with the mutation operator, kicking the ball flips multiple bits one time rather than just one bit.

5.3.12 Implementation

The simulation model and the associated validation framework are implemented by RePast. RePast is an open source software that facilitates design and implementation of agent-based models. It provides mechanisms for both single and batch run. However, the configuration parameters cannot be changed during either single or batch run. So, a new runner that inherits the default runner of RePast is necessary to dynamically translate the gene to configuration parameters and return the outputs that are required for the computation of fitness.

As shown in Figure 5.10, the abstract GA class implements all other functions except the fitness function. The fitness function is overridden by child classes that drive the simulation model after converting the gene to corresponding configuration parameters. Figure 5.11 is the sequence diagram that captures the dynamic strategy used in the validation framework. The main class invokes the genetic algorithm which in turn invokes the simulation model that derives its fitness value. Because the genetic algorithm may evolve over multiple generations, there is a loop for the genetic algorithm until the termination condition is reached.

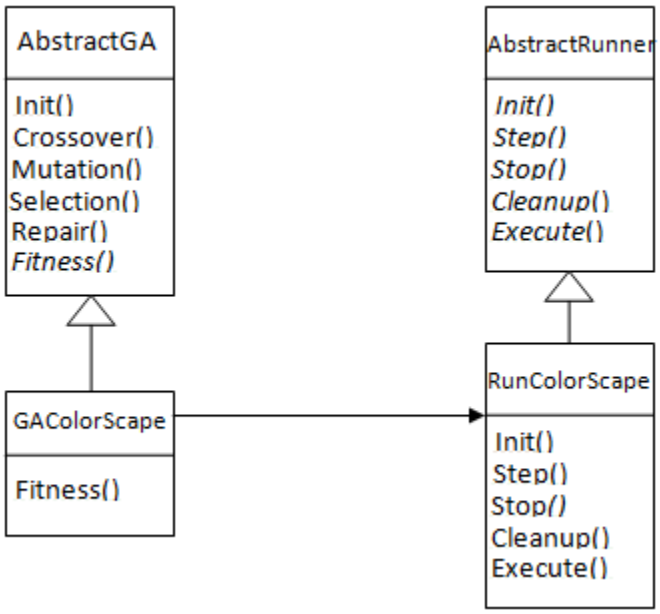


Figure 5.10: Class Diagram of Validation Framework

5.4 Comparison with Overlay Map

Overlay map [75] is a novel tool that presents relationships among disciplines based on citation data.

After 100 generations, the best gene is discovered. Table 5.5 lists all the parameters and their values represented by the gene:

Table 5.5: The Best Configuration against Overlay Map

Name	Value
Carrying Capacity	50
Startup Funding	2
External Resource	1
Tolerance	0.6
Reorganization Tendency	0.1
Receptivity	1
Allocation Strategy	Proportional to Contribution with Technology Transferring
Communication Style	Homophily
Communication Frequency	0.6
Threshold to Grow	0.5

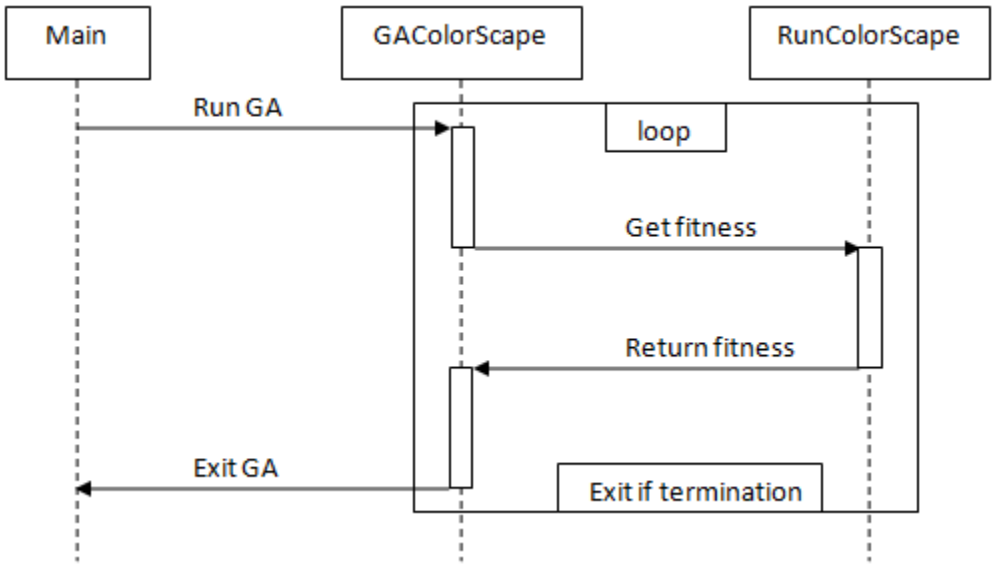


Figure 5.11: Sequence Diagram of Validation Framework

Figure 5.13 is the snapshot of the network generated by the Colorscape model given the configuration parameters shown in Table 5.5.

By comparison, the similarities of Figure 5.12 and Figure 5.13 can be observed as follows:

1. The development levels of communities (fields) are different from each other.

2. Those communities with similar states form clusters.

3. Some communities have more links than others.

The above is the intuitive comparison of network patterns. To gain more confidence, a quantitative comparison is undertaken in terms of six metrics: number of nodes, density, centrality, clustering coefficient, average path, and core/periphery ratio. Table 5.6 presents the comparison of the network metrics generated by the Colorscape model against the corresponding metrics of the overlay map (expected values in the table). Although the confidence intervals of metrics derived from the simulation data do not always contain the corresponding values of overlay map, we can still observe that they are significantly close. In addition, if we reduce the number of target metrics, the Colorscape model is able to generate outputs with confidence intervals containing the expected values.

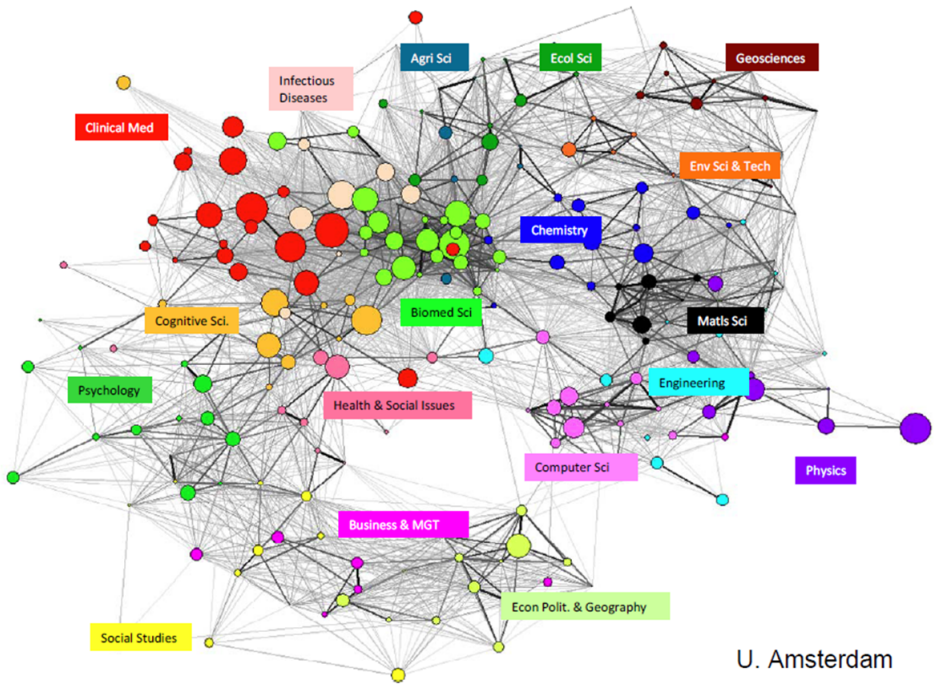


Figure 5.12: Overlay Map [75]

Table 5.6: Simulation Output vs. Overlay Map

Name	Mean Value	Standard Deviation	Confidence Interval of 90%	Expected Values
Number of Nodes	296	79.581	[272, 320]	222
Density	0.214	0.068	[0.194, 0.233]	0.139
Clustering Coefficient	0.574	0.050	[0.559, 0.589]	0.648
Centrality	0.354	0.040	[0.342, 0.366]	0.216
Average Path	1.847	0.086	[1.821, 1.873]	2.415
Core/Periphery Ratio	77.283	36.583	[65.482, 89.084]	73.0

5.5 Comparison with the OBO Domain-Domain Data

Figure 5.14 depicts the actual OBO network, where the nodes with the same color belong to the same group. Because all the groups are the branches of biology, we consider them as a single domain.

After 100 generations, most fit gene is found. Table 5.7 lists all the parameters and their values represented by the gene.

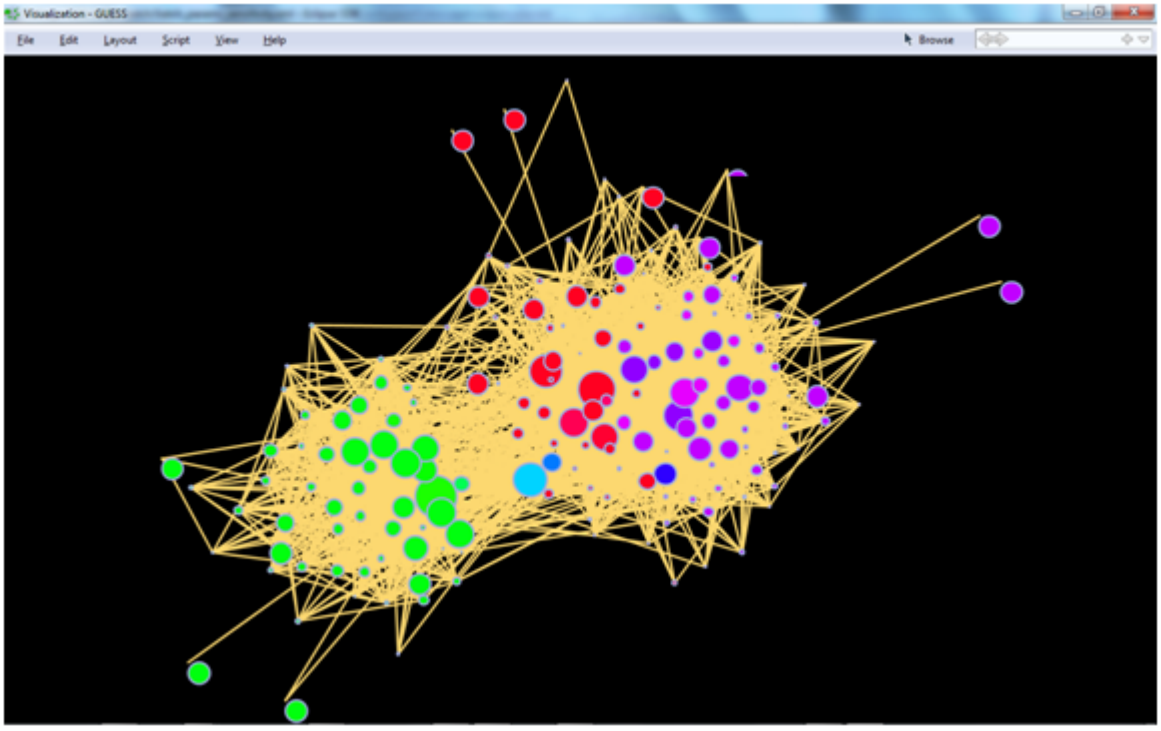


Figure 5.13: Snapshot of the Colorscape Model against Overlay Map

Table 5.7: The Best Configuration against OBO Data

Name	Value
Carrying Capacity	30
Startup Funding	2
External Resource	1
Tolerance	0.6
Reorganization Tendency	0.5
Receptivity	0.9
Allocation Strategy	Uniform allocation with technology transferring
Communication Style	Homophily
Communication Frequency	1.0
Threshold to Grow	0.5

Because the Colorscape model studies the relationships between communities with similar or different domains, all the communities have to be categorized into domains based on their color so as to compare to the OBO network. To illustrate the process, let us observe Figure 5.15 that is a snapshot given the above configuration parameters.

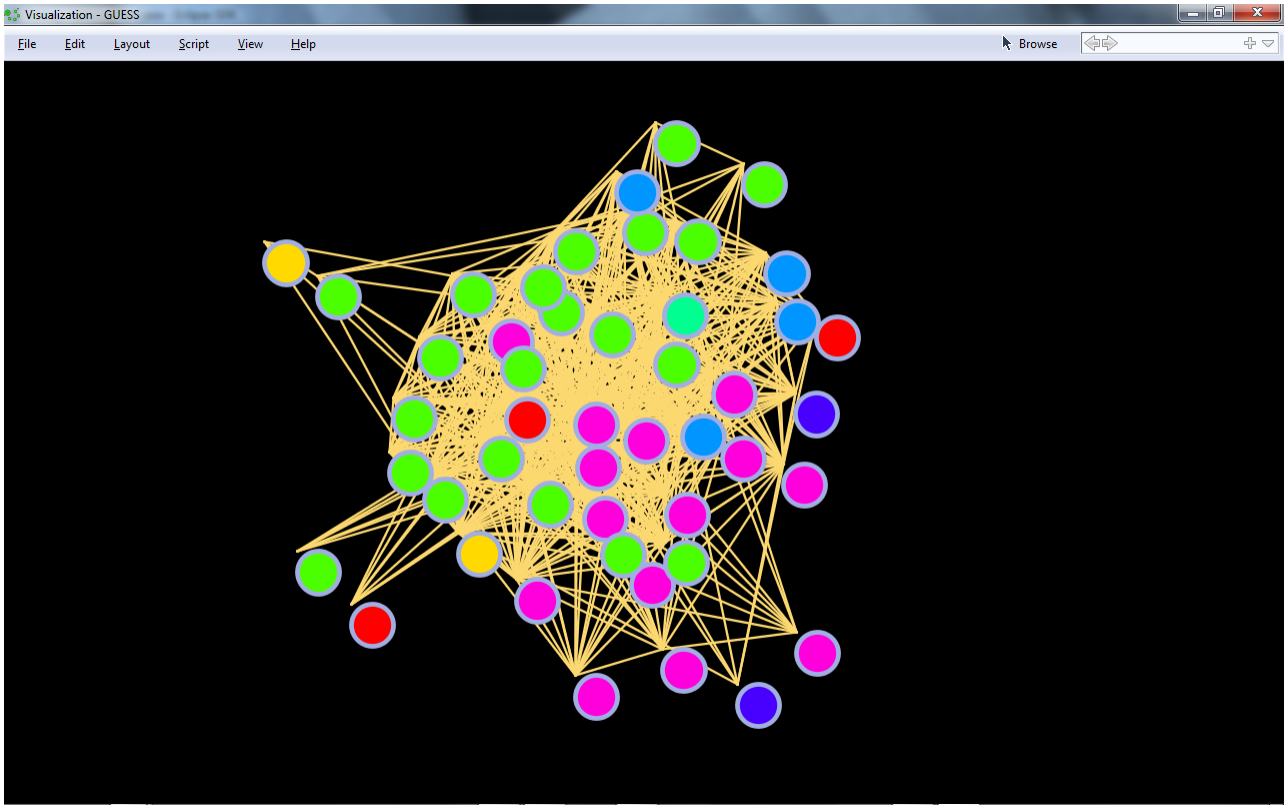


Figure 5.14: OBO Domain-Domain Network

For Figure 5.15, communities can be divided into four domains based on their similarities in terms of their colors, which are shown in Figure 5.16. We compute the metrics for each of these domains and compare the metrics with the metrics of OBO network.

Table 5.8 presents the comparison of network metrics generated by the Colorscape model against the corresponding metrics from empirical OBO data (expected values in the table). Since the confidence intervals of metrics derived from the simulation data contain the corresponding values of the OBO network, we conclude that Colorscape model can generate similar network structures in comparison to OBO.

In addition, the best configuration parameters against the OBO network are recorded in Table 5.7. From the table, we can observe that the best configuration has a medium level tolerance (0.6), high receptivity (0.9), and high degree of communication frequency (1.0). These are indeed the quintessence characteristics of open source science communities.

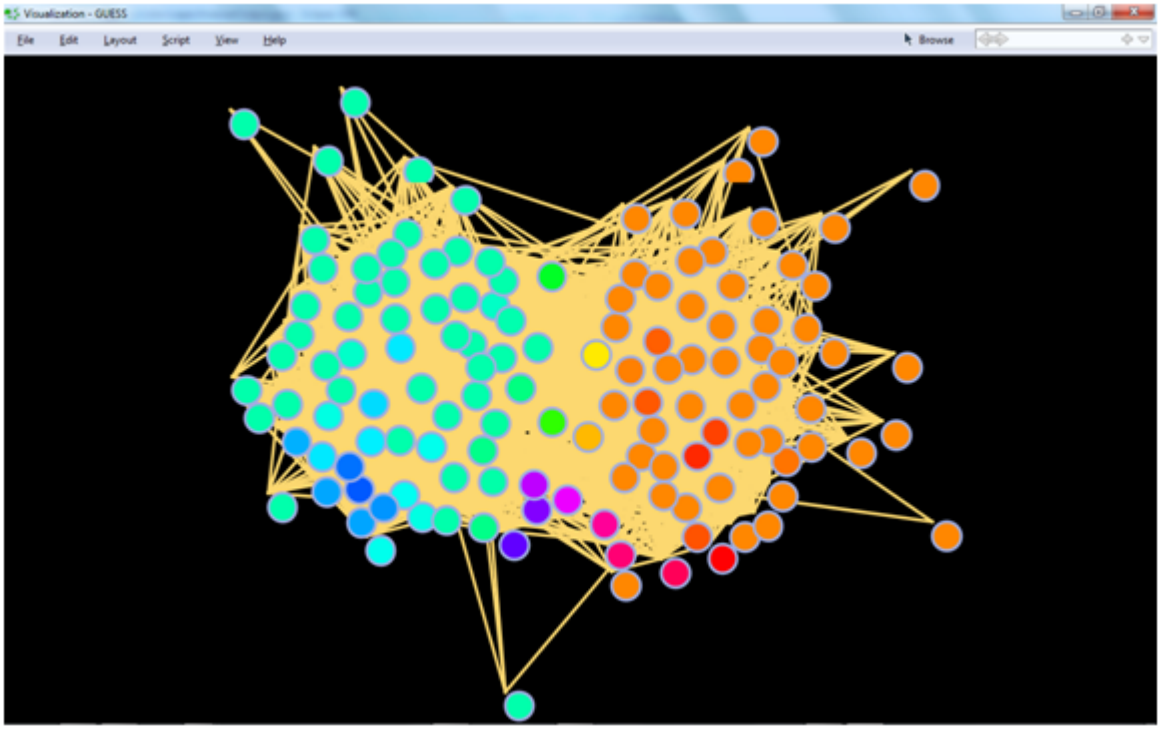


Figure 5.15: Snapshot of Colorscape Model against OBO

Table 5.8: Simulation vs. OBO Data

Metrics	Mean Value	Confidence Interval of 90%	Expected Values
Number of Nodes	55.633	[46.076, 65.190]	49
Density	0.605	[0.521, 0.689]	0.549
Clustering Coefficient	0.846	[0.812, 0.881]	0.880
Centrality	0.355	[0.302, 0.407]	0.405
Average Path	1.404	[1.317, 1.491]	1.406
Core/Periphery Ratio	50.9	[37.2, 64.6]	23.5

5.6 Power Law

When the probability for the occurrence of an event is inversely proportional to its size, power-laws are often expected [66]. Power law appears in many systems, e.g., the distributions of the sizes of cities, earthquakes, forest fires, and people's personal fortunes.

Figure 5.17(a) shows the inequality of communities in terms of resources. Most communities hold the relatively few resources, while a small part of communities hold the relatively many resources. To determine if this can be interpreted by the power law, Figure 5.17(b) shows the relationship between the Logarithm value of number of communities and their

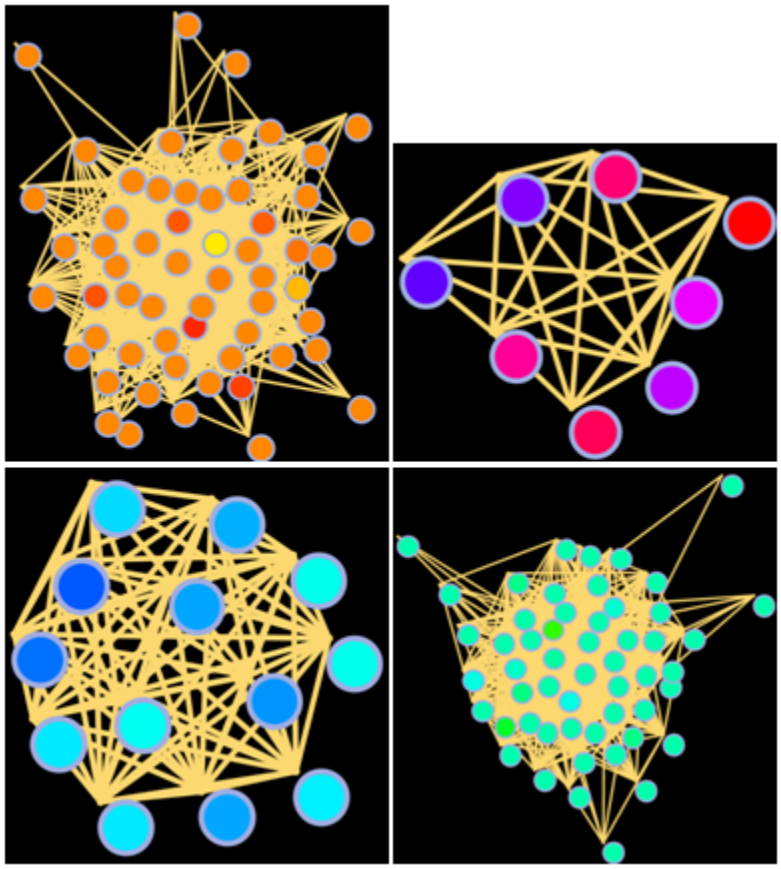
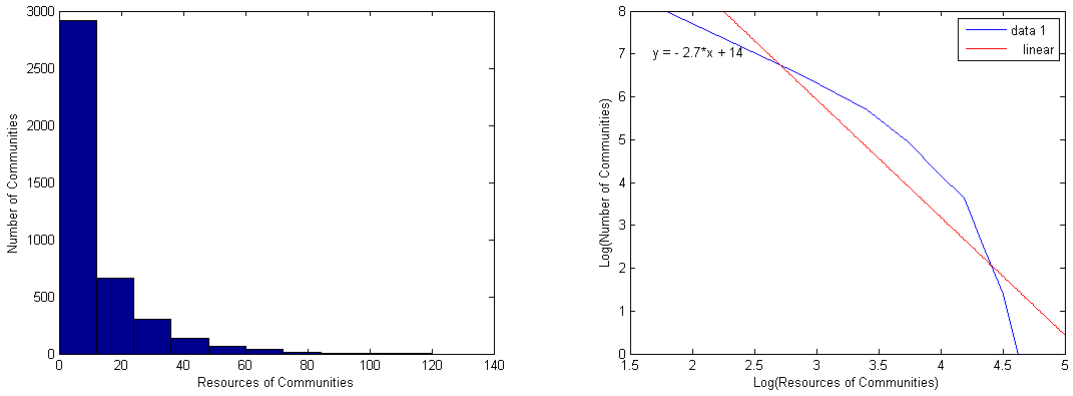


Figure 5.16: Clusters of the Network of Colorscape Model against OBO

resources, as well as the corresponding linear regression curve. Since the R^2 for this fitting is 0.86, there is significant evidence that the Colorscape model can exhibit power-law in resource distribution.



(a) (b)

Figure 5.17: Distribution of Resources in ColorScape Model

Chapter 6

Simulation Results and Evaluation

In this chapter, experiments are conducted to investigate the impact of scientific community traits (i.e., receptivity, flexibility, reorganization tendency) and environmental constraints (i.e., interaction topologies, carrying capacity, resource allocation strategies) on the innovation performance (e.g., diversity and resilience) of GPS.

6.1 Interaction Toplogies

The experiments in this section test five types of interaction topologies and their effects on diversity and resilience of GPS:

1. *One-dimensional grid*: Each community has two neighbors on the left and right side.

2. *Two-dimensional grid*: Each community is embedded in a Von Neumann neighborhood; that is, it has eight neighbors surrounding it. Figure 3.1 includes a snapshot of the 2D grid.

3. *Random network*: The edges between any pair of nodes are created with equal probability.

4. *Random group network*: The nodes within a group have higher probability to build links than those between different groups.

5. *Scale-free network*: The nodes with more links are more likely to be selected to build links. Figure 3.1 includes a snapshot of scale-free network.

6. *Dynamic network*: Communities choose to communicate with other communities based on preferences dictated by the selected social communication theories.

6.2 Measuring Innovation Potential and Performance

Since we are interested in observing potential relations between the structure of the social network and innovation outputs of a community, two types of metrics are considered: innovation metrics and network structure metrics that pertain to integrated differentiation.

We proposed a hierarchy of metrics to examine relations between scientific community traits, structure of communication networks, and innovation performance. The hierarchy shown in Figure 6.1 aims to delineate the relationship between layers of the evaluation framework.

6.2.1 Innovation Metrics

In the evaluation framework shown in Figure 6.1, there are three innovation metrics: robustness, resilience, and interdisciplinarity. Two of these metrics are used in the following experiments: resilience and interdisciplinarity, in which diversity is suggested to be a useful proxy indicator to measure interdisciplinarity [74] [71].

6.2.1.1 Diversity

The process of knowledge creation is based on the combination and elaboration of existing knowledge. Diverse sources of knowledge challenge existing solutions, ignite new ideas, and lead to more impactful solutions [60]. So, diversity is a proxy indicator for innovation potential and capacity. There are three dimensions related to diversity: variety, balance, and disparity [89]. Variety can be computed as the number of clusters of communities of the whole environment. Each cluster is composed of similar communities. To classify communities into clusters, we use the QT (Quality Threshold) clustering algorithm [39]. QT clustering algorithm needs a predefined diameter indicating the maximum difference among members in a cluster. Then a candidate cluster for each community is built by including other communities within the predefined diameter. A cluster with maximum members is selected, and then we recursively run the above steps with the set of communities after

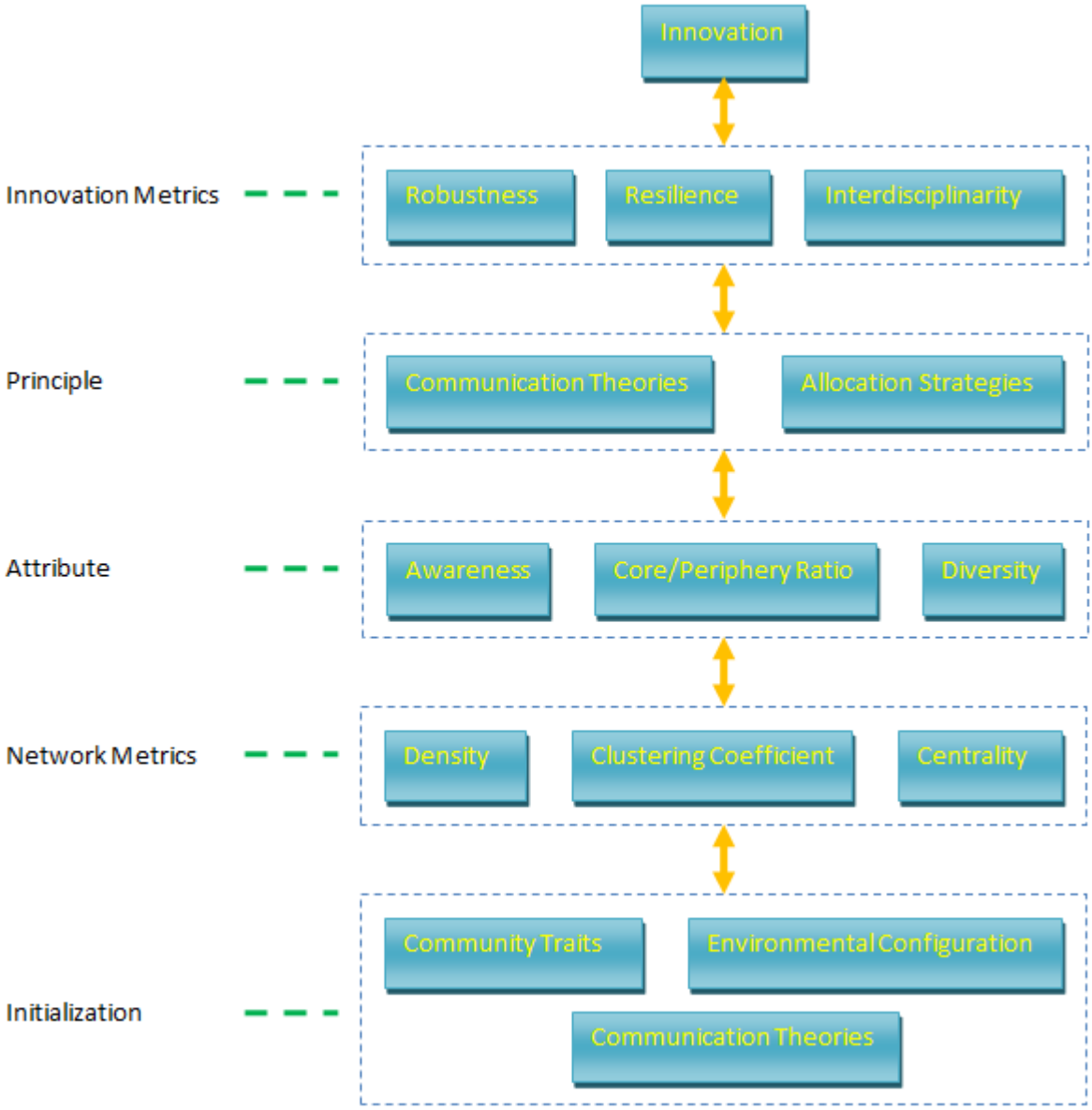


Figure 6.1: The Evaluation Framework

removing communities in the selected cluster. Algorithm 1 is the pseudo code for the QT algorithm:

Balance indicates inequality in terms of resources each community holds. It is calculated using the Gini coefficient [16], which is a measure of the inequality of a distribution, a value of 0 expressing total equality, and a value of 1 maximal inequality [100]. The Gini coefficient is calculated as follows:

Algorithm 1 QTAlgorithm (*Community*[] *communities*, double diameter)

```
Vector<Vector<Community>> result = new Vector<Vector<Community>>();
Community[][] clusterArray = new Community[communities.length][];
for i = 0 to communities.length do
  /*Find cluster for each community*/
  Community[] cluster = findCluster(communities, communities[i], diameter);
  clusterArray[i] = cluster;
end for
int indexMax = findMaxCluster(clusterArray);
result.addAll(clusterArray[indexMax]);
removeCommunities(communities, clusterArray[indexMax]);
if communities.length > 0 then
  /*Recursively call the algorithm with the reduced set*/
  Vector<Vector<Community>> tmpResult = QTAlgorithm(communities, diameter);
  result.addAll(tmpResult);
end if
return result
```

$$G_N = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i}{(n-1)\sum_{i=1}^{n} x_i}, \tag{6.1}$$

where n is the total number of communities. x_i is the resource level of community i.

Disparity refers to the degree of difference of each community, that is, the dissimilarity of communities based on their current color.

6.2.1.2 Resilience

Partly, innovation is the process of finding alternative, more effective ways to address challenges and seize opportunities. On the other hand, resilience is the capacity to adapt, restore in constructive ways while undergoing changes to retain essentially the same function. Hence, innovation is change, but resilience is survival. Due to presence of uncertainty in the evolution of the innovation landscape, resilience is an essential property for a scientific community to sustain its innovation capacity.

Resilience is the capacity of a system to absorb disturbance and reorganize while undergoing changes to still retain essentially the same function, structure, identity, and feedbacks

[94]. Based on this definition, we define resilience as the extent of disturbance of the system that reduces the fraction of active communities to the initial set of communities below a specific threshold.

6.2.2 Network Metrics

Structural properties of networks as they relate to creative output pertain to integrated differentiation [87]. As a general measure of the degree of social interaction, we use density, centrality, and clustering coefficient to determine their potential roles in and relation to innovativeness. Low density and high centrality communities are expected to exhibit higher degrees of innovation capacity [25]. Cliquish networks with low average path lengths are known to be effective in knowledge creation and diffusion [23].

6.3 Simulation Results

Using the ColorScape model, we conducted a series of exploratory experiments to examine how innovation capacity and sustainability of the innovation ecosystem relate to community interaction topologies, connectivity, and resource allocation strategies. Table 3.1 denotes the configuration parameters and their initial values.

6.3.1 Diversity vs. Carrying Capacity

In this experiment, we explore variation of diversity in relation to number of communities within a specific topology. Figure 6.2 evaluates variety, disparity and balance across combination of two factors, number of communities and 1D/2D topology.

In Figure 6.2, we observe that variety and disparity increase with the initial community size, called Carrying Capacity (CC). In the 2D topology, disparity increases with CC up to a critical threshold, after which further increase in dissimilarity diminishes. Computation of variety is based on the QT clustering algorithm based on a pre-selected diameter denoting the maximum difference allowed among members within a cluster. In this experiment, the

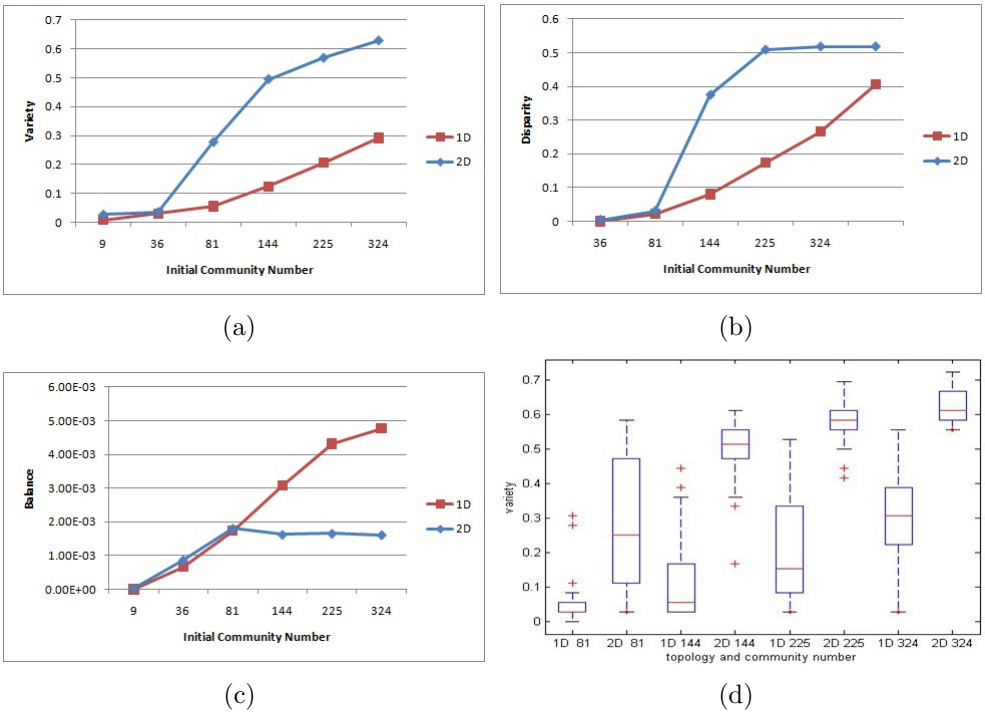


Figure 6.2: Diversity vs. Initial Community Numbers

diameter is set to 10, indicating that the hue difference among communities within a cluster can be up to 10. Therefore, the maximum variety is $360/10 = 36$. That is, diversity cannot increase indefinitely with CC. Based on Figure 6.2(d), the comparison between 1D and 2D suggests that in comparison to 1D topology, the 2D topology is more conducive to fostering variety with a lower degree of uncertainty. Also, the limited sphere of interaction exhibited in the 1D topology inhibits diffusion of influence and hence leads to increased time to reach equilibrium.

Next, to evaluate the impact of neighbor size and hence the sphere of influence within the 1D topology, we gradually increased the interaction window from 2 to 8 neighbors. Observations depicted in Figure 6.3 suggest that interaction window positively affects variety and underlying uncertainty (i.e., dispersion) up to a level, beyond which variety stops improving while uncertainty increases.

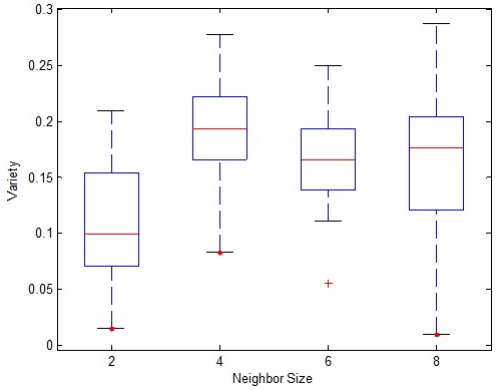


Figure 6.3: Variety vs. Neighbor Size in 1D

6.3.2 Diversity vs. External Resource

The resource allocation strategy used in the baseline model is to distribute all resources uniformly among communities. The total available resource is the sum of contributions of communities and external resources. Figure 6.4 depicts the change in diversity with respect to available external resources.

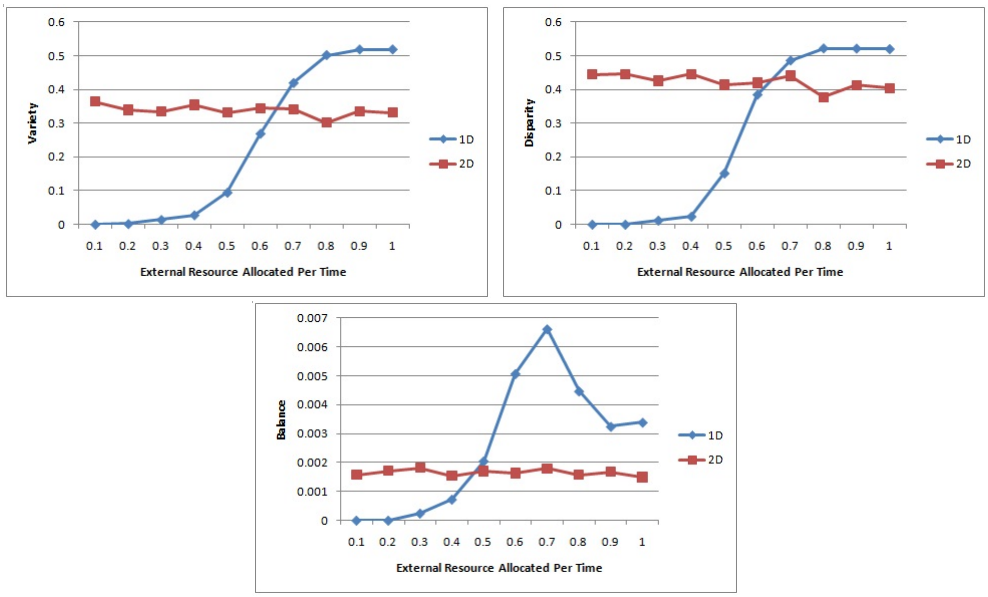


Figure 6.4: Diversity vs. Resource Allocated Per Time

The abscissa indicates the amount of resources allocated to each community per time tick. In the 1D topology, the rate of increase in variety slows and stabilizes over time. On

the other hand, the 2D topology seems to be less sensitive to external resource, indicating higher degree of potential for resilience than 1D.

When external resource is low, a small number of communities can survive. As external resource increases, more and more communities can survive, which leads to increased diversity. This trend increases up to a point beyond which more resources only can increase the number of communities within a cluster rather than the number of clusters. On the other hand, the communities in the 2D topology have more neighbors than those in the 1D topology, which makes communities more likely to form clusters. Communities within a cluster have similar domains, which helps communities improve maturity with less resource consumption during the process of learning discussed in section 3.5. Thus, communities in the 2D topology have higher maturity and more resources left, so that the second part in Equation 3.1 is large enough to sustain all communities.

For policy makers, it is noteworthy that more funds cannot lead to higher diversity. More funds only result in more resources held by communities.

6.3.3 Diversity vs. Reorganization

The experiment in this section aims to find out the relationship between diversity and reorganization. Figure 6.5 depicts change in variety, disparity, and balance against different levels of reorganization tendency.

From Figure 6.5, we can observe that variety and disparity decrease with increasing reorganization tendency, which means that reorganization has negative effects on variety and disparity. On the other hand, specialization has positive effects on variety and disparity. It is consistent with the functionality of specialization and reorganization. Specialization facilitates creation of a new community with a different target color from the current community. However, reorganization involves pulling the target color toward the current color, causing convergence.

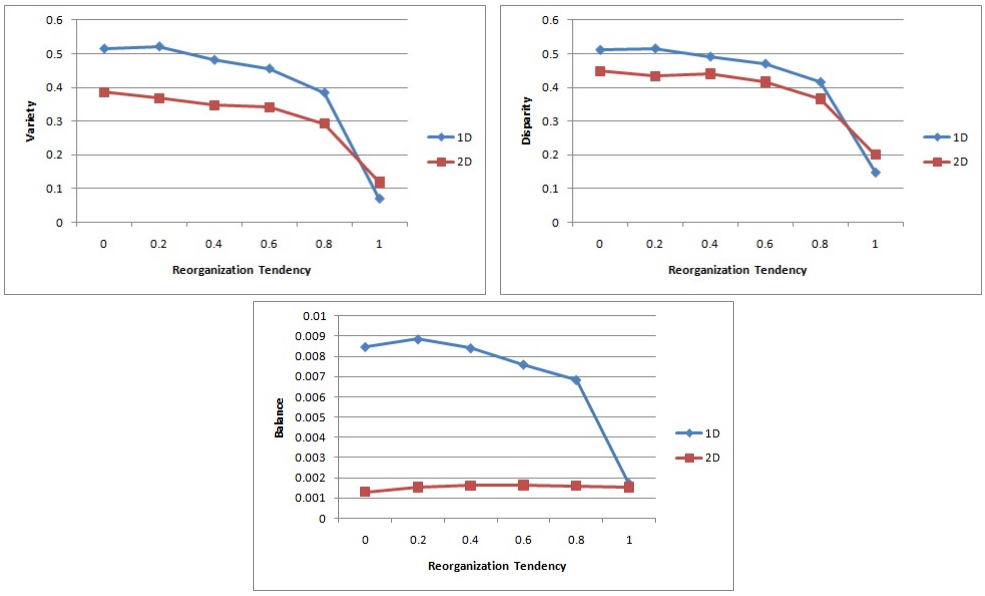


Figure 6.5: Diversity vs. Reorganization Tendency

6.3.4 Diversity vs. Receptivity

In this experiment, we considered alternative interaction topologies (Random and Random Group Network) to discern the relation between variety and community receptivity. Receptivity of a community is defined as the ratio of neighbor influence to inertia. Connectedness is defined as the probability of building links between nodes. Figure 6.6 indicates that there is a critical receptivity threshold, after which the behavior of low and high density communities diverges. Behind this phenomenon, the potential reason is that low receptivity results in few influences from neighbors, which in turn determines context topologies' few effects on variety. Under environments with high receptivity, variety favors low connectivity. The reason is that more communication links cause convergence, which in turn decrease the variety. However, communities with various levels of connectivity converge to the same stable level of variety. Similar patterns are observed in both random and random group networks.

Based on the experimental results, policy makers may encourage communities to be more receptive in a relatively low density environment to reach a high variety. This conclusion is supported by earlier reports and findings [25].

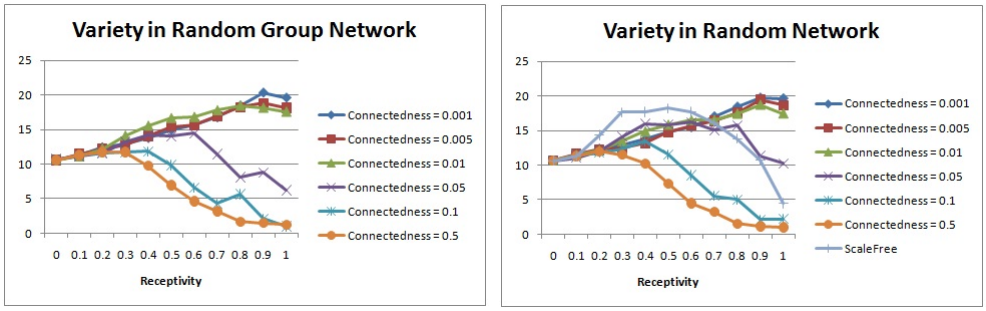


Figure 6.6: Variety in Random and Random Group Network

6.3.5 Resilience of Different Network Topologies

Resilience is defined as the extent of disturbance on the system that significantly reduces the ratio of active communities to CC when external resource is set to maximum [94]. To compute resilience, the number of communities under maximum resource availability (i.e., CC) is set as the base reference level for each topology. Figure 6.7 depicts the number of active communities varying along with external resources in terms of three kinds of network topologies.

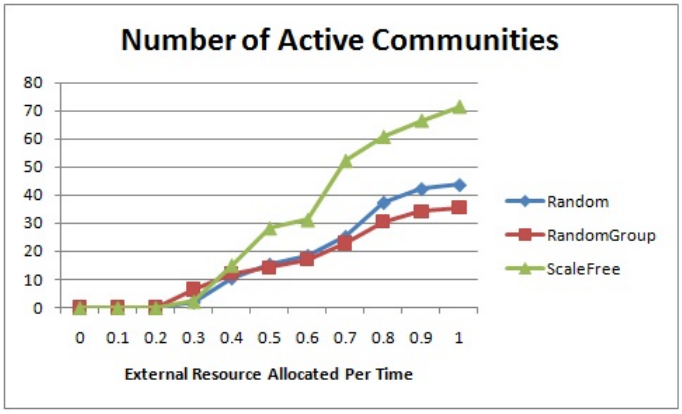


Figure 6.7: Number of Active Communities

As resources are gradually reduced, the ratio (ρ) of number of communities to the CC is computed. The loss ratio is defined as $1 - \rho$ and ranked to identify resilient topologies. According to Table 6.1, scale-free network has the highest resilience, and random group network has higher resilience than random network, because the loss ratio of scale free network is smallest and the loss ratio of random network is largest when external resources

decrease to 0.7. Figure 6.8 confirms that random group network exhibits higher resilience than random network.

Table 6.1: Resilience of Different Network Topologies

	Random		Random Group		Scale Free	
Resources	Number of Communities	Loss Ratio	Number of Communities	Loss Ratio	Number of Communities	Loss Ratio
1	43.77	0	35.57	0	71.43	0
0.9	42.3	0.03	34.27	0.04	66.47	0.07
0.8	37.43	0.14	30.63	0.14	60.7	0.15
0.7	25.33	0.42	22.83	0.36	52.2	0.27

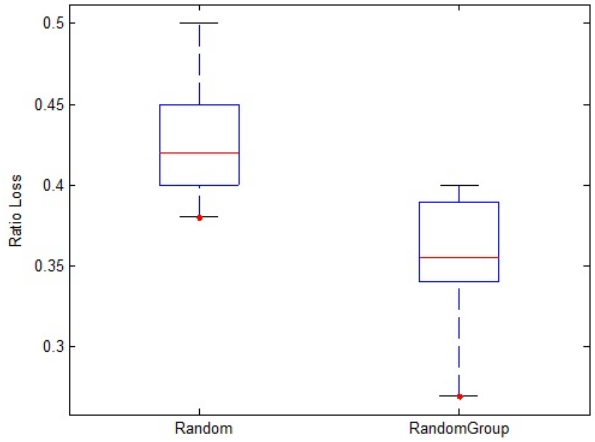


Figure 6.8: Comparison of Random and Random Group Network on Resilience

6.3.6 Relationship between Diversity and Network Metrics

The data to study the relationship between diversity and network metrics are gathered from previous experiments involving sensitivity analysis on receptivity. Each pair of density and variety is classified into buckets that occupy an identical range i.e., 0.1 for each bucket in terms of density. If the density falls into the range of [0, 0.1), then the pair of density and variety belongs to the bucket of 0.1. If the density falls into the range of [0.1, 0.2), then

the pair of density and variety belongs to the bucket of 0.2. After grouping, variety is the average of all pairs in the corresponding bucket.

Figure 6.9 shows that variety increases with density up to a point. After that point, variety decreases with increasing density in both random and random group networks.

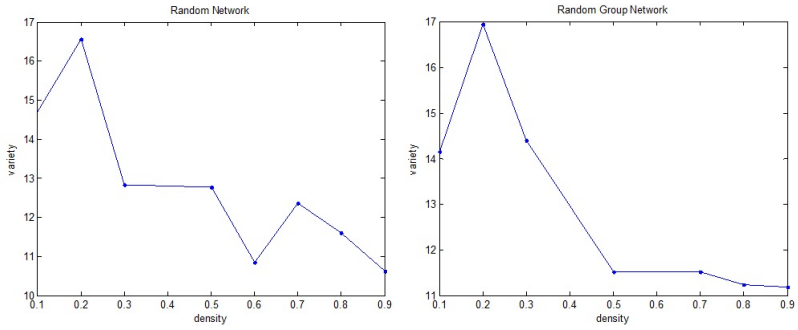


Figure 6.9: Variety vs. Density in Random and Random Group Network

Figure 6.10 plots variety against degree centrality. Variety increases with centrality up to a point. Beyond that point, variety decreases with increasing centrality.

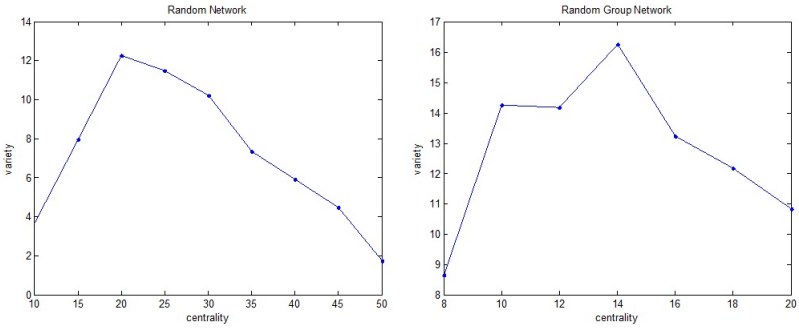


Figure 6.10: Variety vs. Centrality in Random and Random Group Network

In [40], Hohn examines the relationship between species diversity and population density in diatom populations, which is shown in Figure 6.11. Since scientific communities can be viewed as an ecosystem, it is reasonable to compare the phenomena of ecosystem to that of scientific communities. From this figure, we can see species diversity increasing with population density up to a point. As density increases beyond this threshold, diversity starts declining. The density in [40] is defined as the number of individuals per species, which is different from density defined in our research. However, both definitions of density are

related. The more individuals the species has, the more is the dependency among members due to shared, but limited resources.

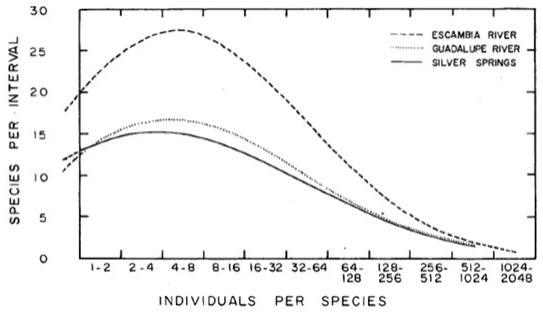


Figure 6.11: Species Diversity vs. Population Density in [40]

In [69], the following proposition about centrality and creativity is presented: individuals with greater centrality are likely to have higher creativity until a level. Beyond this level, greater centrality may constrain creativity. This trend is consistent with our experimental results.

6.3.7 Sustainability, Resource Availability, and Connectedness

In ecology, sustainability refers to the ability of biological systems to remain diverse and productive over time. In the domain of creativity, sustainability can be interpreted as the effectiveness of communities in utilizing resources. So, we relate it to success rate, which measures the extent to which communities are effective in making use of resources to improve their maturity, while maintaining themselves. Success rate is defined as the ratio of the number of active communities remaining at the end of simulation to CC.

Figure 6.12 depicts the relationship between resource availability, interconnectedness, and success rate. The experimental results suggest that if resource availability increases while connectedness is decreased, the success rate increases. Also, when resource is at high level, success rate decreases with increasing connectedness. In addition, when resource is at low level, success rate decreases with decreasing connectedness. A plausible explanation for this observation is that higher resource availability leads to higher variety. Under high

variety, larger connectivity causes each community to be pulled toward multiple different cognitive niches, resulting in lack of focus which in turn costs communities more resources, and hence decreasing the survival rate. On the other hand, lower resource availability leads to lower variety. Under low variety, however, strong connectivity results in more communities sharing similar states, benefiting from each other through a symbiotic relation, which in turn increases the overall survival rate.

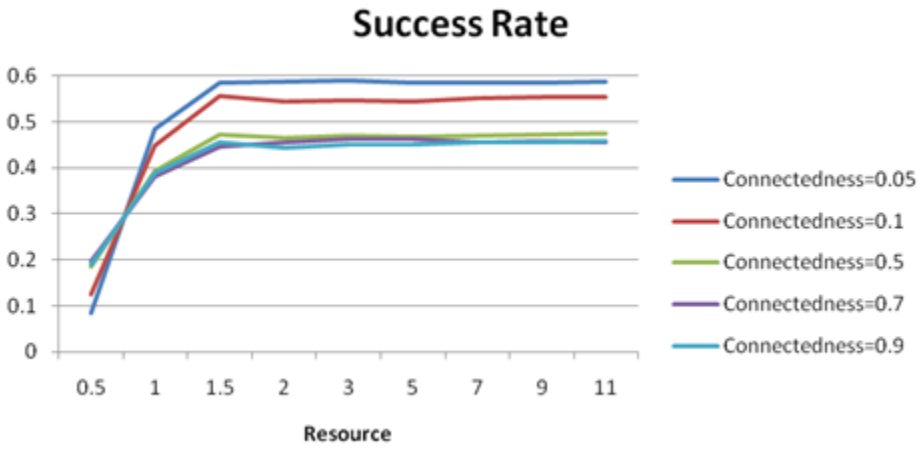


Figure 6.12: Success Rate vs. Resource

Based on these preliminary observations, policy-makers may encourage communities to build highly connected clusters if resource availability is low. On the other hand, under moderate to high-level resource availability, loosely connected clusters may be more effective in promoting an environment conducive to sustainability.

6.3.8 Disparity vs. Resource and Connectedness

Creativity partly involves combination and elaboration of existing knowledge. Therefore, we use diversity as a proxy indicator for collective creativity. As discussed earlier, there are three dimensions related to diversity: variety, balance, and disparity [89].

In this section, we focus on the disparity dimension. Disparity indicates the degree of inequality, which can be measured by the Gini coefficient [100]. The coefficient ranges from 0 to 1, where 0 and 1 refer to perfect equality and extreme inequality, respectively.

Figure 6.13 denotes the relationship between resource availability, connectedness, and disparity. As resource level increases from 0.5 to 1, disparity increases with resource availability when the degree of connectedness is low, since more resources lead to higher success rate, which in turn results in disparity. On the other hand, when the resource level increases further, disparity decreases. This is due to decreased need for interaction for sustainment. This, in turn, decreases inequality. In addition, disparity increases with decreasing level of connectivity, which is possibly due to increased convergence under high connectivity, resulting in decreased disparity.

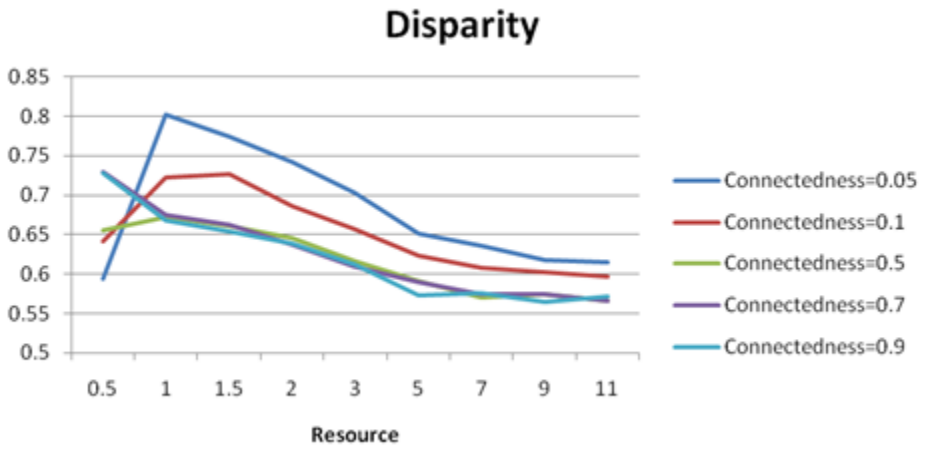


Figure 6.13: Disparity vs. Resource

Based on the previous two experiments, Table 6.2 summarizes how disparity and success rate relate to resource availability and connectedness.

Table 6.2: Success Rate and Disparity

Resource Level	Resource Trend	Connectedness	Disparity	Success Rate
Low	Down	Down	Up	Down
High	Up	Down	Up	Up
High	Up	Up	Down	Down

6.4 Experiments on Resource Allocation Strategy

Understanding the influence of resource distribution across communities is critically important for informed decision-making in science and innovation policy development. The following experiments focus on the impact of resource allocation strategies on diversity. For the allocation strategies, we identify seven options:

1. Allocate resource uniformly among communities.
2. Allocate resources proportional to the contributions of communities.
3. Allocate resources proportional to the size of cluster formed by similar communities.
4. Allocate resources proportional to the importance of domains.
5. Fully competitive allocation.
6. Peer-to-peer (P2P) lending.
7. Random allocation.

In *uniform allocation*, resources are allocated to communities equally regardless of their states. *Resource allocation proportional to contribution* is a reward mechanism. Communities with larger contributions receive more resources. Under the *allocation proportional to cluster size*, the larger the cluster a community belongs to, the more resources the community receives. With *allocation proportional to importance of domains*, disciplines with higher priority receive more resources. The competitive allocation strategy is analogous to the prey/predator model, where resources are distributed among domains, and communities compete for resources. Under random allocation, resources are allocated to a randomly selected set of communities equally regardless of their state. P2P lending involves a contract-bid protocol. The community that invites others for collaboration is called the sponsor. Other communities in the same domain respond with a bid that indicates the ratio of resources the community gets to those resources the sponsor will receive. The community that

answers the call is named as respondent. After receiving all the bids, the sponsor selects a bid with the highest resource gain.

Figure 6.14 presents the class diagram of the resource allocation module, where all classes are inherited from a single class named *ResourceAllocation* that declares two functions implemented by sub-classes. All classes of different allocation strategies only have one single public function i.e., *allocationResources()*. In addition, the common part of total resources is extracted to be a class named *TotalResource* that has two public functions i.e., *fixed()* and *techTransfer()*, which are distinguished by whether or not there is a mechanism to transfer technology.

6.4.1 Design of Resources Allocation Strategies

6.4.1.1 Uniform Allocation

Uniform allocation means that resources are allocated to communities equally regardless with the states of communities. Figure 6.15 represents the process of uniform allocation. Given the total resource (R_T) and total number of communities (N), each community can receive resources (R_i) that amounts to:

$$R_i = R_T \times \frac{1}{N}. \tag{6.2}$$

6.4.1.2 Proportional to Contribution

Resource allocation proportional to contribution is a reward mechanism in that communities with larger contributions receive proportionally more resources. Figure 6.16 represents the process of resource allocation proportional to contributions of communities.

In Figure 6.16, contributions provided by a community (C_i) are moderated by the product of its maturity and resource. This is based on the hypothesis that communities with

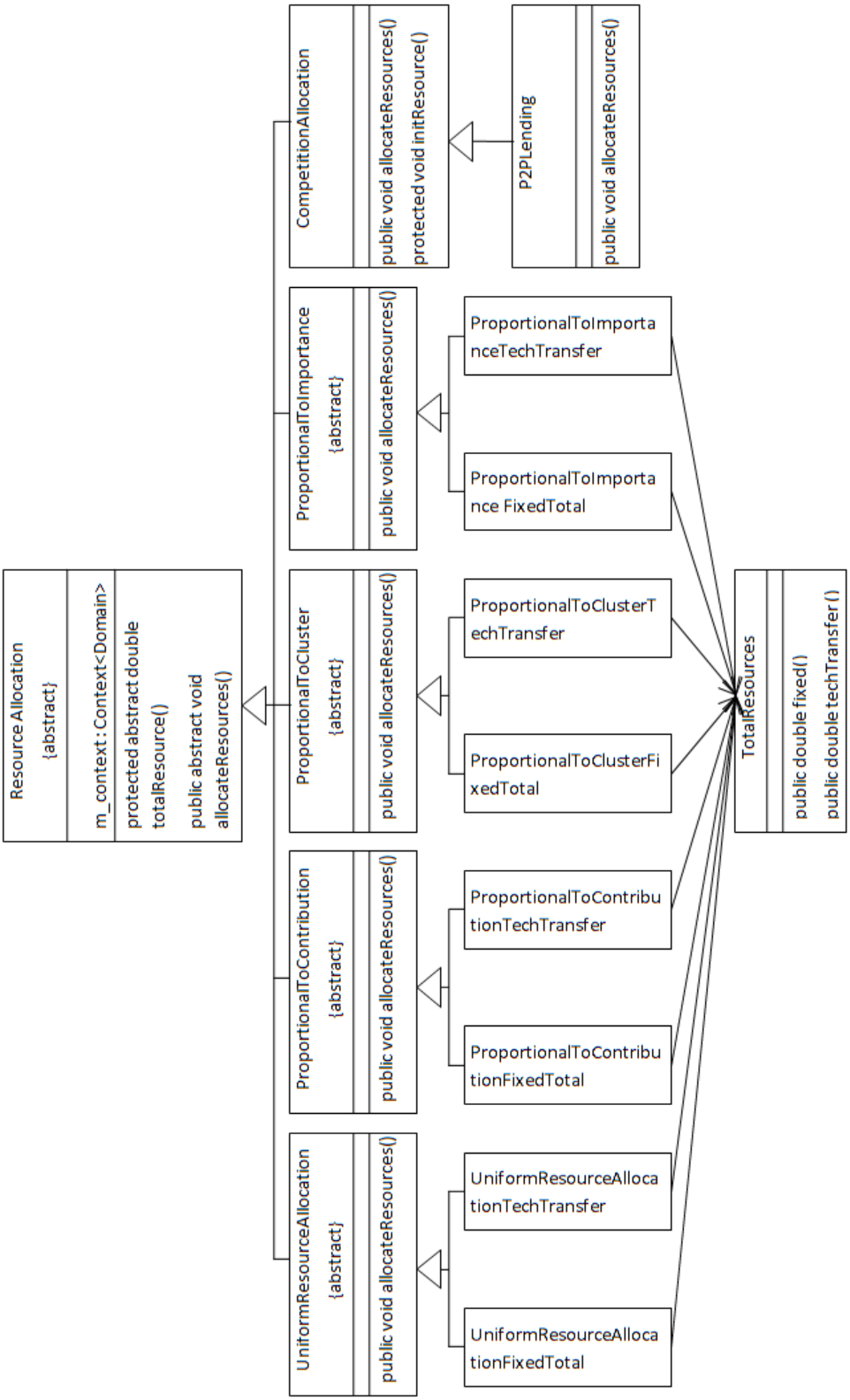


Figure 6.14: Class Diagram of Resources Allocation

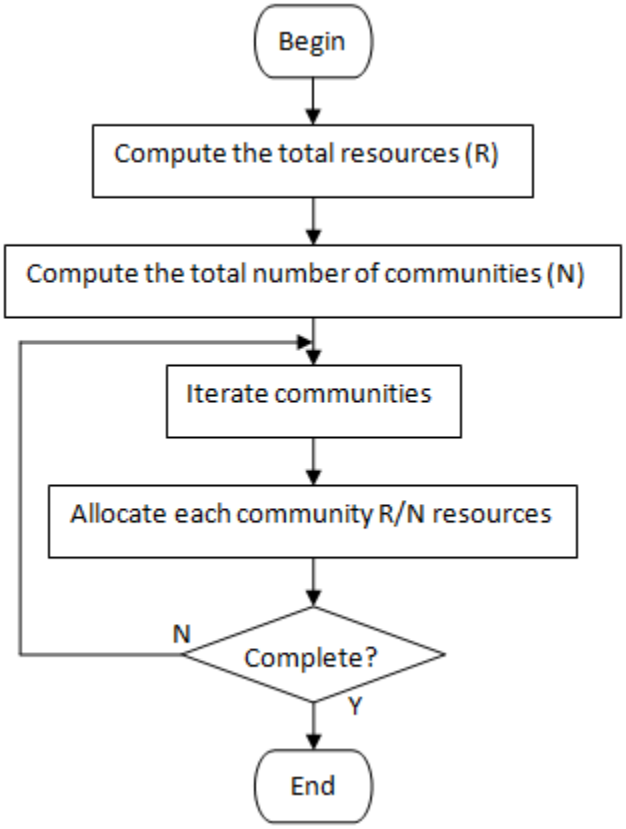


Figure 6.15: Flow Chart of Uniform Allocation

higher maturity and resources are expected to be more productive. Each community can receive resources (R_i) that amounts to:

$$R_i = R_T \times \frac{C_i}{\sum_{j=1}^{N} C_j}, \tag{6.3}$$

where R_T is the total available resources. N is the total number of communities.

6.4.1.3 Proportional to Cluster Size

Allocation of resources proportional to cluster size refers to distribution of resources proportional to size of the cluster to which a community belongs. The purpose of such allocation strategy is to encourage communities to form larger clusters. Figure 6.17 represents the process involved in deciding how much resource to allocate to a community. Given a total resource (R_T), each community gets resources with amount of:

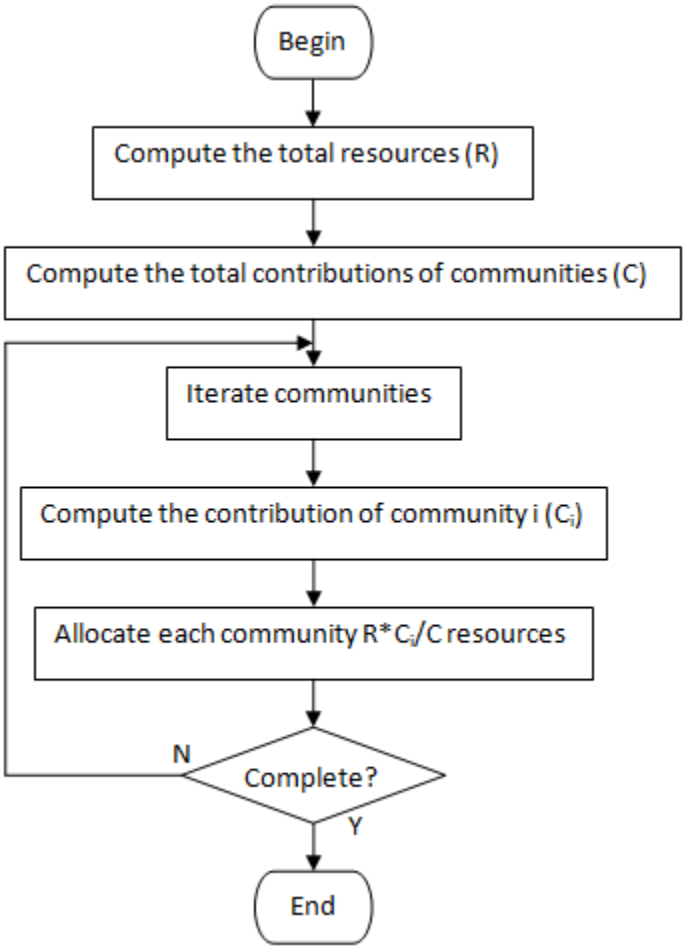


Figure 6.16: Flow Chart of Allocation Proportional to Contribution

$$R_i = R_T \times \frac{S_i}{\sum_{j=1}^{N} S_j}, \tag{6.4}$$

where S_i is the size of cluster community i belongs to, N is the total number of communities. For the sake of illustration, here is an example where there are three communities, among which two communities form a cluster and $R_T = 10$. The community within the cluster gets the resource of 10*2/(2+2+1) = 4. Also, the community without the cluster gets the resource of 10/(2+2+1) = 2. So, the community within the cluster gets resources twice more than the community without a cluster.

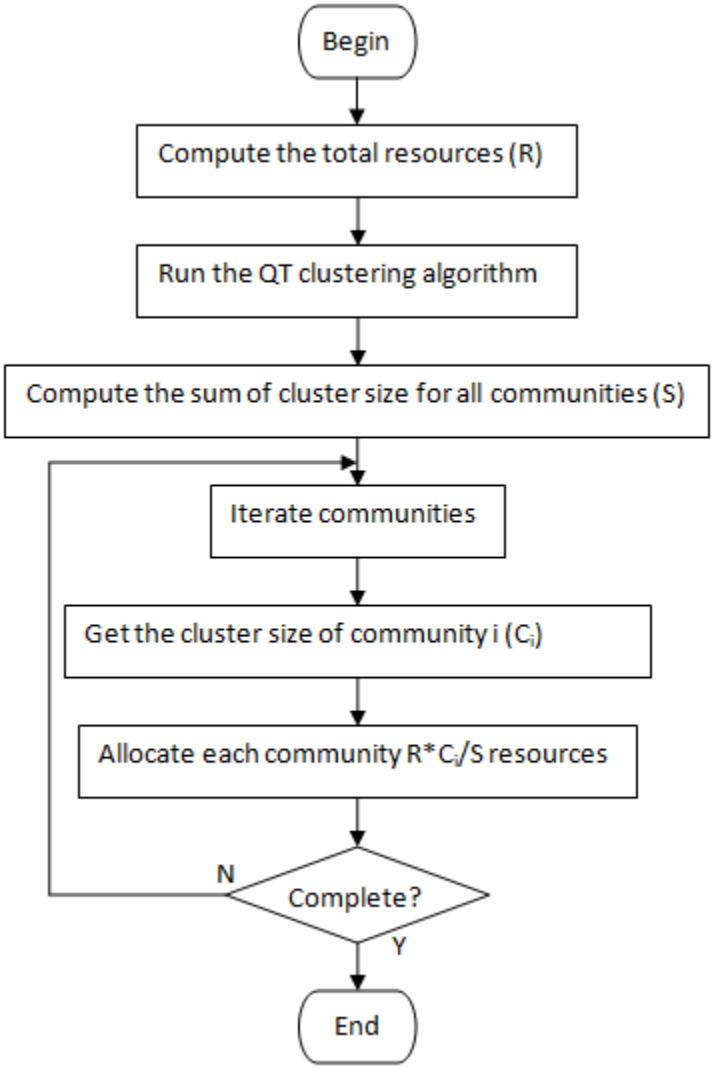


Figure 6.17: Flow Chart of Allocation Proportional to Cluster

6.4.1.4 Proportional to Importance of Domains

Some domains have higher funds than others, e.g., nanotechnology receives more funds than other conventional physics. Figure 6.18 represents the process of allocation proportional to importance of domains. Each community i receives resources (R_i) defined as follows:

$$R_i = R_T \times \frac{W_j}{N_j}, \tag{6.5}$$

where R_T is the total resource. j indicates the domain community i belongs to. W_j and N_j denotes the importance of domain j and total number of communities domain j includes, respectively.

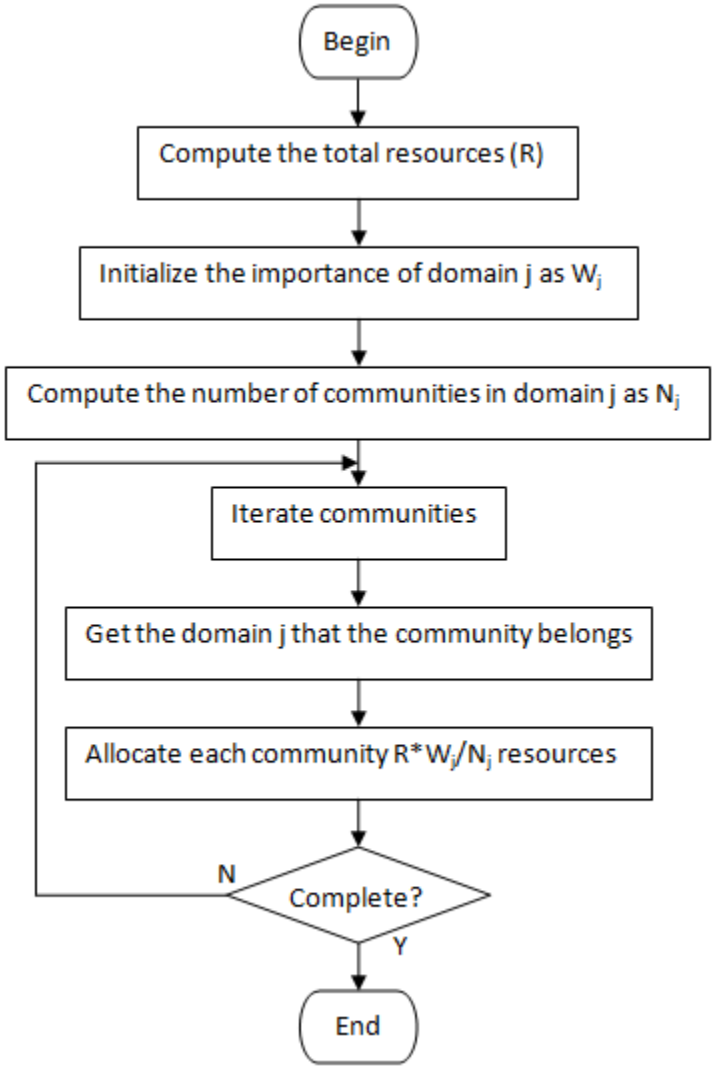


Figure 6.18: Flow Chart of Allocation Proportional to Importance of Domains

For the sake of illustration, the following is an example where the whole range of hue is divided into three domains, that is, [-60, 60), [60,180), and [180, 300), whose corresponding importance is 0.6, 0.3, and 0.1 respectively. If the number of communities in the domain [-60, 60) is 3 and total resource is 100, each community in the domain [-60, 60) can be allocated resources of 100*0.6/3 = 20.

6.4.1.5 Competitive Allocation

The competitive allocation strategy is similar to the prey/predator model, where resources are distributed among domains, and communities receive resources from the domain they belong to. The whole range of hue is divided into 360 domains. Each domain has a fixed amount of resources at the beginning of each time interval. If a community attains the resources within its domain, the resources in that domain become 0. The process is represented in Figure 6.19.

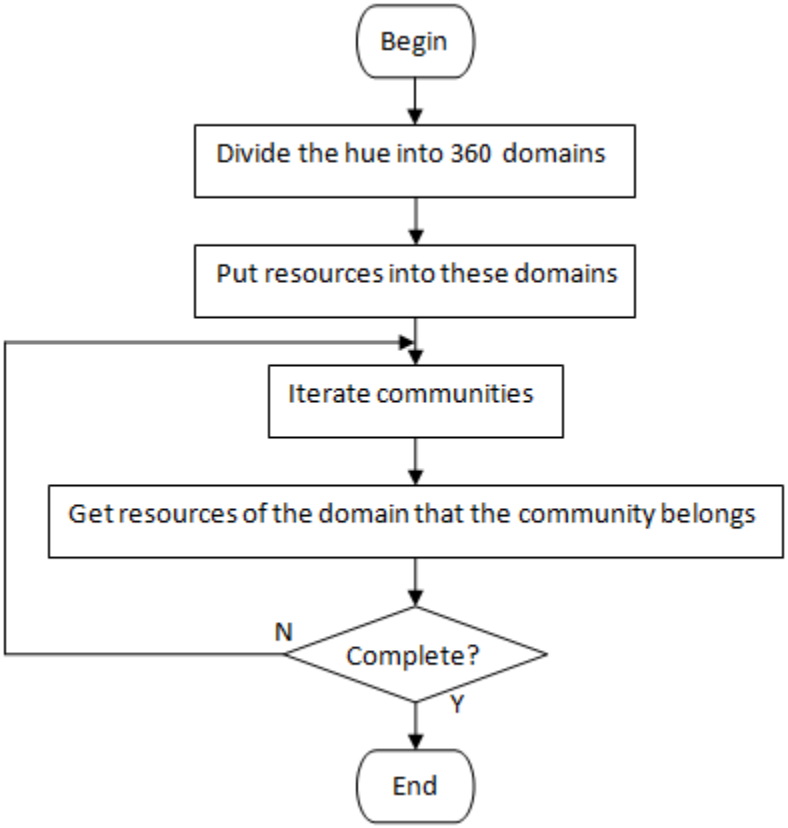


Figure 6.19: Flow Chart of Competitive Allocation

6.4.1.6 P2P Lending

P2P lending introduces the mechanism of calling for proposals on the basis of fully competitive allocation, indicating that a community can request collaboration with other

communities if the domain of the community does not have sufficient resources. The community calling others for collaboration is called *sponsor*. When P2P lending occurs, other communities in that domain respond with a bid that indicates the ratio of resources received by the community to those resources the sponsor receives. The community that answers the call is named as *respondent*. The bid a respondent submits is proportional to resources it holds, that is, the more resources it has, the higher the ratio of profit the community expects to receive. After receiving all the bids, the sponsor community selects a bid with the highest profit. The process is represented in Figure 6.20.

6.4.1.7 Random Allocation

Random allocation involves distributing resources to communities randomly regardless of their state. Figure 6.21 represents the process of random allocation.

6.4.2 Network Pattern vs. Resource Allocation Strategy

In this section, emerging network patterns are qualitatively and visually examined to gain insight about the impacts of resource allocation strategies on diversity. Figure 6.22 depicts network structures generated under allocation strategies 1 (i.e., uniform allocation), 2 (i.e., proportional to contributions) and 3 (i.e., proportional to the size of clusters). We observe that the network under uniform allocation has the highest diversity, while strategies 2 and 3 lead to relatively lower diversity.

In Figure 6.23, the predefined ratios of resources allocated to disciplines indicated by red, green, and blue colors are 60%, 30%, and 10%, respectively. We observe that two types of network patterns emerge under allocation proportional to significance of domains. Figure 6.23(a) depicts that the number of communities in each domain is proportional to their importance. However, the communities with most resources granted may not be as successful as expected, as exhibited in Figure 6.23(b) by relatively small number of red communities. A potential reason is that the cluster of red domains interacts in high frequency with the

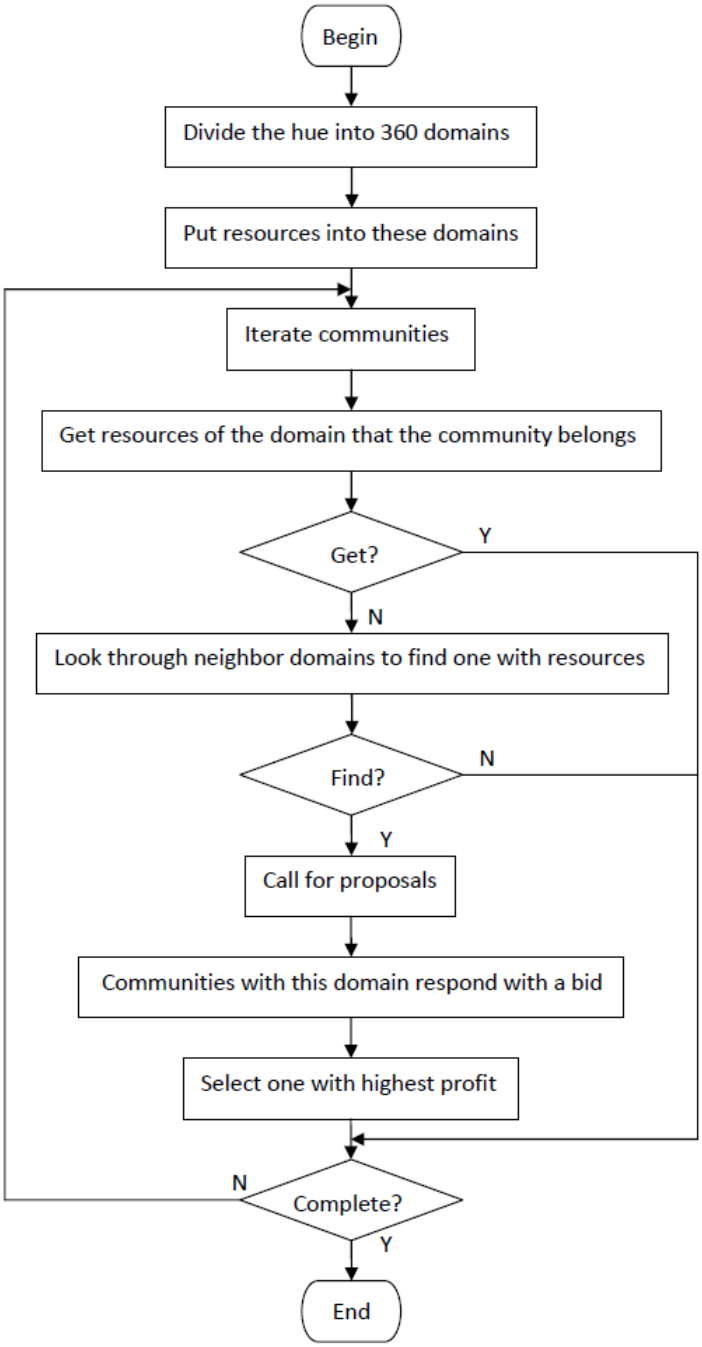


Figure 6.20: Flow Chart of P2P Lending

cluster of green and blue domains. This interaction could have incurred significant resource cost during the learning process, resulting in decreased number of red domains.

To explore the impact of the relation between domains on the final network pattern, we design an experiment where the blue communities and the green communities cannot

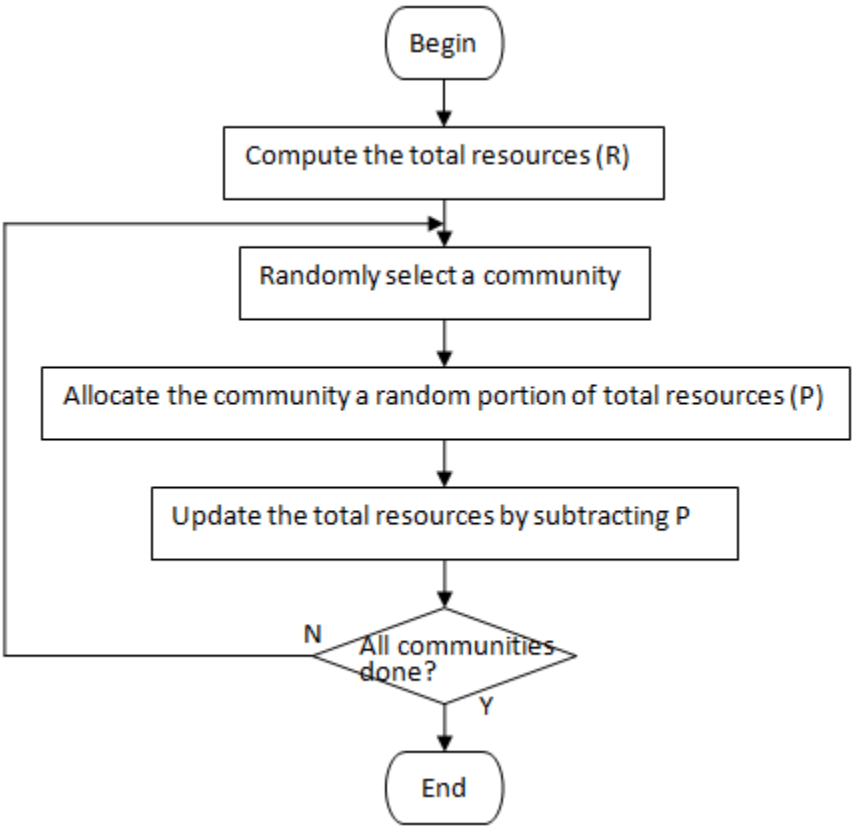
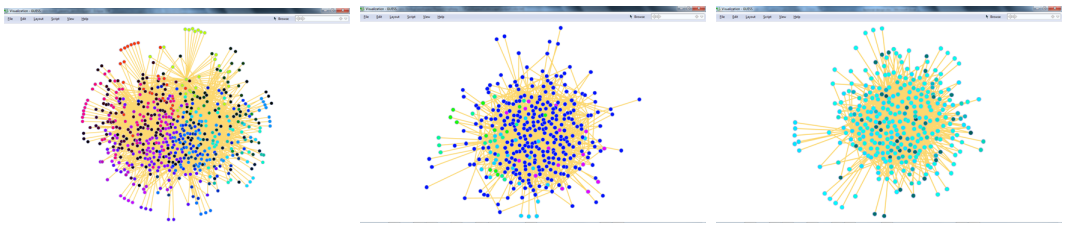


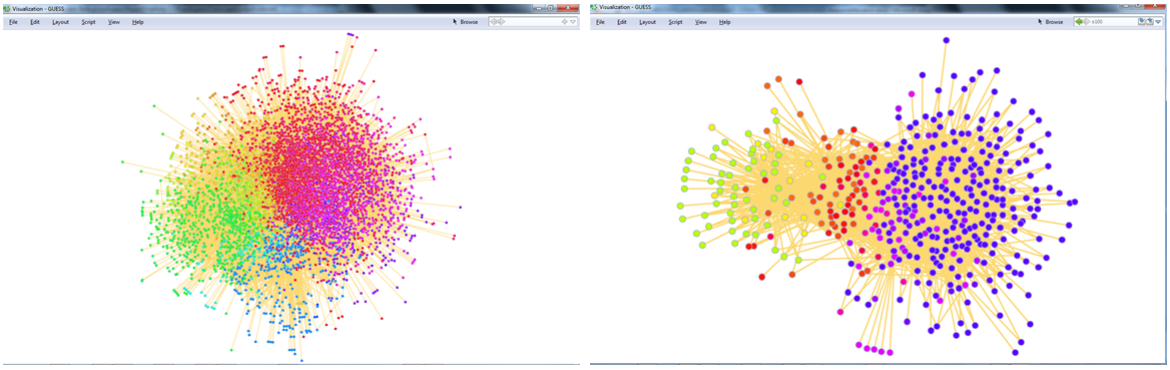
Figure 6.21: Flow Chart of Random Allocation



(a) Strategy 1 (b) Strategy 2 (c) Strategy 3

Figure 6.22: Strategy 1 vs. Strategy 2 vs. Strategy 3

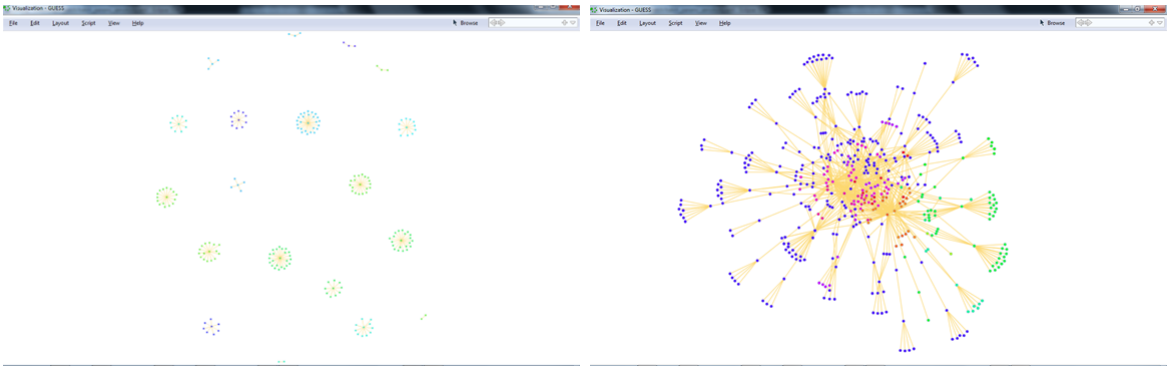
connect with each other so that the red communities are aligned between the green and the blue communities. The simulation model is run thirty times with different random seeds, and the final network patterns can be categorized into two types, which are shown in Figure 6.24. In Figure 6.24(a), all red communities are changed to either blue or green communities. Because blue communities and green communities cannot connect with each other, the final pattern of type 1 is an isolated cluster of blue or green communities. In Figure 6.24(b),



(a) Expected Case (b) Unexpected Case

Figure 6.23: Allocation Proportional to Importance of Domains

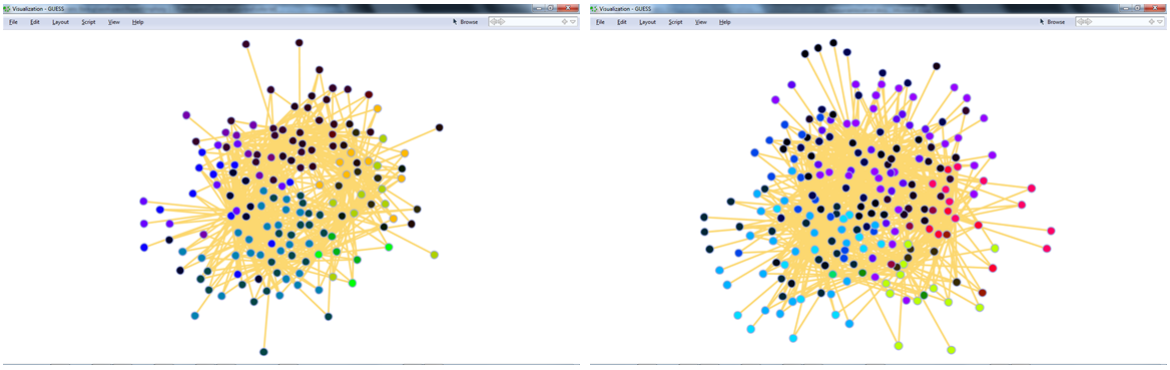
red communities are pulled by both green and blue communities so that more resources are consumed and their maturity is developed slowly. In addition, because the current model is based on Homophily theory, the intensity of influences from peers is proportional to their similarity rather than how many resources the community holds. Therefore, red communities cannot thrive under such network alignment, although most resources are allocated to the red communities.



(a) Isolated clusters of blue or green communities (b) Red communities struggling to thrive

Figure 6.24: Patterns in Network Configuration Experiment

Figure 6.25 depicts the comparison of competitive allocation and P2P lending allocation. We do not observe significant difference on diversity, which suggests that higher success rate, as observed in P2P lending, may not result in significantly higher diversity.



(a) Competitive Allocation (b) P2P Lending

Figure 6.25: Competitive Allocation vs. P2P Lending

6.4.3 Variety vs. Resource Allocation Strategy

For the resource allocation strategy, two aspects are considered. One is the resource size, and the other is the allocation mechanism, which is defined in terms of the seven categories listed in the previous section. For the resource size, two options are examined: fixed amount of total resources and dynamic allocation with technology transferring. As shown in Table 6.3, we examine twelve allocation strategies based on the combination of resource size and allocation strategies (except competitive and P2P lending).

The experiments with each allocation strategy are conducted 30 times and the average variety is recorded as shown in Figure 6.26. Variety can be computed as the number of clusters of communities within the environment. Each cluster is composed of similar communities in terms of their hue.

Based on Figure 6.26, we discern the following:

1. Key area investment with technology transferring (A8) results in the highest variety. This is similar to the case where domains with lower priority still have potential to advance, yet the environment promotes development of domains related to priorities.

2. Uniform allocation (A1, A2) leads to higher variety compared to resource allocation proportional to contributions (A3, A4).

Table 6.3: Allocation Strategies

Symbol	Resource Allocation Strategies
A1	Uniform allocation with fixed external resource
A2	Uniform allocation with technology transferring
A3	Allocation proportional to contribution with fixed external resource
A4	Allocation proportional to contribution with technology transferring
A5	Allocation proportional to cluster size with fixed external resource
A6	Allocation proportional to cluster size with tech transferring
A7	Allocation proportional to importance of domains with fixed external resource
A8	Allocation proportional to importance of domains with technology transferring
A9	Competition allocation
A10	P2PAllocation
A11	Random allocation with fixed external resource
A12	Random allocation with technology transferring

3. Competitive allocation (A9) results in higher variety than P2P lending (A10). The underlying reason is that P2P lending allows communities to share resources, which in turn causes both lender and borrower communities to fade out under limited resources.

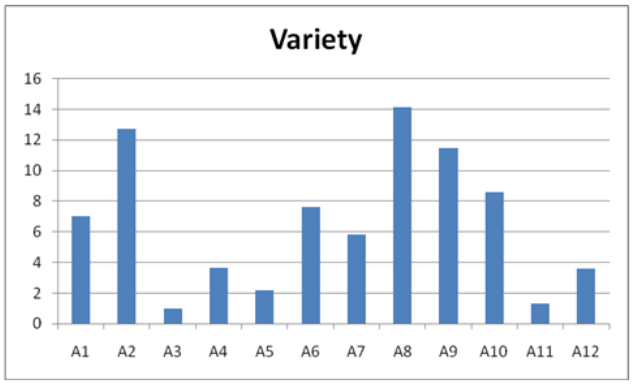


Figure 6.26: Variety vs. Allocation Strategies

Chapter 7

Comparison of Communication Theories in Terms of Innovation Performance

7.1 Introduction

Under the globalization driven by advances in computer and communication technology, the flow of information that transmits through communication networks is independent of space and time, because people can share knowledge and make contributions simultaneously anywhere in the world [61]. Furthermore, the mechanisms for the emergence and evolution of communication networks can be abstracted into several communication theories. Although communication theories describe the internal mechanisms of social communication networks, little research has been conducted to implement computational models using them. Meanwhile, there is no research undertaken for comparison of communication theories in term of their effects on innovation performance.

Communication networks and the organizational forms of the 21st century are undergoing rapid and dramatic changes [32]. There exist theories that focus on the role of interaction mechanisms in explaining the emergence and evolution of communication networks. One advantage of analyzing system dynamics from the perspective of socio-technical networks is the ability of data analysis at various levels such as individual, dyad, triad, organizational, and interorganizational [61]. Homophily, preferential attachment, and exchange theory are mainly about the dyad relationship where a communication tie from community A to community B can be predicated by the communication tie from community B to community A. On the other hand, balance and structural hole theory analyze the triad relationship, where the communication tie between community A and B can be predicated by the third community C that is associated with both A and B. In addition, these theories distinguish with each other in terms of studying internal mechanisms of communication networks from

different perspectives, which include communities' traits, self-interest, and discrepancy in resources. Homophily, preferential attachment, structural hole, exchange, and balance theory analyze communication network at different levels and from different perspectives, hence it is important to use them to model the dynamics between communities and compare them in terms of network patterns and innovation metrics.

7.2 Homophily

Homophily theory explores the emergence of communication networks based on the similarities of network members' traits [61]. Similarity contributes to ease communication, foster trust and increase the predictability of behavior [15]. On the basis of homophily, communities select others who are similar to communicate.

7.2.1 Model Design

The following process is from the viewpoint of a community called the current community. At each time interval, the current community randomly selects another community to communicate based on their similarities, which means that the higher similarity between the current community and the target community results in the higher probability of building communication between them. Figure 7.1 depicts the process of communication guided by the Homophily theory.

The following equations describe how to update influences of neighbors based on homophily.

$$\begin{cases} W_{ji,t} &= W_{ji,t-1} + C_W \times I_{ji,t} \times (1 - W_{ji,t-1}) \qquad if\ I_{ji,t} \geq 0 \\ W_{ji,t} &= W_{ji,t-1} + C_W \times I_{ji,t} \times W_{ji,t-1} \qquad otherwise \end{cases} \tag{7.1}$$

where $W_{ji,t}$ is the influence of neighbor j at the current time. C_W is a number between 0 and 1 and is inversely proportional to inertia (resistance to change in a community). $I_{ji,t}$ is the intensity of change in the influence, which is defined as:

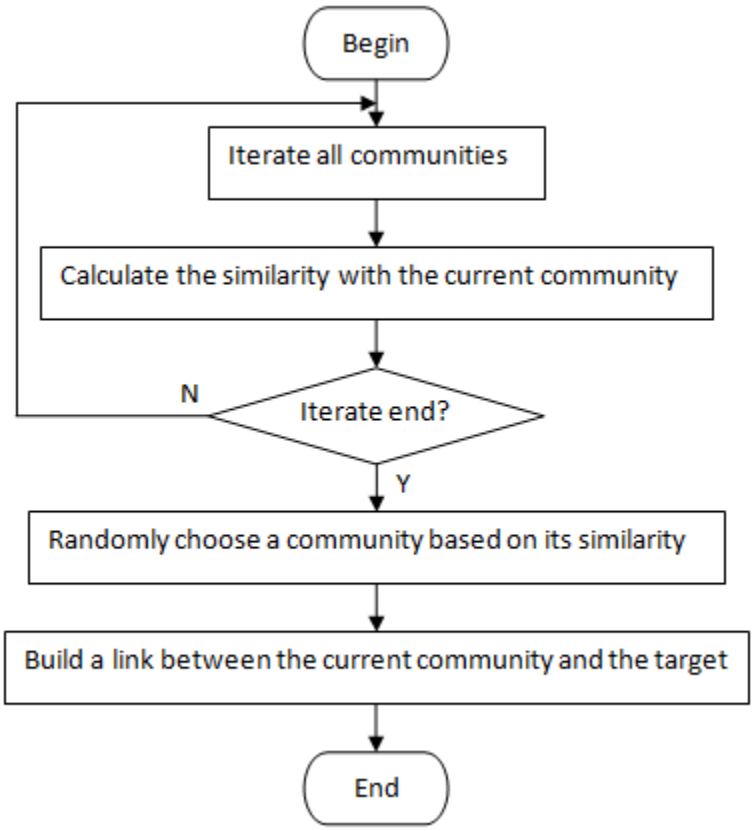


Figure 7.1: Process of Communication using Homophily Theory

$$I_{ji,t} = (1 - D_{ji,t})^4 - (1 - \overline{D_{i,t}})^4, \tag{7.2}$$

where $D_{ji,t}$ is the dissimilarity which is equal to the distance between community i and community j in terms of current hue at the time t whose equation is 3.8. $\overline{D_{i,t}}$ is the average distance between community i and all of the neighbors at the time t. This function grows much faster when dissimilarity between i and j becomes smaller in comparison to average dissimilarity, resulting in higher intensity $I_{ji,t}$.

The equation for the dissimilarity between community i and j is defined as follows:

$$D_{ji,t} = Dissimilarity(H_{i,t}, H_{j,t}), \tag{7.3}$$

where $H_{i,t}$ is the hue of community i at the time tick t. $H_{j,t}$ is the hue of community j at the time j.

$$Dissimilarity(x,y) = \begin{cases} \frac{|x-y|}{180} & if\ |x-y| \leq 180 \\ \frac{360-|x-y|}{180} & otherwise \end{cases} \tag{7.4}$$

7.2.2 Validation

We designed an experiment where there are three communities A, B and C. The similarity between community A and B is 80%, while the similarity between A and C is 20%. The experiment tries to see how likely community A would like to communicate with B or C under homophily theory. The simulation model runs 100 times, among which community A communicates with community B 97 times as shown in Figure 7.2. To amplify the difference of similarity, a square operation is used, which in turn leads community B to have a much higher probability of being communicated by community A than community C.

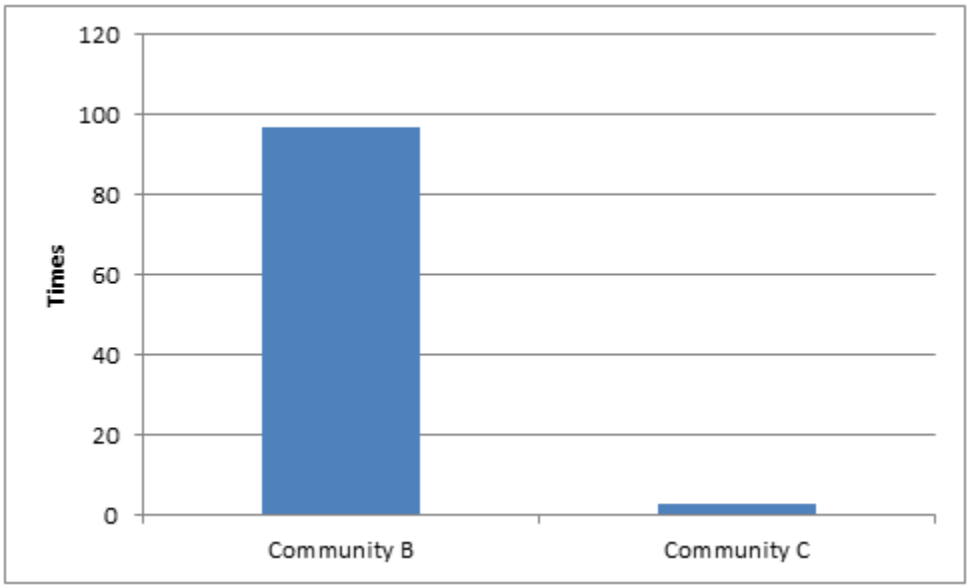


Figure 7.2: Communication Frequency vs. Similarity

7.3 Structural Hole

Structural holes are those places where communities are not connected so that other communities may exploit the places by investing their social capital to indirectly link two or more unconnected communities [61]. The community that fills the structural hole becomes a broker in relationships among others. As shown in an early Italian saying "between two fighters, the third benefits" [17], the community acting as broker can benefit from different knowledge and expertise of other communities. There are two kinds of information benefits for broker identified in [19]: access and timing. Access means getting information that others may not get. Timing refers to getting information earlier than peers.

7.3.1 Model Design

The following process is from the viewpoint of a community denoted as C_0. At each time interval, a community C_1 not connected to C_0 is randomly selected firstly. Then a community C_2 not connected to C_1 is randomly selected. Finally two links between C_0 and C_1, C_0 and C_2 are built respectively. The process is depicted in Figure 7.3.

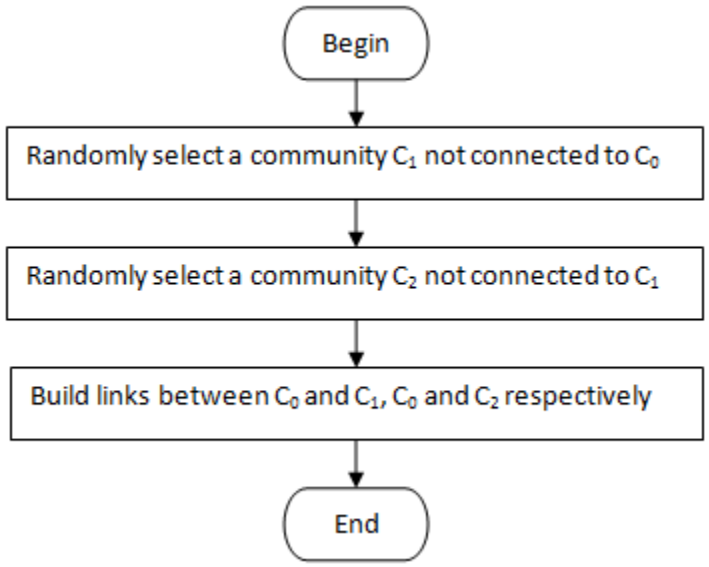


Figure 7.3: Process of Communication using Structural Hole Theory

7.3.2 Validation

Burt in [17] points out that individual's effective network size determines its success potential; that is, individual with larger effective network is more likely to succeed. The ties among a person's network partners attenuate the effective network size, which gets to the max value (i.e., 1) when partners are not connected to one another. On the other hand, the effective network size becomes the min value (i.e., 0) if partners are connected to one another. Meanwhile, clustering coefficient measures how close are the neighbors to being a clique. However, clustering coefficient is equal to 1 if neighbors are fully connected. So, we use *1 - clustering coefficient* to represent network effective size to get 0 for fully connected networks and 1 for isolated networks. In addition, communities' success is reflected by their resources.

The following experiment is to capture the relation between effective network size and resources, where the effective network size and resources of each community are recorded at each time step. Then, the average value of resources is computed with respect to the same effective network size. Based on Figure 7.4, we can observe that resources held by communities increase with communities' effective network size. The potential reason is that larger effective network size means more opportunities around the community, and hence increase in its resource levels.

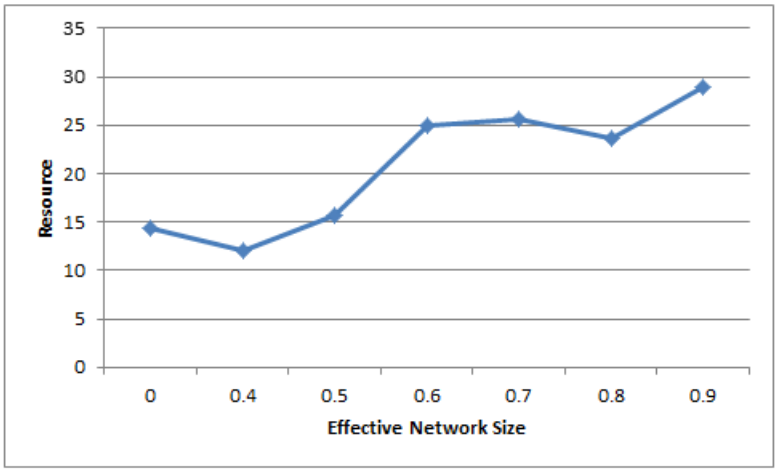


Figure 7.4: Resource vs. Effective Network Size under Structural Hole Theory

7.4 Preferential Attachment

The preferential attachment is a process where resources are distributed among individuals according to how much they already have, i.e., rich get richer. Communities may like to connect to others with more resources in order to steer their own directions to get potentially more resources. On the other hand, communities may intend to connect to peers with larger number of links that indicates a more central position and larger influences within the network. So, there are two branches of preferential attachment theory: preferential attachment based on resources, and preferential attachment based on links.

7.4.1 Preferential Attachment Based on Resources

The preferential attachment based on resources means that the community with more resources are more likely to be communicated. Figure 7.5 shows the process of building connections under the preferential attachment based on resources.

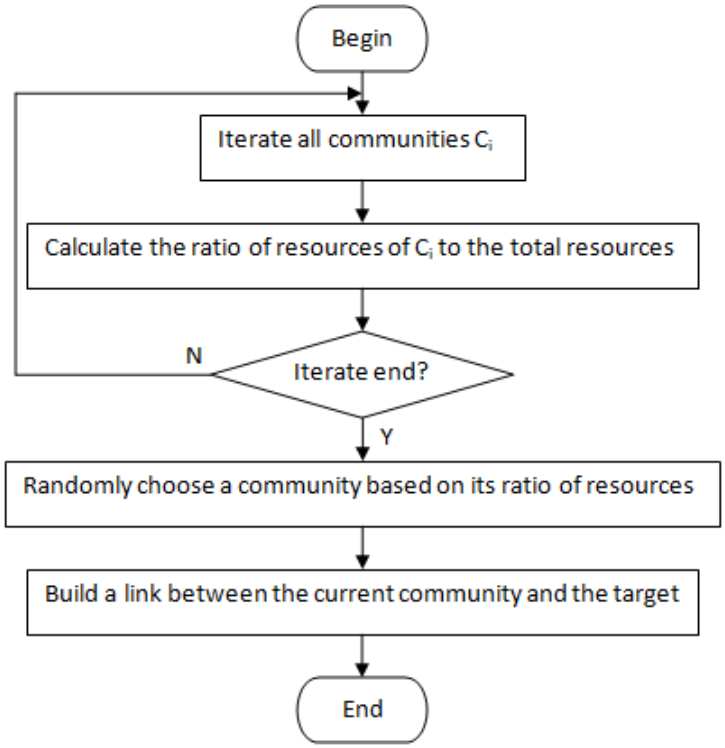


Figure 7.5: Communication Process of Preferential Attachment based on Resources

Under the preferential attachment based on resources theory, communities with more resources have larger influences on others. The following equation describes how to update influences of neighbors using the preferential attachment based on resources.

$$W_{ji,t} = \frac{R_{j,t}}{\sum_{k=1}^{N} R_{k,t}}, \tag{7.5}$$

where $W_{ji,t}$ is the influence of neighbor j on community i at time t. $R_{j,t}$ is the resources of community j. N is the total number of communities.

7.4.2 Preferential Attachment Based on Links

The preferential attachment based on links means that the community with more links are more likely to be communicated. Figure 7.6 delineates the process of using preferential attachment based on links to build connections among communities.

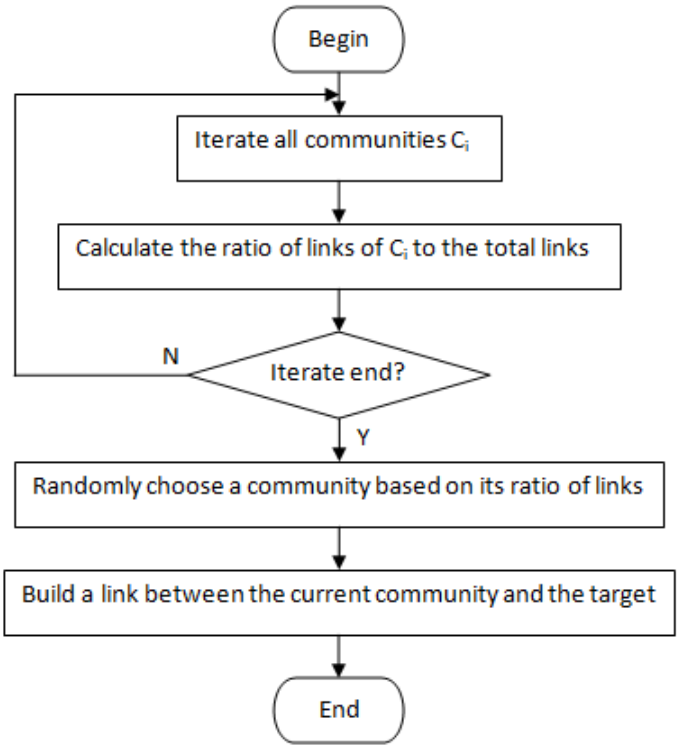


Figure 7.6: Communication Process of Preferential Attachment based on Links

Under the preferential attachment based on links theory, communities with more links have larger influences on others. The following equation describes how to update influences of neighbors using the preferential attachment based on links.

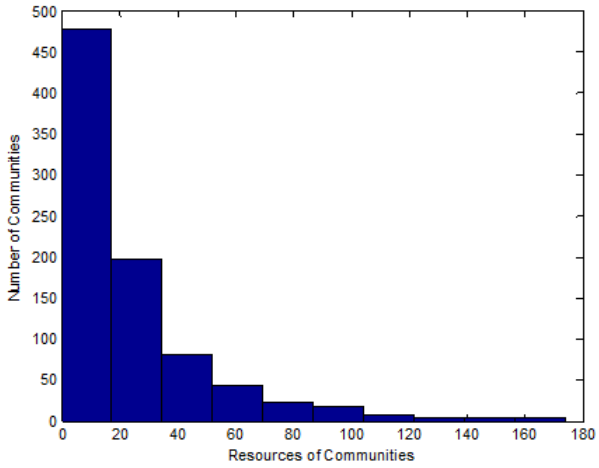
$$W_{ji,t} = \frac{L_{j,t}}{\sum_{k=1}^{N} L_{k,t}}, \tag{7.6}$$

where $W_{ji,t}$ is the influence of neighbor j on community i at time t. $L_{j,t}$ is the number of links of community j. N is the total number of communities.

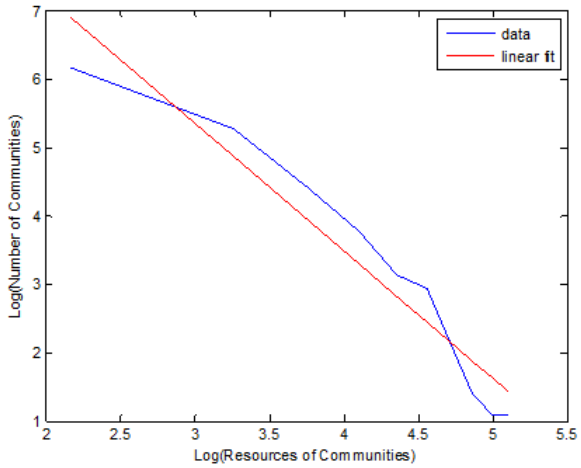
7.4.3 Validation

Under suitable circumstance, preferential attachment can generate power law [106]. For preferential attachment based on resources introduced in section 7.4.1, we run the simulation 30 times and output the resources of communities at the end of each run. Figure 7.7(a) depicts the inequality of communities in terms of resources. Most communities hold the relatively few resources, while a small part of communities hold the relatively many resources. This observation is indicative of the presence of power law. Figure 7.7(b) shows the relationship between the log value of number of communities and their resources, as well as the corresponding linear regression curve. Since the R^2 for this fitting is 0.92, the Colorscape model suggests the presence of power law in resource distribution.

For preferential attachment based on links introduced in section 7.4.2, we run the simulation 30 times and print out the number of links of communities at the end of each run. Figure 7.8(a) depicts the inequality of communities in terms of links. Most communities have the relatively few links, while a small part of communities hold relatively many links. Because the Colorscape model has a mechanism of specialization where a new community is generated and a link between original and new community is built, the creation of this link does not follow preferential attachment theory, which in turn results in the communities with two links are more than those with one link. This observation is indicative of the presence of power law. Figure 7.8(b) shows the relationship between the log value of number of



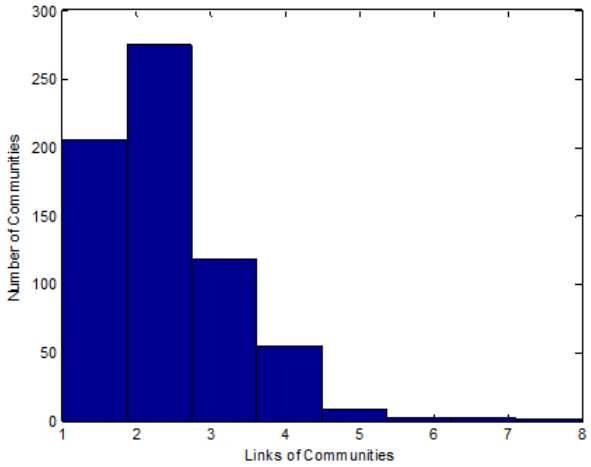
(a) Histogram of Communities' Resources



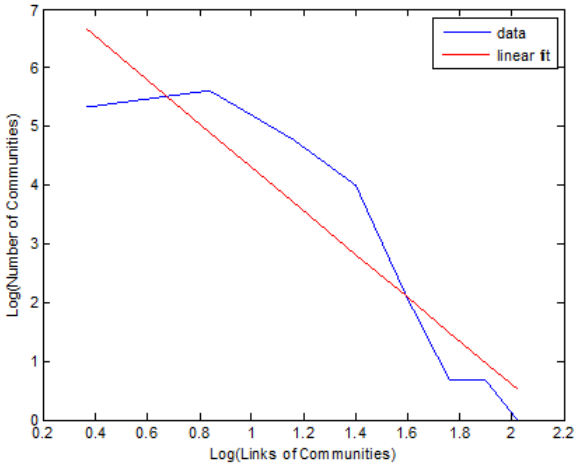
(b) Linear Regression of Logarithmic Value of Resources

Figure 7.7: Communities' Resources

communities and their links, as well as the corresponding linear regression curve. Since the R2 for this fitting is 0.85, the Colorscape model suggests the presence of power law in link distribution.



(a) Histogram of Communities' Links



(b) Linear Regression of Logarithmic Value of Links

Figure 7.8: Communities' Links

7.5 Balance Theory

7.5.1 Model Design

Heider's balance theory [38] states: "my friend's friend is my friend; my friend's enemy is my enemy; my enemy's friend is my enemy; my enemy's enemy is my friend", which means friends have similar attitudes, while enemies have different opinions on the third object. As scientific communities keep creative, they may desire to communicate with neighbors including both similar and dissimilar communities in order to maintain a highly diverse environment. When using balance theory to study the activities of scientific communities, communities try to keep the interactions balanced in terms of disciplines among communities that are communicated with.

To illustrate the triad relationship that balance theory focuses on, let us consider the following example. From the perspective of community A, the probability of building link between A and B is determined by the peer communities associated with A. If there are more neighbor communities of A similar to A than neighbors dissimilar to A and the similarity between A and B is lower than the average, then it is more likely for A to communicate with B. On the contrary, if there are less neighbor communities of A similar to A than dissimilar neighbors and the similarity between A and B is lower than the average, then it is less likely for A to communicate with B. These relations are described in Table 7.1.

Table 7.1: Illustration of Building Links based on Balance Theory

Similarity between A and B	**Number of Neighbors Similar to A**	**Probability of Building Link**
High	More	Low
High	Less	High
Low	More	High
Low	Less	Low

The similarity between communities is determined by their disciplines, which is defined as Equation 7.3 and 7.4.

When using balance theory to select peers to communicate, if more communities are above the average dissimilarity, either decrease the influence of a community with above average dissimilarity or randomly select another community with below average dissimilarity in order to reach balance. On the other hand, if more communities are lower than the average dissimilarity, either decrease the influence of a community below average dissimilarity or randomly select another community with above average dissimilarity in order to reach balance. The process of using balance theory to build the communication network is shown in Figure 7.9.

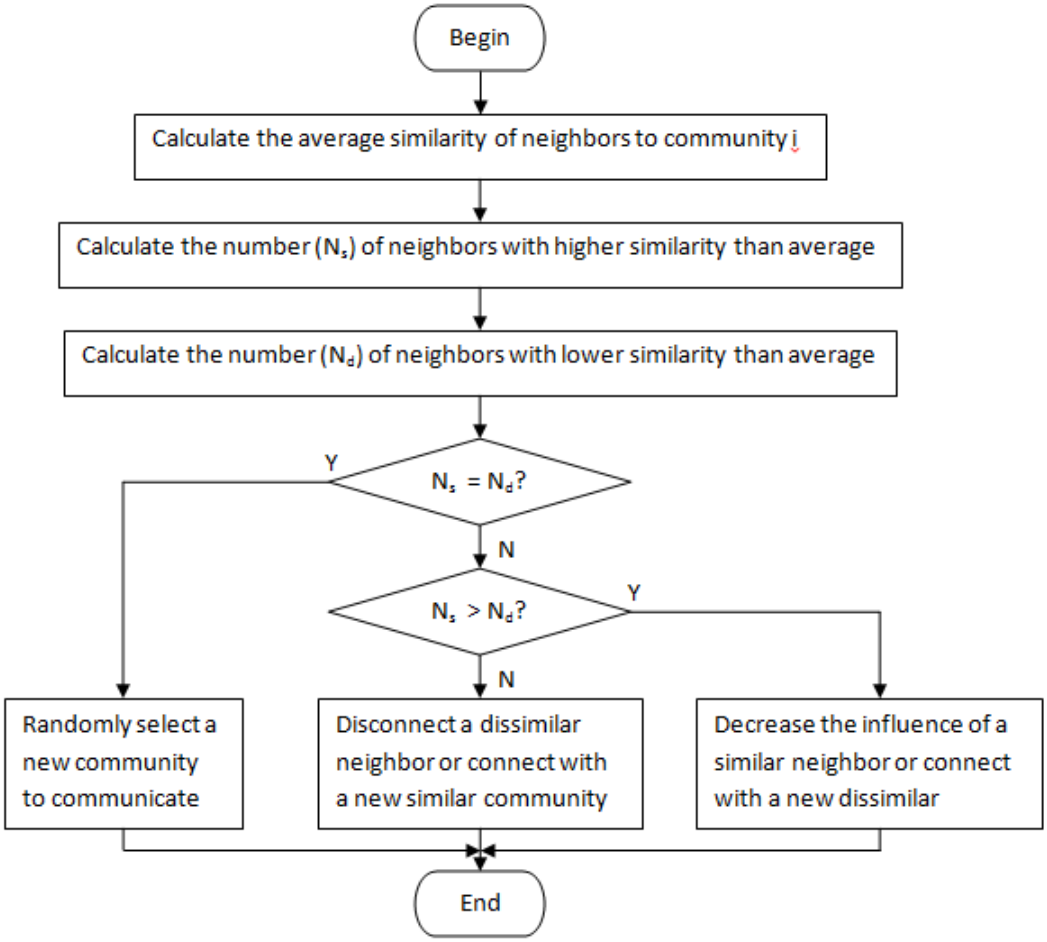


Figure 7.9: Process of Communication using Balance Theory

The following discussion delineates the process of updating influence of neighbors using balance theory. The balance is defined as the equilibrium where the discipline of communities

converge to. In cases where there are differences in opinion between communities, their needs for balance motivate them to increase their communication frequency with one another to reach an agreement. Given the assumption that communities' opinions are based on their discipline, the larger difference of discipline results in the larger difference of opinion. So, communities would like to communicate more with communities with larger difference in order to reach the balance. The Colorscape model is based on boundary processes that drive interacting communities to move toward each other, which in turn reduces their differences. Such interactions guided by balance theory and boundary processes cause the dissimilarities between communities to change dynamically. During each interaction, if the dissimilarity of a community's neighbor j is above the average between the community i and all community $i's$ neighbors, the communication frequency between community i and j increases. Otherwise, their communication frequency decreases. Equation 7.7 describes how the influences of neighbors are updated.

$$\begin{aligned} \Delta W_{ji,t} &= sin(\frac{\pi}{2}(D_{ji,t} - \overline{D_{i,t}})), \\ W_{ji,t+1} &= W_{ji,t} + \Delta W_{ji,t}, \end{aligned} \tag{7.7}$$

where $D_{ji,t}$ is the dissimilarity between community i and its neighbor j at time t. $\overline{D_{i,t}}$ is the average dissimilarity between community i and all its neighbors. $W_{ji,t}$ and $\Delta W_{ji,t}$ are the influences of community j on community i and the increment of such influences, respectively.

In the extreme case, the maximum $D_{ji,t}$ is 1 and $\overline{D_{i,t}}$ is close to 0, then the maximum of $D_{ji,t} - \overline{D_{i,t}}$ is 1. Under this case, $\Delta W_{ji,t} = sin(\frac{\pi}{2}(1-0)) = 1$. On the other hand, the minimum $D_{ji,t}$ is 0 and $\overline{D_{i,t}}$ is close to 1. Then the minimum $\Delta W_{ji,t} = sin(\frac{\pi}{2}(0-1)) = -1$. Figure 7.10 depicts change in $\Delta W_{ji,t}$ over $D_{ji,t}$ given different $\overline{D_{i,t}}$.

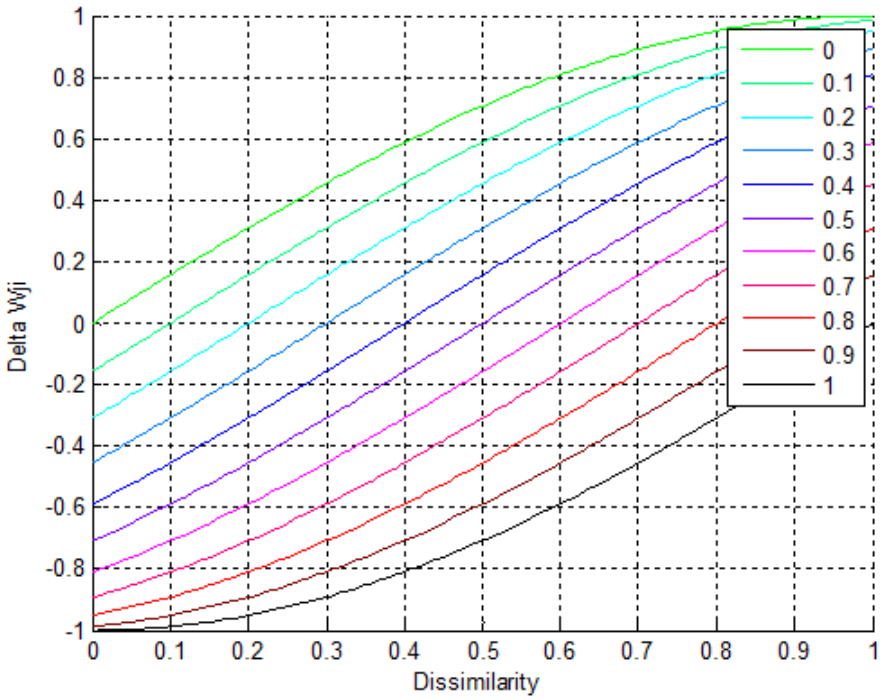
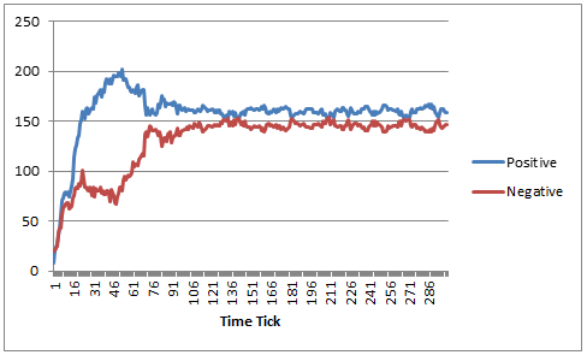


Figure 7.10: Influences Change with Dissimilarity

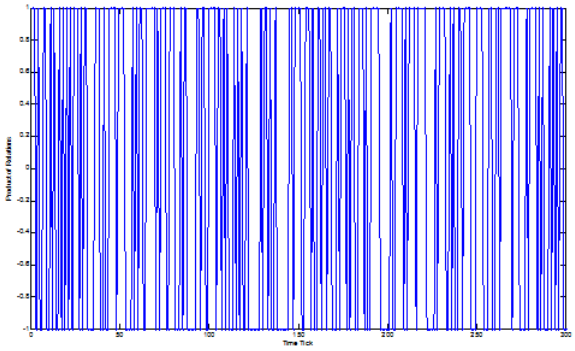
7.5.2 Validation

In [54], Lane posits that the relationship between organizations can be categorized into positive and negative. It is pointed out in [107] that the condition for a network to be balanced is that the product of relation ties between organizations is positive. According to homophily theory, communities would like to communicate with those communities similar to them. So, for community i, the communication ties between community i and others with higher similarity than average are viewed as positive relationship. On the contrary, the communication ties between community i and others with lower similarity than average are viewed as negative relationship. We plot positive and negative relationship, which correspond to the number of communities with higher and lower similarity respectively, which is shown in Figure 7.11(a). Also the product of all relation ties is plotted, where each positive tie is represented by 1 and each negative tie is represented by -1, which is shown in Figure 7.11(b).

Based on these two figures, we can observe that the product of all ties oscillates between -1 and 1. In complex adaptive systems, there are three fundamental kinds of attractors: fixed



(a) Positive and Negative Relations Change over Time



(b) The Product of Relations

Figure 7.11: Relations under Balance Theory

point attractor, limit cycle attractor, and chaos attractor. Figure 7.11(b) shows there is a limit cycle attractor existed, in which the exact state of the system cannot be predicted, although we know it will be either -1 or 1. It means that the system reaches a dynamic balance compared with the fixed balance in [107], where the number of positive and negative ties keeps the same dynamically, shown in Figure 7.11(a). Such a dynamic balance makes communities satisfy with their status, and the communication network formed by them is balanced.

7.6 Exchange Theory

7.6.1 Model Design

According to exchange theory, the necessary condition for the realization of a network tie is the discrepancy in resource. In the Colorscape model, the discrepancy in resource between communities is reflected by the brightness component of the HSB color model. When a community cannot achieve enough resources by solving problems in its own domain, it tries to solve inter-disciplinary problems by collaborating with peers. Once community i finds an inter-disciplinary problem with potential funds, community i will ask other communities for collaboration. Another community j who can solve the problem may be willing to collaborate. Then a link between community i and j is created. In such cases, what are exchanged

between communities are resources based on the knowledge and skills of communities. Different from the balance theory, the exchange theory interprets communication at the dyad level, which means the communication link is determined by the two parties involved in the collaboration.

There is a website named *InnoCentive* [45] that uses challenge-driven innovation mechanism to bridge companies that have problems to be solved and users who would like to capitalize their knowledge. When a company has a problem to be solved, the company may try to post the problem on this website. Those users who are interested in this problem may submit their solutions, one of which will be selected by the sponsor company. The author of the selected solution is rewarded.

P2P lending process is defined as follows:

1. Divide the discipline into 36 domains (i.e., 10 degree per domain), and put some problems in each domain.

2. If the domain a community inhabits has problems, then the community receives the corresponding funding up to a threshold, i.e., the maximum value a community can achieve per time step.

3. If the domain a community inhabits does not have problems, the community looks through neighbor domains until the community finds a domain with problems. Then the community calls for a proposal to collaborate.

4. All other communities will receive the invitation. Only those communities within the domain will respond with a bid that shows the ratio of funding the community gets to those resources the sponsor gets.

5. The community who initializes the call for proposals chooses the bid with the highest ratio.

There are three kinds of collaborations that may occur during the P2P lending process: two parties within discipline, two parties cross discipline, and triple parties. Type I collaboration takes place when the sponsor community receives reply from communities within the domain where the problem exists. Type II collaboration is interdisciplinary collaboration, which happens when the domain of the problem is just between the sponsor and the responder communities. Type III collaboration occurs when there are two responders whose domains are just adjacent neighbor of the problem. These collaborations are illustrated in Figure 7.12, where the purple and blue circle represent sponsor and responder respectively.

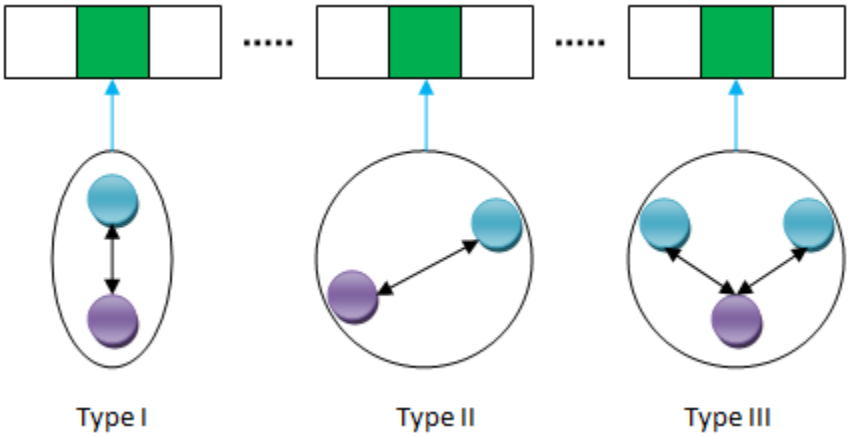


Figure 7.12: P2P Collaborations

If there are several communities that can collaborate, a community may choose one that can help reduce the dependency on other communities. In order to reduce the dependency, a community i seeks to forge links with communities not connected with community i. In [61], Monge points out the network extension, which means that organizations can seek to increase the number of exchange alternatives by creating new network links.

Each communication theory has two functions, one of which is to build a communication network. The other is to update the weights of neighbors. When using the exchange theory to update links connecting to neighbors, the weights increase when an exchange occurs during a time interval. Otherwise, the weights of links connecting to neighbors decrease if no exchanges happen between them during a time interval.

7.6.2 Validation

7.6.2.1 Resource Accessibility

Brass [13] claims that the organization's access to resources is reflected by closeness centrality that refers to the extent to which people, group, and organizations can reach all others in a network through a minimum of intermediaries. It means that higher closeness will have more resources. Further, Brass [14] found that the measure of centrality correlated with reputational measures of power, which in turn influences the organization's ability to achieve resources.

The closeness centrality is calculated as shown in Equation 7.8 [65].

$$C_i = \frac{N-1}{\sum_{j=1}^{N-1} d_{i,j}}, \tag{7.8}$$

where $d_{i,j}$ is the minimum distance between community i and j. N is the total number of communities in this network.

We design an experiment where the simulation of the Colorscape model was replicated 30 times. At the end of each single run, the closeness centrality and resources of each community are recorded. For each level of centrality, the average level of resources of corresponding communities are computed. Figure 7.13 depicts the average resources of communities changing along with the closeness centrality of these communities.

Based on Figure 7.13, we can observe that the average resource level increases along with closeness centrality. The underlying reason may be that higher closeness centrality due to direct connections with peers, results in larger number of similar communities due to boundary processes. Thus, the community with higher closeness centrality is likely to survive, because it has more opportunities for collaboration with similar peers.

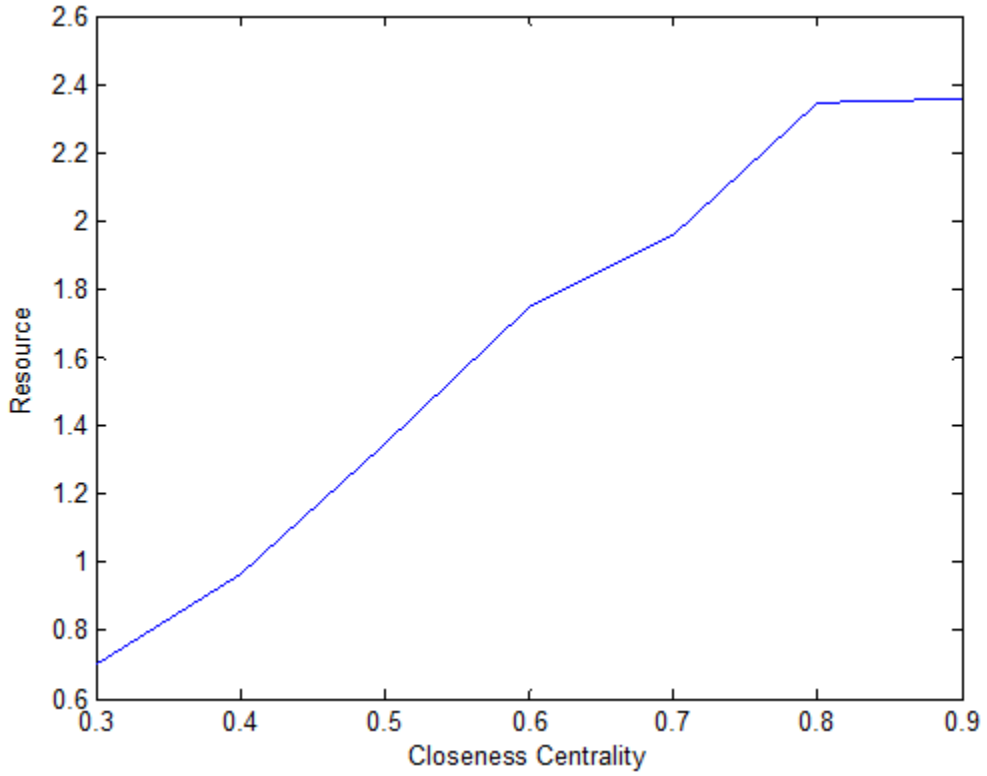


Figure 7.13: Resource Availability along with Closeness Centrality

7.6.2.2 Law of N-Squared

Krackhardt [51] identified the constraint of "Law of N-Squared", which simply notes that the number of potential links in a network organization increases geometrically with the number of people.

We increase the total number of communities from 10 to 190, and then count the number of communities that may build collaborations with respect to each community in the P2P lending. For example, considering that the total number of communities is 10, if the number of potential target communities for each community i is x_i, then the total number of potential target communities is $\sum_{i=1}^{10} x_i$. For each case, simulation model is run 30 times to get the average value. Figure 7.14 shows that the number of target communities geometrically increases with population.

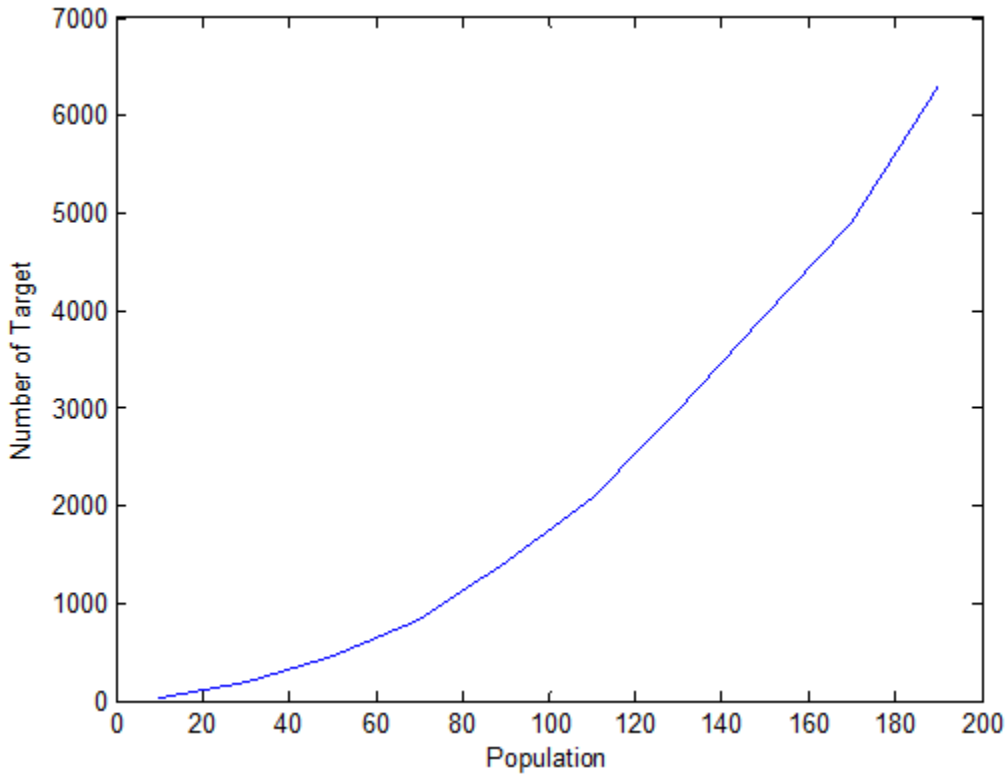


Figure 7.14: Number of Target Communities vs. Population

7.6.2.3 Iron law of Oligarchy

The other constraint identified by Krackhardt [51] is the "Iron law of Oligarchy", which is the tendency for groups and social systems, even fervently democratic ones, to end up under the control of a few people.

Figure 7.15 depicts how the network guided by the exchange theory evolves over time. The initial number of communities is 20. At the beginning, communities start communicating with each other. More and more communications happen over time so that communities tightly connect with each other. Furthermore, the boundary process pulls communities to move toward each other in terms of their domain. Because each cell in the resource landscape can only afford one community, communities within the same cell have to search for collaboration. Although such collaboration may happen, communities still fade out due to small portion of resources the collaboration community likes to share. Thus, the total

number of communities decreases until the last winner communities stay within their resource cells.

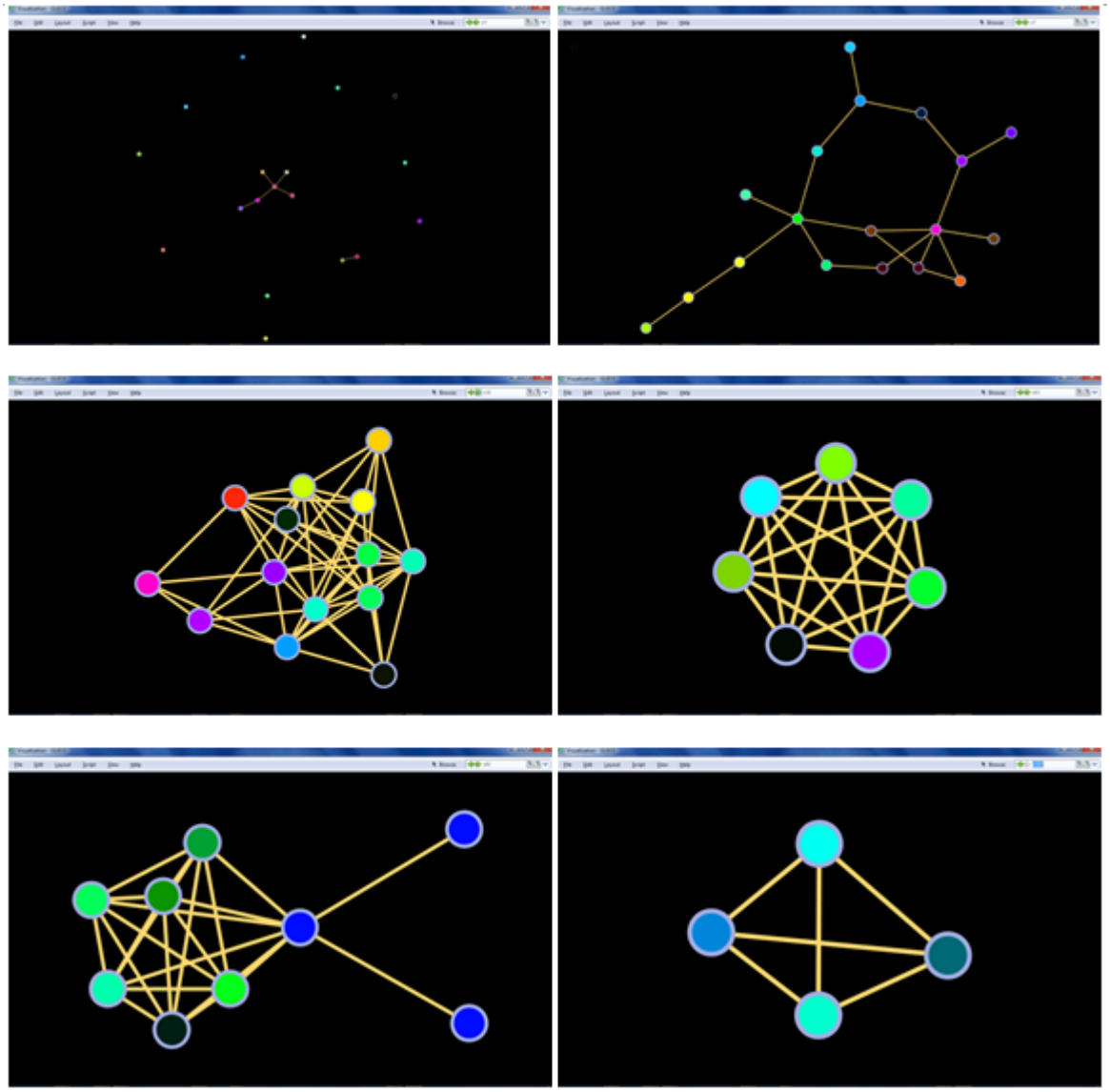


Figure 7.15: Emergent Networks over Time

7.7 Experiments on Communication Theories

In this section, experiments are conducted to investigate the impact of scientific community traits (i.e., receptivity, flexibility) and environmental constraints (i.e., external resources, communication strategies) on the innovation potential and performance of GPS.

Table 7.2 lists all the parameters that could be changed in the following experiments.

Table 7.2: Experimental Parameters

Name	**Default Value**
Carrying Capacity	60
Startup Funding	2
External Resource	2
Tolerance	0.6
Reorganization Tendency	0.5
Receptivity	0.5
Allocation Strategy	P2PAllocation
Communication Style	Balance
Communication Frequency	0.4
Threshold to Grow	0.5

7.7.1 Variety vs. External Resource

Figure 7.16 depicts how diversity changes with respect to the size of external resources injected into the environment. The abscissa indicates the amount of resources allocated to each community per time tick. For all the communication theories, the variety increases along with external resources. Also, the scale of variety is really similar, indicating that these communication theories do not have significant differences on the effects on variety under the P2P allocation strategy.

After setting the communication frequency to 0.1, we depict the change in variety over external resources in Figure 7.17. From this figure, we observe that variety is less sensitive to the external resources, since variety almost remains unchanged. Based on observations denoted by Figures 7.16 and 7.17, policy-makers need to be cognizant that increasing funding does not always help increase variety, especially for those communities with relatively low communication frequency.

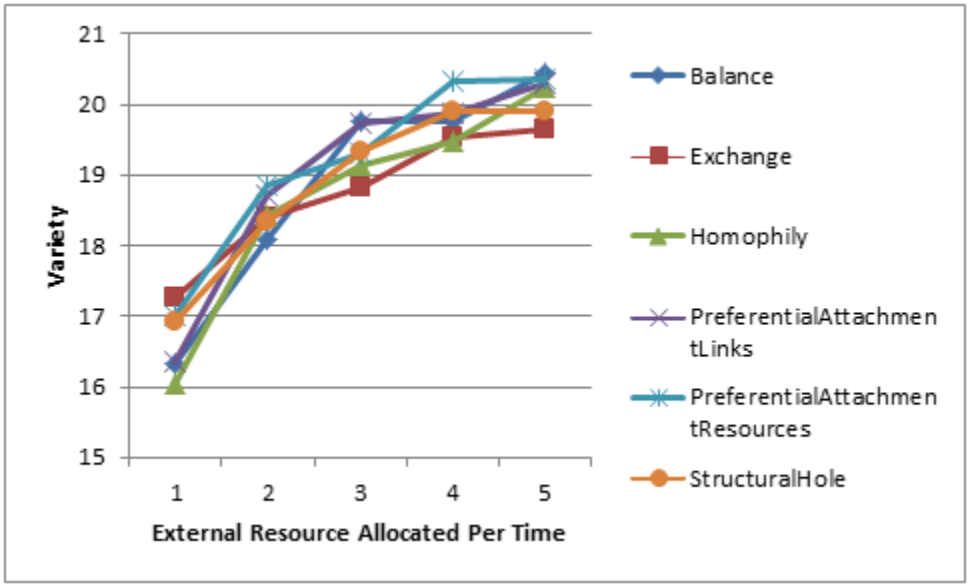


Figure 7.16: Variety vs. External Resources at Moderate Communication Frequency

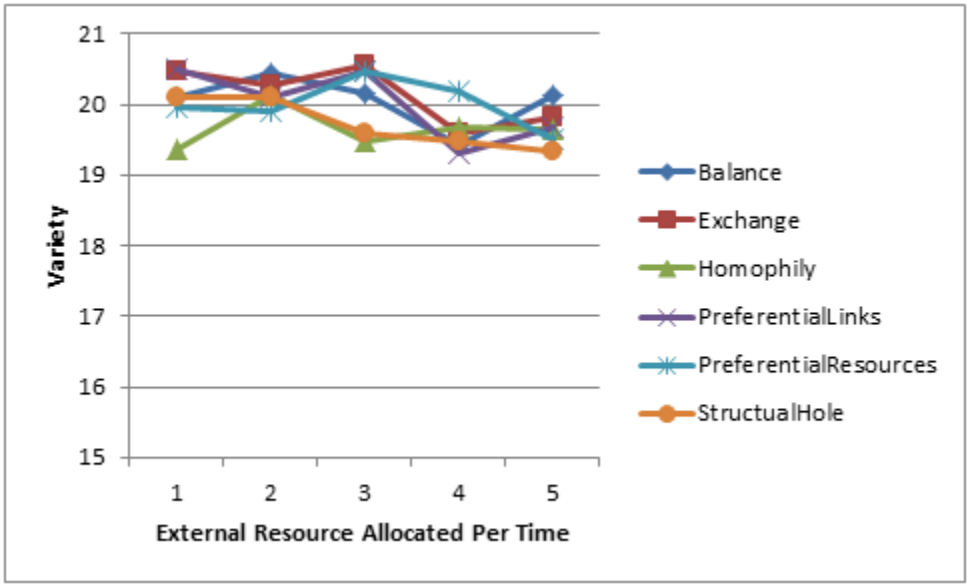


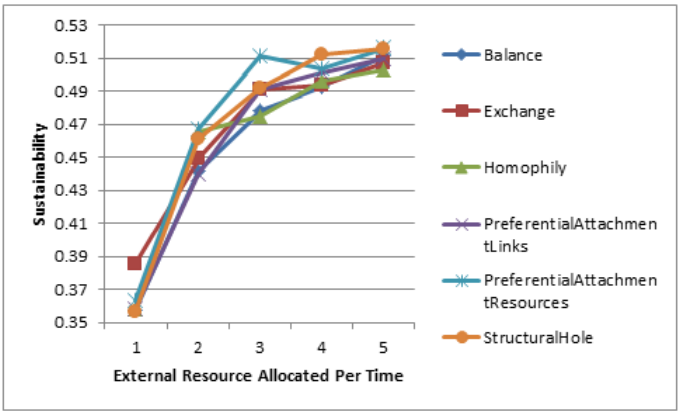
Figure 7.17: Variety vs. External Resources at Low Communication Frequency

7.7.2 Sustainability vs. Resource Availability

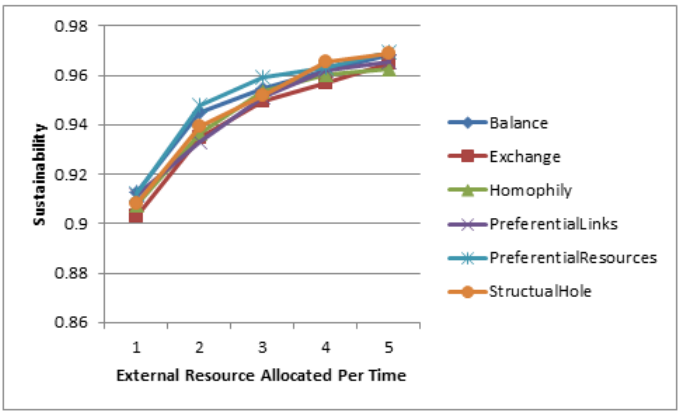
In ecology, sustainability refers to the ability of biological systems to remain diverse and productive over time. In the domain of creativity, sustainability can be interpreted as the effectiveness of communities in utilizing resources. So, we relate it to success rate, which measures the extent to which communities are effective in making use of resources to

improve their maturity, while maintaining themselves. Success rate is defined as the ratio of the number of active communities remaining at the end of simulation to carrying capacity.

Figure 7.18(a) shows that sustainability increases with external resources, while its rate of increase gradually decreases with increasing external resources. This suggests the presence of an asymptote, toward which sustainability moves with increasing external resource levels. If communication frequency is decreased to 0.1, the change in sustainability over resources is as depicted in Figure 7.18(b), in which similar trends are observed but with relatively large scale; that is, sustainability increases with decreasing communication frequency. As scientific communities can be viewed as artificial ecosystems, the communication frequency is similar to the evolution frequency. Lower evolution frequency leads to fewer species to be eliminated.



(a) High Communication Frequency

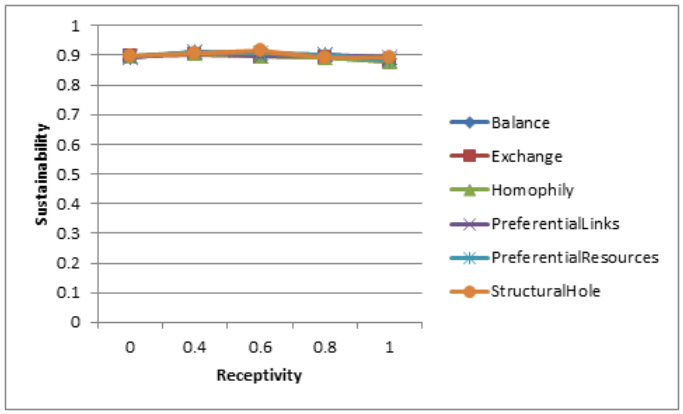


(b) Low Communication Frequency

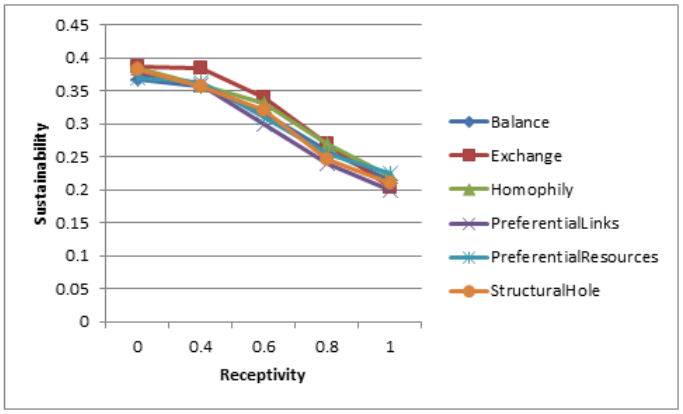
Figure 7.18: Sustainability vs. External Resources

7.7.3 Sustainability vs. Receptivity

Figure 7.19(a) shows the change in sustainability with respect to varying levels of receptivity at low communication frequency. Receptivity of a community is defined as the ratio of neighbor influence to inertia. We observe that sustainability almost does not vary with receptivity at low communication frequency. When the communication frequency increases to 0.7, sustainability vs. receptivity is depicted in Figure 7.19(b), which shows sustainability decreasing with increasing receptivity. Based on this comparison, decision-makers may develop policies to encourage communities to be more independent, if there are too many interacting activities in the domain.



(a) Low Communication Frequency



(b) High Communication Frequency

Figure 7.19: Sustainability vs. Receptivity

7.7.3.1 Variety vs. Receptivity

Figure 7.20 shows the change in diversity with respect to varying levels of receptivity. Receptivity of a community is defined as the ratio of neighbor influence to inertia. This figure shows that variety increases with increasing receptivity for all the theories. When receptivity is low, these communication theories lead to similar variety, because communication theories will not have effects on interactions between communities if communities have little influence on each other. In addition, in comparison to other theories, the exchange theory is less sensitive to receptivity in terms of variety. The potential reason is that the communication guided by exchange theory is based on distribution of resources on the innovation landscape, which is not directly related to variety. In addition, variety under balance, homophily, and structural hole theory increases monotonically with receptivity, the reason behind which is that these three communication theories build connections based on traits of communities that are directly related to variety.

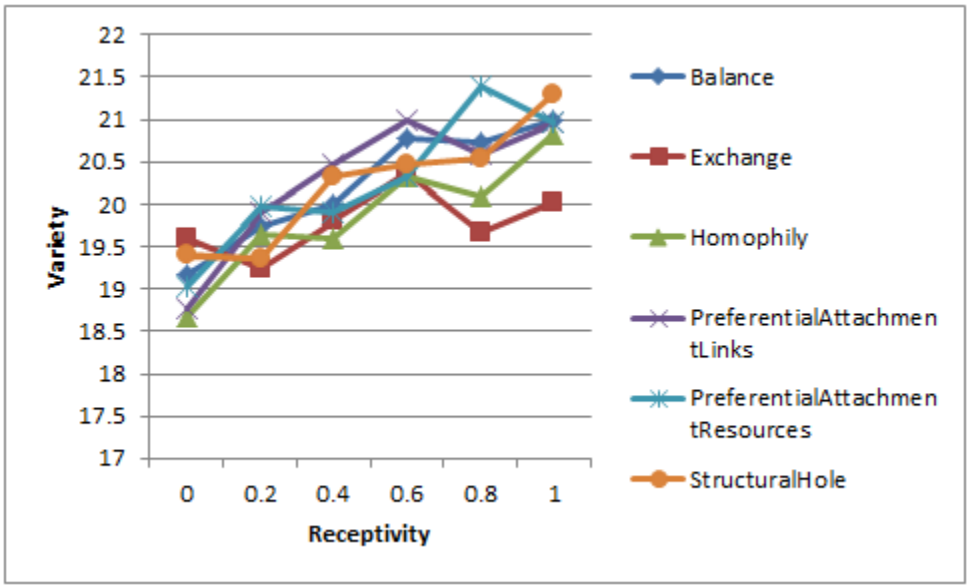


Figure 7.20: Variety vs. Receptivity under Low Communication Frequency

Figure 7.20 shows the relation between variety and receptivity under low communication frequency. When communication frequency is increased to 0.7, the relation between variety and receptivity is depicted in Figure 7.21, based on which, we can observe the opposite

trend depicted in Figure 7.20. When the communication frequency is low, variety increases along with receptivity. On the other hand, variety decreases with increasing receptivity under high communication frequency. This comparison suggests that decision makers may consider promoting policies that encourage communities to be more open, if the scientific domain has relatively fewer communication activities. On the contrary, decision makers may encourage communities to increase inertia if more communication activities occur in this domain.

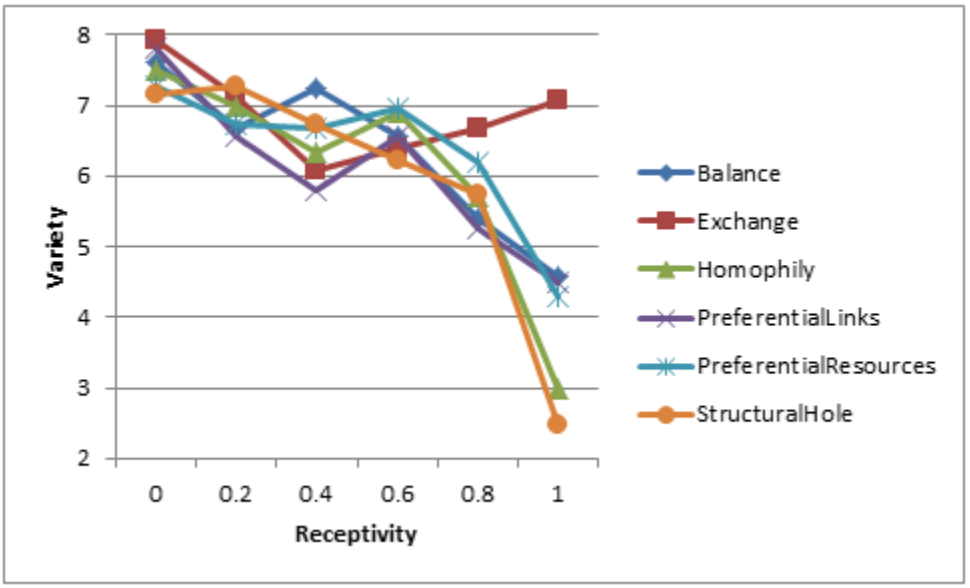


Figure 7.21: Variety vs. Receptivity under High Communication Frequency

7.7.3.2 Innovation Potential

It is shown in [25] that communities with low density and high centrality are expected to exhibit higher innovation potential. Figure 7.22 depicts change in density and centrality over receptivity under various communication theories.

Based on these figures, we can observe that all communication theories except preferential attachment based on links lead to the emergence of communication networks with small density and large centrality along with increasing receptivity. When communities are guided by preferential attachment based on links, communities are always willing to connect

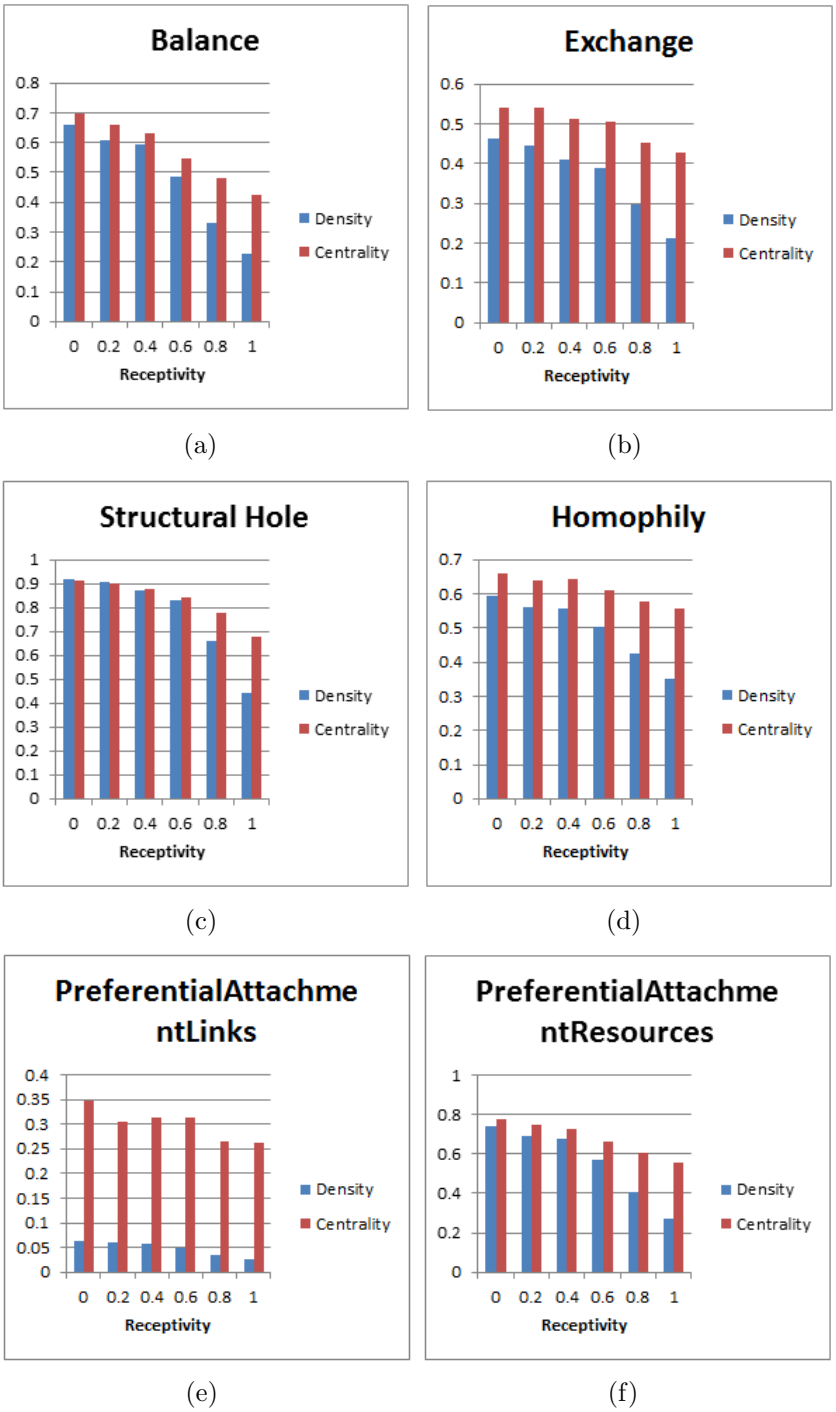


Figure 7.22: Innovation Potential

to those with larger number of links, which directly determines the communication network structure. So, receptivity under preferential attachment based on links theory does not play

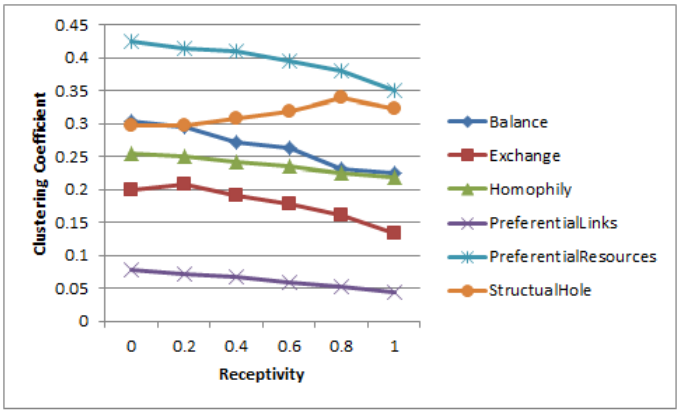
a significant role as it does under other theories in terms of density and centrality. In addition, under the homophily theory, density significantly decreases with increasing receptivity, while centrality almost remains the same, which demonstrates that receptivity is a positive factor that improves innovation potential.

7.7.3.3 Knowledge Diffusion Efficiency

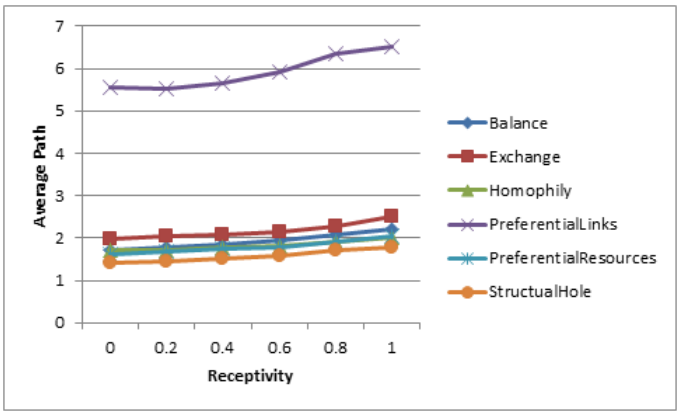
Cliquish networks with low average path lengths exhibit the small-world phenomena and are known to be effective in knowledge creation and diffusion [23]. It is proved that the small world structure is an efficient architecture for new knowledge to diffuse [22]. Small world network structure is identified by a high clustering coefficient and a shorter average path. Figure 7.23 shows the change in clustering coefficient and average path length over receptivity under different communication theories, with the communication frequency set to 0.4. From these figures, we can observe that the communication network guided by balance, exchange, homophily, preferential attachment based on links, preferential attachment based on resources results in decreased clustering coefficient and increased average path length, along with increasing receptivity. It demonstrates that receptivity is a negative factor for communities under these theories to form a small world. In addition, balance theory leads to the highest knowledge diffusion efficiency, since it results in the highest clustering coefficient and one of the shortest average path length. Based on this experimental result, decision-makers may encourage communities to keep self-centering to form a small world network when communication frequency is moderate.

7.7.3.4 Network Patterns

When the communication frequency is high, emergent networks generated by balance and exchange theories always lead to a high density. So, we decrease the communication



(a) Clustering Coefficient vs. Receptivity

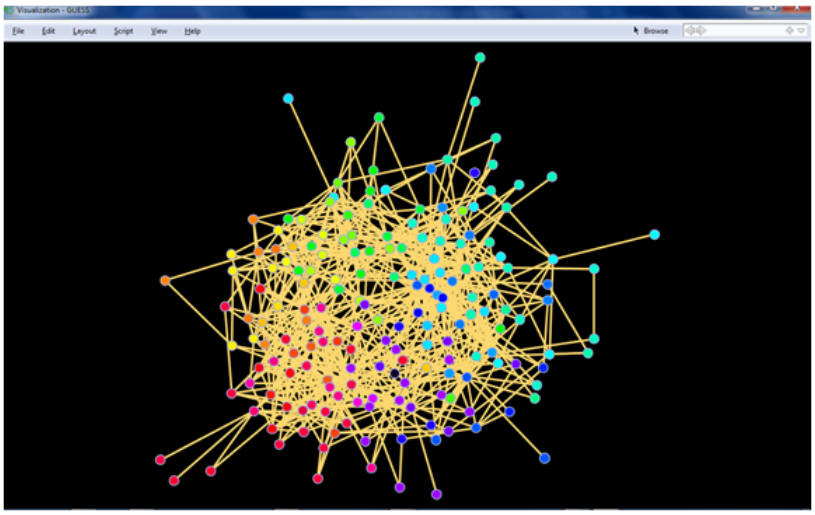


(b) Average Path vs. Receptivity

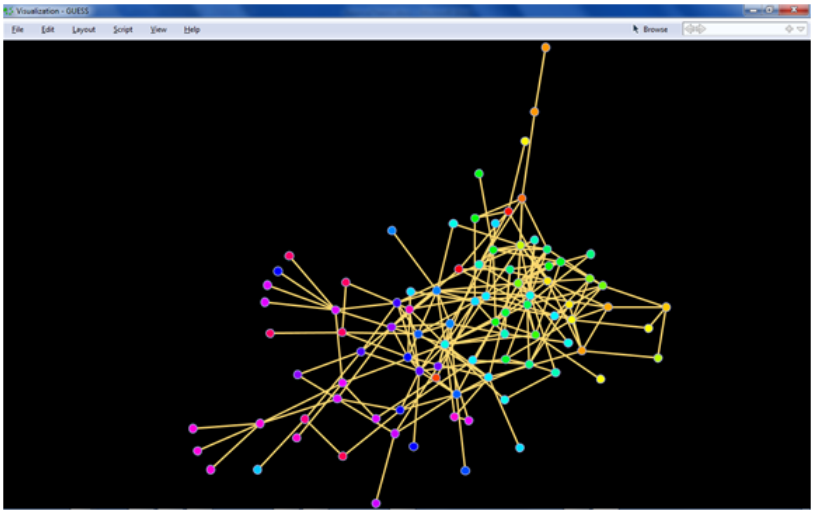
Figure 7.23: Knowledge Diffusion Efficiency

frequency to 0.1; that means each community has 0.1 probability of undertaking communications in each time interval. The following figures (Figure 7.24, 7.25, and 7.27) show network patterns formed by communities under communication theories including homophily, structural hole, preferential attachment, balance, and exchange, respectively. Based on the comparison, the networks under homophily and exchange theory have clusters of similar communities in terms of hue emergence, which are depicted in Figure 7.24. In addition, the network (shown in Figure 7.25) guided by preferential attachment based on links exhibits the property of scale-free network, since its link distribution (shown in Figure 7.26) follows a power law with $R^2 = 0.81$. Moreover, networks under balance, structural hole, and preferential attachment based on resources theories demonstrate a core/periphery network, where a core with highly connected communities emerges, which are shown in Figure

7.27. The core/periphery ratio of these three networks are 4, 11, and 6 respectively, indicating that more core communities are surrounded by fewer periphery communities. Similar phenomenon is exhibited in the OBO network as shown in Table 5.8.



(a) Homophily Theory



(b) Exchange Theory

Figure 7.24: Networks Generated under Homophiliy and Exchange Theory

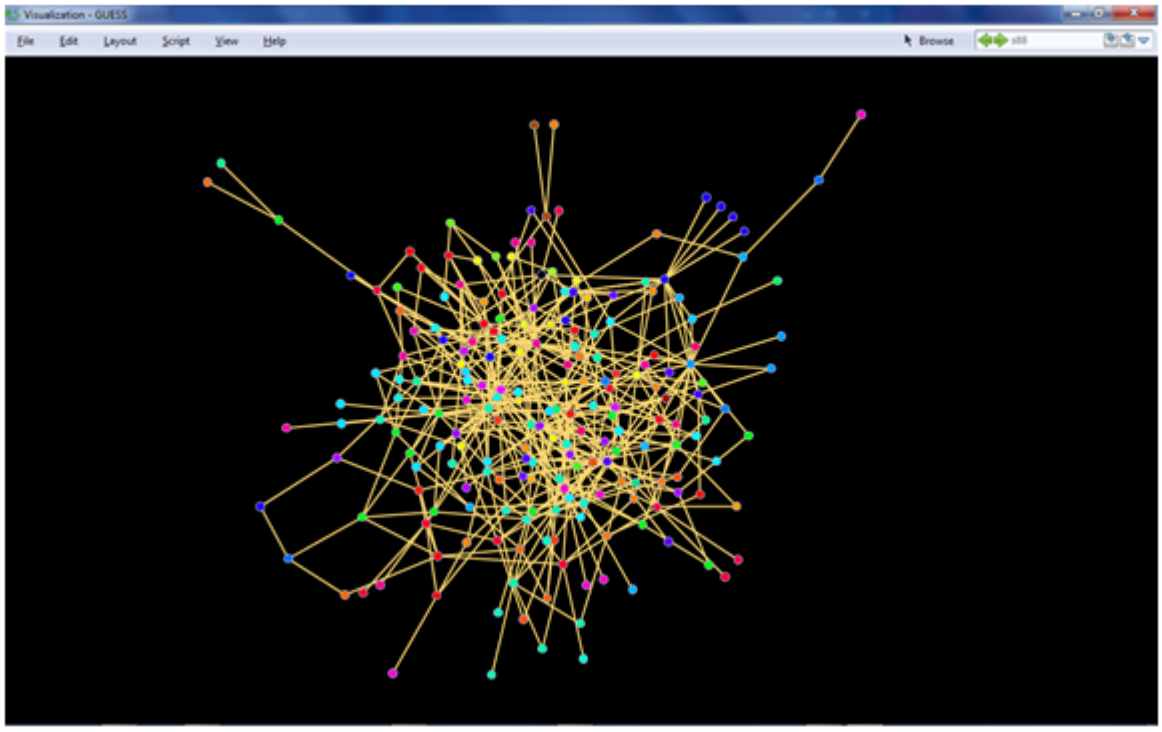
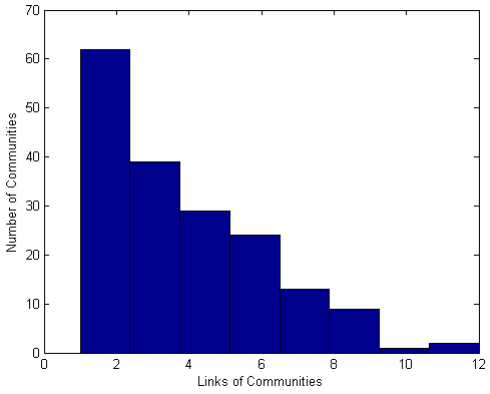
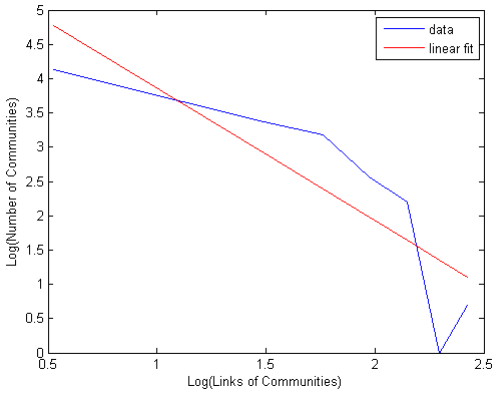


Figure 7.25: The Network Generated under Preference Attachment based on Links Theory

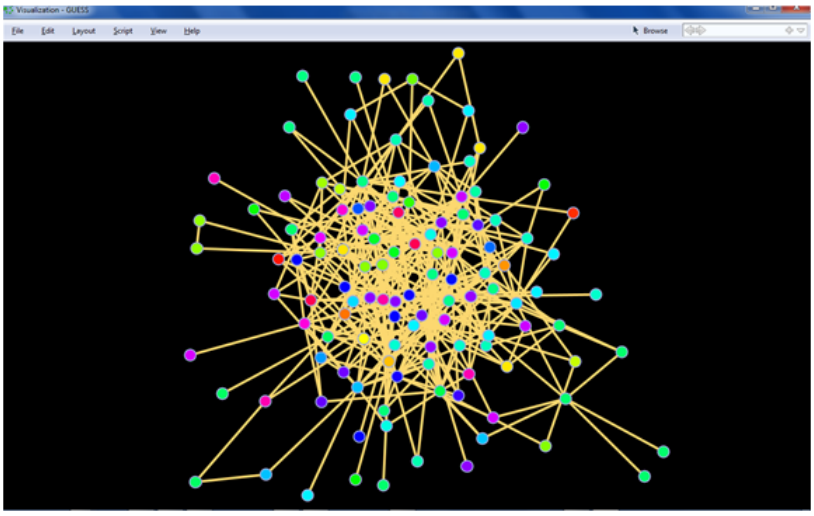


(a) Histogram of Communities' Links

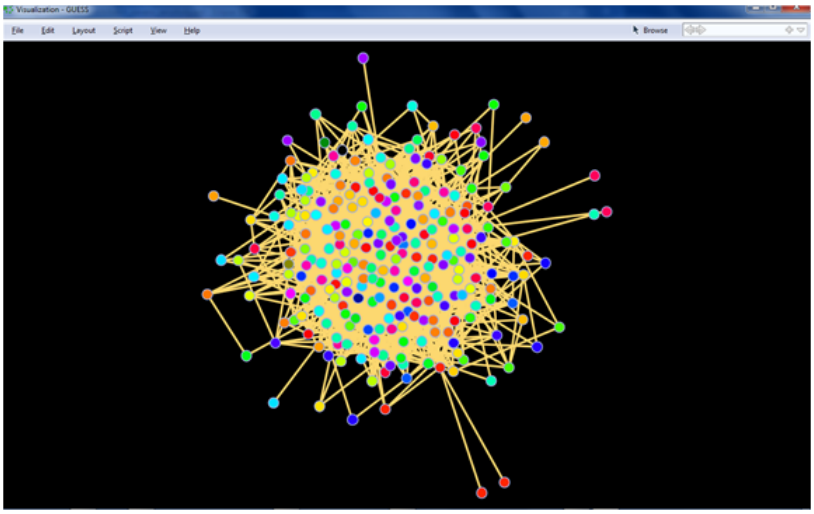


(b) Linear Regression of Logarithmic Value of Links

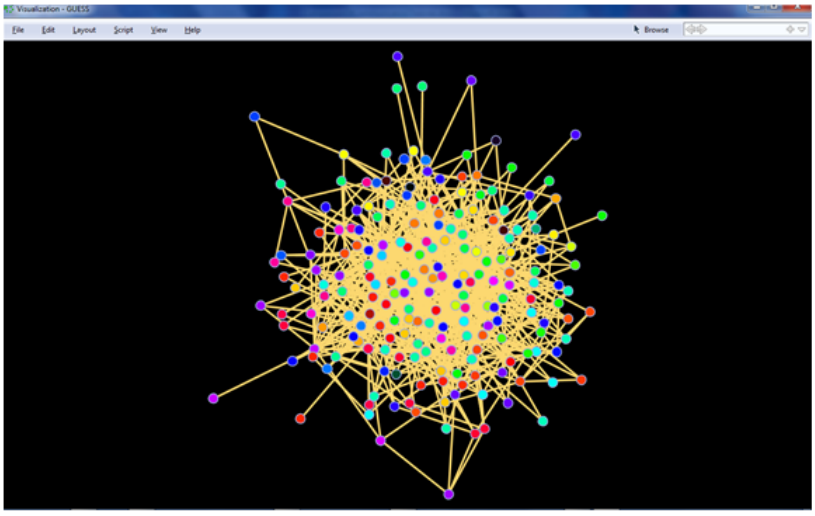
Figure 7.26: Communities' Links



(a) Balance Theory



(b) Structural Hole Theory



(c) Preference Attachment based on Resources Theory

Figure 7.27: Networks under Balance, Structural Hole, and Preference Attachment based on Resources Theory

Chapter 8

Conclusions

In this chapter we outline our findings and discuss them in the context of collective creativity in global participatory science. Also, future research avenues for extending the current model to resolve its limitations are delineated.

8.1 Findings and Discussion

In this study, we conceptualized and simulated the growth and development of scientific communities in terms of a complex adaptive communication system that follows the principles of creative artificial ecosystems.

Using social communication theories as behavioral rules of agents can facilitate development of a new layer over the existing Agent Communication Language (ACL) [31] framework that is based on the speech-act theory [7]. The new layer enables specification of communication mechanisms over the basic primitives provided by ACL. The communication protocol manages connections between agents from different perspectives including communities' traits, self-interest, and discrepancy in resources. In addition, the RGV (Robust Generative Validation) strategy presented in this dissertation can help researchers address Verification and Validation (V&V) challenges of ABM, e.g., counterintuitive emergent behavior, as well as structural and parametric uncertainty. The usefulness of the RGV framework is examined by the validation process of the ColorScape model against the empirical OBO data and the science overlay map.

This research provides a computer-aided tool for science policy development, which is a theme that aims to provide a scientifically rigorous quantitative basis that can be used

by policy makers to assess the impact of their decisions on the growth and development of scientific fields.

The main objective of this research is to explore the impact of scientific community traits (i.e., receptivity, flexibility, reorganization tendency) and environmental constraints (i.e., interaction topologies, resource allocation strategies, socio-technical communication preferences) on the innovation potential (e.g., diversity, sustainability, and resilience) of GPS. Based on the experiments conducted with the ColorScape model, we draw the conclusions discussed in the following sections.

8.1.1 ColorScape: A General Purpose Model

The ColorScape model introduced in this dissertation is a general-purpose creative artificial innovation ecosystem model that can mimic the behavior of both traditional and open innovation communities. The model is conceptually grounded and validated in terms of its capability to generate similar metrics against the science overlay map [75] and the empirical OBO network [86].

8.1.2 Community's Traits vs. Diversity

In low density networks, increasing levels of receptivity improves diversity up to a level. On the contrary, diversity decreases with increasing receptivity in highly coupled networks. Under environments with high receptivity, the reason that diversity favors low connectivity is that presence of dense communication channels causes convergence, which in turn decreases diversity. Experimental results suggest encouraging communities to be more receptive in relatively low density environments to attain higher levels of diversity.

Reorganization adversely affects diversity. On the other hand, specialization has positive effects on diversity. Reorganization and specialization strategies help communities adapt to their environment, when the community cannot meet the expectations

of its members. This observation is consistent with the functionality of specialization and reorganization. Specialization facilitates creation of a new community with a different target color from the current community, while reorganization involves pulling the target color toward the current color, causing convergence.

8.1.3 Environmental Constraints vs. Diversity, Sustainability, and Resilience

The size of the carrying capacity of the knowledge ecosystem has positive effects on diversity. Yet, there is a point of diminishing returns. Increasing the number of communities improves the probability of forming more clusters comprised of similar communities. But there is maximum diversity given a fixed scientific spectrum and the maximum difference within clusters. So, diversity cannot increase indefinitely with the carrying capacity. By the same token, increasing external resources leads to increased diversity up to a point, beyond which more resources can only increase the number of communities within a cluster rather than the number of clusters. For policy makers, it is noteworthy that neither external resources nor initial community number can keep diversity increasing, i.e., there is a tradeoff between the available funding and the expected level of diversity.

Disparity increases with resources up to a point. Beyond that point, disparity decreases with increasing resources. At the same time, lower connectedness cause higher disparity. Disparity increases with the increasing level of resource availability, because more resources lead to higher success rate, which in turn results in disparity. On the other hand, beyond that point, disparity decreases because of the decreased need for interaction for sustainment, which in turn decreases inequality. In addition, disparity increases with decreasing level of connectivity; this is due to decreased convergence under low connectivity. To meet the desired level of disparity, policy-makers need to consider the connectedness of the social communication network when making decisions on allocating external funding.

The 2D topology is more resilient than the 1D topology, and scale-free networks have higher resilience than random and random group networks. The 1D topology corresponds to relations between upstream and downstream organizations. Generally speaking, upstream organizations lead the frontier research and determine the direction of future research in their domains. Downstream organizations mainly transfer the technology developed by upstream organizations into products. The 2D topology adds the collaborations between organizations at the same level in the organizational ecological chain.

Communities with highly connected clusters under low level of resource availability can experience high levels of sustainability. On the contrary, under moderate level of resources, loosely connected clusters are more likely to survive. A plausible explanation for this observation is that higher resource availability leads to higher variety. Under high variety, larger connectivity causes each community to be pulled toward multiple different cognitive niches, resulting in lack of focus which in turn costs communities more resources, and hence decreasing the survival rate. On the other hand, lower resource availability leads to lower variety. Under low variety, however, strong connectivity results in more communities sharing similar states, benefiting from each other through a symbiotic relation, which in turn increases the overall survival rate. Based on these observations, policy-makers may encourage communities to build highly connected clusters if resource availability is low.

8.1.4 Network Metrics vs. Variety

One goal of this research is to identify a metric hierarchy, two layers of which are network metrics and attributes. Network metrics include density and centrality, while attributes include diversity. Little research is undertaken to make clear the relation between these two layers. Based on the experiments conducted with the Colorscape model, we observe that **variety increases with density and centrality up to a point, beyond which variety is inhibited.** Similar results are also found in [40][69]. According to these observations,

researchers may estimate expected levels of variety based on the density and centrality of social networks; that is, neither low nor high density/centrality results in high variety. High variety occurs at moderate density/centrality.

8.1.5 Allocation Strategies vs. Variety

Key area investment with technology transferring results in the highest level of variety. This is similar to the case where domains with lower priority still have potential to advance, yet the environment promotes development of domains related to priorities. Additionally, the communities with most resources granted may not be as successful as expected, if the domain with the top priority is located between several significantly different domains. A potential reason is that the domain with top priority is pulled toward several different directions, which could have incurred significant resource cost during the learning process, resulting in decreased number of communities with most resources granted. For policy-makers, they should also consider the interaction networks around the domain with the top priority, when making decisions about funding allocation.

8.1.6 Communication Strategies vs. Diversity, Sustainability, and Innovation Potential

Examined communication strategies are not significantly different from each other, especially in regard to the relation between variety and external resources under the P2P allocation strategy. Increasing funding does not always help increase variety, especially for those communities with relatively low communication frequency. Communication theories change the strategy of communities in selecting targets, based on which local niches are emerged. Different formation of local niches results in different local diversity. But the local diversity does not influence the global diversity significantly. In addition, low communication frequency results in fewer interaction activities, causing fewer resources to be consumed. So, external resources have little effect on variety at

low communication frequency. As scientific communities can be viewed as artificial ecosystems, the communication frequency is similar to the evolution frequency. Lower evolution frequency leads to fewer species to be eliminated. This suggests that sustainability favors low communication frequency. Based on this observation, policy-makers may discourage inter-organizational activities when resources are limited.

Under low communication frequency, openness and receptivity lead to higher variety. On the contrary, variety decreases with increasing receptivity under high communication frequency. The potential reason is that higher receptivity results in more communities sharing similar states under low communication frequency, benefiting from each other through a symbiotic relation, which in turn increases the overall survival rate. On the other hand, under high communication frequency, higher receptivity results in the convergence of communities, which in turn lead to more communities inhabiting within the same domain. Under the P2P allocation strategy, one domain can only sustain fixed number of communities so that the survival rate decreases with increasing receptivity. This is also why **sustainability decreases with increasing receptivity at high communication frequency.**

Receptivity is a positive factor that improves innovation potential for communities under high communication frequency. At high communication frequency, higher receptivity results in lower survival rate, which in turn leads to lower density and lower centrality. However, density decreases at a higher rate than centrality does, resulting in significant difference between levels of density and centrality. Networks with low density and high centrality are attributed with higher innovation potential [25]. In comparison to the previous finding that sustainability and variety decrease with increasing receptivity at high communication frequency, there is a tradeoff between sustainability, variety, and innovation potential. Decision-makers have to take this observation into consideration to develop policies that balance these three indicators.

Networks governed by the homophily and the exchange theories yield clusters of similar communities. Under the homophily theory, communities communicate with similar peer communities. Under the exchange theory, the transaction occurs when communities solve problems by collaboration, during which collaborating communities become similar. So, both of these theories help local clusters with similar communities to emerge. Based on this finding, policy-makers need to be cognizant that local niches are likely to exist in networks of communities guided by the homophily or the exchange theory.

8.2 Extensions

The main emphasis of the Colorscape model presented in this dissertation is the interconnection among communities. Hence, the model can be used to simulate the behavior of networks formed by communities, among which dynamic relationships exist.

These types of communities include the following [22]:

- **Shared Instrument**: The main objective of such communities is to increase access to a scientific instrument. Shared Instrument collaboratories often provide remote access to expensive scientific instruments such as telescopes. For such communities, the Colorscape model can help discern effective strategies for improving the collective use of expensive instruments.

- **Virtual Community of Practice** is a network of individuals who share a research area and communicate online. Virtual Communities may share news of professional interest, advice, techniques, or pointers to other resources online.

- **Virtual Learning Communities** aim to increase the knowledge of participants, but not necessarily aimed toward conducting original research.

- **Distributed Research Centers** are similar to a university research center, but they are operated at a distance. It is an attempt to aggregate scientific talent, efforts, and resources beyond the level of individual researchers.

For networks comprised of the types of communities listed above, the Colorscape model needs to be slightly modified according to the specific characteristics of relations between each type of communities, so that it can simulate the dynamics of emergent networks and facilitate the analysis of the output data.

8.3 Limitation and Future Research

One limitation of the Colorscape model is its inability to generate network patterns similar to the science overlay map, although the structural indicators such as density, centrality, clustering coefficient are sufficiently similar. The potential reason is that communities select target communities to communicate globally. It may be valuable to limit the scope of potential targets that communities can select. This strategy aims to encourage more local niches to emerge, which is observed in contemporary research development with relatively high-coupled clusters and fewer international connections.

Besides communication theories already implemented in the Colorscape model, there are other theories that can also be embedded in the model.

Public goods theory [81] explains the economics of collective ownership such as public bridges, parks, and libraries, which are distinguished from the private ownership. Two characteristics of public goods are noteworthy: impossibility of exclusion and jointness of supply. There is a determining factor in generating public goods named critical mass [67], which is defined as the minimum interest that drives the majority of people to realize the public good.

Cognitive social structure [50] is to characterize individual community's perceptions of the social network. The theory can be used to build the community's understanding of the network, which is partial and may be different from the real network. The communities in the partial network cognized by a community are its candidate objects for future communication. Two steps are needed, one of which is to build the partial network. The other is to select communities to communicate within the partial network.

Cognitive consistency theory [42] argues that communities are satisfied with their positions in the communication network if their associated peer communities are connected with one another. Assuming the set of neighbors of a community A is S, then the influence of a community B in S on A is proportional to the number of $B's$ links to other communities in S.

This research provides a computer-aided tool (i.e., the ColorScape Model) for science policy development, so that decision-makers can assess the impact of their decisions on the growth and development of scientific fields in advance. By undertaking experiments presented in Chapters six and seven, decision-makers can alter communities' traits, resource allocation strategies, and socio-technical communication preferences to examine their impacts on innovation potential and performance. In addition, the resource allocation module and the communication preferences module can be extended and replaced by other modules that decision-makers are interested in. Hence, the ColorScape model can be customized to facilitate conducting abstract thought experiments for exploring effective strategies, while allowing informed decision-making for science and innovation policy.

Bibliography

[1] Missile Defense Agency. Department of defense documentation of verification, validation & accreditation (vv&a) for models and simulations. Technical report, Missile Defense Agency, 2008.

[2] P. Ahrweiler, A. Pyka, and N. Gilbert. A new model for university-industry links in knowledge-based economies. *Journal of Product Innovation Management*, 2010.

[3] H. Aldrich. *Organizations evolving*. Thousand Oaks.

[4] Teresa M Amabile, Sigal G Barsade, Jennifer S Mueller, and Barry M Staw. Affect and creativity at work. *Administrative Science Quarterly*, 2006.

[5] Martyn Amos. *Theoretical and Experimental DNA Computation*. Springer, 2005.

[6] W. Brian Arthur, Steven N. Durlauf, and David A. Lane. *The economy as an evolving complex system*. Addison-Wesley, 1997.

[7] John Langshaw Austin. *How to Do Things With Words*. Harvard University Press, 2005.

[8] Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press.

[9] Jerry Banks. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. Wiley-Interscience, 1998.

[10] Arjun Bhutkar. Synthetic biology: Navigating the challenges ahead. *The Journal of Biolaw & Business*, 2005.

[11] Aharon Blanka and Sorin Solomon. Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). *Physica A: Statistical Mechanics and its Applications*, 287:279–288, 2000.

[12] 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*, 2002.

[13] Daniel J. Brass. Being in the right place: a structural analysis of individual influence in an organization. *Administrative Science Quarterly*, 29:519–539, 1984.

[14] Daniel J. Brass. Technology and the structuring of jobs: Employee satisfaction, performance, and influence. *Organizational Behavior and Human Decision Processes*, 35:216–240, 1985.

[15] Daniel J. Brass. A social network perspective on human resources management. *Research in personnel and human resources management*, 1995.

[16] Iain Buchan. Calculating the gini coefficient of inequality. Technical report, Northwest Institute for BioHealth Informatics, 2002.

[17] R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, 1992.

[18] Ronald S. Burt. Social contagion and innovation: Cohesion versus structural equivalence. *The American Journal of Sociology*, 1987.

[19] Ronald S. Burt. The gender of social capital. *Rationality and Society*, 1998.

[20] James S. Coleman. *Individual Interests and Collective Action: Studies in Rationality and Social Change*. Cambridge University Press, 1986.

[21] James S. Coleman. *Foundations of Social Theory*. Belknap Press of Harvard University Press, 1998.

[22] R. Cowan. The dynamics of collective invention. *Journal of Economic Behavior & Organization*, 52:513532, 2003.

[23] R. Cowan and N. Jonard. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 2004.

[24] Susan Cozzens. A deeper look at the visualization of scientific discovery in the federal context. Technical report, Georgia Tech, 2008.

[25] C. Dhanaraj and A. Parkhe. Orchestrating innovation networks. *Academy of Management Review*, 2006.

[26] Virginia Dignum, Frank Dignum, and Liz Sonenberg. Design and analysis of organization adaptation in agent systems. In *Agent-Directed Simulation and Systems Engineering*. Wiley, 2009.

[27] B. Edmonds and E. Chattoe. When simple measures fail: Characterising social networks using simulation. Technical report, Social Network Analysis: Advances and Empirical Applications Forum, 2005.

[28] Vernon J. Ehlers. The future of u.s. science policy. *Science*, 279:302, 1998.

[29] Joshua M. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press, 2007.

[30] Henry Etzkowitz. The triple helix of university-industry-government relations implications for policy and evaluation. Technical report, Retrieved from http://ssi.sagepub.com/cgi/doi/10.1177/05390184030423002, 2002.

[31] FIPA. Fipa acl message structure specification. Technical report, Foundation for Intelligent Physical Agents, 2002.

[32] J. Fulk and G. DeSanctis. Articulation of communication technology and organizational form. In *Shaping organizational form: Communication, connection, and community*. Thousand Oaks, 1999.

[33] Paul J. Gemperline. Statistical evaluation of visualization methods. Technical report, Department of Chemistry, East Carolina University, 2007.

[34] N. Gilbert. A simulation of the structure of academic science. *Sociological Research Online*, 2, 1997.

[35] N. Gilbert, P. Ahrweiler, and A. Pyka. Learning in innovation networks: some simulation experiments. *Innovation in complex social systems*, 2010.

[36] Peter A. Gloor, Maria Paasivaara, Detlef Schoder, and Paul Willems. Finding collaborative innovation networks through correlating performance with social network structure. Technical report, MIT, 2007.

[37] Lance H. Gunderson and C. S. Holling. *Panarchy: Understanding Transformations in Human and Natural Systems*. Island Press, 2001.

[38] F. Heider. Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112, 1946.

[39] L.J Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Research*, 1999.

[40] Matthew H Hohn. The relationship between species diversity and population density in diatom populations from silver springs, florida. *Transactions of the American Microscopical Society*, 1961.

[41] J. H. Holland. *Hidden Order: How Adaptation Builds Complexity*. Perseus Books, 1995.

[42] Paul W. Holland and Samuel Leinhardt. The statistical analysis of local structure in social networks. *NBER Working Paper Series*, w0044, 1974.

[43] George Caspar Homans. *Social behavior: Its elementary forms*. Harcourt, Brace & World, 1961.

[44] C. Hovland and E. Hunt. The computer simulation of concept attainment. *Behavioral Science*, 5:265–267, 1960.

[45] InnoCentive. Challenge driven innovation. http://www.innocentive.com/seekers/challenge-driven-innovation, 2011.

[46] Alex Kacelnik. Timing and foraging: Gibbons scalar expectancy theory and optimal patch exploitation. *Learning and Motivation*, 33, 2002.

[47] David Kaiser, Vincent Lepinay, and David Jones. Predictive modeling of the emergence and development of scientific fields. Technical report, Massachusetts Institute of Technology, 2010.

[48] David Klahr and Herbert A. Simon. The dynamics of collective invention. *Psychological Bulletin*, 125:524–543, 1999.

[49] Genevieve J. Knezo. Federal research and development: Budgeting and priority-setting issues, 109th congress. Technical report, Resources, Science, and Industry Division, 2006.

[50] D. Krackhardt. Cognitive social structures. *Social Networks*, 9:109–134, 1987.

[51] David Krackhardt. Constraints on the interactive organization as an ideal type. *The Post-Bureaucratic Organization*, pages 211–222, 1994.

[52] V. Krebs and J. Holley. Building sustainable communities through network building. Technical report, 2002.

[53] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University Of Chicago Press, 1996.

[54] D.R. Lane. Spring 2001 theory workbook. http://www.uky.edu/~drlane/capstone/persuasion/bal.htm, 2001.

[55] Michael W. Macy and Robert Willer. From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 2002.

[56] Jon McCormack. Artificial ecosystems for creative discovery. Technical report, Centre for Electronic Media Art Faculty of Information Technology, Monash University Clayton 3800, Australia, 2007.

[57] R. McDermott. Learning across teams. *Knowledge Management Review*, 1999.

[58] John H. Miller and Scott E. Page. *Complex Adaptive Systems, An introduction to computational models of social life*. Princeton University Press, 2007.

[59] Melanie Mitchell and Charles E. Taylor. Evolutionary computation: An overview. *Annual Reviews*, 1999.

[60] Susan A. Mohrman and Caroline S. Wagner. 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, Center for Effective Organizations Marshall School of Business University of Southern California, 2008.

[61] Peter R. Monge and Noshir S. Contactor. *Theories of Communication Networks*. Oxford, 2003.

[62] nanoHUB.org. Network for computational nanotechnology. `http://nanohub.org`, 2010.

[63] NEESgrid. Network for earthquake engineering simulation. `http://it.nees.org/`, 2010.

[64] M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98:404–409, 2001.

[65] M. E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27:39–54, 2005.

[66] M. E. J. Newman. Power laws, pareto distributions and zipfs law. *Contemporary Physics*, 46:323 351, 2005.

[67] P.E. Oliver. Formal models of collective action. *Annual Review of Sociology*, 19:271–300, 1993.

[68] Tim Pawlenty, Edward G. Rendell, and Raymond C. Scheppach. Higher education, mandates and unintended consequences: An analysis of the moe mandate in hr 4137. *National Governors Association*, 2008.

[69] Jill E. Perry-Smith. The social side of creativity: A static and dynamic social network perspective. academy of management review. *The Academy of Management Review*, 2003.

[70] Peter Pirolli. An elementary social information foraging model. *Proceedings of the 27th international conference on Human factors in computing systems*, 2009.

[71] Alan L. Porter and Ismael Rafols. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 2009.

[72] Daniel I. Prajogo and Pervaiz K. Ahmed. Relationships between innovation stimulus, innovation capacity, and innovation performance. *R&D Management*, 2006.

[73] Andreas Pyka, Nigel Gilbert, and Petra Ahrweiler. Agent-based modelling of innovation networks - the fairytale of spillover. *Complexity*, 2009.

[74] I. Rafols and M. Meyer. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 2009.

[75] Ismael Rafols, Alan L. Porter, and Loet Leydesdorff. Science overlay maps: a new tool for research policy and library management. Technical report, Science and Technology Policy Research, University of Sussex, 2010.

[76] John H. Reed, Gretchen Jordan, and Edward Vine. Impact evaluation framework for technology deployment programs. Technical report, US Department of Energy, 2007.

[77] Repast. Repast. `http://repast.sourceforge.net/`, 2010.

[78] V. Riss and Hans Friedrich Witschel. What is organizational knowledge maturing and how can it be assessed? *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, 2009.

[79] Jr Rykiel. Testing ecological models : the meaning of validation. *Ecological Modeling*, 1996.

[80] A. Saltelli, Andres Ratto, M., Campolongo T., Cariboni F., Gatelli J., D. Saisana, M., and S. Tarantola. *Global Sensitivity Analysis*. The Primer, John Wiley & Sons.

[81] P. Samuelson. The pure theory of public expenditure. *Review of Economics and Statistics*, 36:387–389, 1954.

[82] National Science and Technology Council. National nanotechnology initiative: The initiative and its implementation plan. Technical report, Retrieved from http://ssi.sagepub.com/cgi/doi/10.1177/053901840304230002, 2000.

[83] Scott Shane. Encouraging university entrepreneurship? the effect of the bayh-dole act on university patenting in the united states. *Journal of Business Venturing*, 2004.

[84] J. Shrager and P. Langley. *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufman, 1990.

[85] SKIN. Skin. `http://cress.soc.surrey.ac.uk/SKIN`, 2010.

[86] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 2007.

[87] R. Stankiewicz. Technology as an autonomous socio-cognitive system. *Dynamics of Science-Based Innovation.*, 1992.

[88] Craig R. Scott Steven R. Corman. Perceived networks, activity foci, and observable communication in social collectives. *Communication Theory*, 1994.

[89] Andrew Stirling. Diversity and ignorance in electricity supply investment: Addressing the solution rather than the problem. *Energy Policy*, 22:195–216, 1994.

[90] Richard Swedberg. *Entrepreneurship: The Social Science View*. Oxford University Press, USA, 2000.

[91] Bill Valdez and Julia Lane. The science of science policy: a federal research roadmap. Technical report, National science and technology council, 2008.

[92] Stephen Vincent. *Input Data Analysis*. Compuware Corporation, 1998.

[93] Caroline S. Wagner. *The new invisible college, Science for development*. Brookings Institution Press, 2008.

[94] Brian Walker. Resilience, adaptability and transformability in socialecological systems. *Ecology and Society*, 2004.

[95] Karl E Weick. *The Social Psychology of Organizing*. McGraw-Hill Humanities/Social Sciences/Languages, 1979.

[96] Etienne Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1998.

[97] Etienne Wenger. Communities of practice and social learning systems. *Organization*, 7:225–246, 2000.

[98] M.A West and J.L. Farr. *Innovation and Creativity at Work*. John Wiley& Sons, 1990.

[99] Wikipedia. Complex system. `http://en.wikipedia.org/wiki/Complex_system`, 2010.

[100] Wikipedia. Gini coefficient. `http://en.wikipedia.org/wiki/Gini_coefficient`, 2010.

[101] Wikipedia. Innovation. `http://en.wikipedia.org/wiki/Innovation`, 2010.

[102] Wikipedia. Integration testing. `http://en.wikipedia.org/wiki/Integration_testing`, 2010.

[103] Wikipedia. Minimum viable population. `http://en.wikipedia.org/wiki/Minimum_viable_population`, 2010.

[104] Wikipedia. Resource allocation. `http://en.wikipedia.org/wiki/Resource_allocation`, 2010.

[105] Wikipedia. Scientific community. `http://en.wikipedia.org/wiki/Scientific_community`, 2010.

[106] Wikipedia. Preferential attachment. `http://en.wikipedia.org/wiki/Preferential_attachment`, 2011.

[107] Wikipedia. Social balance theory. `http://en.wikipedia.org/wiki/Social_balance_theory`, 2011.

[108] Levent Yilmaz. Dynamics of collective creativity and open innovation in scientific commons, complex adaptive systems perspective. Technical report, Auburn University, 2008.

[109] Levent Yilmaz. An agent simulation study on conflict, community climate and innovation in open source communities. *International Journal of Open Source Software and Processes*, 2009.

[110] Levent Yilmaz. On the synergy of conflict and collective creativity in open innovation socio-technical ecologies. *International Conference on Computational Science and Engineering*, 2009.

[111] Levent Yilmaz, Guangyu Zou, and Osman Balci. A robust evolutionary strategy for generative validation of agent-based models using adaptive simulation ensembles. *2011 IEEE/ACM Winter Simulation Conference*, 2011.

[112] Guangyu Zou and Levent Yilmaz. A computational model of collective creativity and innovation in virtual open source science networks: What distinguishes innovative virtual communities? *2010 IEEE/ACM Winter Simulation Conference*, 2010.

[113] Guangyu Zou and Levent Yilmaz. Dynamics of knowledge creation in global participatory science communities: open innovation communities from a network perspective. *Comput Math Organ Theory*, 2010.