Exploring the Impact of Socio-technical Communication Styles on the Robustness and Innovation Potential of Global Participatory Science

by

Özgür Özmen

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 5, 2013

Keywords: Complex Adaptive Systems, Agent Based Simulation, Self-Organization, Communication, Collective Action, Innovation, Diversity, Robustness, Resilience, Social Network, Knowledge Creation

Approved by:

Jeffrey Smith, Professor of Industrial and Systems Engineering
Levent Yilmaz, Associate Professor of Computer Science
Alice Smith, Professor of Industrial and Systems Engineering

Abstract

Emerging cyber-infrastructure tools are enabling scientists to transparently co-develop, share, and communicate in real-time diverse forms of knowledge artifacts. In this research, these collaborative environments are modeled as complex adaptive systems using collective action theory as a basis. Communication preferences of scientists are posited as an important factor affecting innovation capacity and resilience of social and knowledge network structures. Using agent-based modeling, a complex adaptive social communication network model is developed. By examining the Open Biomedical Ontologies (OBO) Foundry data and drawing conclusions from observing the Open Source Software communities, a conceptually grounded model mimicking the dynamics in what is called Global Participatory Science (GPS), is presented. Social network metrics and knowledge production patterns are used as proxy metrics to infer innovation potential of emergent knowledge and collaboration networks. Robust communication strategies with regard to innovation potential are questioned by exploring different parameter and mechanism configurations. The objective is to present the underlying dynamics of GPS in a form of computational model that enables analyzing the impacts of various communication preferences of scientists on innovation potential of the collaboration network. The ultimate goal is to further our understanding of the dynamics in GPS and facilitate developing informed policies fostering innovation capacity.

Acknowledgments

I would like to express my gratitudes to the advisory committee of this dissertation. Without their support, suggestions, criticisms, and credences, this work would not be possible. Special thanks to Dr. Levent Yilmaz for introducing me to the Complex Adaptive Systems, countless creative discussions, encouragement, and his never-ending confidence on me. Other special thanks to Dr. Jeffrey Smith for mind-stretching arguments, his expertise, and his guidance. Thanks to Dr. Alice Smith for recruiting me to the Auburn family and her valuable comments on my dissertation. Thanks for their patience at the times i procrastinated. I wish i could be a better student.

Additionally, thanks to the National Science Foundation (NSF) for the partial support as authorized by the contract number NSF-SBE-0830261. The research is also partially funded by the Industrial and Systems Engineering Department assistantships/fellowships at Auburn University.

I would like to dedicate this hard-work to my family. To my favorite people who love regardless...

Table of Contents

List of Illustrations

List of Tables

Chapter 1

INTRODUCTION

Science is becoming increasingly global and participatory due to online collaboration opportunities such as e-mailing, web-based social networking, and open-access collaboration platforms. Hence, scientists interact not only locally but also globally by constructing self-organizing collaboration networks. Wagner (2008) states the following regarding the emergence of these fluid networks:

> They constitute an invisible college of researchers who collaborate not because they are told to but because they want to, who work together not because they share a laboratory or even a discipline but because they can offer each other complementary insight, knowledge, or skills.

One of the most significant problems in organizational scholarship is to discern how social collectives govern, organize, and coordinate the actions of individuals to achieve collective outcomes (O'Mahony and Ferraro, 2007). The first phase of this research explores micro-level (inter-scientist) socio-technical processes and mechanisms that explain emergent behaviors observed in scientific communities that collaborate over the cyber-infrastructure. Scientific knowledge creation in such communities is called Global Participatory Science (GPS) (Zou and Yilmaz, 2011). First, based on the views advocated by Wagner (2008) and Monge and Contractor (2003), the structure and behavior of GPS are interpreted as complex adaptive systems (CAS). Second, recent ethnographic studies (Nielsen, 2010; Ostrom and Hess, 2007), which suggest that GPS is a collective action undertaken by autonomous self-organizing scientists, are leveraged. The first question of interest is "Which interaction mechanisms in the literature explain operational behavior of GPS and its underlying socio-technical processes?" Then the focus of this research is on "How we can specify and implement these mechanisms in the form of a computational model to gain empirical insight and perform exploratory analysis?"

It is demonstrated by Wagner (2008) that science is complex because researchers interact in both competitive and cooperative ways, with no imposed blueprint. Furthermore, she states that it is adaptive because scientists respond to environmental changes such as funding preferences and new discoveries . In this work, *Information foraging*, *preferential attachment*, and *population dynamics* are conceptualized as the underlying self-organization mechanisms of knowledge creation in GPS.

There are simulation studies that explore knowledge creation processes in science (Shrager and Langley, 1990; Cowan and Jonard, 2004; Gilbert, 1997). However, in these models social interactions are not taken into account. Using the *collective action* theory, which includes models of *self-interest* (based on knowledge gain), *exposure* (based on social influence), *cognitive burden* (based on expertise of scientists), and *tension* (based on complexity of the projects) in scientific knowledge generation, theoretically-grounded conceptual model of scientist behavior is developed.

The understanding of CAS is more likely to arise with the help of computer-based models (Holland, 1996). Agent Based Modeling (ABM) provides us with the opportunity to directly identify individual entities along with their relationships and capabilities. Hence, simulation of these mechanisms using the ABM worldview is a powerful method that is adopted in this study.

ABM involves rationally bounded human agents. Therefore, validation of these models is a challenging task, and the assumptions of the models should be explicitly explained. But also computational laboratories are not supposed to be, indeed *should* not be, exact replication of reality (Burton, 2003). For validation purposes, the presence of scale free networks, adaptive/renewal activity cycles, and network formation phases (e.g. core/periphery) are investigated. These are known characteristics peculiar to collaboration networks and GPS.

Communication among agents in a CAS has an intense effect on the system level behavior (Shlesinger, 2007). Wagner (2008) indicates that if we can discern identifiable patterns and mechanisms of communication among the scientists, then understanding can lead us to determine how the scientific endeavor operates and how policymakers can effectively influence its evolution and growth. In an NSF workshop in 2006[1], participants are called for new theoretical models that foster understanding innovation through computational and cognitive models of creativity (Yilmaz, 2008b). The second phase of this research focuses on implementing computational mechanisms of selected social communication theories: (1) *Homophily* theory, (2) *Social capital* theory, (3) *Human capital* theory, (4) *Social exchange* theory. Then the the goal is to evaluate the evolution of the network. The question to address is: “Which social communication mechanisms among scientists are more effective in fostering innovation potential?”

Generative mechanisms of social capital, human capital, homophily, and social exchange theories, which are relevant social communication theories applicable to the problem domain, are implemented. There are studies that discuss diversity (Dhanaraj and Parkhe, 2006; Powell et al., 1996) and network connectivity (Pyka, 2009; Burt, 1995) as potential indicators of innovativeness. Diversity of the emergent knowledge and collaboration network structures is measured and used as an indicator of interdisciplinarity. Analysis of core/periphery ratio, small-world phenomena (based on clustering coefficient and average path length), degree centrality, and density of emergent collaboration networks are also conducted to assess innovation potential.

[1] NSF/SRS Workshop on Advancing Measures of Innovations: Knowledge Flows, Business Metrics, and Measurement Strategies, 2006

Additionally, the utility of additional activity metrics (the number of active members, distribution of expertise) are evaluated as proxy metrics of innovation potential.

The OECD and European Commission encourage the research on innovation metrics, unanticipated consequences, and unrealized opportunities emerging from the complex adaptive environments. Scientific networks cannot be controlled but can be guided by the policymakers to influence collaboration environments. Recognizing uncertainty and the lack of knowledge about the environment in science-based innovation systems such as GPS, designing robust systems is important as opposed to searching for optimal system design. The third phase of the research is focused on the question: "How to explore different parameter and mechanism configurations to seek and identify more robust communication strategies in terms of variance of innovation potential metrics?"

In order to address this question, an exploratory modeling tool is developed which consists of coupled genetic algorithm and metamorphic relations module to create models with distinct scenarios. Hence, decision-makers (policy-makers in this research) can explore different parameter and mechanism set-ups. Mean absolute errors (MAE) of different metrics are used to measure robustness performance of each point in the scenario space. A feedback mechanism is constructed based on robustness performance and metamorphic relations to facilitate further generation and exploration of the scenario space. More robust scenario space is investigated, while the average robustness behavior for each communication mechanism under various conditions (over generations) is monitored. Additionally, the levels that parameter values converge at more robust landscapes are evaluated. The tradeoffs between *robustness* and *innovation potential* are delineated by social network metrics.

The National Academy of Engineering report indicates that leadership in innovation is essential to U.S. prosperity and security.[2] By developing agent-based models

[2] http://www.nae.edu - As of 4.07.2013

and conducting the analysis described above, the goal is to further our understanding of innovation in GPS and to support policy-makers in nurturing open scientific environments.

In the following Chapter, a summary of the literature relative to the topic is provided. Chapter 3 outlines the methodology adopted and then research problems are introduced in detail. In Chapter 4, base-model design is discussed along with verification and validation efforts conducted, and Chapter 5 includes the socio-communication model development and sensitivity analysis. Chapter 6 introduces the exploratory modeling algorithm that is designed for policy-makers to explore different mechanism and parameter spaces for discovery of robust strategies. In Chapter 7, the future work and concluding remarks of the research are summarized.

Chapter 2

LITERATURE REVIEW

This chapter provides a background for the research conducted. It starts with general concepts on science and research followed by the definitions that help us to understand the environment of inquiry. Consequentially, the governance mechanisms exist in GPS are summarized supported by the ethnographic studies on Open Source Software (OSS) communities. Thereafter, GPS is examined from CAS perspective and discussion of how to study CAS is done. Domain knowledge is concluded with an introduction to ABM and simulation studies relevant to this research. Then theoretical basis of the computational models implemented in this research is delineated. Subsequently, literature review about collaboration networks such as GPS from innovation and social network perspective is revealed. In the last section, several concepts about robustness and resilience of socio-technical systems are introduced.

2.1 General Concepts on Science

There is no precise definition of science that is widely-accepted. For example, Feynman (1998) has a range of definitions for science from a special method of finding things out, to knowledge arising from the things found out as well as new things brought by the things found out or doing of new things. In another study, science is addressed as a *mode of inquiry*(Epstein, 2008). Besides different interpretations of science, this research is interested in the creativity process and the environmental changes occurred during the creation of knowledge in scientific environments. Therefore, an introduction to traditional science and research activities is summarized in the following section.

2.1.1 Science and Research

In his influential work, Kuhn (1996) discusses how scientific fields develop as a result of cyclic revolutionary efforts. In normal science, researchers study based on past scientific accomplishments that are accepted by the community they study in, and they find themselves practicing parallel to the foundations of that particular community, at the same paradigm. Paradigm is the word for avenues of questions and the style of conducting research. It emerges through time based on the previous successes of the paradigm fellows' scientific efforts. None of the paradigms can explain a phenomenon exactly, but a widely-accepted paradigm survives, forming disciplines and professions. Scientists learn to practice through disciplinary laws, concepts, and their implementations. Since a paradigm cannot explain all the facets of a phenomenon, new phenomena and anomalies occur. These anomalies can result in two outcomes: novelty of fact or novelty of theory. Consequentially, these novelties can cause refinement of the paradigm or failure of it. Failure means crises that would end up with a revolution (new paradigm) or an exception left to be handled for prospective researchers.

Adaptive cycles are a good way to explain this progress in science. Walker et al. (2004) describes four states of an adaptive cycle:"growth and exploitation phase, conservative phase, chaotic collapse, and innovation phase." The growth is cumulative but up to a certain extent. After a paradigm reaches the conservative phase, there needs to be reorganization or a new paradigm shift because it cannot explain the phenomenon any longer. Chaotic collapse brings new opportunities to compete against each other and then innovation phase starts a new cycle with the fittest novel strategy in the environment.

On the other hand, research is defined as puzzle-solving (Kuhn, 1996; Arndt, 1985). The aim of the traditional research is to explain what is known to be there in advance. Although they contribute to the existing paradigm, the provided answers are questionable. Research is cumulative while scientific revolutions are non-cumulative, and science improves by paradigm shifts as a result of the evolutionary process, in which the fittest survive. While contradictory definitions exist about the difference between research and science, Latour (1998) expresses a noteworthy observation to explain the scientific endeavor emerging today:

> Science is certainty; research is uncertainty. Science is supposed to be cold, straight, and detached; research is warm, involving, and risky. Science puts an end to the vagaries of human disputes; research creates controversies. Science produces objectivity by escaping as much as possible from the shackles of ideology, passions, and emotions; research feeds on all of those to render objects of inquiry familiar.

Scientists form hierarchically-structured teams or groups and synchronously conduct traditional science activities in a similar environment. Traditional science has a high-entry threshold, static structure in terms of turnover of the participants, and has finished products, which are usually in the form of publications. Complex traits of the new scientific activities differ from the traditional scientific activities regarding the communication styles, proximity and mobility of the actors, organizational hierarchies, and products of the collaboration.

Gibbons et al. (1997) call the traditional way of doing science Mode-1 knowledge production. It is in a disciplinary framework, hierarchical, and institutionalized while Mode-2 science as a new production process of knowledge is transdisciplinary. Mode-2 science includes many diverse scientists or participants in the process. It has a deeper quality control because the product is socially accountable. It is more formed around an application area or an artifact (any form of information such as document, code, vocabulary) including contributors from diverse ranges of disciplines, and the product is transdisciplinary, which means it is not reversible to the contributing disciplines. Mode 2 threatens the existence of Mode-1 knowledge production, but Mode-1 and Mode-2 live simultaneously.

2.1.2 Open Science Paradigm

Open Science is defined as a mode of knowledge production which has the disclosed knowledge of the earlier participants as an input to future researchers (Mukherjee and Stern, 2009). David (1998) defines the force of universalist pattern of open science as providing entry into scientific artifacts and open discussion by all participants, while promoting "openness" in regard to new findings. Carayol and Dalle (2007) explain open-science phenomenon as significant freedom of scientists to choose whatever they want to do and ho ver they want to do it. It is a Mode-2 way of knowledge production and un-institutionalized.

With the increasing use of the Internet and collaboration tools, terms *e-science* (The FANTOM Consortium, 2005), and *service-oriented science* (Booth et al., 2004) are coined referring to scientific research enabled by networks of loosely connected communicating services. Cofundos (Auer and Braun-Thürmann, 2011) as a stakeholder driven research platform, the openscience project[1], creativecommons[2], and innocentive[3] are some of the web platforms fostering innovation and creating governance mechanisms in open science. Open science summit[4] is another initiative that gathers open science endeavor to discuss the future of this emerging way of doing science. Open Science GRID (OSG), Enabling Grids for eScience in Europe (EGEE), Biomedical Informatics Research Network (BIRN), Network for Earthquake Engineering Simulation (NEES) of National Science Foundation and Open Biomedical Ontologies (OBO) in Sourceforge are some examples of open science initiatives (Foster, 2005).

Merton (1979) described four basic elements of a community: "universalism (a shared interpretation), communism (information sharing), disinterestedness (having

[1] www.openscience.org - As of 4.07.2013

[2] http://creativecommons.org/science - As of 4.07.2013

[3] http://www.innocentive.com/ - As of 4.07.2013

[4] http://opensciencesummit.com/ - As of 4.07.2013

objective scientific inquiry), and organized skepticism (proof and review process)." Open Source Software Development (OSSD) communities having the aforementioned features are also producing science in a form of software. OSSD governance frameworks are definitely beneficial to explain open science. GNU project[5], Apache software foundation[6], and Linux operating system[7] are successful OSSD communities that are still active.

Nowotny et al. (2001) describe the main features of Mode-2 science environment (*agora*), in which science and the public meets. Agora has diverse participants from various disciplines with different interests and expertise. Participants produce not only reliable but also more socially robust knowledge because of continuos review mechanisms in agora. They are self-authorizing, and the expertise is socially distributed. Agora is complex and full of uncertainty, which fosters further innovations co-evolving with the society.

In Sourceforge[8], OBO Foundry[9] can be defined as open source science development platform that has diverse scientists who are setting principles for interoperable ontology creation and are forming the shared terminology in biomedical domain. There is no top-down leadership in communities and it consists of different communities focused on different subjects. These communities are divided into domains which are basically smaller communities. Artifacts and emailing are the main collaboration tools. Artifacts can be perceived as any form of knowledge (e.g. document, code, bug report) and are created and elaborated by the community members evolving through a consensus.

The actors of open science are academics, not only software developers as in the OSS communities. Intellectual product is not only software but can be papers or

[5] http://www.gnu.org/ - As of 4.07.2013

[6] http://www.apache.org/ - As of 4.07.2013

[7] http://www.linux.com/ - As of 4.07.2013

[8] http://sourceforge.net/ - As of 4.07.2013

[9] http://www.obofoundry.org/ - As of 4.07.2013

datasets in digital form (as artifacts). Ostrom and Hess (2007) list the differences between OSSD and traditional scientific practices as sharing not only the research product (e.g. software, paper) but also the research process, which is different than the traditional journal publishing. Others outside the organizational borders can also participate in different scientific research projects. Open commons do not hold full copyright and they foster the speed of publishing the ideas and innovations compared to standard peer-reviewed journal process.

2.1.3 Governance of Open Scientific Communities

One of the most significant problems in organizational scholarship concerns how social collectives govern, organize, and coordinate the actions of individuals to achieve collective outcomes (O'Mahony and Ferraro, 2007). Jensen and Scacchi (2010) aim to develop understanding for how to characterize the ways and means for affecting governance within and across OSS projects, as well as the participants and technologies that enable these projects and the larger communities of practice, in which they operate and interact.

An important concern that arises from the Open Science concept is what kind of incentives should be created to encourage scientists and academics to participate. Even though journals have not paid authors for articles since the beginning of scientific era, there are intangible rewards that authors want their work to be known, read, built upon, used, and cited (Ostrom and Hess, 2007). Ostrom and Hess (2007) also state that the authors can write journal articles without considering what idea is the best seller or what would be noticed by the widest audience. So academic freedom and gaining reputation can be seen as some of the incentives in open science but there needs to be more incentives for better participation. Ostrom and Hess (2007) exert attention on:

> (1) How to license digital content that is not computer software, (2) how to work within the existing norms and incentive structures faced by most scientists and academics in their workplace today, (3) how to govern such a collaboration, and (4) how to finance such an endeavor.

In addition to the previously described governance studies for OSS, Preece and Shneiderman (2009) offer the *reader-to-leader* framework for technology-mediated social participations. They assert that all users first become aware of the participatory and become a reader, some then become contributors, then collaborators, and later on possibly leaders. They claim that people read because (1) they benefit from it, (2) recognition is an important driving force of contributing, (3) interest triggers a contribution which may turn into collaboration, (4) trust plays a role, and (5) altruism is identified as a major motivator for encouraging contribution and collaboration.

Lattemann and Stieglitz (2005) aim to examine central structures and coordination patterns in open source communities. They see intrinsic motivation, group identification processes, learning, and career concerns as the key drivers for a successful cooperation among the participants. In their study, it is assumed that all member groups are partially intrinsically and partially extrinsically motivated in every stage of the life cycle (introduction, growth, maturity and decline or revival) of the community. It can be perceived as positive and negative feedback mechanisms. They point out that conventional control mechanisms are not usable in systems based on volunteer work because there is no possibility to penalize or reward members financially. They advocate that the adequacy of governance tools is related with the motivation, and motivation is related with the different member groups and life cycle stages of a community.

Jensen and Scacchi (2010) illustrate an alternative perspective offering a multi-level analysis and explanation for the governance of OSSD as below, in which this research is interested in the micro-level analysis:

> In particular, Open Source Software Development (OSSD) projects can be examined through a *micro-level* analysis of (a) the actions, beliefs, and motivations of individual OSSD project participants, and (b) the social or technical resources that are mobilized and configured to support, subsidize, and sustain OSSD work and outcomes. Similarly, OSSD projects can be examined through *meso-level* analysis of (c) patterns of cooperation, coordination, control, leadership, role migration, and conflict mitigation, and (d) project alliances and inter-project socio-technical networking. Last, OSSD projects can also be examined through *macro-level* analysis of (e) multi-project OSS ecosystems, and (f) OSSD as a social movement and emerging global culture.

Ostrom and Hess (2007) remind us that the analysis of different types of motivations fall into three groups: "technological, sociopolitical, and economic." They present the main technological reason for someone to participate as the need for software that is unavailable or too expensive and the main socio-political motivations as the belief in the social or political movement and the desire to participate in a broader community with shared interest as well as *passion* to contribute. In their study, economic motivations in OSS communities are building human capital through learning by reading the existing software code and peer-review process, and signaling the ability as an expert (earning reputation).

Understanding governance and the relationships between the actors in these communities are essential to foster the conducive behaviors that steers these communities towards a desired goal. Scientific knowledge creation in such communities is stated as Global Participatory Science (GPS) (Zou and Yilmaz, 2011). Which tools are used to explain this phenomenon are explained in the following sections.

2.2 GPS from Complex Adaptive Systems Perspective

It is demonstrated that science is complex because researchers around the world interact in both competitive and cooperative ways, with no imposed blueprint, and it is adaptive because both the science and participating scientists respond to environmental changes such as funding preferences or new discoveries (Wagner, 2008). Participants of GPS learn from the knowledge repository or through communication with each other. The knowledge network and the social network evolve over time influencing each other and forming macro-level structures.

Complex Adaptive Systems (CAS) can be described as a framework to understand the world around us. Complexity itself can be perceived in two different ways; qualitatively and quantitatively. Standish (2008) asserts that qualitatively, complexity is related with the ability to understand a system or object while quantitatively, complexity is used to define something being more complicated than another. Yam (2005) describes CAS as:

> A new approach to science, which studies how relationships between parts give rise to the collective behaviors of a system and how the system interacts and forms relationships with its environment."

CAS are formed of elements that have wide range in both form and capability (Holland, 1996). Shlesinger (2007) describe CAS as "composed of interacting thoughtful (but perhaps not brilliant) agents." The phrase *not brilliant* raises concerns about *bounded rationality.* Axtell and Epstein (2006) discuss the empirical data, which demonstrate that all individuals should not necessarily be rational to produce efficiency in macro-level outcomes of a system in CAS. Given that individual rationality is bounded, they explore how much rationality should exist in a system to generate macro-level patterns. In CAS, big changes can generate small outcomes, while small perturbations can cause big emergent behaviors. Yam (2005) defines emergence as the interdependence between details and the larger view of a system.

In addition to *bounded rationality*, Monge and Contractor (2003) describe the main elements of complex systems in terms of the network of agents, their attributes or traits, the rules of interaction, and the structures that emerge from these micro-level interactions. Authors list typical classes of agent traits as location, capabilities, and memory. Communication among the agents, as a main interaction mechanism, has an intense effect on the system level behavior of a complex adaptive system (Shlesinger, 2007). Yang and Shan (2008) depict that agents use belief-desire-intention framework to guide their behavior.

In another seminal work *Hidden Order*, Holland (1996) describes 7 basics (four properties and three mechanisms) that are common in CAS. Basic principles that are also adopted by this research are listed below:

- *Aggregation (a property):* In one sense, it means defining similar things in the same class. In another sense, it is emergent macro-level behaviors caused by aggregate micro-level relationships.
- *Tagging (a mechanism):* It results in selective interaction between the agents.
- *Nonlinearity (a property):* There are nonlinear interactions among the components of the system that means a change on a component non-linearly effects the state of another component.
- *Flows (a property):* There is a flow of components and information through the system.
- *Diversity (a property):* There are heterogenous agents and components in the system.
- *Internal Models (a mechanism):* It refers to anticipation among the agents.
- *Building Blocks (a mechanism):* There is a repetition of novel situations or structures that are emerged through building blocks (re-usable categorical parts).

Van Aardt (2004) views the OSS development communities as complex adaptive systems. Muffatto and Faldani (2003) represent the OSS community actors with their interaction flows and depict them in terms of three fundamental processes that

Axelrod and Cohen (2001) identify in complex adaptive systems: variation, interaction, and selection. Scacchi and Jensen (2008) state that recent empirical studies of OSS projects reveal that OSS developers often self-organize into organizational forms characterized as evolving socio-technical interaction networks (STINs). They are self-organizing because usually without extrinsic leadership, they are formed and led.

2.3 Understanding CAS by Agent Based Modeling

Interdisciplinarity and computer-based thought experiments are common features of CAS studies (Holland, 1996). As the purposes of our models diversify and types of the inputs range, models formed become more interdisciplinary. Standard models of others and recombination of mathematical algorithms enrich the process of modeling CAS. Shlesinger (2007) defines the models as maps to develop scientific understanding of the system of interest involving combination of theory, practice, and art. The main purpose of CAS studies is to understand the underlying relationships between parts while mostly people think that the problem is in the parts (Yam, 2005).

The understanding of CAS is more likely to arise with the help of computer-based models (Holland, 1996). Axelrod (1997a) states that applications of simulation in social science are really diverse so that it has no natural home as a field. There are two methods to explore CAS. Agent based modeling (ABM) is bottom-up and Method of Systems Potential (MSP) is top-down approaches which derive CAS properties analytically (Yang and Shan, 2008).

ABM gives us the opportunity to directly identify the entities along with the relationships and capabilities of them. ABM captures emergent phenomena because it has a holistic approach that perceives a system as more than the sum of its constituent parts. The system level behavior cannot be explained by the properties of the units in the system. Since ABM is used more with the behavioral entities, it provides an opportunity to model more realistically.

Holland (1996) states that the abstractions of the reality, which are agent based models in this research, are metaphorical representations and are actually “the world as it might be, not the world as it is.” In computer based simulations, accuracy is expected but the model is not the exact reality. In the following section, some examples are summarized regarding the use of ABM that is related to this research.

2.3.1 Examples of Agent Based Modeling

The use of agent based modeling as an explanatory tool is widespread among disciplines and is gaining more popularity in the last decades. Epstein (2006) asks the question “Can you grow it?” instead of “Can you explain it?” for observed social phenomenon. He gives examples of agent-based models generating social interactions and emergent behaviors in socio-cultural contexts. He calls agent-based computational models as scientific instruments.

There are many inspiring implementations of agent based simulation models that explain different systems and create understanding for different contexts. Carlson and Doyle (2002) discuss *highly optimized tolerance* in a statistical physics environment. Their forest-fire model aims to show connection between micro-level mechanisms and the macro-level failures by building self-organized criticality and robustness barriers. In another study, abstract computational forest-fire models are developed to gain an understanding of mechanisms underlie different ecosystems and ultimately, different fire management strategies are evaluated (Moritz et al., 2005).

This research focuses on socio-technical processes and use of ABM in socio-technical environments. As a good example, Axelrod' s *disseminating culture* model (1997b) examines the relationship between *local convergences* and *global polarization*, and builds social influence mechanisms of individual and group differences in terms of traits and features. In another study, Axelrod (2006) questions how the states form and dissolve from small political actors providing new thinking on the policy making of the real world in order to maintain more sustainable political structures. In their remarkable work on the effect of individual rationality on macro-level behavior, Axtell and Epstein (2006) point out that imitative behavior and the social interactions are not well considered in economic models. Therefore, they develop an agent-based model for timing of retirement incorporated with social interactions and social imitation process to investigate how desired (optimal) behavior converges even with relatively small amount of rational agents.

Epstein et al. (2006) also examine the epidemic case of smallpox in a *county-level* context with different scenarios (vaccination, population, etc.) and feed the simulation with real data to question when and how much vaccination is needed to stop the epidemic. In another study, Epstein (2006) grows an artificial hierarchical management mechanism for a company to measure the adaptability of the employee hierarchies. Then he creates an objective function representing the trade-off between maintenance costs of the managerial layers and costs of missed marketing opportunities. At the end, he calculates the optimum hierarchy of the management for different initial set-ups.

Cui et al. (2009) create a hypothetical open source software development environment presenting a stigmergy approach and investigate whether they can validate the network structures and collaboration frequency among scientists that are emerged from micro-level stigmergy preferences. Yilmaz (2009) develops an understanding of the coordination of OSS communities focusing on governance and the conflict management strategies and measures the performance in terms of collective creativity. Dron and Anderson (2009) perceive the technology not only as a tool but also as a component (processes and rules). They claim that people should design systems considering the significant components, and predictability of the behavior is not likely so people need to make the systems adaptable. McCormack (2007) creates an artificial ecosystem, in which he builds the metaphor between adaptive individual agents and colors exploring novel discovery processes.

2.3.2 Simulation Models of Science

Different scholars use simulation to study scientific domains. For instance, Gilbert (1997) introduces a model to determine whether it is possible to reproduce observed regularities in science using a small number of simple assumptions. His model generates knowledge structures consistent with observed Zipf distributions involving scientific articles and their authorship, but it does not consider social processes as a mechanism. Naveh and Sun (2006) continue their analysis on top of Gilbert' s model and explore how different cognitive settings may affect the aggregate number of scientific articles produced. They argue that using cognitively realistic models in simulation may lead to novel insights in academia, but they just consider implicit and explicit learning.

In the context of collective knowledge creation and diffusion, Cowan and Jonard (2004) simulate the knowledge exchange process to examine the relationship between network performance and the network architecture. Shrager and Langley (1990) perceive science as problem solving including machine learning techniques. However, in these studies, the social interactions are not taken into account. Socio-technical modeling of science and representing knowledge generation as a social phenomenon draw significant attention among researchers. Similarly, in this research, the focus is on micro-level (inter-scientist) behaviors and developing a plausible socio-technical model of GPS.

2.4 Communication and Collective Action in GPS

Communication preferences and opportunities are important interaction mechanisms embedded in the process of knowledge generation in science. Wagner (2008) states that if we can discern identifiable patterns and mechanisms of communication among the scientists then that can lead us to understand how this scientific endeavor works and how policymakers can influence its evolution and growth. So, science can also be perceived as a complex communication network consisting of individual scientists who communicate, form partnerships, create opportunities, share their findings and adapt to new constraints in their environment.

Little is known about the role of the grounding theoretical mechanisms of communication in GPS. Table 2.1 categorizes the social communication theories that are widely known in *Psychology* domain.

Olson (1974) argues in his seminal work *The Logic of Collective Action* that "unless the number of individuals in a group is quite small, or unless there is coercion or some other special device to make individuals act in their common interest, rational, self-interested individuals will not act to achieve their common or group interest." It is essential to understand Olson' s collective action dynamics in order to explain the merits and nature of GPS. A group of people may all benefit greatly from a collective action, yet be unable to act together to achieve it.[10] Ostrom and Hess (2007) also state that the challenge in FOSS commons is how to achieve collective action to create and maintain commons or public good.

Collective action is focused mainly on mutual interests and the possibility of benefits from coordinated action (Monge and Contractor, 2003). There is also a social dilemma introduced by Hardin (1982), in which he asserts that the mutual-interest and individual-interest conflicts resulting in dissolving of the collective action. The dilemma between mutual and self interest is essential.

[10] http://michaelnielsen.org/blog/the-logic-of-collective-action/ - As of 4.07.2013

Table 2.1: Social Communication Theories

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 or 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)

Monge and Contractor (2003)

One of the terms employed by Bonacich (1990) is communication dilemma, which stresses the conflict between individual communication preferences and the organizational needs of communication. Self-interest theories explain some of these communication preferences of the scientists. They postulate that people make what they believe to be rational choices in order to acquire personal benefits. These personal benefits can be human capital, social capital or reputation. In this research, mechanisms for self-interest, exposure, preferential attachment, and communication theories are developed along with the collective action as an underlying mechanism.

2.5 Understanding GPS as an Innovation and Collaboration Network

Kuhn (1996) approaches science from a paradigmatic point of view and perceives it as a collective innovation. It is collective because it builds upon the past achievements of the others and innovation is essential because science causes cyclical revolutions that occur periodically, resulting in formation of new paradigms and iteratively adopting inventions. Even though the definition of innovation is dependent on the system of inquiry, with a more abstract approach, it can be expressed as a critical event that destabilizes the state of the system and leads to a self-organizing new state (Pyka, 2009).

As a systemic sense in science, the emergence of new knowledge structures, new channels of communication and new network topology can be described as innovation. It is known that most of the outputs of an innovation system are the number of publications or patents and the inputs are resources allocated; However, the process that transforms inputs into outputs is a black-box (Milbergs and Vonortas, 2006). The next generation innovation metrics are more focused on emergence. In GPS, plausible underlying socio-technical mechanisms that lead to emergence of desirable macro-level behaviors can be described as the processes that Milbergs and Vonortas (2006) address. Emerging social-network structures and emerging diversity in the topology can be perceived as innovation indicators.

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

Saviotti (2009) perceives the scientific product of an economic system as a knowledge network and introduces network interactions between the knowledge base of the firms. He synthesizes network science, complex systems, innovation, and knowledge networks approaches in his model and analyzes the network connectivity to discuss innovation. Diaz-Guilera et al. (2009) model propagation of innovations analyzing a spread of stimulus among a network in terms of connectivity, and they use complex interaction mechanisms such as punctuated equilibrium, self-organized criticality under the assumption that the cost of connectivity is stable.

Similar to Saviotti (2009) and in addition to network model preferences, the underlying assumptions of social network models created by Gilbert (2006) accepts the maintenance of the network as costless. Thus, Lynne and Gilbert (2009) postulate to limit the size of personal networks because of the costs of keeping the network alive.

Monge and Contractor (2003) list the four characteristics which are critical to the creation of a public good: interests, resources, benefits, and costs. Udehn (1993) states that only self-interest is inadequate and must be replaced by an assumption of mixed motivations. "What is the mix of these motivations?" is the question awaiting for further exploration.

Social network analysis, as one of the lately developed fields, has recently attracted increasing attention among the scientists. Social network analysis constructs networks from social relations and their functions in society (Wasserman, 1994b). Pyka (2009) represent some empirical results on the trends in innovation networks:"(i) The emergence of novelty tends to create new but poorly connected nodes, thus temporarily reducing the connectivity of the system. (ii) The subsequent diffusion of the innovations establishes new links and raises again the connectivity of the system. (iii) As a result of (i) and (ii), the connectivity of the system is likely to fluctuate around a given value." Dhanaraj and Parkhe (2006) present *Hub firms* in an innovation network to manage *knowledge mobility, innovation appropriability*, and *network stability.* They regard the network and the members of the network as coupled and dependent on one another.

All intelligible ideas, information, and data that can be delivered or gathered in a format can be referred to as knowledge (Ostrom and Hess, 2007). Innovation results from the recombination of knowledge held by the collaborators, and the extent to which agents' knowledge complements each others' is an issue of cognitive integrity (Cowan and Jonard, 2004). The introduction of new ideas through weak ties can foster innovation and development of the system (Wagner, 2008). In addition to the artifacts, GPS has interactive communication outputs (Monge and Contractor, 2003). In other words, connectivity of the members (the network itself) and communality can be identified as the goods of the collective action.

Lynne and Gilbert (2009) suggest four different types of network models: regular lattice, small-world, scale-free, and random. Watts (1999) describes four characteristics of a small-world phenomenon. He argues that a small-world network consists of large number of actors which are connected to relatively small numbers of actors. There are no central actors, and the network is sparse. Relationships among actors overlap; that is, friends of friends are more likely to be friends too.

Scale-free networks have a degree distribution that follows a power-law. Albert and Barabási (2002) state that "Most real networks, however, exhibit preferential attachment, such that the likelihood of connecting to a node depends on the degree of the node." The preferential attachment mechanism creates power-law distribution, in which the ones with high level of resources, attract more resources.

Lynne and Gilbert (2009) argue that social networks are not random since people connect with others who are similar to themselves. Scale-free networks are not realistic because people do not only use preferential attachment, in which people connect to the ones, who already have many links. Because, people do not necessarily know who has the most number of connections. Newman and Watts (2006) postulate: "the small-world model is not in general expected to be a very good model of real networks," because small-world models do not produce nodes with high degrees of connectivity. Hence, Lynne and Gilbert (2009) conclude that social network models need to fall somewhere in-between scale-free and small-world, which is a new challenge in the modeling.

2.5.1 Social Network Metrics and Innovation Potential

De Nooy et al. (2005) explain that “the main goal of social network analysis is detecting and interpreting patterns of social ties among actors.” Social networks represent the complexity of human interactions (Wasserman, 1994b) and their topologies are represented by sets of people or social actors and the set of peer-to-peer relationships among them. Social distance mathematically presents a degree of closeness and acceptance that these actors or group of actors feel towards each other (Boguna et al., 2004). Boguna et al. (2004) also discuss three specific issues in social networks: “transitivity of the relationships between peers (clustering), correlations between the number of acquaintances (vertex degree) of peers, and the presence of a community structure with patterns.”

The degree centrality for each actor is measured in order to capture degree distributions. The degree centrality of an actor is calculated as the proportion of possible ties that exist for that particular actor. Another metric that can be considered is ego density a term coined by Burt (1982). The term refers to the proportion of existing ties that includes the actor as a peer. It is a useful metric to assess which nodes or actors are more likely to spread knowledge and innovation (Wasserman, 1994b). Additionally, density is another metric, which is averaged standardized degree in the whole network. Higher density suggests a higher connectivity and group cohesion (Blau, 1977). The variability of individual indices can be quantified so that the degree centrality of a network is calculated as a measure of variability among degrees of actors.

Regarding innovation potential, Yilmaz (2008a) argues that higher density networks have a better mobility of knowledge, which is desirable for innovation; however, higher density also diminishes the positive effects of diversity on innovation by creating shared norms and skills. Therefore, it is essential to measure density along with the diversity of a network. Yilmaz (2008a) identifies high centrality and low density networks as another indicator of innovation potential that leads to more structural holes and transformation of knowledge. Both Yilmaz (2008a); Burt (1995) discuss the importance of high centrality and fewer structural holes as a competing preference.

In order to support the given hypothesis, more metrics need to be explored. Distance metrics between and among the groups such as Euclidian distance, Manhattan distance, Mahalanobis distance, and Hamming distance are some of the most important common metrics in use.[11] Primarily, geodesic distance as a form of Euclidian distance is used in social network metric calculations, whereas the other distance metrics are outside of the scope of the social networks context. However, Hamming distance is useful when analyzing diversity among scientists.

Wasserman (1994b) suggests that closeness centrality and betweenness centrality are indicative of cohesion within the network. Closeness centrality of an actor states how close an actor is to all other actors. There are different approaches for measuring group closeness (Freeman, 1979; Bolland, 1988). The measure is between 0 and 1, and lower values indicate better dissemination of information. Betweenness centrality of an actor is the number of shortest paths that pass through a node divided by all of the shortest paths within the network. This metric is useful for determining the nodes where the network can fall apart. When the normalized metric for the group is calculated, the higher values indicate it is easier to destruct the connectivity in the network because connectivity is highly dependent on a few actors. Additionally, eigenvector centrality and information centrality metrics capture who is connected

[11] http://www.statsoft.com/textbook/cluster-analysis/ - As of 4.07.2013

to the most popular (central) nodes and who is connected to best information paths (in the case of valued links), respectively (Wasserman, 1994a). These metrics can be used in resilience analysis, but they are computationally costly to calculate.

The clustering coefficient is the most important metric of interest for capturing clustering tendencies within the network. It is the density of a node in its neighborhood, calculated from the average of coefficients for all actors. It is indicative of the presence of different communities or groups within the network (Schank and Wagner, 2005). Higher values might indicate sparsely clustered groups or a high connectivity in the whole network as a structure. Therefore, mathematically, the average of all shortest paths between nodes in the network can help to distinguish which structure is present. Also, it is discussed that the longest of the shortest paths in the network can be useful for indicating the diameter in the network (as higher values reflect more sparseness).[12] As a consequence, calculating average path length along with the clustering coefficient would allow us for distinguishing the *high centrality-fewer structural holes* hypothesis.

After how to measure clustering and degree distributions are discussed, the aforementioned third issue in social networks is the fractal-like network structures. One structure to measure is the small-world phenomenon which indicates a higher clustering coefficient and relatively short average path length. These networks are clustered, but there are also many bridges and structural holes between clusters. The small-world phenomenon can be measured by the ratio between the clustering coefficient and the average path length. Greater values indicate a better small-world structure (Uzzi and Spiro, 2005). Additionally, degree centrality and density metrics can be indicative of scale-free structures that have few highly central actors as opposed to the majority of the actors that have small degree centrality.

[12] http://www.slideshare.net/gcheliotis/social-network-analysis-3273045 - As of 4.07.2013

Core/periphery ratio is a relatively new metric, and there is no consensus on how to calculate it. Core/periphery ratio is calculated by simply dividing the number of core members to the number of periphery members. It is a measure of innovation and the larger periphery is better as an indicator of diffusion of innovations (Krebs and Holley, 2002). A well-known technique to identify core and peripheral nodes is the recursive method that removes the nodes with a smaller number of degrees than a predetermined number until there is no node remaining to remove (Boyd et al., 2006). Then, remaining nodes are counted as core nodes while the removed ones are counted as the peripheral nodes.

2.6 Diversity and Innovation Potential

Uzzi and Spiro (2005) analyze small-world phenomenon through innovation in Broadway musicals. They indicate that quality of the show's performances increases with small-world network up to certain extent, after which there is a diminishing effect on performance. Diversity needs to be spurred within the network. Badis et al. (2009) also state that if we observe the companies in a market and ecosystems in the nature, we can see a diminishing return of similarity. At some point, having more similarity things diminishes the rate of benefits. There are also studies that discuss diversity in the population (Dhanaraj and Parkhe, 2006; Powell et al., 1996) and network connectivity as an indicator of innovativeness of the network (Pyka, 2009; Burt, 1995). Interdisciplinarity as a form of diversity is desirable in GPS, and emergent knowledge and collaboration network structures can be used as proxy metrics of innovation potential.

In this research, both interdisciplinarity and the connectivity in the network reveal patterns that allow us to discuss on innovativeness based on those foundations described in the previous section. A detailed summary of diversity metrics that are measured in this research is made in Chapter 5.

2.7 Robustness and Resilience in Socio-technical Systems

It is worth mentioning that robustness has different definitions in different problem domains. In ecology, robustness refers to preservation of diversity in a population, while in medicine, it refers to healing and compensation. In cell biology, robustness refers to how the cell fate decisions are consistent (Krakauer, 2006). Flack et al. (2005) focus on a pigtailed macaque society by removing leaders and observing how the society reacts to this perturbation in terms of conflict management. It is highly related to the self-organization and to the levels of interactions between the individuals in the system.

Pavard et al. (2006) define a robust system as one that adapts its behavior to the unexpected outcomes and perturbations in the environment. Robustness refers specifically to the ability of a system to operate in a desired way when that particular system faces a wide range of operational condition (Sheard and Mostashari, 2008).

Resilience and robestness can have similar definitions in different domains. Resilience is more related with how long does it take for a system to regain a desired output after a perturbation. Smith and Stirling (2008) describe resilience as "the dynamic persistence of a regime under episodic shocks" and robustness as "system maintenance under cumulative stress." There is a need to acknowledge the uncertainty and the lack of knowledge about GPS. So, robustness analysis is considered an essential method for testing since the simulation models are likely to give variable outputs, and because there is a need to capture the robustness of a mechanism through distinct scenarios and parameter set-ups. Robustness is valuable to explore as opposed to only searching for an optimal behavior in terms of a fitness function in an unchanging environment.

In the following Chapter, the stakeholders of this research are introduced, methodology and research questions are briefly summarized.

Chapter 3

RESEARCH PROBLEMS AND METHODOLOGY

In this chapter, the stakeholders in the research and the environment of interest are introduced. Then, significance of the research problems and the contributions that the research are described.

3.1 Stakeholders of the Research

The Science of Science Policy (SoSP) is "an emerging interdisciplinary field aiming to provide scientifically rigorous basis from which policy makers can assess the impacts of scientific enterprise, improve the understanding of its dynamics and assess the likely outcomes."[1] There are three themes of SoSP:

- Understanding Science and Innovation
- Investing in Science and Innovation
- Using the Science of Science Policy to address National Priorities

Innovation is at the core of SoSP themes because the National Academy of Engineering report states that it is critical for US prosperity[2] and is a desirable outcome of all collaboratories. Collaboratories are described as "a computer-supported system that allows scientists to work with each other, facilities, and databases without regard to geographical location" (Finholt and Olson, 1997), which are observed in GPS. From the perspective of SoSP, scientific exercises can be conducted in traditional science environments or GPS environments. Considering its knowledge creation process,

[1] $http : //scienceofsciencepolicy.net$ - As of 4.07.2013

[2] http://www.nae.edu - As of 4.07.2013

science is a collective action taken by diverse, autonomous individuals. Additionally in GPS, scientists are self-organizing all over the globe, collaborating on the same projects regardless of their physical proximity, learning from each other, and doing research without an imposed blueprint. The governance mechanisms of individuals not only affect the individual gains from scientific activities but they also affect emerging macro-level patterns. With the increasing importance of research on scientific enterprise, SoSP created a roadmap for guidance of research on these scientific communities. SoSP asks three fundamental questions relative to this research[3]:

- What are the behavioral foundations of science and innovation?
- How and why do communities of science and innovation form and evolve?
- Is it possible to predict discovery?

Traditional science and GPS activities differ at different aspects of scientific environments. Table 3.1 summarizes the comparison of the features by which both scientific enterprises can be described.

[3] The Science of Science Policy: A Federal Research Roadmap, 2008

Table 3.1: Traditional Science vs. Global Participatory Science

Criteria	Additional Criteria		Traditional Science Teams	Open Science Community
Distribution	Space		Co-located	Distributed
	Time		Synchronous	Asynchronous + synchronous
Communication			Informal	Formal (structured electronic communication)
Organizational	Structure		Hierarchical	Networked
	Style		Team/Formal Group	Community/Market
Openness	Product	Access is	Push-driven	Pull-driven
		Granularity of transparency	Complete product	Incomplete product
		Integration of contributions	Pre-production decisions	Pre and post-production review and use
	Process	Decision-making	Closed	Open/transparent
		Authority	Centralized/hierarchical	Decentralized
Mobility	Entry Threshold		High	Low
	Turnover rate		Low	High

3.2 Research Problems

There is uncertainty regarding the theoretical foundations of science communities. Also it is stated in SoSP roadmap that "theoretical and computational models of science and innovation must be developed!" In the light of the needs stated by SoSP, the initial research questions to be explored in this research are:

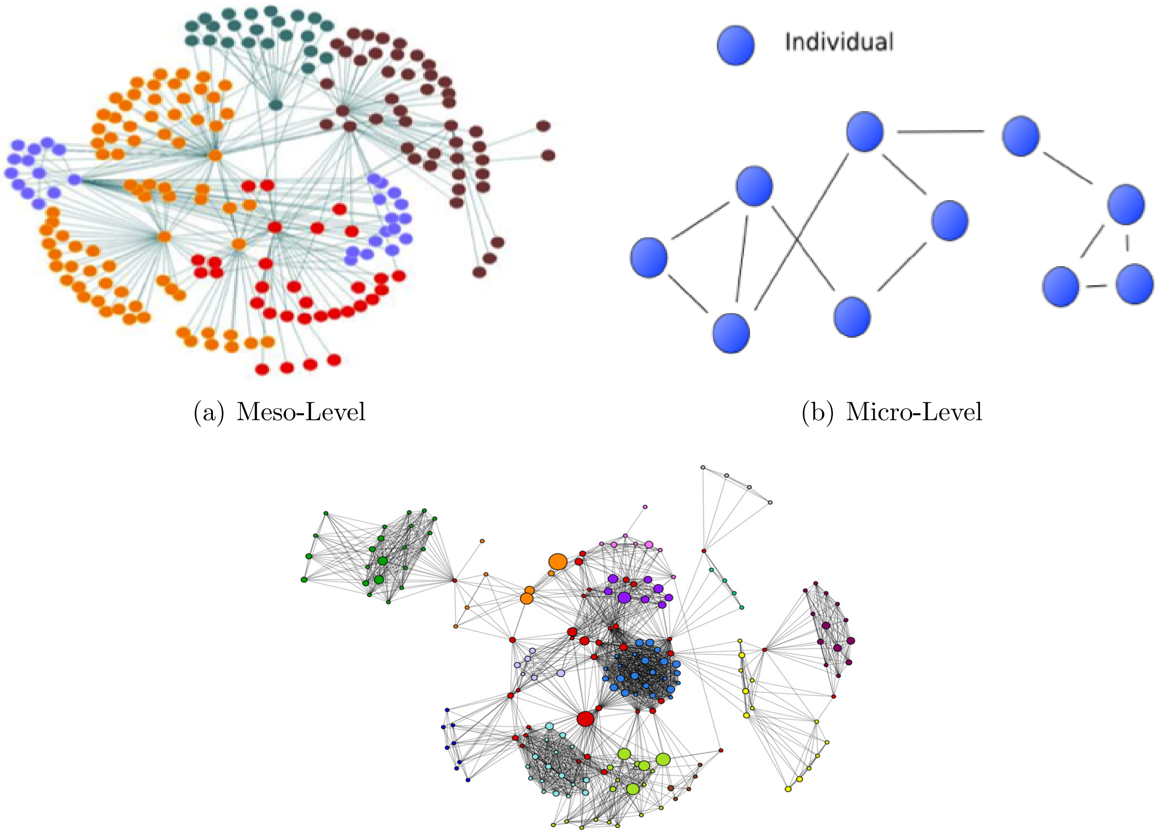
- Which interaction mechanisms in literature explain operational behavior of GPS and its underlying socio-technical processes?

- How we can specify and implement these mechanisms in the form of a computational model to gain empirical insight and perform exploratory analysis?

There are three levels that science can be studied from: micro-level (inter-scientist interactions), meso-level (interactions among communities), and macro-level (communities of communities structures - ecosystem level. Figure 3.1 is introduced for better interpretation of the network topologies. In meso-level analysis, nodes can be perceived as communities, while in micro-level analysis, the nodes can be scientists. In the macro-level analysis nodes can be domains that include many different communities.[4] The links between the nodes can be interpreted as any kind of relationship (i.e. collaboration, social, funding) and this interpretation is based on the intention and the purpose of the model developer.

[4] http://www.cliquecluster.org/content/research-program - As of 4.07.2013

Figure 3.1: Network Visualizations of Micro-level, Meso-level, and Macro-level Science Studies



(a) Meso-Level

(b) Micro-Level

(c) Macro-Level

The CAS characteristics observed in GPS provide for the opportunity to study underlying micro-level (inter-scientist) behaviors in order to develop plausible explanations for this phenomenon. *Collective action* theory is selected an underlying theoretical base for addressing the operational behavior of GPS. More details about model development and implementation of *collective action* theory are described in the following chapters.

Following the development of the base-model, the next step is to explore the emerging network structures in order to understand them. Therefore, policymakers can benefit from their unanticipated opportunities and eventually manage the evolution of GPS networks. Communication among agents, which is present within collaboration networks, has an intense effect on the system level behavior. Theoretically grounded explanations of communication behaviors among scientists provide for an opportunity to explore different communication preferences and their effects on social network structures. The research question of interest in this research is:

- Which social communication mechanisms among scientists are more effective in fostering innovation potential?

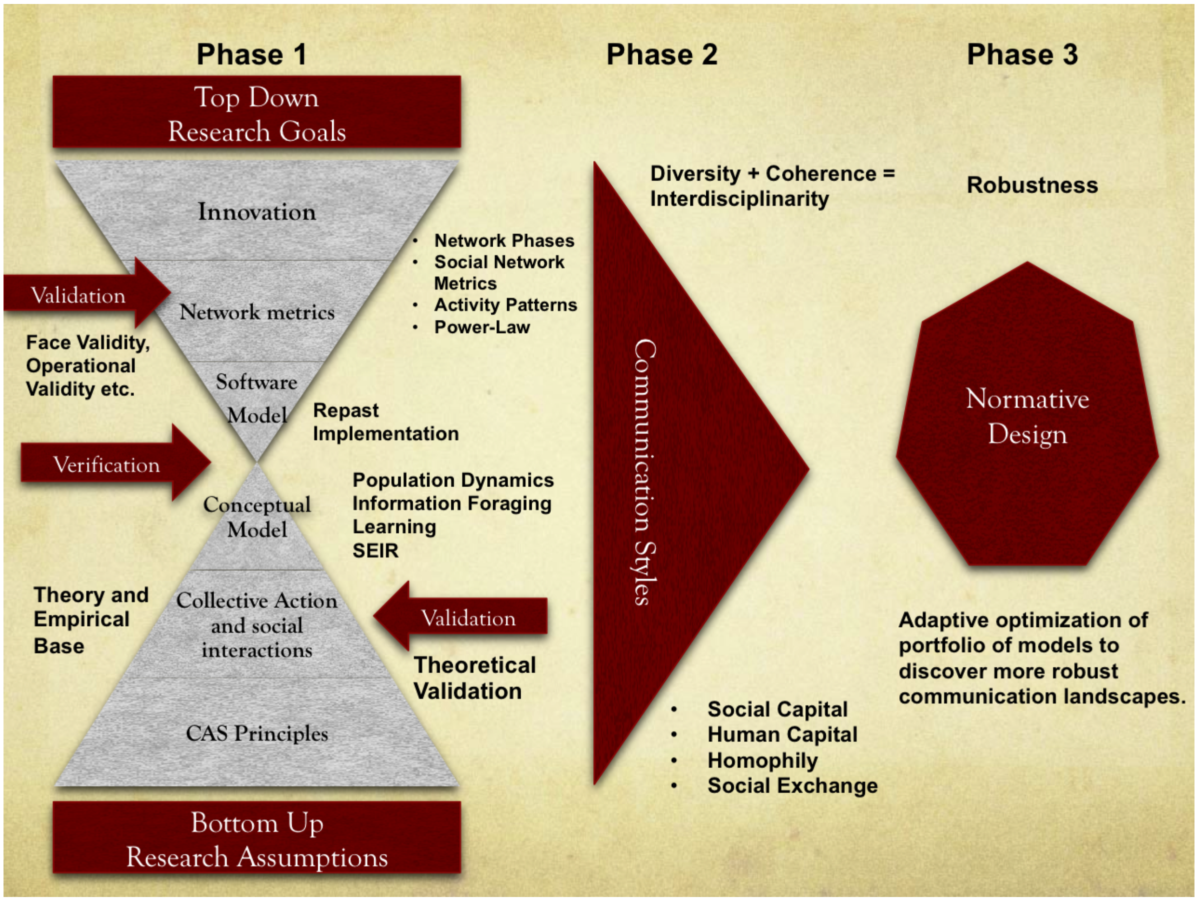
Ultimately, considering the uncertainty of the mechanisms and bounded rationality that exists in the environment (as global information does not exist among agents), variable outcomes are likely to occur. Robustness can be identified as the level of variability the system exhibits under various environmental and intrinsic conditions. Less variable outcomes are indicative of more robust landscapes. Since robust system design is more important than finding an optimal behavior of a single scenario, the research question in this study is:

- How to explore different parameter and mechanism configurations to seek and identify more robust communication strategies in terms of variance observed in innovation potential metrics?

3.3 Methodology Chart

In this research, a bottom-up approach is adopted that has top-down guidance as articulated by the objectives of the study. The methodology is described in Figure 3.2.

Figure 3.2: The Chart of Objectives and Methodology



The base-model is grounded on theories and operating principles derived from observations on the system of interest. First, the base-model is built with CAS principles in mind. In the model, interpretation of *collective action* theory and social interactions complement CAS principles (i.e., tagging, information-flows, diversity, non-linear interactions). Along with the theory base, the information foraging mechanism, which is inspired from food foraging in nature (Pirolli and Card, 1999), is built and preferential attachment mechanism is designed as essential interaction process. SEIR metaphor and population dynamics are introduced which conclude the conceptual model development. Conceptual model development is followed by the implementation of it as a computer simulation.

The second phase of the study builds on the base-model developed in phase one. Different communication preferences of the scientists are implemented. Thus, different communication mechanisms and their effects on innovation potential are sought. The third phase consists of the discovery of more robust configurations by a tool developed to help decision-makers to explore different mechanism and parameter set-ups.

3.3.1 Phase 1 - Conceptual Model Theory Base

Scientists join or leave a problem domain on the basis of problems to be explored and projects to be accomplished, and their position in the scientific enterprise depends upon their knowledge, levels of interest, popularity, personal learning objectives, resources, and commitments (Hollingshead et al., 2002).

In this research, Olson' s *collective action* theory is identified as the socio-cognitive interaction mechanism in GPS. It basically asserts that when the benefits an individual gains are greater than the costs he or she is burdened with, then that individual will join the collective action. GPS is perceived as a collective action because artifacts as a product of the collaboration are *public goods* which are owned by the community and have features such as *jointness of supply* and *impossibility of exclusion.* Because, knowledge produced is open; all may benefit from the knowledge, and benefits of another do not diminish the benefits that can be gained by others.

In GPS, scientists are always in close proximity to one another so long as they collaborate on the Internet. Scientists use web services and platforms to collaborate and socialize. Although traditional science is institutionalized and has certain incentive mechanisms for scientists such as tenure, journal publishing, and funding, these mechanisms do not exist in open science environments. Scientists participate because of an altruistic belief in the action; they believe that collective innovation is necessary in the interest area of that action, they desire to gain greater knowledge, and they want to broadcast their skills and expertise to fellow scientists by weaving a social network.

Basically, scientists follow their self-interests on the theme. But self-interested people are more likely to acquire what they want without paying a great price. Exploitation causes an inevitable free-riding problem, but it does not destruct the value of the work done in GPS. So, interested and self-motivated scientists keep contributing to the collective. The participation in GPS is not compulsory, but the social pressures existing within the open science community leave scientists exposed to the groupthink present in collective behaviors. This phenomenon creates *exposure* to the mutual-interest. The dialectic interaction between mutual-interest and self-interest is essential. While mutual-interest in an action drives an individual to participate, self-interest might cause avoidance from participation, or vice-versa.

Through the development process of base-model, verification and validation studies are also conducted. At each level of conceptual model development, the mechanisms are conceptually grounded on sound theories, based on the others' work, and empirical findings. The detailed summary of verification and validation efforts are listed in Chapter 4.

3.3.2 Phase 2 - Social Communication Model and Innovation

In this phase, social communication theories that are relevant to GPS environment are selected. Generative mechanisms for selected theories are interpreted and recommendations for their implementation are given. Then, sensitivity analysis is conducted to measure innovation potential for different simulation set-ups. Selected communication theories are listed below:

- *Human Capital* mechanism states that scientists have broadcasted information about the expertise of others and try to connect with the other scientists who have higher expertise than themselves.

- *Social Capital* mechanism states that scientists will attach themselves to scientists with high or terminal degrees in the social network.

- *Homophily* mechanism states that scientists can perceive the interest information of others and will try to connect to scientists who are familiar to them.

- *Social Exchange* mechanism states that scientists are going to have the information about what other scientists know and their degree of expertise. Then, a scientist will connect with scientists who are experts in an area in which he or she is not familiar in order to strengthen his or her own expertise.

- *Random* mechanism only allows scientists to connect to other randomly selected scientists.

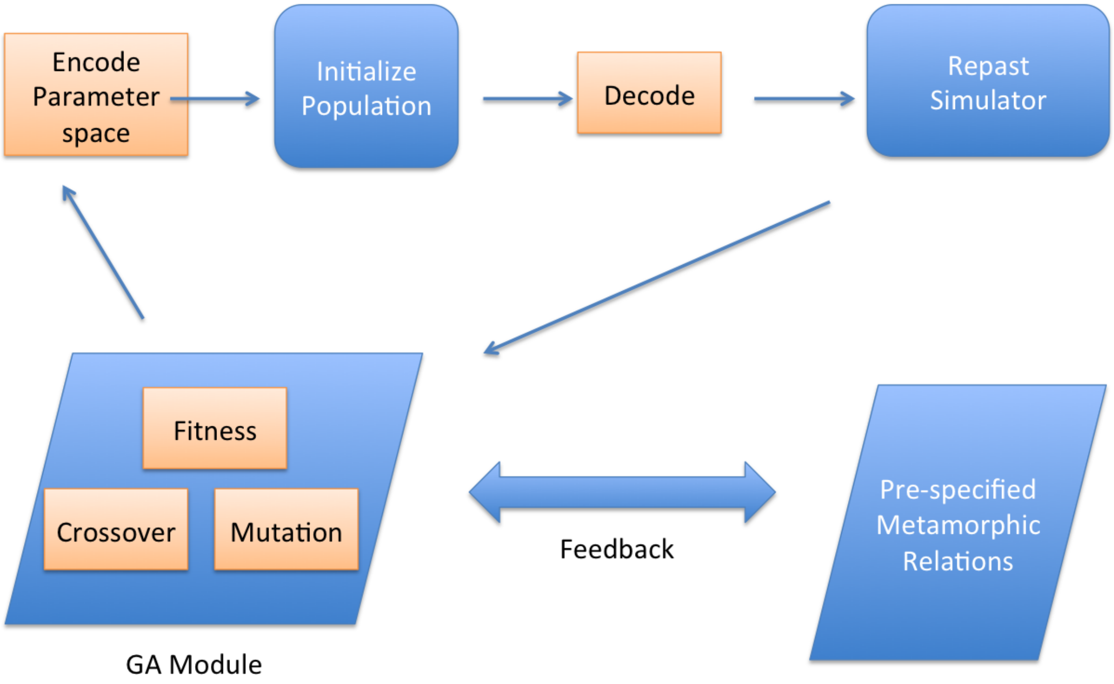
- *Mixed communication* is a scenario that randomly assigns one of the aforementioned five theories to scientists. Each scientist then behaves according to that particular theory. Also, probabilities that are assigned to each theory are parameterized, and further analysis on population dynamics can be conducted.

  Following the implementation, sensitivity runs are conducted. Diversity, interdisciplinary, and social network metrics are measured to be able to distinguish more innovative communication behaviors. The results are presented to policy-makers, so that they can promote desirable behaviors in open science environments.

3.3.3 Phase 3 - Robustness Analysis

In this phase, an algorithm is created that consists of a search algorithm that explores different parameter values and a module that creates plausible simulation scenarios. These two parts are integrated via a feedback mechanism. The results of the search algorithm are measured by an objective function that indicates variability of selected innovation potential metrics. Less variable results are assumed to be more robust. The policy-maker (decision-maker) can generate new plausible scenarios by observing the fittest parameter configurations. The ultimate goal is to find a strategy that behaves more robustly than the others regarding the variability of the performance metrics and to measure the robustness of different communication mechanisms under various conditions. In an environment that has a high level of uncertainty, the more robust strategies should be considered for implementation as opposed to optimal but not robust strategies, because the most innovative strategy might create an unsustainable, highly fragile environment. In Chapter 6, communication between the components of the exploratory software is described in detail. In order to illustrate the high-level structure of the exploratory software, Figure 3.3 is illustrated below. In the next chapter, base-model and its components are introduced in detail along with validation and verification studies conducted.

Figure 3.3: Exploratory Software Coupling GA and Metamorphic Relations



Chapter 4

BASE-MODEL COMPONENTS AND VALIDATION

There is no consensus over what the term *conceptual modeling* means. Robinson et al. (2010) identify some key factors of conceptual modeling in their book. They claim that it starts from problem situation and moves through the questions about modeling such as,"What do we require to model?","What do we model," and "How do we model?" The questions are iterative, and there is continuous feedback with revisions. The conceptual model is simplified. It is not the code or software model, and it considers client-side perspective as much as modeler's. Robinson, Brooks, Kotiadis, and Van Der Zee' s precise definition is:

> The conceptual model is a non-software-specific description of the computer simulation model (that will be, is or has been developed), describing the objectives, inputs, outputs, content, assumptions, and simplifications of the model.

The following sections give information on the conceptual model of the base-model including the assumptions, grounding theories, and interaction mechanisms. A detailed conceptual model description is followed by verification and validation efforts, and analysis to determine the initial conditions of simulation experiments.

4.1 Base-Model Mechanisms

In GPS, scientists participate in artifacts or create new ones without a central authority and meso-level (community-level) governance. In the formulation of base-model, there is no enculturation or entrance threshold for a scientist who is willing to become active and contribute to an artifact. Let us imagine a web tool, in which motivated scientists, who believe in the collective action, can browse the list of open artifacts, select one of them, contribute to it, and thus learn from the artifact. In the following sections, the micro-level interaction mechanisms of the base-model are introduced.

4.1.1 Artifact Selection

It is mentioned in the previous sections that scientists browse the web-tool, which is metaphorically a grid in the model. But, not all scientists are equal in terms of time spent in browsing the online tool. Some scientists browse more titles while some browse fewer. That means the environment is heterogenous regarding the width of scopes among scientists. Each scientist has a scope that is bounded (they do not have the perfect information about the whole environment) and scientists can only operate within that scope while searching for an artifact. The selection process is based on the calculation of three dimensions:

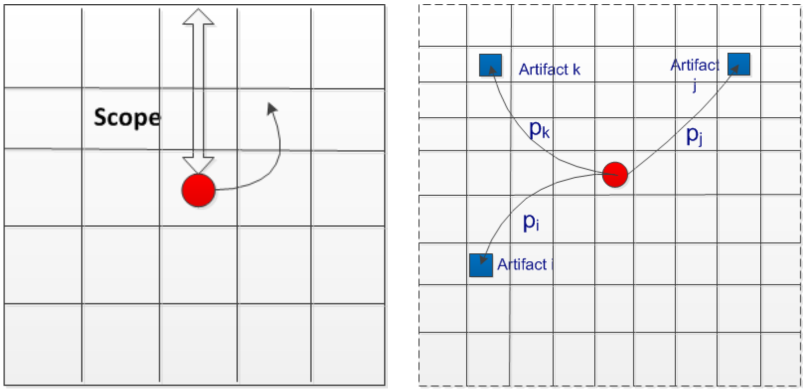
- *Popularity:* Scientists might select an artifact according to the artifact popularity; the more elaborated the artifact is, the more likely it is to be selected ($0 < pa < 1$).
- *Self-interest:* Scientists are more likely to select familiar artifacts ($0 < si < 1$).
- *Imitation:* Artifacts with greater number of active members are more likely to be selected ($0 < im < 1$).

Each dimension has a weight that signifies its importance in selection process. Initially, each weight is equal and $w_{pa} + w_{si} + w_{im} = 1$. Each artifact j has an incentive $P_j = w_{pa} \times pa + w_{si} \times si + w_{im} \times im$. In the case of being exposed to more than one artifact, a roulette wheel selection algorithm is used to assign probability p_j to each artifact j and select one of them based on the assigned probabilities.

$$p_j = \frac{P_j}{\sum_{i=1}^{N} P_i} \tag{4.1}$$

where N is the total number of artifacts that are within the scope of a scientist. Principally, in this research, the roulette wheel algorithm is used in all selection processes since people in real life do not select the most likely decisions but they generally satisfice. That means, instead of selecting the choice that has the best value, they give their decision probabilistically relative to each decision element or criteria. Rationality in decisions are probabilistic based on clues collected from the environment. This idea is supported by Ariely (2008) in *Predictably Irrational*. Figure 4.1 shows the representation of the moving and artifact selection processes on a grid.

Figure 4.1: Moving and Artifact Selection Processes in the Model



(a) Moving of a scientist

(b) Artifact selection of a scientist

Selection process is an important component in the model that encourages some artifacts to be preferred more than the others. Preferential attachment is also an essential mechanism that precipitates the power-law distribution as a CAS hallmark (Holland, 1996). Another motivation that preferential attachment process originates in is what Barabasi (2002) states in his influential work *Linked.* He points out that the *random universe* idea is good for mathematical representation of networks. However, real-social networks are more than random as they have a selection process (Lynne and Gilbert, 2009; Barabasi, 2002).

In the fusion of these three dimensions, the goal is to capture the effects of three different kinds of information perceived by scientists while browsing. There are the aforementioned three kinds of available information, because in the base-model there is a bounded-rationality assumption. Scientists perceive the environment but all information (all parameter and variable values that belong to other agents) is not available for them. This is supported by logic derived from observations gained by the ethnographic analysis on OBO. The logic is: before scientists join an artifact, they can only interpret what the artifact is about, the number of active members contributed recently, and how many posts or contributions are there on that particular artifact. This assumption can also be derived by observing forum websites. So, scientists are not aware about the complexity of artifacts or the degree information of the fellow scientists. In the base-model, scientists associate with an artifact, read, contribute, and learn from it, subsequently they get familiar with that information.

The formulation of the selection process is an additive model. An additive model represents the combined effects of the explanatory variables and their interaction is equal to the sum of their separate effects. This research aims to capture the intensity of these three dimensions by weights that are associated with every single dimension. Subsequently, setting different weights for each dimension can reflect the different importance each scientist attributes to available information in their decision process, which can be considered in the future work. But in the base-model, the weights of each dimension are set initially and are the same among all population members. An advantage of the linear combination method that is described above is flexibility in creation of different scenarios. But there is a question of concern regarding which weighting schema is best (Wu, Bi, 2009). An additive model with a simple weighting is used, which is used by Axtell and Epstein (2006) and Yilmaz and Hunt (2010).

The additive model is easy to implement and can be updated easily. The model is also suitable for calculating average effect. Major difficulties with the model are the definition, assessment, and interpretation of the weights (Belton and Stewart, 2002). Other options are summarized in Wu et al. (2009) such as the Borda Count Method (voting for the values), the Probabilistic Method (instead of looking at the extremes, it focuses on the average), and the Correlation Methods; however, they are outside the scope of this study. Additionally, because of computational complexity, simple weights satisfy the intended purpose of this research.

Another disadvantage of additive model is the assumption that dimensions are independent of each other. It can be argued that the number of recent active members affects the total number of contributions in an artifact. So, these two variables are not independent of each other. An alternative method is to calculate the correlation coefficient for the interdependent dimensions. Accepting that there is no global interpretation of the multi-criteria evaluation described, the relative advantages of different methods vary in different contexts (in this research it is collective behavior) and decision makers (Wood, 2009). The main intentions of additive model are appropriateness of the method in a social process, applications in previous models (Axtell and Epstein, 2006; Yilmaz and Hunt, 2010), flexibility in creating new scenarios, and the simplicity of implementation.

4.1.2 Collective Action Mechanism

There are four major attributes in collective action (Monge and Contractor, 2003). The interpretations of each attribute in GPS are described below:

- *Resources:* Scientists have time and expertise acting as resources devoted to the collective action. Metaphorically, in the base-model, the browsing area of a scientist can be considered to be the time resources that the scientist devotes to collective action.

- *Interest:* Scientists have many scholarly interests and desire to participate in activities that match their interests. Each artifact, which is a product of the collaboration, has a theme. Another characteristic, *altruism*, can be perceived as belief in the collective action.

- *Cost:* Scientists have to bear a *cognitive burden* and *tension* related with the artifact to make a successful contribution. Cognitive burden is a variable that is related to the cognitive difficulties a scientist faces while trying to contribute to an artifact. Tension can be defined as how easy it is to give direction to the artifact and elaborate on it. Tension is related to the phases of the project life cycle in open science communities.

- *Benefit: Familiarity* and *exposure* are the driving forces for participation. Scientists gain benefits such as a growth in the social ties formed with other scientists, also known as social capital and learning.

There are two driving forces for scientists in collective action: self-interest and mutual-interest. Scientists are more likely to benefit from familiar topics (Monge and Contractor, 2003) and as a form of imitation they are more likely to follow the crowd (exposure mechanism). *Familiarity* is the parameter of self-interest and is the average similarity of two lists: interest ($I_k[i]$) of scientist k and the theme ($T_j[i]$) of artifact j. Theme is basically what subjects the artifact is about. Both interest and theme are lists of binary variables. Familiarity $F_{k,\,j}$ for scientist k to artifact j is calculated in equation 4.2, where N is the total number of interest or theme areas. So i is the index of the interest and theme lists.

$$F_{k,\,j} = \frac{1}{N} \sum_{i=1}^{N} Min(I_k[i], T_j[i]) \tag{4.2}$$

Multiple exposure mechanisms are considered in this study. The first examined only the proportion of active scientists in the artifacts. But this strategy does not properly capture the effects of negative feedback (Holland, 1996). An alternative mechanism is to calculate the proportion of active scientists in the whole community; however, scientists do not have the complete information about other scientists. Therefore, the exposure mechanism is built around scientists' individual social networks. The goal is to capture the general trend in the environment through activeness in his or her social network. The implementation described by Axtell and Epstein (2006) drives this approach.

Also, the influence of some scientists is greater than that of other scientists. The weights of the social ties between the scientists are calculated by the collaboration intensity between pairs. The more two scientists collaborate, the stronger the tie between them. Each time two scientists collaborate, the weight of the tie is incremented by certain amount so long as they have a pre-existing social tie between them (initially 0.1). Consequently, weight information is used as the intensity of influence between two scientists. The more two scientists collaborate together, the greater their shared influence.

Exposure is the weighted influence of active scientists in the whole social network of a single scientist. Exposure is defined in Equation 4.3, where $X_{k,\,t}$ is the *exposure* for scientist k at time t, and $A_{k,\,t}$ is binary variable, which is "1" if scientist i is active at time t. w_{ki} is the measure of collaboration intensity between scientist k and scientist i, and N is the total number of scientists k is connected to at time t.

$$X_{k,\,t} = \frac{\sum_{i=1}^{N} w_{ki} \times A_{k,\,t}}{\sum_{i=1}^{N} w_{ki}} \tag{4.3}$$

The *cognitive burden* of a scientist is dependent on two lists: the expertise ($E_k[i]$) of scientist k, and the complexity ($C_j[i]$) of artifact j. Both are defined as a list of real numbers between 0 and 1. For the sake of simplicity, each scientist k is assumed to have a minimum cognitive burden $minB_k$. Cognitive burden of a scientist k for artifact j is the following, where N is the total number of areas and $C_j[i]$ is the complexity of the artifact j on theme i:

$$B_{k,\,j} = minB_k + \frac{\sum_{i=1}^{N} Max(0, C_j[i] - E_k[i])}{N} \tag{4.4}$$

$S_{j,\,t}$ is the maturity of artifact j at time t, which is defined as the average complexity:

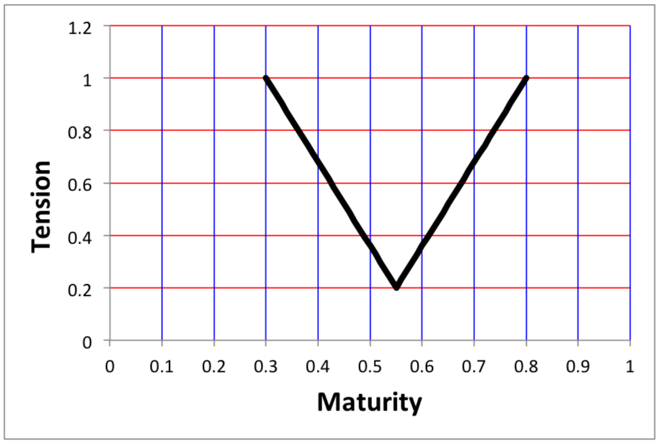
$$S_{j,\,t} = \frac{1}{N}\sum_{i=1}^{N} C_j[i] \tag{4.5}$$

Tension is related to the artifact's maturity and is higher at the beginning of the artifact' s lifetime, since at the early stages of a project it is difficult to have contributions as there exists a tension among contributing scientists in expanding the scope of the artifact theme and complexity. Tension decreases with increasing numbers of collaborations and goes up again when the artifact becomes more mature. Wynn' s project life cycle approach is the underlying assumption here. But for simplicity, the interpretation of tension $(\sigma_{j,\,t})$ in artifact j at time t is a V-Shaped function, in which $Min\sigma_j$ is the minimum tension artifact j has (it is different for each artifact), $S_{j,\,t}$ is the maturity of artifact j, and θ is the mid-point of the maturity range that artifact j can take. The maturity range extends between initial-maturity and completion-maturity of an artifact. Each artifact has a different completion-maturity, at which point the artifact is closed and concluded.

$$\sigma_{j,\,t} = \begin{cases} (1 - \frac{1-Min\sigma_j}{\theta}) \times S_{j,\,t} & \text{if } S_{j,\,t} \leq \text{initial-maturity} + \theta \\ (Min\sigma_j + \frac{1-Min\sigma_j}{\theta}) \times (S_{j,\,t} - \theta) & \text{otherwise} \end{cases} \tag{4.6}$$

For better illustration, Figure 4.2 describes the shape of the introduced function that updates *Tension.* As a note, in Figure 4.2, the completion maturity of the artifact is 0.8, initial maturity is 0.3, and minimum tension is set to 0.2.

Figure 4.2: Shape of the Function that Updates Tension



Some scientists believe in the necessity of scientific collaboration within GPS more than others. The *altruism*, an independent variable, explains belief in collective action. A scientist's decision to become active is based on Olson' s statement in the case of shared costs, which states,"if the benefit is more than the costs of an action, people will participate" (Olson, 1974). There is an analogy between *benefit* and multiplication of self-interest and exposure, and between *cost*, and multiplication of tension in an artifact and the cognitive burden of a scientist. The condition to become active is below:

$$B_{k,\,j} \times \sigma_{j,\,t} - F_{k,\,j} \times X_{k,\,t} \leq Altruism \tag{4.7}$$

where *altruism* is a value which is fixed throughout the simulation and is different for each scientist.

4.1.3 Learning and Influencing Processes

Scientists interact via the flow of information through the artifacts. When a scientist considers cost/benefit analysis (as described above), he or she contributes to repository of the artifact in productive ways; for instance, by commenting, posting a solution, or writing code. The monotonic transfer mechanism as described in Page (2010) regarding information flows is interpreted in GPS and is implemented. If the expertise of the scientist is greater then complexity of the artifact, then a contribution results in:

$$C_j[i] = C_j[i] + (1 - C_j[i]) \times E_k[i] \times \omega_j \tag{4.8}$$

where i is a randomly selected area, j is the contributed artifact, k is contributing scientist, and ω_j is the elasticity of the artifact. The greater elasticity of an artifact indicates that it is easier to elaborate on that particular artifact. This transfer mechanism also refers to faster growth in complexity if the expertise of scientist k is higher than the complexity of artifact j. The transfer mechanism indicates slower growth when complexity approaches to its maximum value "1."

The transfer mechanism for influencing the *expertise* level of the contributor, or learning process, is articulated below. This mechanism indicates slower growth than the influencing process in order to make it harder to gain expertise throughout time:

$$E_k[i] = E_k[i] + (1 - E_k[i]) \times (C_j[i] - E_k[i]) \times B_{k,\, j} \tag{4.9}$$

where i is a randomly selected area, j is the contributed artifact, and k is the contributing scientist. The higher the challenge or cognitive burden is, the higher is the learning of the contributor. Learning is only justified when the complexity of the artifact is greater than the expertise of the scientist. Both transfer mechanisms assume a monotonic increase in the expertise levels of scientists and the complexity levels of artifacts over time. These assumptions are also based on our observations and logic derived from OSSD communities. Expertise is, logically, something that is non-decreasing within the same context and without relative evaluation. Additionally, complexity is observed to increase along with the number of contributions in OSS environments, which makes it harder for fellow scientists to follow and understand the artifacts.

Other than through transfer mechanisms, a contribution may cause a change in the theme of an artifact or it may cause a change in a scientist's interests. The higher the expertise of a scientist, the more likely there is to be a change in an artifact's theme. The higher the maturity of an artifact, the less likely it will be to change the interest level of a scientist. In both cases, a random area on both the interest and theme lists are equalized to demonstrate the influence of the contribution. Additionally, there is a mutation mechanism on an interest area of a scientist with a certain probability (e.g., 0.01) at each time tick since a person's interests are subject to change through time.

4.1.4 Information Foraging Mechanisms

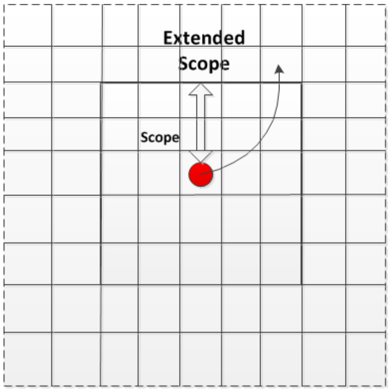
Metaphorically, scientists can be viewed as predators. Predators are expected to abandon their current territory (e.g., domain) when the local capture rate (e.g., success of problem solving) is lower than the estimated capture rate in the overall environment (Bernstein et al., 1988). Information foraging theory, developed by Pirolli and Card (1999), assumes that people, if they have an opportunity, will adjust their strategies or the topology of their environment to maximize their rate of information gain. In this study, scientists join or abandon artifacts based on perceived cues about their performance in attaining the desired outcome.

Every scientist has a different instrumentality, meaning that they have different levels of expectations for the amount of time they should spend on their research until they have a successful contribution. Each scientist has a different initial expectation, which is called *timeToContribution* and shown as $TC_{k,\,t}$ for scientist k at time t. Each scientist k has a memory-factor α_k, which postulates that lower level memory encourages conservative behavior which guides the scientist to maintain their previous expectations. Scientists modify their expectations as following after every successful contribution:

$$TC_{k,\,t} = (\alpha_k \times TP_{k,\,t}) + [(1 - \alpha_k) \times TC_{k,\,t-1}] \tag{4.10}$$

where $TP_{k,\,t}$ is the number of time ticks passed without having a successful contribution for scientist k at time t. If the amount of time passed is more than the modified expectations, then the scientist forages. In foraging, the scope is expanded (e.g., two times) and scientist moves to a different area in the expanded scope on the grid.

Figure 4.3: Information Foraging Behavior



In food foraging, Charnov (1976) states that a forager should leave a territory if the rate of gain (in terms of energy) within the territory that the forager resides in drops below the rate of gain that can be achieved by traveling to a different territory. In Charnov' s Marginal Value Theorem, the gain starts after a certain time t where t is the amount of time forager spends traveling to a new territory. Analogically, in GPS, the amount of time spent for traveling to another territory is almost instantaneous. Therefore, the tradeoff between time spent in traveling and the expected rate of gain is not valid.

In the base-model, the described basic foraging mechanism is used. Also, a second foraging mechanism is developed: optimal foraging a term inspired by Pirolli (2007) and Charnov (1976). Optimal foraging checks the rate of return in terms of expertise a scientist gains from the environment. During a time window, if the rate of return drops consecutively below the maximum rate of return achieved, then the scientist forages. In both strategies, every scientist has a different expectation regarding the amount of time that will pass until the success criteria is achieved.

4.1.5 Population Dynamics

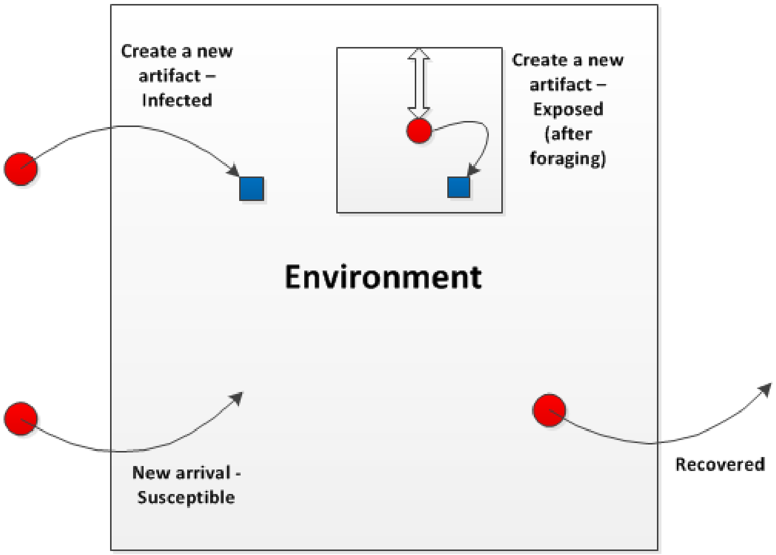
Yilmaz (2008a) states that innovation communities are self-organizing complex adaptive systems; however, not all complex systems are self-organizing (Monge and Contractor, 2003). A system is self-organizing when the network is self-generative (e.g. new arrivals), there is mutual causality between parameters, energy is imported into the system (e.g., creating new artifacts and opportunities), and the system is not in an equilibrium state.

The simulation environment in this research is not a closed system. Like web platforms in real life, the model has new user arrivals. There is no recruiting process for scientists in order to maintain simplicity. New scientists, who start to browse the system, are created in context with a certain arrival rate. At each time tick, with a certain probability (e.g., 0.2), a new arrival enters the system, either creating a new artifact (with probability of 0.05) or just browsing the environment. Figure 4.4 illustrates the new scientists arriving at the system.

4.1.6 SEIR Metaphor

The SEIR model is a widely known epidemiology model (Newman, 2010). It stands for four states of an individual's transition:

Figure 4.4: Population Dynamics in the Environment


Create a new artifact – Infected
Create a new artifact – Exposed (after foraging)
Environment
New arrival - Susceptible
Recovered

- *Susceptible(S)* state describes the initial state. All individuals are susceptible to the collective action or collaboration in the community.

- *Exposed(E)* state represents the interaction with an activity, virus, or idea.

- *Infected(I)* state that describes the influence on an individual by an activity, virus, idea, or some sort of knowledge.

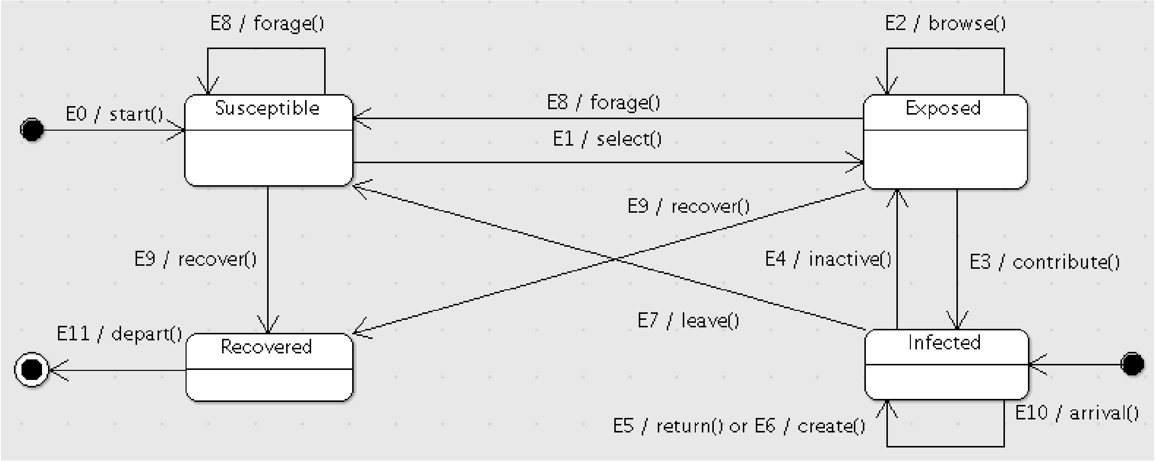
- *Recovered(R)* state is an inactive state. Scientists become inactive and leave the environment in the *Recovered* state.

The metaphor built from the SEIR model is described through state machine formalism in Figure 4.5. A state machine performs actions when a certain event occurs.[1] The state machine illustrated in Figure 4.5 is idle except that the times events are realized. The actions that cause the transitions between the states can be described below:

- *E0/start()*: Initial population and majority of the new arrivals (the ones arriving without creating new artifacts) are initialized.

[1] http://www.agilemodeling.com/artifacts/stateMachineDiagram.htm - As of 4.07.2013

Figure 4.5: SEIR Model



- *E1/select():* Initial population and new arrivals start the simulation at *Susceptible* state. If they find an artifact in their scope, they switch to *Exposed* and move on to the location of the artifact.

- *E2/browse():* After evaluation of the selected artifact, scientists might decide not to become active. In that case, they browse their scopes searching for further opportunities, changing the selected artifact or residing on the same.

- *E3/contribute():* If a scientist decides to become active after the collective action mechanism is evaluated, then he or she transitions to an *Infected* state.

- *E4/inactive():* At the next time tick after a successful contribution, there is a 20% chance that a scientist might change his or her artifact preference switching to one of the past contributed artifacts. Scientists keep a list in mind consisting of the artifacts they contributed in the past and probabilistically select one of them. The more recent an artifact is contributed, the more likely it is to be selected again.

- *E5/return():* An *Infected* scientist evaluates the collective action mechanism at each time tick, decides whether or not to become active in the following time-tick, and transitions to *Exposed.*

- *E6/create():* If a scientist cannot become active for certain amount of time and has foraged for a while, then he or she can create a new artifact (5% chance) which is related with his or her interest areas and become active on it.

- *E7/leave():* Active scientists might not be able to find an artifact in their memory list to study on (e.g. all past artifacts might be closed), so they leave the contributed artifact and transition to a *Susceptible* state.

- *E8/forage():* If a scientist cannot become active during the time that it takes them to have their expectations fulfilled, then the scientist forages, expanding the scope by a factor (initially 2).

- *E9/recover():* If the expertise of a scientist is over a certain threshold (e.g. 0.5, 0.7), he or she can not have a successful contribution, and keep foraging in the environment for a certain amount of time (e.g. 3), then with a certain probability (e.g. 0.2), the scientist becomes *Recovered* and leaves the environment.

- *E10/arrive()*: A proportion of the new arrivals (5% chance) start the system, create a new artifact, and they start to work on that particular artifact in *Infected* state.

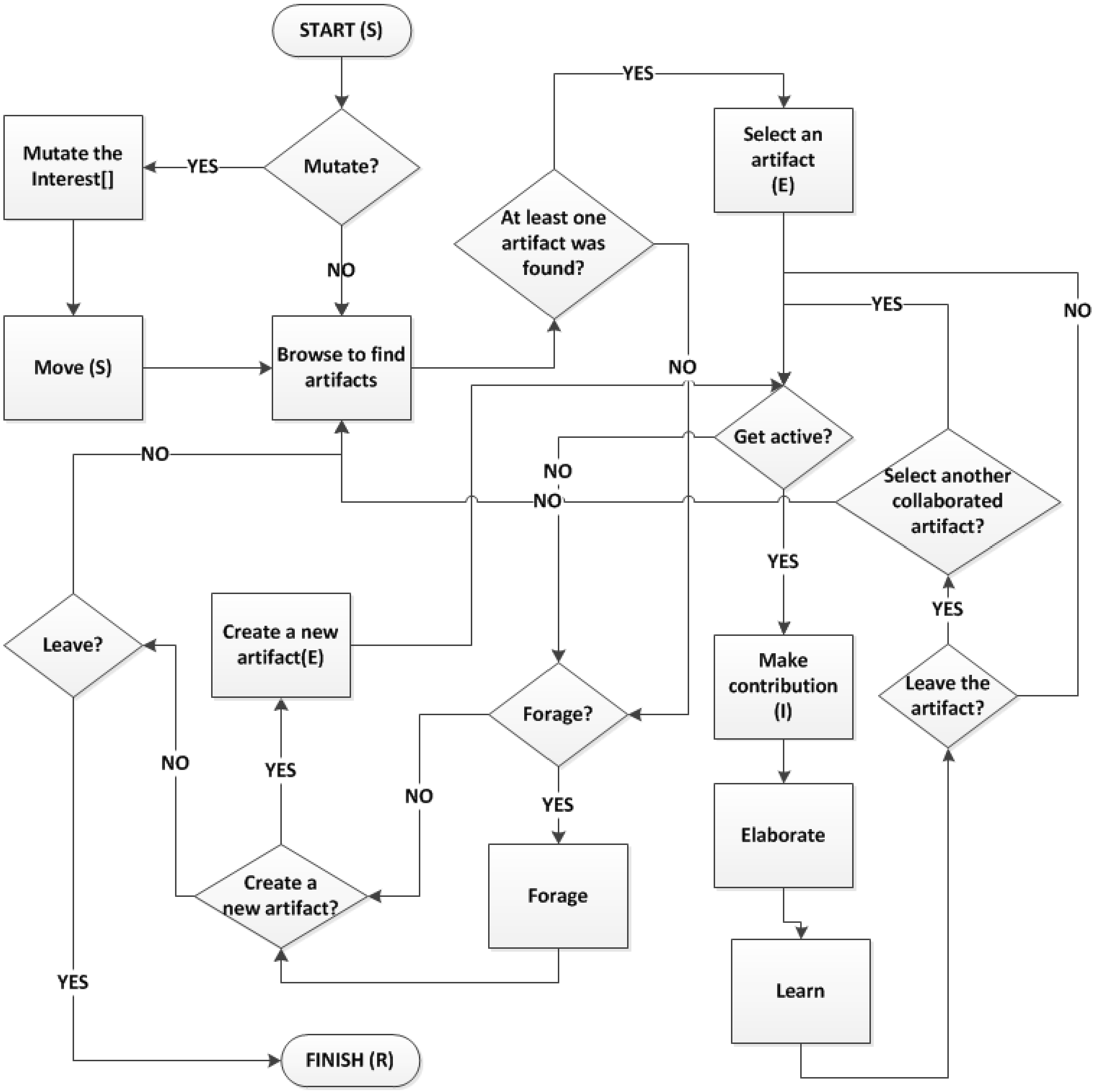
- *E11/depart()*: *Recovered* scientists leave the environment. While in the base-model, the nodes and ties of recovered scientists are retained, whereas in Phase-2, scientists dissolve their social ties and disappear.

The activity flow chart of a scientist in Figure 4.6 represents the flow of the mechanisms as a whole. In flow charts, the processes are associated with the vertices and, when it is on a node, it executes activities.

4.1.7 Conceptual Model Validation

Sargent (2005) defines conceptual model validity as the following:

Figure 4.6: Activity Flow Diagram of a Scientist



(1) The theories and assumptions underlying the conceptual model are correct, and (2) the model representation of the problem entity and the models structure, logic, and mathematical and causal relationships are reasonable for the intended purpose of the model.

If the model outputs are tested by real world data, it is known as black-box validity. Typically in socio-technical environments, providing real world data is a luxury. In order to increase the white box validity, or credibility of a model, certain questions should be answered such as, “How well is the grounding theoretical base of the model?” and “How realistic are the inputs?”

Since the conceptual model is the abstraction in the modeler' s mind, model representation is critical for ensuring better communication between clients, stakeholders, and the modeler. Balci (1990) describes the communicative model, or representation of the conceptual model and identifies six forms: “(1) structured, computer assisted graphs, (2) flowcharts, (3) structured English and pseudocode, (4) entity-cycle (or activity cycle) diagrams, (5) condition specification, and (6) other diagraming techniques.” Good representation of the conceptual model can provide an easier assessment process and better understanding of the model as well as increased credibility. In this research, to delineate the base-model components, the research provides flowcharts, state-charts, visual snapshots, mathematical formulas, and structured English. Additionally, pseudo-codes of some important mechanisms in the implementation are added to the Appendix A. Table 4.1 summarizes the grounding theory base and the assumptions in the base-model.

Table 4.1: Conceptual Mechanisms and Assumptions

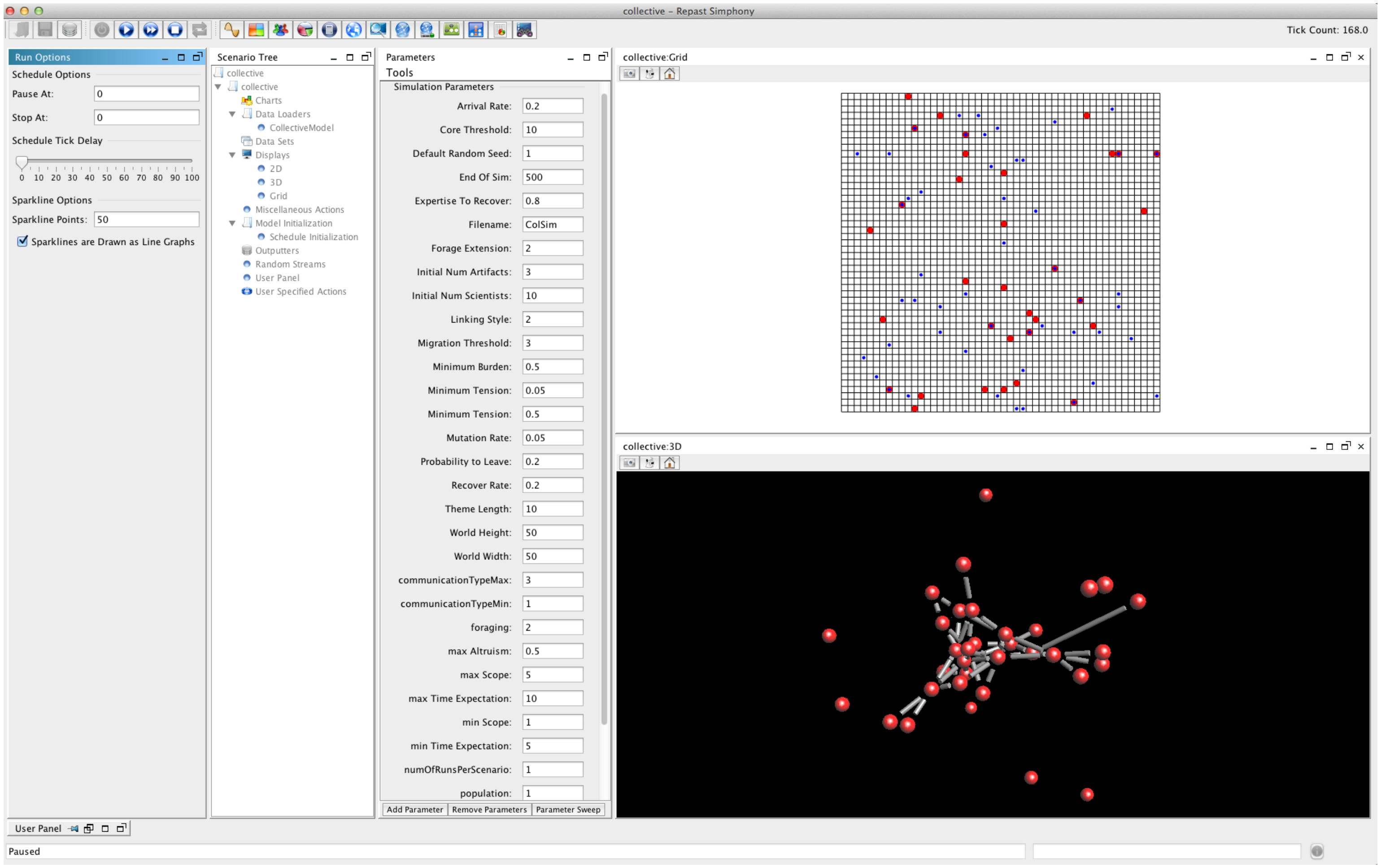
Mechanism/Assumption	**Grounding Theory/Study**
Artifact Selection Mechanism	*Preferential Attachment* Process (Barabasi, 2002)
Bounded Rationality Assumption	Axtell and Epstein' s model (2006)
Roulette Wheel Algorithm	Inspired by *Predictably Irrational* (Ariely, 2008)
Information Foraging Mechanisms	Metaphors in (Pirolli, 2007) and *Marginal Value Theorem* from Charnov (1976)
Collective Action assumption	Olson' s *Collective Action* Theory (1974)
Exposure to Mutual-interest	Axtell and Epstein' s model (2006)
Tension within the Projects	Evolution of Project Life-cycles in OSSD Communities (Wynn, 2003)
Population Dynamics	*Self-Organization* Principles (Camazine, 2003; Monge and Contractor, 2003)
Learning and Influence Mechanisms	Information Flows Process in Page (2010)
CAS Principles (e.g. Tagging, Information Flows)	Holland' s *Hidden Order* Book (1996), Assumptions in Yilmaz (2008a) and Wagner (2008)

4.2 Computational Model and Repast Implementation

Repast (Recursive Porous Agent Simulation Toolkit) is the toolkit used to create the simulation environment. It is an open source simulation tool that allows for development of multi-agent simulation models in Java. Though it has a powerful framework that supports developers while building the context, it is open source and it has a shallow learning curve. The number of demo models and documentation need more detailed explanations. Even though it is possible to encounter maintenance problems related with Repast, it has a highly active community that answers the submitted inquiries and helps users to fix the software bugs.

In this research, the main interaction context is a grid. Both scientists and artifacts are assigned to a cell. Cells are multi-occupancy, which means a cell can have more than one agent. Figure 4.7 represents a snapshot of RePast API. The right-top quadrant is the grid environment, and the right-bottom quadrant is the 3D social network representation. On the far left column, users can schedule the length of the simulations and change their run speed. In the parameters column, users can explore the mechanism and parameter space by entering the values representing a desired scenario. On the grid, blue nodes are artifacts, red ones are the scientists.

Figure 4.7: RePast API



collective – Repast Simphony

Tick Count: 168.0

Run Options

Schedule Options

Pause At: 0

Stop At: 0

Schedule Tick Delay

Sparkline Options

Sparkline Points: 50

☑ Sparklines are Drawn as Line Graphs

Scenario Tree

- collective
  - collective
    - Charts
    - Data Loaders
      - CollectiveModel
    - Data Sets
    - Displays
      - 2D
      - 3D
      - Grid
    - Miscellaneous Actions
    - Model Initialization
      - Schedule Initialization
    - Outputters
    - Random Streams
    - User Panel
    - User Specified Actions

Parameters

Tools

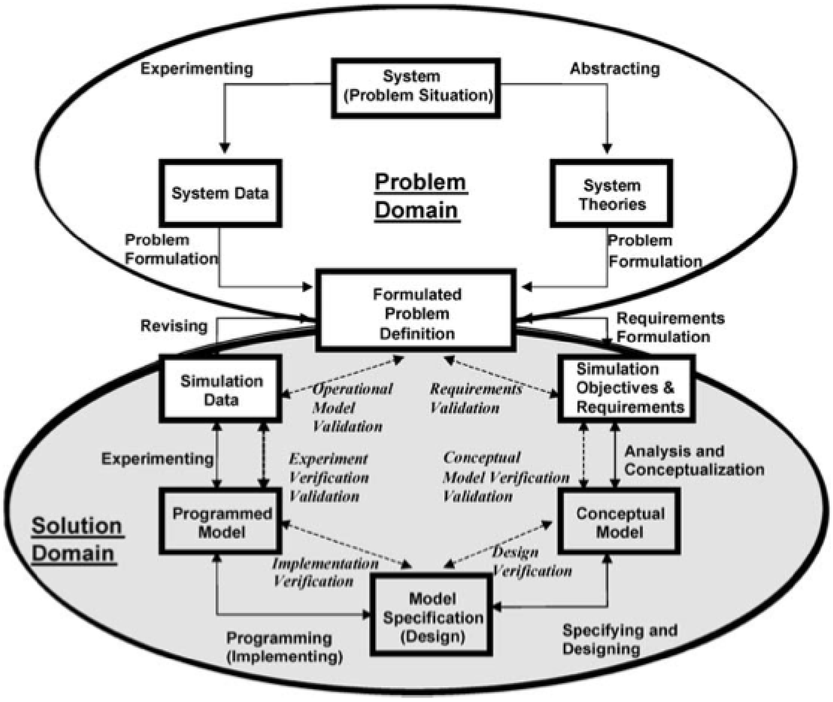
Simulation Parameters

Arrival Rate:	0.2
Core Threshold:	10
Default Random Seed:	1
End Of Sim:	500
Expertise To Recover:	0.8
Filename:	ColSim
Forage Extension:	2
Initial Num Artifacts:	3
Initial Num Scientists:	10
Linking Style:	2
Migration Threshold:	3
Minimum Burden:	0.5
Minimum Tension:	0.05
Minimum Tension:	0.5
Mutation Rate:	0.05
Probability to Leave:	0.2
Recover Rate:	0.2
Theme Length:	10
World Height:	50
World Width:	50
communicationTypeMax:	3
communicationTypeMin:	1
foraging:	2
max Altruism:	0.5
max Scope:	5
max Time Expectation:	10
min Scope:	1
min Time Expectation:	5
numOfRunsPerScenario:	1
population:	1

4.3 Model Verification

Verification indicates how correctly the implemented model represents the conceptual model. When the conceptual model becomes more complex, the magnitude of complexity of the computational model increases significantly. Yilmaz (2006) incorporates model verification and validation in a life cycle of a simulation study.

Figure 4.8: Life Cycle of a Simulation Study


Experimenting
System (Problem Situation)
Abstracting
System Data
Problem Domain
System Theories
Problem Formulation
Problem Formulation
Formulated Problem Definition
Revising
Requirements Formulation
Simulation Data
Operational Model Validation
Requirements Validation
Simulation Objectives & Requirements
Experimenting
Experiment Verification Validation
Conceptual Model Verification Validation
Analysis and Conceptualization
Solution Domain
Programmed Model
Conceptual Model
Implementation Verification
Design Verification
Model Specification (Design)
Programming (Implementing)
Specifying and Designing

Yilmaz (2006)

The end result of verification is technically not a verified model, but rather a model that has passed all the verification tests.[2] The verification tests conducted in this research are listed below:

- Eclipse[3] is used as an editor and a development environment. *Debugging* is used to detect anomalies at each implemented module.[4] Step by step, each piece of

[2] http://jtac.uchicago.edu/conferences/05/resources/V&V_macal_pres.pdf - As of 4.07.2013

[3] http://www.eclipse.org/ - As of 4.07.2013

[4] http://www.ibm.com/developerworks/web/library/wa-debug/index.html - As of 4.07.2013

code that updates variables is checked by the oversight in debugger through reasonabless analysis.

- *Unit-test* principles[5] ensure that every algorithm and method are checked individually at extreme conditions (e.g. the behavior of certain outputs when there is only one scientist in the network) before implementation (Extreme Programming). Flow diagrams are used to verify the code.

- Performance metrics are manually calculated for small populations (typically a network of 5 scientists) and outputs of the simulations are compared with manual calculations.

- Input parameters are printed throughout the simulation to check any inconsistencies that might have been caused.

- The code is self-documented. Every piece of code that is assumed to be important has comments attached to it.

- Data interchange files (gdf) are generated for further verification in the format that records the node and edge data created by the simulations. Then, the network visualization tool *Scibrowser*, written in Python and developed in the Auburn University Simulation and Systems Engineering Lab parses the *gdf* files and calculates the network metrics. Calculations of the implemented Java code are compared with *Scibrowser* outputs.

- OBO Foundry collaboration data is parsed in the Auburn University Simulation and Systems Engineering Lab. This data is explored to determine parameter and output ranges that can be encountered in GPS.

[5] http://geosoft.no/development/unittesting.html - As of 4.07.2013

4.4 Model Validation

Simulation, as a simple definition, generates a model of a system with suitable inputs and observes the outputs (Bratley et al., 1987). Simulation models can be described as abstractions of real-world systems or proposed real-world systems, so they cannot be expected to have every feature of a real system represented in the model (Robinson et al., 2010). From this definition, the question to be asked is: "Do we have a consensus between the model we build and what we intend to do?" The problem formulation and definition are reflected highly on credibility and qualitative performance of a model. Sargent (2005) defines validation as:

> Model validation is usually defined to mean substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.

The approach of Silverman and Bharathy (2011) to validity assessment considers the life cycle of the entire simulation study and assesses the validity under the following four dimensions: "(1) methodological validity, (2) internal validity, (3) external validity, and (4) qualitative, causal and narrative validity." In methodological validity, authors consider modeling process and software process adequacy while obtaining inputs. Internal validity refers to the theoretical base of the behaviors in the model, and external validity examines how reasonable the output data is. Qualitative analysis consists of cross-validation techniques such as face validation, comparison of graphics, and visual analogies.

Regarding methodological validity, the ethnographic analysis is conducted in this research that occurs through the observation of the environment of interest. OBO data is analyzed to understand the possible outcomes and parameter ranges while determining initial conditions of the simulation runs. The simulation modeling process is followed and supported in order to avoid initialization bias, to support terminating state decisions, and to establish the number of replications. In the previous sections, the theoretical basis of the conceptual model was described for internal validity concerns. External validity is checked by observing the variability of outputs for different scenarios. Further, a single scientist is followed in two methods: debugging and visual tracking on Repast API to compare behavior against expected regularities. Qualitative analysis is performed in the next sections, presenting different macro-level emergent patterns for validation purposes.

According to Yilmaz (2006), there are two classifications of validation studies: traditional and holistic/pragmatic approaches. The traditional approach sees the model as either valid or invalid with regard to its application area. The pragmatic/holistic approach does not value the definite correctness or incorrectness of the model. The traditional view supports a division of our model into its parts (reductionist) in order to examine whether parts are representative of the real system or not. The traditional view ensures that the predictive capability of the model is relevant to the real system. But in complex systems, the holistic approach is more dominant, meaning that the system is more than the summation of its parts. The ability of a simulation model to generate an anticipated emergent behavior or to mimic the data does not necessarily mean that it is good representation of reality (Yilmaz, 2006). Silverman and Bharathy (2011) claim that models are frequently evaluated by their capability to estimate an observed phenomenon over a specified range that means each model has a fitness of use.

To what extent the model should be perceived as credible is also another question of concern. Sargent (2005) postulates that validation is usually too costly in terms of time and resources to determine the absolute validity of the model regarding the domain and purpose of the study. While more effort might be better, reasonable enough effort on validation can be satisfactory.

Additionally, validation studies of socio-technical system simulations can be problematic. The lack of data and high-level abstraction in terms of assumptions of individual behaviors make it difficult to assess validity. Klügl (2008) lists the primary obstacles for the validation of the proposed agent-based models including: transient dynamics in the model, non-linearity, amount of effort, and availability of data. Data can be used to train the model and calibrate it before conducting sensitivity analysis. Tuning the model parameters using a meta-heuristic can increase the credibility of the model by representing its ability to mimic real world cases. However, if the model aims for a generalization ability and is desired to be used for exploratory analysis, then creating what-if scenarios and tuning the model would cause an over-fitting problem.

4.4.1 OBO data and Over-fitting Problem

In statistics, over-fitting is the violation of parsimony that means including more terms, variables, and/or procedures than necessary in the model (Hawkins, 2004). Experimenters explore the relationships between the measures. Complicated models are not easy to interpret, which results in an over-fitting problem. The realism achieved by mimicking the real world data and components in great detail may make the model inappropriate and may impair the ability of the model to answer the questions of interest (Laine, 2006). Grunwald (2005) points out the dangers of over-fitting:

> If you over-fit, you think you know more than you really know. If you under-fit, you do not know much but you know you do not know much. In this sense, under-fitting is relatively harmless, but over-fitting is dangerous.

In general, over-fitting happens when a model learns to describe noise in addition to the real dependencies between input and output.[6] Cawley and Talbot (2010)

[6] https://alliance.seas.upenn.edu/ cis520/wiki/index.php?n=Lectures.Overfitting#toc1- As of 4.07.2013

suggest separating model testing and model fitting processes from one another while training the model with a wider range of available data. Reunanen (2003) also suggests dividing available data into training and test sets. This suggestion, however, assumes that the data is identically distributed. In the case of social network metrics in OBO, the time series data has trends and high variability, especially during the early stages of the network. Distinguishing different network phases in the data and dividing each phase into subsets could be a solution to the problem of generating identically distributed data. However, the sample data sets of OBO are not large enough for this practice. Cawley and Talbot (2010) state that it is possible to overcome the over-fitting problem by regularization, early stopping or ensemble algorithms. Early stopping first suggests us separating data into training and test subsets, then training the model with the training set, pausing at times to test the model with the test data. The training is stopped if the test results start to become less significant at which point the validation is optimal. Regularization penalizes the complexity of the model in the fitness function to avoid fitting the noise. An ensemble is defined as a collection of models whose predictions are combined by weighted averaging or voting (Caruana et al., 2004).

In OBO, scientists form communities and domains related to different areas of health sciences while collaborating on the ontology data to standardize the shared terminology. It is a *Sourceforge* style science development activity. In OBO data, the assumption is that if two scientists collaborated on the same artifact in the same month, then they are connected. OBO log-data (between 2000 - 2009) is parsed from *Sourceforge* and the social network data is generated.

Over-fitting is likely to be undesirable when the sample data is small, which is the case for OBO data. There is not a significant amount of data, and additional data is not available. The validation method conducted in this research implements a genetic algorithm that evolves the model parameters, thereby minimizing the absolute difference between simulation outputs and the OBO data. Since there are few communities with a significant amount of data, statistically estimating the distribution of certain metrics became unrealistic. Additionally, fitting the simulation parameters to a single community data would destruct the generalization ability. Hence, OBO data is only used for model calibration (oversight for the level of activity among artifacts and scientists) and validation of emergent patterns. Three macro-level patterns (CAS specific and domain specific) are sought in simulation outputs and OBO data for validation purposes in the following sections.

4.5 Initial Conditions and Terminating State Decision

Naturally, in complex adaptive models, the outputs are highly probable to have cyclic structures. In theory, these systems never stop running and usually do not reach equilibrium, because there are phase transitions in long run. The analysis of steady-state simulations is more difficult than terminating simulations. In this research, the sensitivity analysis is considered to be conducted by measuring the point-estimators at the terminating-state. However, high variability among point-estimators is observed at a terminating state. So, instead of observing point estimators, the average of performance metrics are measured for last 100 time ticks before the terminating state. In order to justify the terminating state decision, preliminary runs are conducted, which have a variety of scenarios. The list of actions taken and conclusions derived from the output behaviors are as the following:

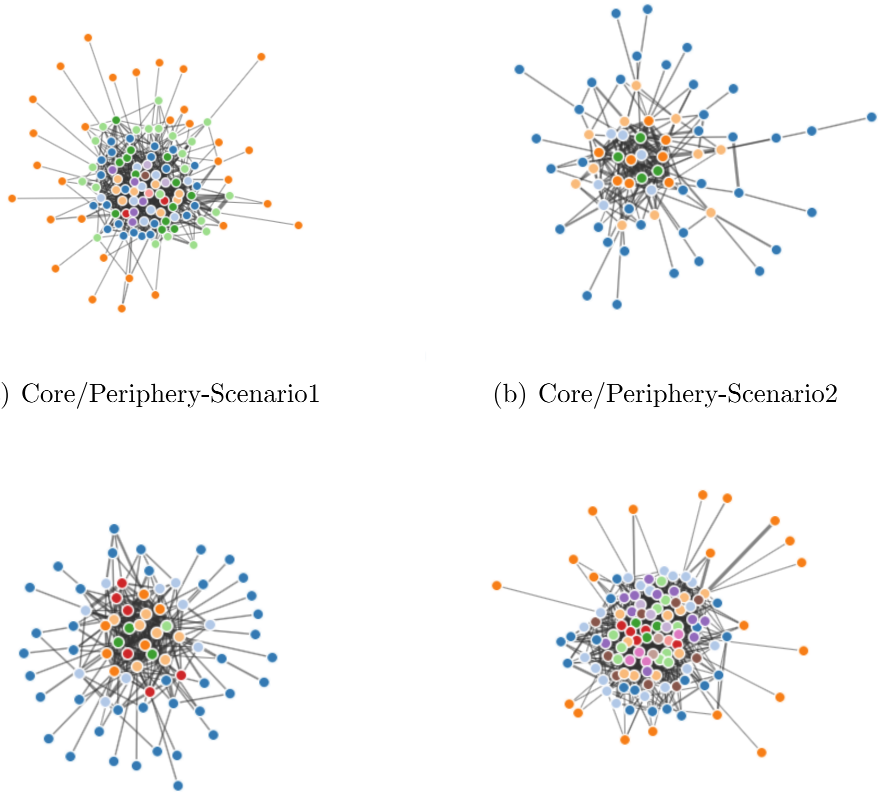
- As a result of the computational complexity, it became impossible to run simulations for thousands of time ticks and to output the time series data for each performance metric. Therefore, the simulation run-length is set to 1250 time ticks for each scenario (30 replications for each), which is observed to be sufficient to discern patterns of different output metrics in the long run. The main goal is to measure the trends of time-series data to identify different stages of network evolution.

- A warm-up period for the simulation runs is not needed, that eliminates the concern of initialization bias.
- Preliminary analysis is conducted for random scenarios, and time series for different performance metrics are plotted. Plots are simply eyeballed for qualitative analysis purposes.
- No terminating state works perfectly for each scenario and each metric, following the stochastic nature of this study. While some scenarios converge to the core/periphery stage after 100 time ticks, some scenarios may need 400 time ticks to converge. Therefore, the terminating state can be determined that is sufficient enough to observe core/periphery stage under various scenarios.

In OBO scenarios before 500 time ticks, variances decrease and the metrics seem to be fluctuating around the same values. In *random* connection scenarios, it is qualitatively discernible that the metrics start to fluctuate around the same values before 500 time ticks, but later time ticks it is possible to observe data trends due to the dissolution of members from the social network. In this study, the goal is to analyze the outputs at the core/periphery stage and to evaluate the network at that stage.

Furthermore, the justification for the terminating state can be supported by the following network snapshots. Figure 4.9 illustrates network snapshots at 400 time ticks for four scenarios extracted from the preliminary runs. These snapshots represent distinct scenarios incorporating high-level of differences between parameter values. The core periphery stage can be observed in the snapshots that has less variable network metrics values. These networks are knitted closely in the core and are more resilient. Hence, the terminating-state of the model is set to 500 time

Figure 4.9: Sample Core-Periphery Structures at time 400



(a) Core/Periphery-Scenario1

(b) Core/Periphery-Scenario2

(c) Core/Periphery-Scenario3

(d) Core/Periphery-Scenario4

ticks, which can be perceived metaphorically as a 10-year collaboration period (by comparing the output to OBO). The analysis are conducted over last 100 time ticks (from 400 to 500 time ticks). Appendix A has the snapshots of different performance metrics for different scenarios in order to illustrate the behavior of time series data.

4.5.1 Initial Conditions

Selected initial parameter values that are determined after the preliminary results were observed are listed in Table 4.2.

Table 4.2: Initial Settings of the Model

Parameter Name	**Initial Value**	**Purpose**
Weigh of Familiarity	0.34	Weight assigned to *Familiarity* in preferential attachment mechanism
Weight of Imitation	0.33	Weight assigned to *Imitation* in preferential attachment mechanism
Weight of Popularity	0.33	Weight assigned to *Popularity* in preferential attachment mechanism
Arrival rate	0.2	Probability that a new scientist arrives in the system at each time tick
End of simulation	500	Indicates the time tick to stop the simulation
Initial number of artifacts	10	Initial number of artifacts created on the context
Initial number of scientists	25	Initial number of scientists created on the context
Probability to leave	0.2	It is a turnover rate for a scientist to leave an artifact
Maximum Altruism	0.5	Maximum level of *Altruism* a scientist can take
Maximum Scope	5	The maximum number of cells that a scientist can browse
Minimum Scope	1	The minimum number of cells that a scientist can browse

Continued on next page

Table 4.2 – *Continued from previous page*

Parameter Name	**Initial Value**	**Purpose**
Maximum Time Expectation	10	Maximum number of time ticks until a reward/contribution
Minimum Time Expectation	5	Minimum number of time ticks until a reward/contribution
Minimum Tension	(0, 0.5]	Range of values that lower bound of *Tension* takes
Minimum Cognitive Burden	(0,0.5]	Range of values that lower bound of *Cognitive burden* takes
Elasticity	(0, 1]	A value closer to one indicates easiness of stretching the complexity of an artifact
Completion Threshold	(Initial_Maturity, 1]	Higher values mean relatively longer life-cycle for an artifact
Memory Factor	(0, 1]	It is the weight given to previous estimation as opposed to new experience in foraging behavior
Core Threshold	5	Number of connections a scientist should have to be core in Core/Periphery calculations
Theme Length	10	Number of bits in *Interest, Theme, Complexity,* and *Expertise* arrays
Forage Extension	2	Multiplier that expands the scope in foraging mechanisms

Continued on next page

Table 4.2 – *Continued from previous page*

Parameter Name	**Initial Value**	**Purpose**
Recover Rate	0.2	Probability of getting in *Recovered* state if the conditions are occurred
World Width	50	Number of cells the grid has horizontally
World Height	50	Number of cells the grid has vertically
Foraging Mechanism	2	Basic (2) or Optimal Foraging(1) mechanisms
Migration Threshold	3	Number of migrations to be experienced before artifact creation and leaving
Artifact Creation Rate	0.05	The probability that new arrival or a scientist who passed migration threshold creates a new artifact
Mutation Rate	0.01	Probability that a scientist will change his/her interest at each time tick

4.5.2 Bonferroni Analysis

After the length of runs has been established to avoid initialization bias and to set the conditions for sensitivity analysis, the next step is to decide on the number of replications (n) that will be used. Output metrics are not normally distributed in the models of this research and the normality assumption can not be drawn. The lack of distribution is inconsequential because if n replications for each scenario are conducted, and are repeated for r times with different random number seeds, then the mean of each replication batch is expected to be normally distributed. This is a result of the Central Limit Theorem.[7]

Sampling variability is a primary concern in the assignment of number n. In this research, Bonferroni Analysis is conducted to determine the number of replications because multiple performance measures are observed. The so-called problem of *multiple comparisons* should be mitigated. There may be a number sufficient for estimating an output metric with a given confidence interval. But for different scenarios and different metrics, the number of replications could vary. Regardless, Bonferroni inequality states that "all intervals should contain their performance measure simultaneously." Relative to this concern, the Bonferroni inequality addresses an overall probability of at least $1 - \alpha$ that the confidence interval of all k metrics contain their own expected performance measures. If the confidence interval of metric s is 1-α_s, then Bonferroni inequality states:

$$P(\text{All intervals contain their respective performance measure}) \geq 1 - \sum_{s=1}^{k} \alpha_s \quad (4.11)$$

Fourty-four scenarios (8 OBO and 36 Random) and five performance metrics are used for the Bonferroni analysis. α_s is set to 0.02 for each metric. t-statistics are used to determine the minimum number of n that would assure that all metrics fall between their respective confidence interval with overall confidence of $1 - 0.10$, simultaneously. The initial number n is set to 30 replications. The decision of the half-width to assure is set to 10% of the mean. So, the educated guess of n is found as:

$$t^2_{n-1,1-\alpha_s/2} \times \frac{s^2}{h^2} \quad (4.12)$$

[7] http://www.math.csusb.edu/faculty/stanton/probstat/clt.html - As of 4.07.2013

where $t_{n-1,1-\alpha_s/2}$ is t statistics, s^2 is the variance, and h^2 is the square of the half-width. The analysis is done for each scenario by recording the mean and standard deviation at 500 time ticks. Table 4.3 below represents the maximum n that is found among fourthy-four scenarios for each metric. *Density* is excluded from the analysis because of the high coefficient of variation. Additionally, scenarios with low arrival rates (e.g. 0.1) and small initial populations are excluded from the analysis due to the significant effects of node removal on the variability of social network metrics in smaller (population) social networks. The reason for the exclusions was to avoid the impact of high variability on the n determined by Bonferroni Analysis. Also, it is observed that under the initial conditions, these exclusions do not have an impact on n that is determined.

Table 4.3: Maximum n Values Found

	Avg. Path Length	**DC**	**CC**	**DiversityS**	**DiversityN**
Mean	2.75	0.27	0.28	0.26	0.47
Standard Deviation	0.35	0.08	0.10	0.08	0.01
Half-width	0.275	0.027	0.028	0.026	0.047
Maximum n	9.91	48.60	**78.14**	62.18	0.62

($t_{29,1-0.01} = 2.462$)

A conservative approach is taken as the number of replications is set to 100, which is assumed to be sufficient in order to avoid sampling variability. With regard to the computational complexity and time constraints, conducting 100 replications provides an acceptable performance. As a note, in Bonferroni analysis, the metrics such as *diversity among artifacts* and *diversity among links* are excluded because they have patterns similar to the other *diversity* metrics that are used in the analysis. It is also revealed that the scenarios with small populations and arrival rates should be analyzed with large numbers of replications (around 500) or else they should be excluded from the sensitivity analysis to be able to capture statistical significance.

4.6 Activity Time Series

Fluctuating time series are familiar representation of dynamical systems (Kendall, 2001). Complex adaptive systems, which include non-linear dynamics, exhibit attractors. Attractor can be subset of the states that system can phase through emerging from typical initial conditions.[8] It is common to observe more than one attractor in complex adaptive systems. Chaotic attractors seem random but actually they constitute a complex order that does not repeat, which represents that a dynamical system behaves within certain ranges of possible behaviors (Goldstein, 2011). Kendall (2001) indicates that nonlinear population dynamics represent chaotic attractors. Hence, the activity diagrams of simulation runs and OBO data for a randomly chosen community are plotted.

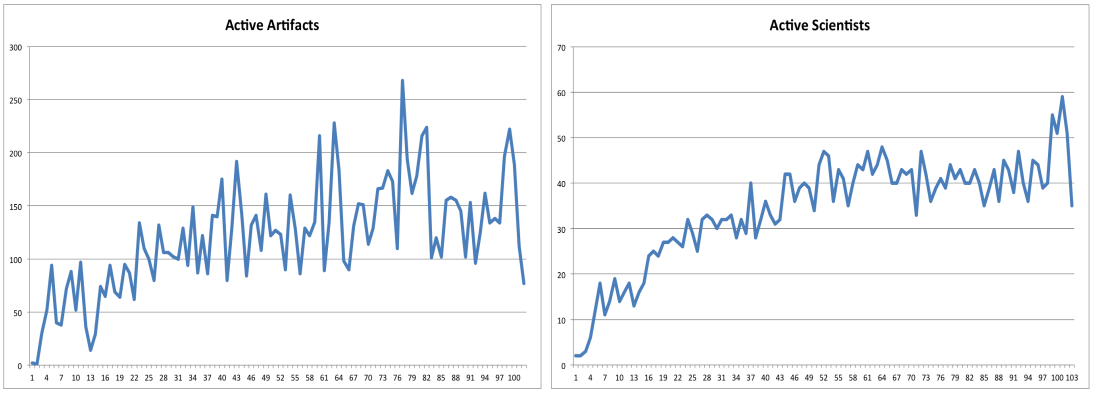
In Figure 4.10, activity diagrams of a single community in OBO are illustrated. Figure 4.11 also presents the representation of simulation results for a single run to illustrate similar fluctuating structures. This pattern is observed at each run in the simulation outputs. In both simulation and OBO data, chaotic attractors are observed as a hallmark of complex adaptive systems. These findings suggest that the CAS assumption is legitimate in the model and the real world data (OBO) also exhibits the same CAS patterns.

4.7 Power-Law Distributions

Another phenomenon this research explores is *scale-free* network structures, which creates power law distribution (Blank and Solomon, 2000). Network topology is expected to have a small number of highly central users with a substantial number of links to others while most of the network members have small number of links. The contribution data is suspected to have the same behavior, which means that a small number of scientists have high number of contributions while most other

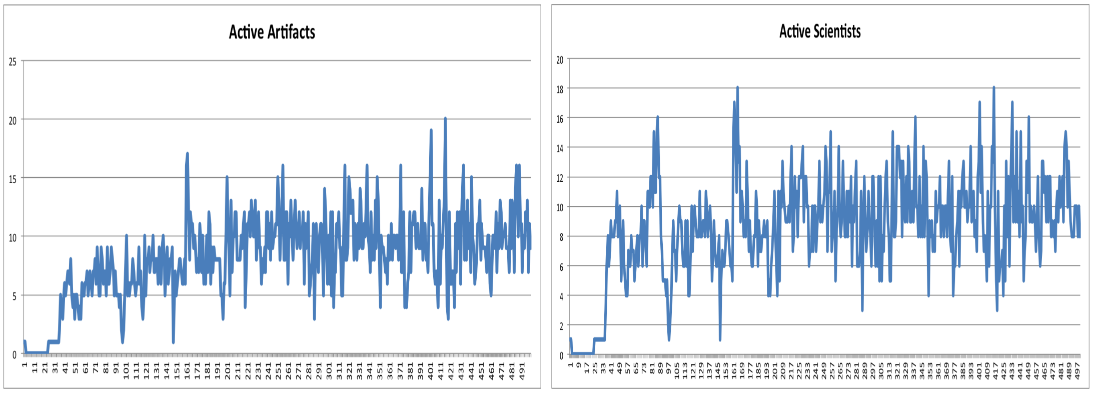
[8] http://www.scholarpedia.org/article/Basin_of_attraction - As of 4.07.2013

Figure 4.10: Number of Active Artifacts and Active Scientists Over Time - OBO



(a) Active artifacts (b) Active scientists

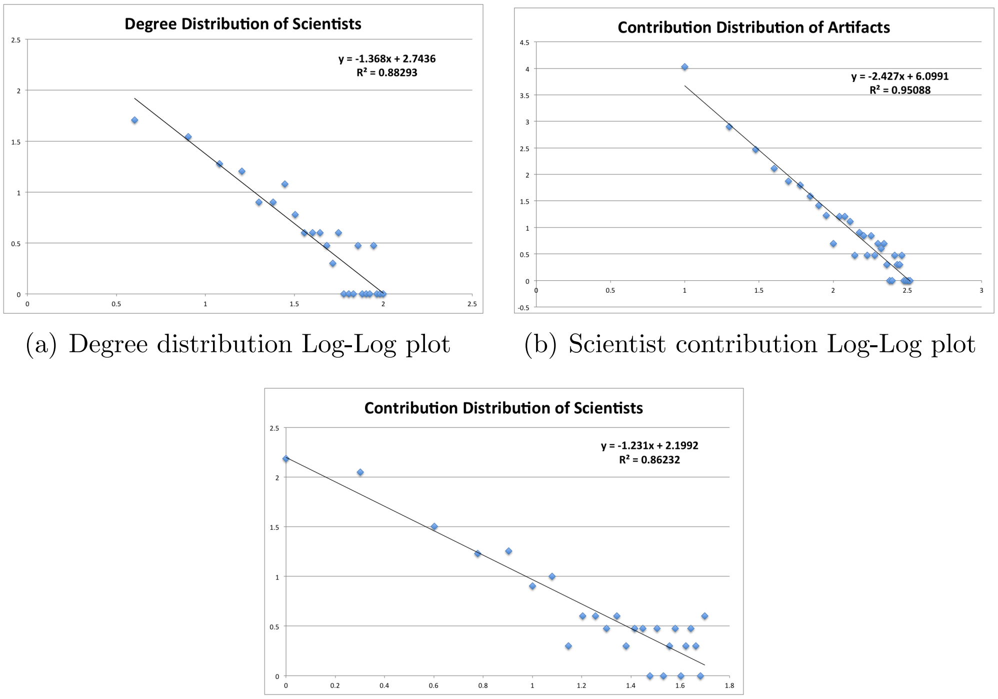
Figure 4.11: Number of Active Artifacts and Active Scientists Over Time - Simulated



(a) Active artifacts (b) Active scientists

scientists have smaller numbers of contributions. The observation of power law distributions is also peculiar to CAS. Figure 4.12 shows Log-Log plots of degree and contribution distributions of cumulative OBO data.

Figure 4.12: Log-Log Plot of Degree Distribution and Contribution Distributions of OBO



(a) Degree distribution Log-Log plot

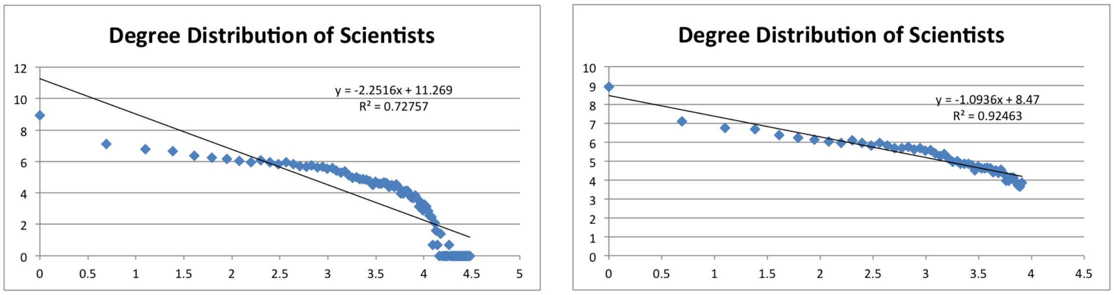
(b) Scientist contribution Log-Log plot

(c) Scientist contribution Log-Log plot

Power law distributions indicate that the magnitude of a phenomenon is inversely proportional to the frequency of that particular phenomenon. Fitness can be analyzed by linear regression of log-log space. Because of multiple observations of the same value, there is a noise in the tail. There are two ways to create bins of data. The first way is to have equal width for each bin, and the second way is to normalize the widths of bins such as logarithmic bins. If there is not a good fit, the data can be fitted from a minimum value or until a maximum value since power law distributions are sometimes mixed with another distributions.

The Log-Log diagrams of the contribution distribution among scientists, distribution of degree information, and contribution distribution of artifacts are represented below. In order to generate more data and better illustrations, the simulations with the initial parameters are run for 200 times, after which the data is accumulated. Each graph contains bins of equal width.

Figure 4.13: Log-Log plot of Degree Distribution of Scientists - Simulated



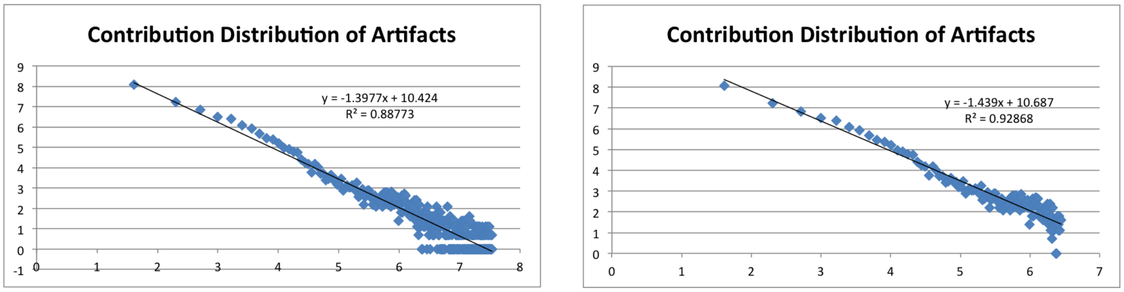
(a) Degree distribution Log-Log plot

(b) Degree distribution Log-Log plot - After cutting off the tail

It is not possible to confidently derive a power law distribution from Figure 4.13 part (a). The main mechanism that causes a power law distribution in the base-model is *preferential attachment.* A reason for having an exponential decrease in the tail is the lack of highly central members in the community. That is why outliers are excluded after the bin that falls between 55 and 60 connections in Figure 4.13 part (b) so that a good fit is observed, which indicates that the power law distribution exists until certain degree values have been reached. The contribution distribution for artifacts is represented in Figure 4.14.

In Figure 4.14 part (a), all data containing that has a high noise in the tail is included. Figure 4.14 part (b) shows the data that is cut off from a certain point, at which the continuity of histogram data (bins) starts to disconnect. The contribution distribution of artifacts is highly indicative of power law distribution, which is expected because of the artifact selection mechanisms implemented. In Figure 4.15 there is also a good fit for the contribution distribution of scientists. The power law

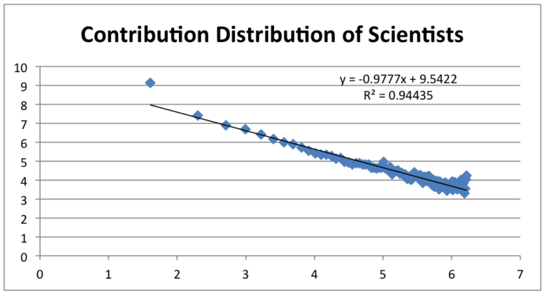
Figure 4.14: Log-Log Plot of Contribution Distribution of Artifacts - Simulated



(a) Contribution distribution Log-Log plot of artifacts

(b) Contribution distribution Log-Log plot of artifacts - After cutting off the tail

Figure 4.15: Log-Log Plot of Contribution Distribution of Scientists - Simulated



(a) Contribution distribution Log-Log plot of scientists

results support the CAS assumption and validity of the CAS principles implemented in this research.

4.8 Collaboration Network Phases

Creation of a fractal-like structure in the course of evolution is another CAS emergent property (Yang and Shan, 2008). Regarding collaboration networks, there are four stages through which a network evolves. The network topology is observed and animated through time to discern scattered, one hub, multi-hub, and core/periphery structures consecutively (Krebs and Holley, 2002). These four main stages that collaboration networks phase through are:

- Scattered Clusters: The stage where the community starts with emergent clusters isolated from each other.

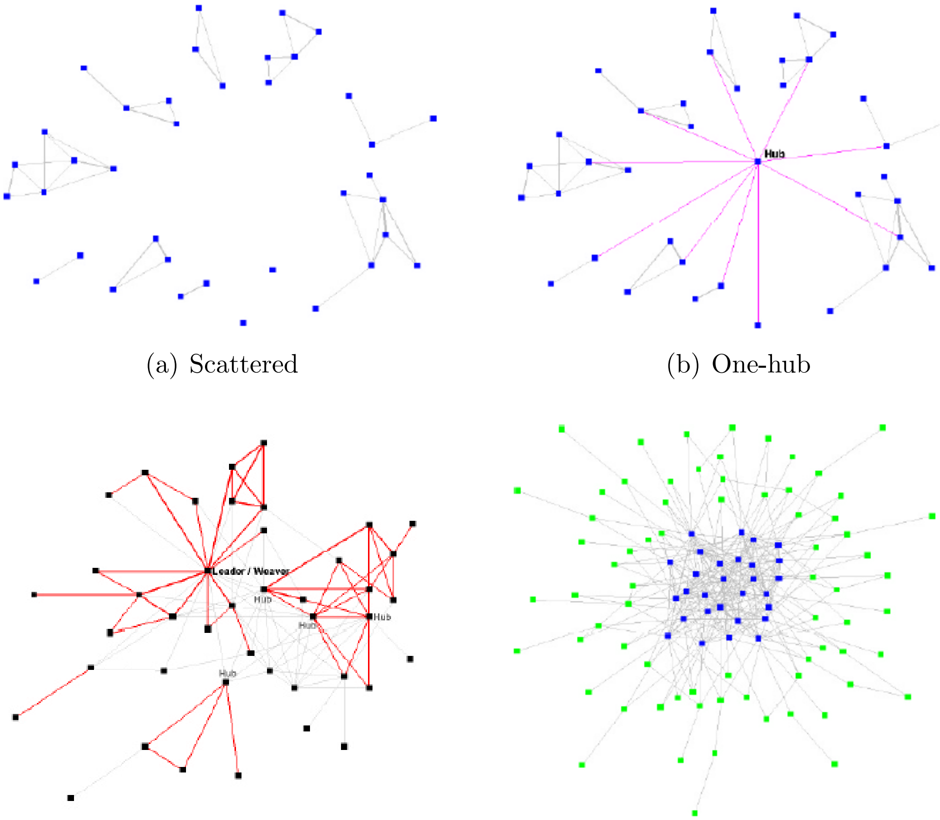
- Single Hub-and-Spoke: The stage where a hub or single actor begins to connect different clusters.

- Multi-Hub, Small-World Network: In this stage, other hubs start to emerge that are connected by weak ties.

- Core/Periphery: This stage emerges after a long period of weaving by the hubs. It is stable and easy to maintain.

As described in the theoretical model (Figure 4.16), the four phases of collaboration networks are discernible in OBO data (Figure 4.17). The snapshots in Figure 4.18 are taken from a single simulation run to illustrate the generated collaboration networks. In contrast to the simulation data, OBO has star-like structures which indicates a connection between core members and a significant number of inactive users who have only one connection to that particular star. Due to the OBO, some core members close or conclude the artifacts that are inactive; as a result, they form a connection with the creators who never elaborate or create any artifacts. The phase transitions can also be detected in the simulation data, and the snapshots may be used for face-validity purposes.

Figure 4.16: Emergent Network Patterns Over Time - Theoretical Model



(a) Scattered

(b) One-hub

(c) Multi-Hub

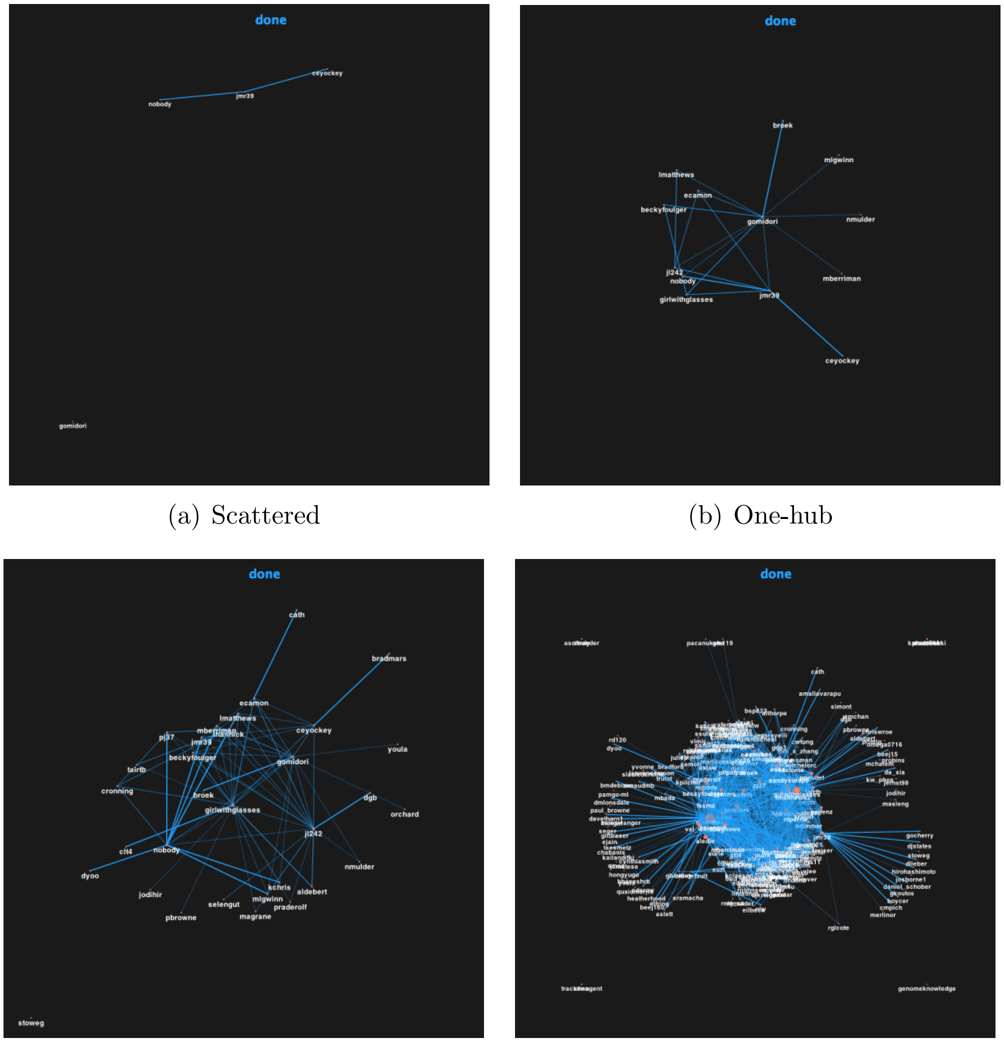
(d) Core/Periphery

Krebs and Holley (2002)

The snapshots reveal that the simulated networks exhibit the theoretical stages through which a collaboration network evolve. This finding supports the model validation. The same patterns are also supported by OBO data observations.

In Chapter 5, socio-communication model is introduced along with the definitions of the output metrics. Subsequently, sensitivity analysis are conducted for various scenarios and innovation potential is discussed.

Figure 4.17: Emergent Network Patterns Over Time - OBO Data



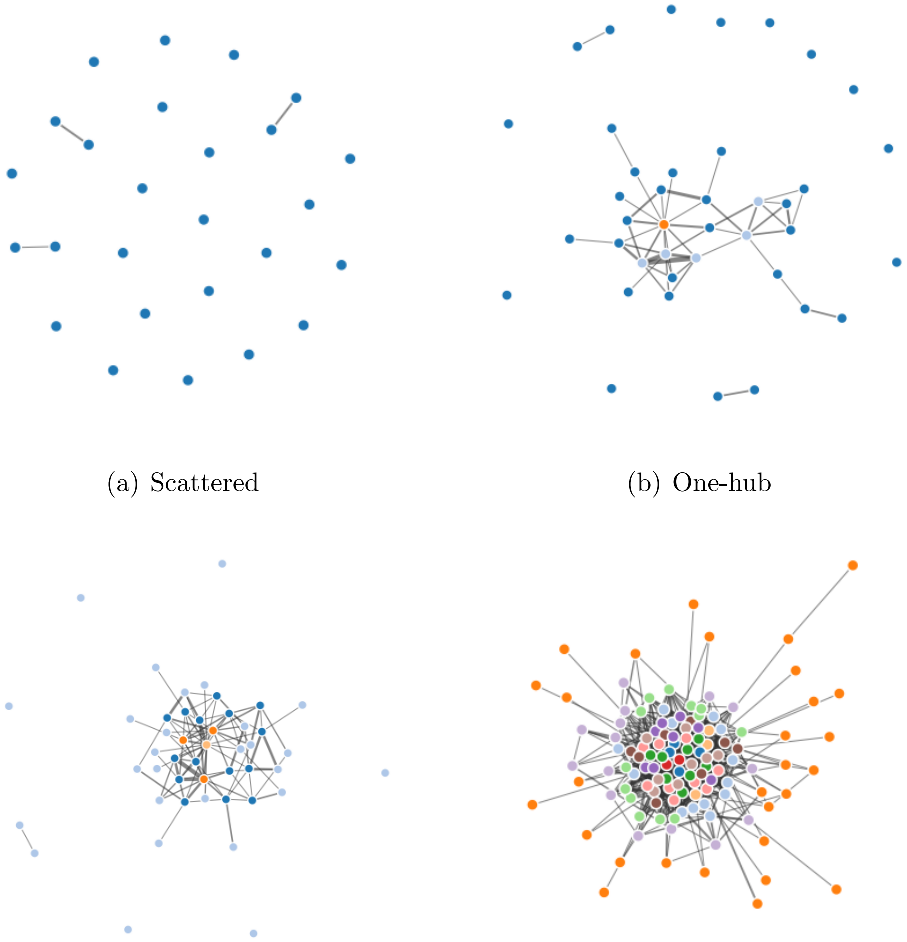
(a) Scattered

(b) One-hub

(c) Multi-Hub

(d) Core/Periphery

Figure 4.18: Emergent Network Patterns Over Time - Simulated



(a) Scattered

(b) One-hub

(c) Multi-Hub

(d) Core/Periphery

Chapter 5

SOCIO-COMMUNICATION MODEL AND EXPLORATORY ANALYSIS

The assumption in the base-model states that if two scientists contribute to an artifact at the same time period, they are then connected. This assumption on how to connect scientists is changed before different communication theories are introduced to the base-model. In the socio-communication model, scientists attempt to attach themselves with other scientists based on their communication preferences. In the following sections, the shared assumptions among different communication mechanisms and implementations of particular communication theories in GPS are discussed.

5.1 Communication Preferences

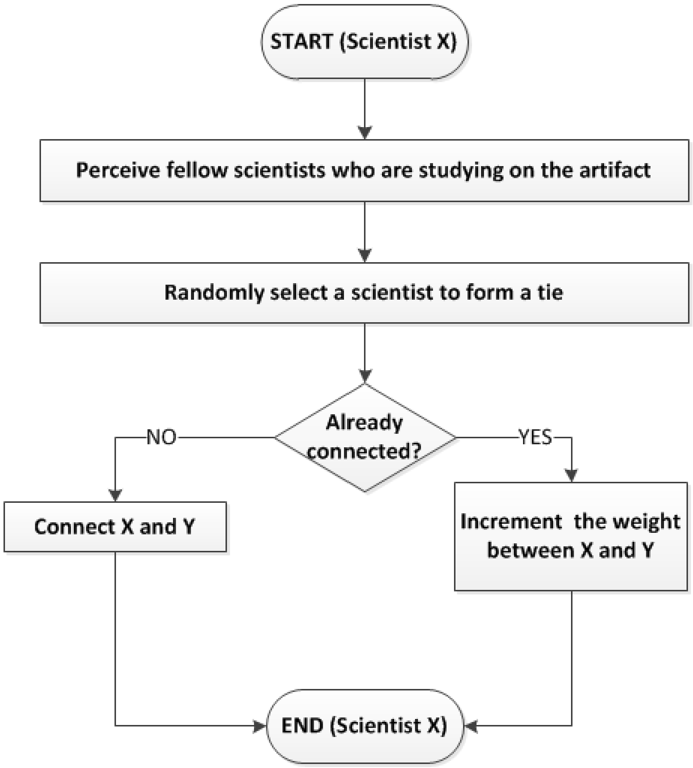
Scientists are familiar with the people who study the same artifacts that they themselves study. Scientists may only consider others who elaborated on the artifact themselves are active on, at the same time tick. Also, at each time tick, an active scientist can try a connection only once. This is a matter of *resources* scientists have. As an underlying assumption, the scientists are homogeneous and have the same amount of resources. If the connection/interaction request is accepted and the target scientist is already a member of his or her network, then the weight of the tie between them is increased incrementally. *Reciprocity* is an important concern when forming a tie with another. Even though scientists are willing to connect to the ones with higher expertise or higher degree, these scientists might not want to connect in response. In order to address the reciprocity issue, the tendency and motivations of the other scientist, who is selected to be connected, should be taken into account.

If forming a tie is a decision of both parties, then successful connection should also be based on how the source scientist is perceived by the target scientist who was selected. The mechanisms for each selected communication theory are described in the following sections.

5.1.1 Random Connection

In this mechanism, a scientist selects another scientist to create a random connection. The flow diagram describes the random process below:

Figure 5.1: Activity Flow Diagram of Random Connection Mechanism

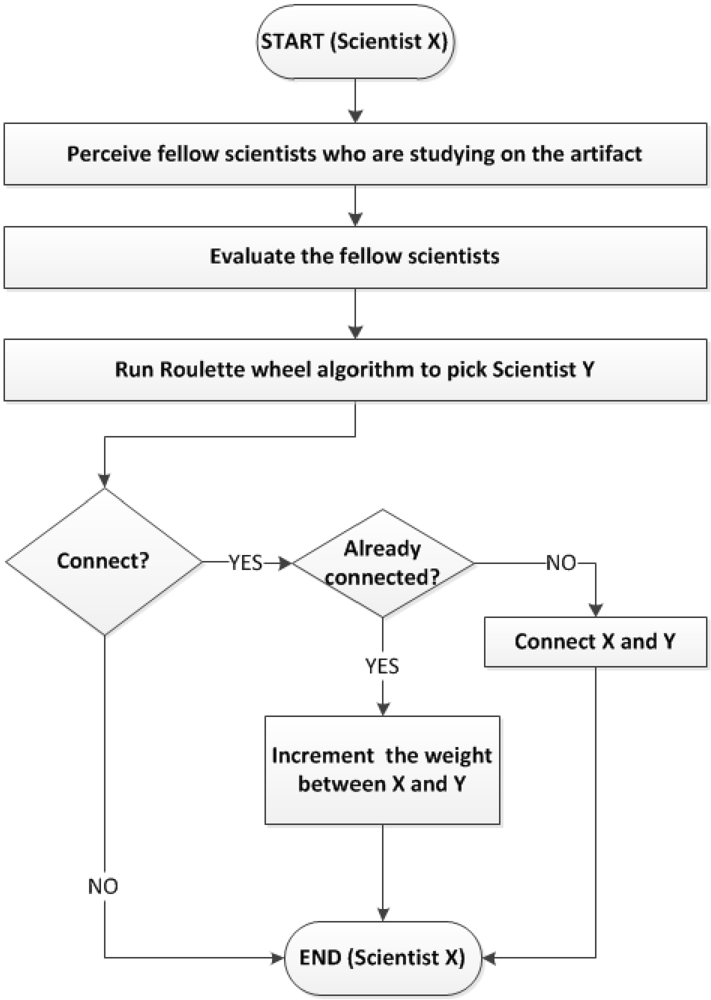

START (Scientist X)
Perceive fellow scientists who are studying on the artifact
Randomly select a scientist to form a tie
Already connected?
NO
YES
Connect X and Y
Increment the weight between X and Y
END (Scientist X)

5.1.2 Human Capital

Human capital is defined in terms of attributes and characteristics that one has, such as reputation and knowledge. The theory of human capital explains that people who have greater numbers of attributes and features gain more advantages in the network (Becker, 1978).

In this research, the *human capital* mechanism states that scientists have the broadcasted expertise information of others and they attempt to connect to other scientists with higher expertise. The general activity flow diagram of communication mechanisms that are implemented are described in Figure 5.2. In *human capital* mechanism, the evaluation is based on the expertise levels of the fellow scientists who study on the same artifact. The greater the expertise of scientists, the more likely they are to be selected by others.

Figure 5.2: Activity Flow Diagram of Communication Mechanisms


START (Scientist X)
Perceive fellow scientists who are studying on the artifact
Evaluate the fellow scientists
Run Roulette wheel algorithm to pick Scientist Y
Connect?
YES
Already connected?
NO
Connect X and Y
YES
Increment the weight between X and Y
NO
END (Scientist X)

Decision criteria P_{ij} for scientist i to select scientist j is the average expertise level of scientist j:

$$P_j = \frac{\sum_{n=1}^{M} E_j[n]}{M} \tag{5.1}$$

where M represents the total number of expertise areas and $E_j[n]$ is the expertise level of scientists j on area n. In the following formula below, p_{ij} is the probability of i to select j from candidate scientists.

$$p_{ij} = \frac{P_j}{\sum_{n=1}^{N} P_n} \tag{5.2}$$

where N is the total number of scientists who actively study on the same artifact. Roulette wheel algorithm is used to determine which scientist is selected as j by scientist i. Regarding reciprocity, the probability that the connection will be successful is represented by the Equation 5.3 below.

$$p_{ji} = \frac{P_i}{\sum_{n=1}^{N} P_n} \tag{5.3}$$

5.1.3 Social Capital

Social capital is the sum of the resources that the virtual or actual ties of a person has in a network. An example of social capital is the theory of *structural holes* (Burt, 1995). The theory asserts that people invest in social opportunities from which they expect to profit. Structural holes are the non-connected actors in the network that create opportunities to be filled by others. When the non-connected actors form ties, better information sharing can be generated. In this work, the sum of the number of actual ties is interpreted as the social capital among scientists.

The social capital mechanism indicates that scientists possess the others' degree information, and that these scientists try to forge connections with other scientists that have higher degrees than themselves. Bounded rationality is a concern, because scientists may not know online connections of others, necessarily. Scientists can not perceive the perfect information about the condition of whole network (such as who is connected to who and if they are in different cluster). It is observed that cyber-infrastructures typically broadcast the number of connections each user possesses and that information is readily available. The more connections a scientist has online, the more likely for them to be selected by other scientists to form a link.

In addition to degree information, closeness centrality and betweenness centrality can be candidate attributes for use in social capital mechanisms. However, scientists have limited information, and data items such as the number of central actors between clusters and the proximity of a scientists to all clusters of the network cannot be processed. Another reason for selecting degree information as an attribute is the computational ease to its calculation. Both betweenness and closeness need computationally costly algorithms to be calculated; since these calculations are performed on each time tick for each scientist, the run-time of the simulations are severely altered.

Regarding *social capital* mechanism, in Figure 5.2, the evaluation process is based on the degree information of the fellow scientists. The selection process is conducted based on p_{ij}:

$$p_{ij} = \frac{DC_j}{\sum_{n=1}^{N} DC_n} \tag{5.4}$$

where N is the total number of scientists in the same artifact and DC_j is the degree centrality of scientist j. The reciprocal response from scientist j to scientist i is based on the following probability:

$$p_{ji} = \frac{DC_i}{\sum_{n=1}^{N} DC_n} \tag{5.5}$$

5.1.4 Homophily Theory

Homophily theory states that there is a stronger tendency to form social ties with others who are regarded as similar to one's self than with someone perceived as being different.[1] People are more likely to communicate with the others who have similar attributes. Shared attributes may include one' s interests, or any other variable related to human capital. Effectively, people are likely to communicate with people similar to themselves because no effort is required to build mutual understanding.[2]

The implemented homophily mechanism states that scientists have information about others' interests, and that they forge connections with those who share the same interests as themselves. In Figure 5.2 the evaluation is based on the interest levels of flow scientists.

$$P_j = \frac{\sum_{n=1}^{M} Min(I_i[n], I_j[n])}{N} \tag{5.6}$$

where M is the total number of interest areas. Below, p_{ij} is the probability of i to select j from candidate scientists.

$$p_{ij} = \frac{P_j}{\sum_{n=1}^{N} P_n} \tag{5.7}$$

where N is the total number of scientists who are active on the same artifact. With regard to reciprocity, the probability that the connection will be successful is the same:

$$p_{ji} = p_{ij} \tag{5.8}$$

[1] http://faculty.ucr.edu/ hanneman/soc157/18_Homophily.html - As of 4.07.2013

[2] http://jcmc.indiana.edu/vol11/issue4/yuan.html - As of 4.07.2013

5.1.5 Social Exchange Theory

Cognitive consistency theory focuses on the individual's perceptions of their network. People aspire to balance the attitudes in their network. *Social exchange* theory is the reverse of cognitive consistency, the condition in which unbalanced network members exist, the theory encourages a person to exchange information and resources. The theory also has a conflict with self-interest theories because self-interest theories focus only on maximizing the individual value (Monge and Contractor, 2003).

The social exchange mechanism states that a scientist have information about what others know and the level of their expertise. Scientists will attempt to connect with other experts in order to balance their own expertise level. In Figure 5.2, the evaluation process is based on the expertise gap between the fellow scientists.

$$P_j = \frac{\sum_{n=1}^{M} Max(E_j[n] - E_i[n], 0)}{N} \tag{5.9}$$

where M is the total number of interest areas. Below, p_{ij} is the probability of i to select j from candidate scientists.

$$p_{ij} = \frac{P_j}{\sum_{n=1}^{N} P_n} \tag{5.10}$$

where N is the total number of scientists in the same artifact to who scientist i is not connected. Regarding reciprocity, the probability that the connection will be successful is the same as follows:

$$p_{ji} = p_{ij} \tag{5.11}$$

5.2 Learning Process

To reiterate, information flows are represented between artifacts and scientists in the base-model. In order to measure the effects of different communication preferences on diversity within the network, it must be assumed within the communication model that scientists can learn from one another through the revision of their interest and expertise levels. The influence process introduced in the base model also describes a learning process. The following formula represents the transfer mechanism for learning of a scientist. When a scientist is connected to a new scientist, with a certain probability (0.80), their expertise level will be revised according to the formula below.

$$E_k[i] = E_k[i] + (1 - E_k[i]) \times (E_j[i] - E_k[i]) \tag{5.12}$$

where i is a randomly selected area, j is the scientist who is connected to, and k is the scientist requesting communication process. When a connection is not realized, a randomly selected scientist will learn from a randomly selected scientist in their network. Learning is only justified when the expertise of scientist j is greater than the expertise of the scientist k. The interests of a scientist are also influenced by other scientists. The higher the expertise gap, the more likely they are to be influenced. This mechanism only updates a randomly selected area i on interest of scientist k:

$$I_k[i] = I_j[i] \tag{5.13}$$

5.3 Social Network Metrics

It is previously mentioned in the literature summary that the innovation potential is likely to be captured by measuring various social network metrics. The relative social network metrics that are identified as important in this study and the summary of their definitions are listed below:

- *Network Density:* The calculation of the proportion of possible ties that exist in the network (Rowley, 1997). *Density* is $\frac{2|E|}{N(N-1)}$, where $|E|$ is the total number of edges in a network and N is the total number of nodes. High density is an indicator of mobility, which means increased connectivity and transfer of ideas within the network.

- *Diversity in the network:* The number of different people connected in the network. Diversity can be measured through different skills, expertise, resources, and reputation.

- *Diversity in the population:* The perception of differences in others within the population over time. Difference can also be in terms of skills, expertise, resources, and reputation. How the diversity metrics are measured is defined in the following sections.

- *Degree:* An actor's total number of connections.

- *Degree centrality of an actor:* According to Wasserman (1994b), people with the greatest number of ties are the most central actors in a network. It is the proportion of possible ties that exists for an actor. In this research, the *degree centrality* of scientist i is $DC_i = \frac{Degree_i}{N-1}$, where N is the total number of nodes in the network.

- Degree Centrality of a Network: The range of variability among degrees of the actors. *Degree centrality of a network* is $DC_{Network} = \frac{\sum_{i=0}^{N} DC_{max} - DC_i}{N-2}$, where N is the total number of nodes and DC_{max} is the maximum degree centrality a scientist has in the network.

- *Clustering Coefficient:* The number of edges in a neighborhood divided by the maximum possible number of edges that could exist in that neighborhood. The coefficient provides information about how actors in a network tend to cluster together. For each scientist i, a neighborhood is defined. The proportion of possible ties that exist between neighbor nodes is measured, assuming that the neighborhood is a network itself. The clustering coefficient of the whole network is the average of the clustering coefficients of individual scientists.

- *Average Path Length:* The average number of steps along the shortest paths for all possible pairs of network actors (Albert and Barabási, 2002). It can also be stated as the average of the shortest paths from every scientist i to scientist j. Lower values indicate higher cliquishness and fewer structural holes.

- *Core/Periphery Ratio:* The ratio of the number of core actors to the number of periphery actors. Boyd et al. (2006) state that "individuals in a group belong to either the core, which has a high density of ties, or to the periphery, which has a low density of ties." The *core-periphery ratio* calculation removes less central nodes recursively until only central nodes remain in the network. The remaining nodes are counted as core nodes and the removed nodes are counted as peripheral members (Borgatti and Everett, 2000).

In this study, the high centrality and fewer structural holes argument is adopted to measure innovativeness in a network (Burt, 1995). This hypothesis promotes moderate level average path length, higher degree centrality, and moderate level clustering coefficients in the network while it may decrease the density, which fosters mobility and diffusion of ideas. Since there is high turnover of the scientists within the environment, the high centrality with fewer structural holes argument is critical even though high density networks are known to be innovative. However, high density is still presented to decision-makers for comparison of different scenarios. Considering the knowledge generation, the total maturity of the artifacts, the distribution of expertise levels of scientists, activeness in the population are also discussed as collective creativity metrics.

5.4 Diversity and Interdisciplinarity

In this research, a *variation among a type* strategy is adopted (Page, 2010) in order to measure diversity in the environment. Basically, diversity can be defined as combination of three properties:

- *Variety:* It defines "how many types of things are there?" The number of different scientists, themes or interests can be perceived as variety depending on the metric. All else equal, the greater the variety, the greater the diversity.

- *Balance:* It defines "how much of each type of thing are there?" The greater equality of the balance, the better the diversity. For Balance formulation, *Shannon entropy* (Stirling, 2007) is used in which p_i indicates the proportion of members of given type in the total population:

$$-\sum_{i=1}^{N} p_i \times \ln p_i \tag{5.14}$$

- *Disparity:* It defines "how different from each other are the types of things?" Greater disparity is beneficial for diversity. The theme of artifacts or interests of scientists are used as attributes to calculate disparity in the population.

Although *diversity* is advocated as a unique quality in policy-making, the interpretation of diversity is context dependent and relative to the intention of the decision-makers. The challenge is how to accommodate threefold understanding and how to aggregate them in a metric. For that purpose, Stirling (2007) introduces an effective heuristic used in different studies (Benhamou and Peltier, 2010; Rafols and Meyer, 2010):

$$D = \sum_{ij} (d_{ij})^{\alpha} (p_i \times p_j)^{\beta} \tag{5.15}$$

where D is the diversity in the network or population, d_{ij} is the disparity between type i and j, and $i \neq j$. While p_i and p_j are defined as proportions of type i and type j, α and β are the relative weightings that are assigned to each component in the formula.

There are four diversity metrics calculated in this study. For diversity in the population of scientists and artifacts, d_{ij} is defined as dissimilarity between interest and theme arrays, respectively. While measuring link diversity in the network, every node is accepted as a kind, where p_i and p_j are the proportion of ties in the network that scientist i and scientist j have respectively, and d_{ij} is the dissimilarity between the individual networks of scientist i and scientist j. In order to calculate node diversity of the network, d_{ij} is calculated as the dissimilarity between scientist i and j based on their interest arrays as it is used in the diversity among population of scientists.

Additionally, interdisciplinary is known to be a desired output of scientific activities (Rafols and Meyer, 2010). The authors define interdisciplinarity by two aspects: diversity and coherence, or the extent to which two things are related. Coherence can be defined in terms of the network of relations; however, what types of relations are sought is the question of interest. Thagard' s *explanatory coherence* (1989) is another method in order to capture different types of coherence metrics within the environment. In this research, the diversity metric created to measure diversity among the nodes, density of the network, and the average path length are observed to discuss about interdisciplinarity that exists in the network.

5.5 Sensitivity Analysis

Before sensitivity analysis are conducted, the response surface of the model is explored. Screening experiments are run with the parameter values that are identified as important factors and expected to be effective factors on the outputs. The goal of applying *response surface methodology (RSM)* is to observe the effects of various parameters values on various performance metrics and to identify parameters that can be used in sensitivity analysis. Identified parameters and their respective levels are listed in Appendix A.

5.5.1 Response Surface Analysis

RSM includes use of statistical and mathematical techniques to develop, improve, and eventually optimize processes (Carley et al., 2004). Regarding simulation outputs, a performance metric can be called *response*. Independent variables can be model inputs, which are environmental or mechanistic parameters in the model. Practically, while applying response surface analysis, an approximation model of response space is created. Hence, in this research, first-order multiple linear regression model is developed, which is also called *main effects* model.

Subsequently after the simulation runs, parameter values and corresponding performance metrics for each scenario are imported in IBM-SPSS tool. *All possible regression method* is not used, since the number of equations to be examined increases exponentially as the number of candidate variables increases. Therefore, *backward elimination* method is adopted, which is known to be a good variable selection procedure, when the effects of all candidate variables on performance metrics are desired to be observed. *Backward elimination* basically starts with all variables in the model and then F-statistic is calculated for each variable as if it were the last variable to enter the model. If p-value is more than desired level, then that variable is removed. This procedure continues until there is no variable to remove. In Appendix A, the tables of IBM-SPSS results are presented. By examining the RSM results, the following conclusions are outlined, which can also be used to support internal validity of the model.

- Maximum *altruism* level is highly effective on all metrics.

- Minimum *cognitive burden* that is associated with the environment (perceived as the difficulty of the problem domain) arises lower activity that triggers less density in the network.

- Minimum *tension* that exists in the environment is related with elasticity of the problem domain or transparency of the community. It is an altering factor on outputs. Higher level of lower bound for the tension has similar effects as minimum cognitive burden.

- Maximum *altruism*, minimum *cognitive burden*, and minimum *tension* are expected to be effective on the performance metrics, because they directly affect the *collective action* mechanism and activeness in the population.

- Communication preferences of the scientists are observed to have important effect on the output. Since, the purpose of phase two is to measure the effects of different communication preferences on innovation potential, communication type is coupled with various variables to observe their combined effect on the simulation outcome. The results are presented in Appendix A for the design of further sensitivity analysis in the future.

- Standardized β (coefficients of variables) values are observed to avoid effect of scale. Also p-values are observed and the the p-values with less statistical significance are printed in bold characters in Appendix A.

- *Forage* extension and minimum *time expectation* are not as effective as expected on the outputs. Apparently, they just stretch or compress the timeline of the activity time series. The metrics converge to similar results at the terminating state.

- All parameter values that are associated with population dynamics (arrival rate, probability to recover, probability to leave, and expertise level to recover) are effective on the performance metrics. The reason for that is considered as the mechanism that dissolves the inactive scientists from the network.

- *Migration* threshold as an environmental parameter, which is related with the *patience* among community members, is also effective on the outputs.

Under the lights of these observations, various scenarios can be created by mutating effective parameters. In this study, simulation experiments to discern more innovative communication preferences are conducted under base-model parameter set-up as relative to the intention of this research. Sensitivity analysis are summarized in the following sections.

5.5.2 Communication Preferences

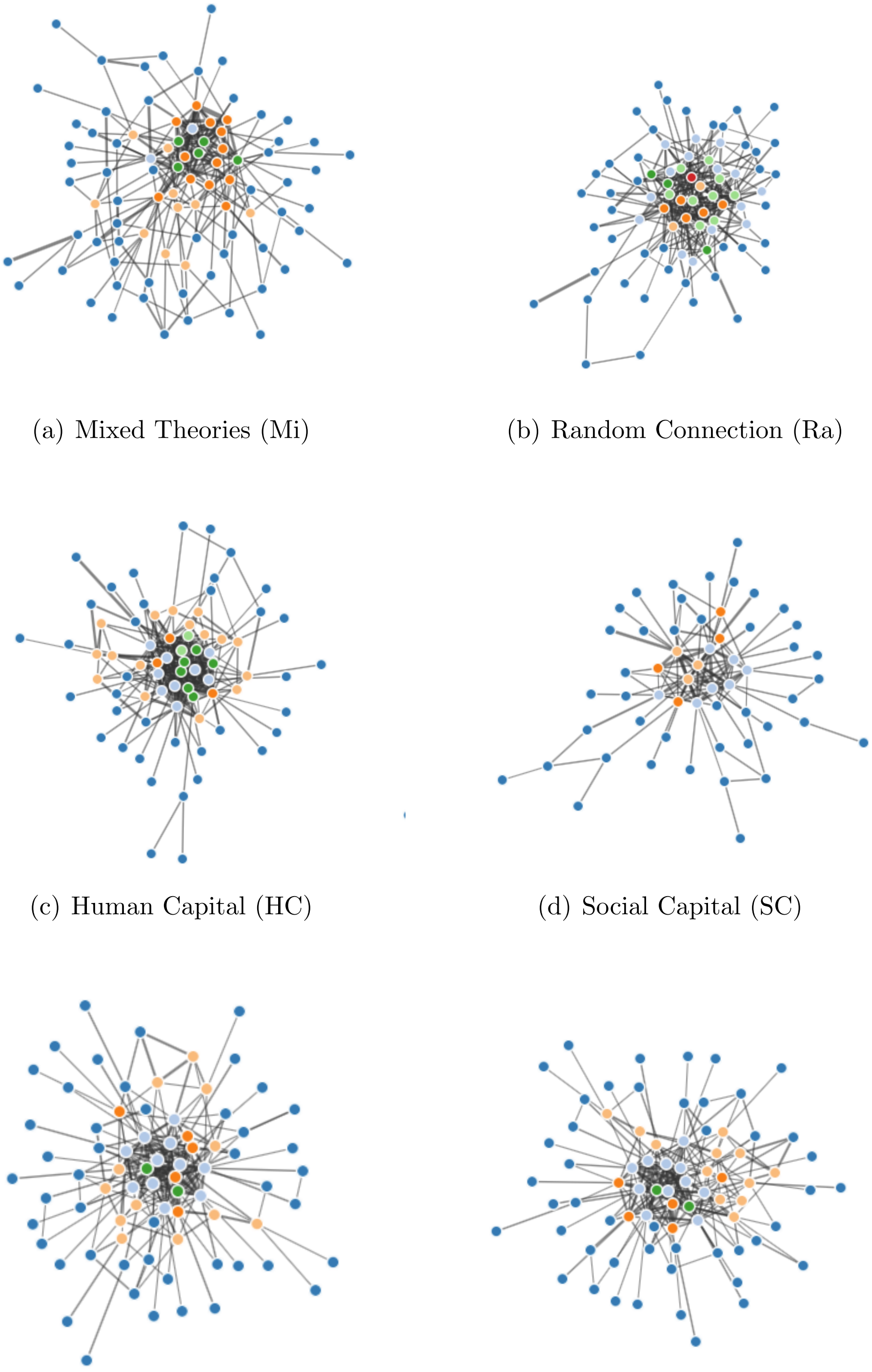
First, the simulation runs are conducted at initial conditions for each communication mechanism. Figure 5.3 illustrates emerging network topologies for different communication mechanisms at time tick 500. As a note, darker colors indicate higher levels of expertise.

Figure 5.4 represents error bars of density with 95% confidence interval for different communication mechanisms. Especially, human capital and social capital mechanisms promote connections between central and peripheral nodes, since more central or expert scientists are more likely to be selected to form ties. Therefore, this process results in more connections among different clusters in the network.

Figure 5.5 represents error bars of degree centrality with 95% confidence interval for different communication mechanisms. Random, human capital, and social capital mechanisms seem to have significantly higher degree centrality in the network, which means there are highly central scientists and the variance of degree levels among scientists is higher.

Figure 5.6 represents error bars of clustering coefficient with 95% confidence interval for different communication mechanisms. Random, human capital, and social

Figure 5.3: Network Visualizations at Terminating State



(a) Mixed Theories (Mi)

(b) Random Connection (Ra)

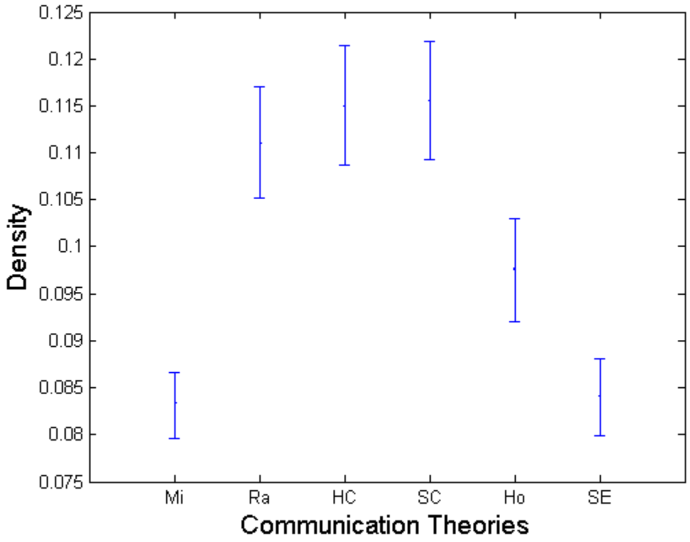
(c) Human Capital (HC)

(d) Social Capital (SC)

(e) Homophily (Ho)

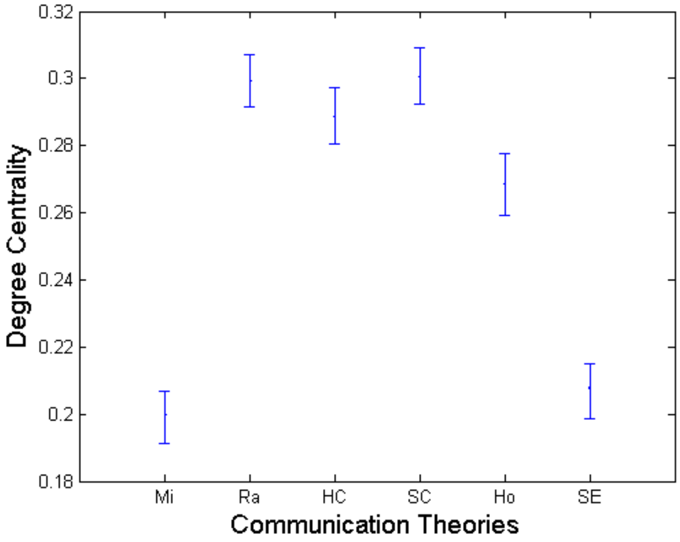
(f) Social Exchange (SE)

Figure 5.4: Density



Mi: Mixed Communication, Ra: Random Communication, HC: Human Capital, SC: Social Capital, Ho: Homophily, SE: Social Exchange

Figure 5.5: Degree Centrality



capital mechanisms result in higher clustering coefficients. Clustering coefficient is also used in small-world calculations in the following analysis.

Figure 5.6: Clustering Coefficient

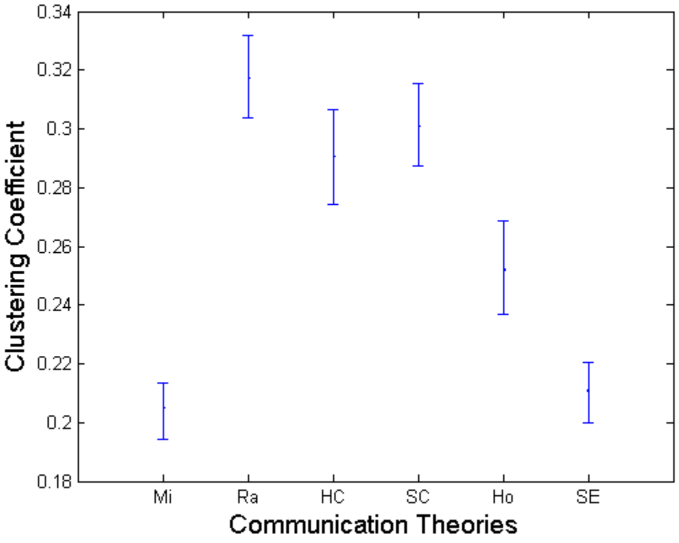


Figure 5.7 represents error bars of average path length with 95% confidence interval for different communication mechanisms. Social exchange and mixed communication mechanisms promote higher average path length and more dissolved network structures.

Figure 5.7: Average Path Length

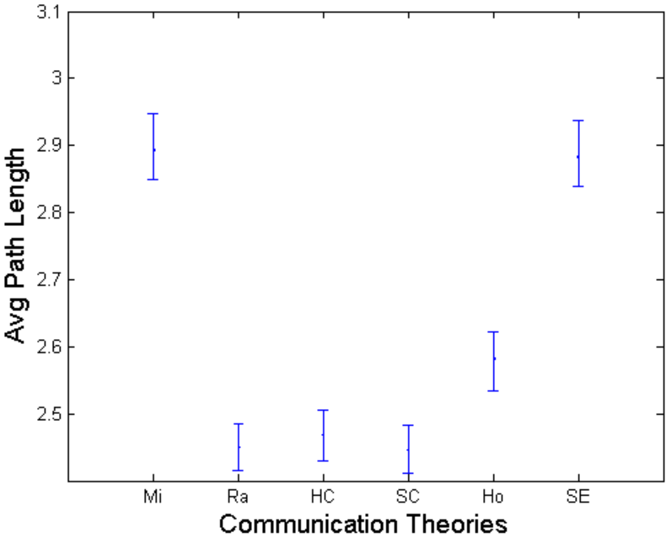


Figure 5.8 represents error bars of core/periphery ratio with 95% confidence interval for different communication mechanisms. Mixed communication and social exchange mechanisms result in better core/periphery structures generating greater number of periphery members to diffuse innovation.

Figure 5.8: Core/Periphery Ratio

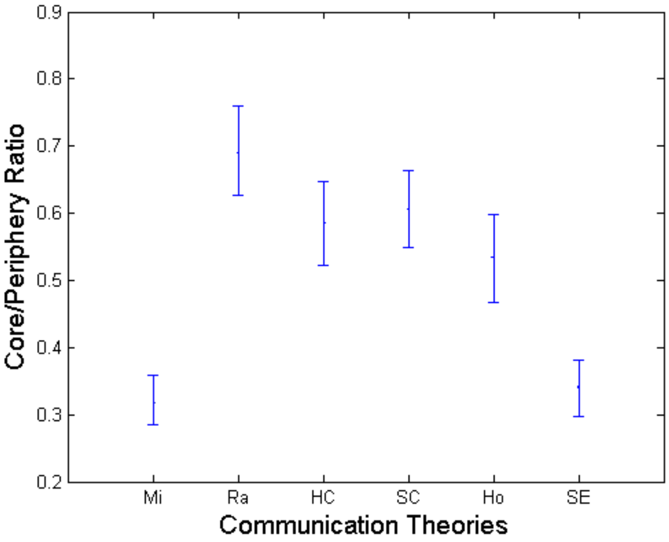


Figure 5.9 and Figure 5.10 represent error bars of diversity among scientists and diversity among artifacts with 95% confidence interval, respectively. In both graphs, it can not be argued that different mechanisms cause significantly different levels of diversity. The underlying reasons creating this phenomenon are the high variety and convergence of disparity among the population to 0.50, which is thought to be caused by *binary* interest and theme arrays.

Figure 5.11 represents error bars of diversity among links with 95% confidence interval. Social capital mechanism results in more diverse social networks based on links.

Figure 5.12 represents error bars of diversity among nodes with 95% confidence interval. Random, human capital, and social capital mechanisms cause more diverse

Figure 5.9: Diversity Among Scientists

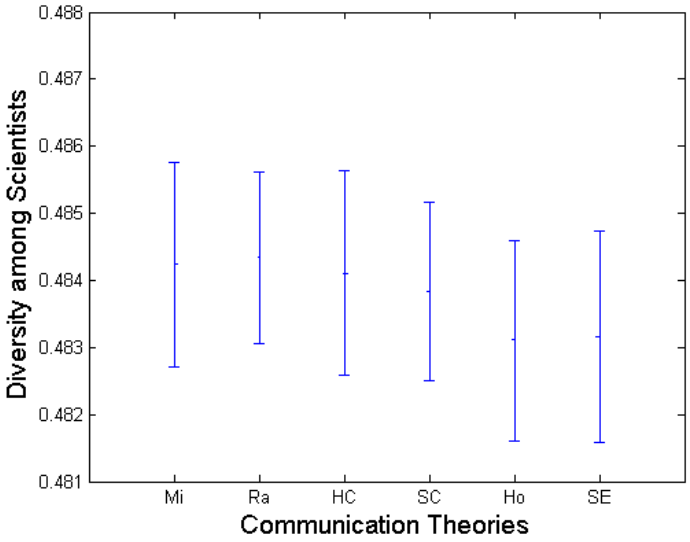
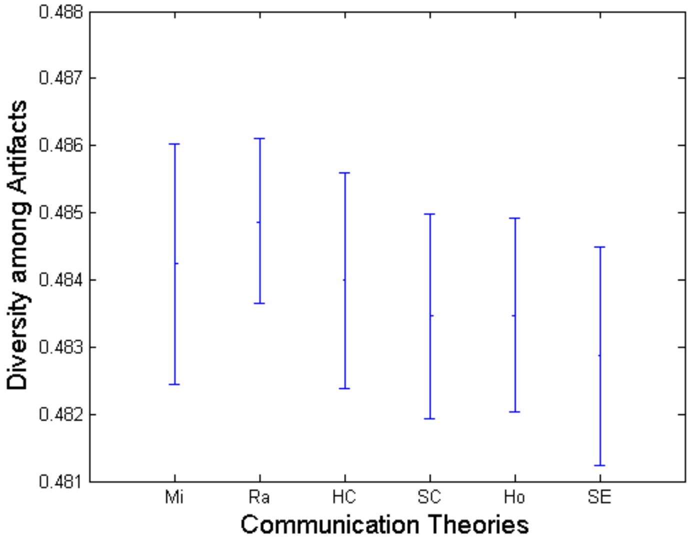


Figure 5.10: Diversity Among Artifacts



social networks regarding how much dissimilar scientists are connected based on their interests.

Figure 5.13 represents expertise distribution among scientist population. In the figure, there is a peak at bin that represents expertise levels between 0.80 and 0.90. It is caused by *expertiseToRecover* value. After expertise level of 0.80, if scientists

Figure 5.11: Diversity Among Links

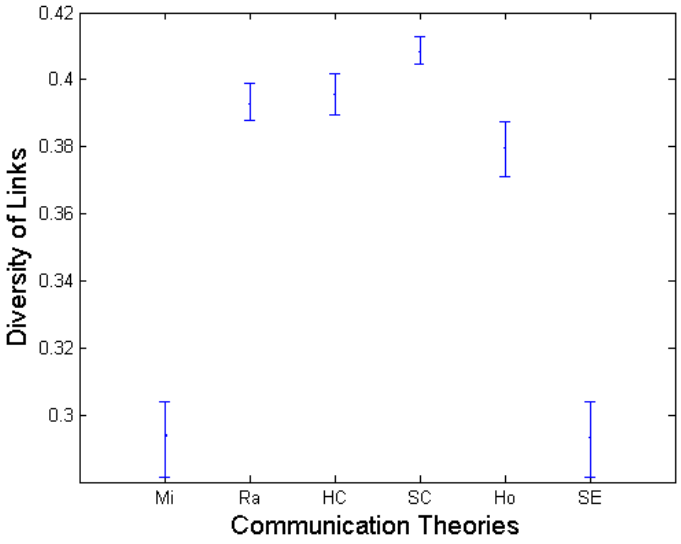
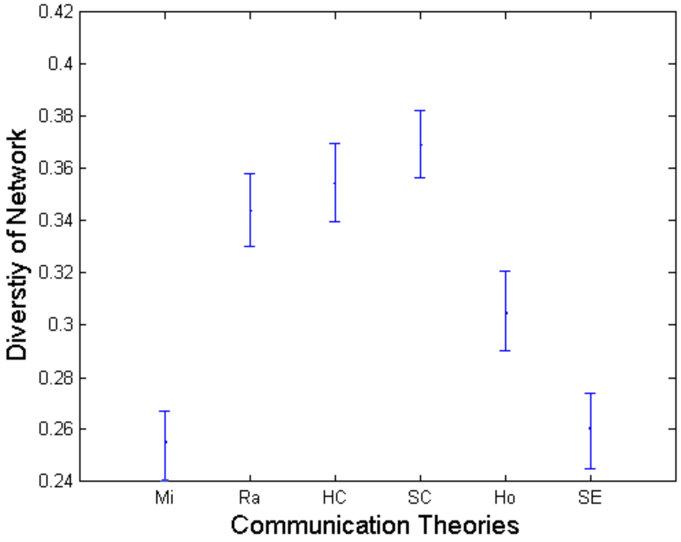
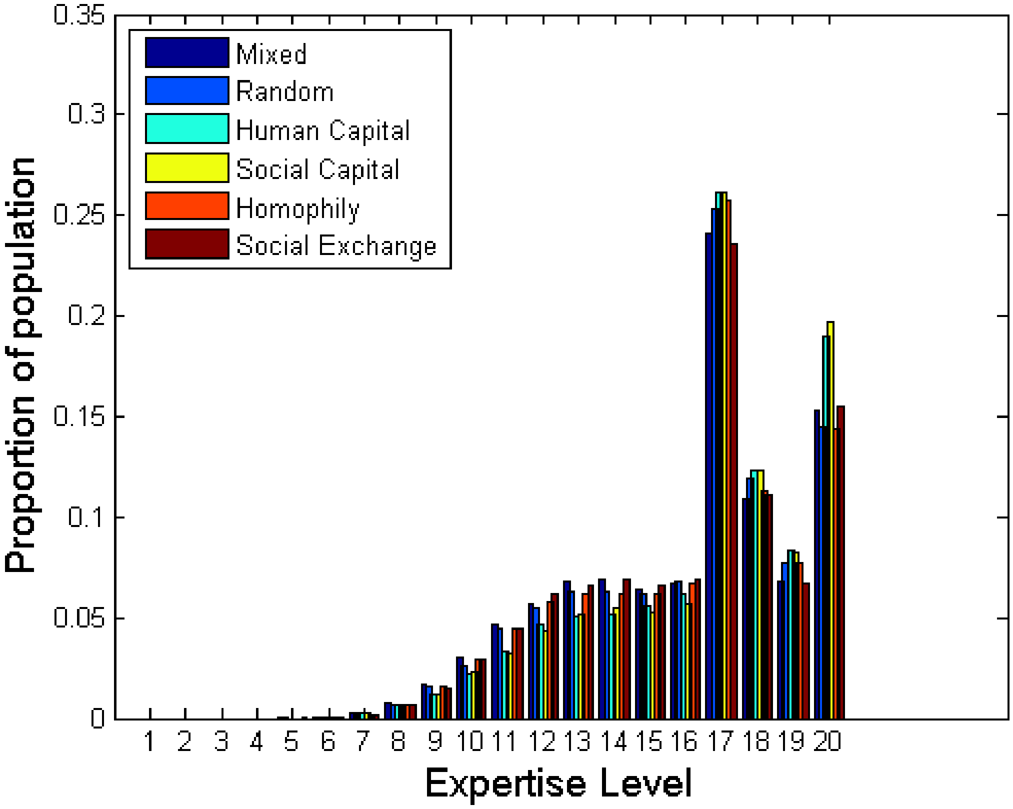


Figure 5.12: Diversity Among Nodes



can not become active, they recover and leave the environment. So, there are more number of scientists who fall in the bin between 0.80 and 0.85. The last peak in the graph represents the scientists who are highly active and central in the network. Even though it is hard to distinguish, the levels between the bars of different colors indicate that highly expert scientists are more likely to occur in human and social capital mechanisms.

Figure 5.13: Expertise Distribution

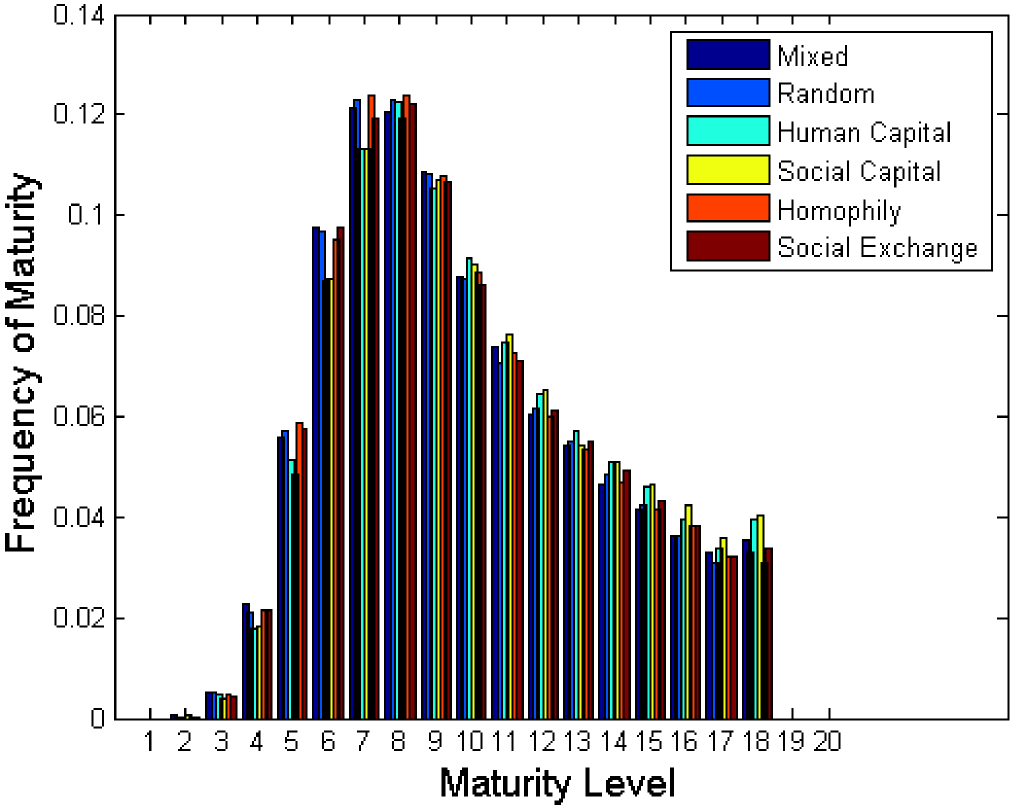


X-axis is the bin number. Each bin has width of 0.05. Y-axis is the proportion of the population whose expertise levels fall on respective bin

Figure 5.14 represents maturity distribution among the artifact population. *Human capital* and *social capital* mechanisms create artifacts that have higher levels of maturity, which is a result of information flows between scientists and artifacts.

In order to discuss through how disparity emerges for different communication theories, Figure 5.15 and Figure 5.16 represent disparity distributions of links and nodes, respectively. As a result, homophily mechanism generates cliques, therefore it creates more disparate individual networks. Disparity among the nodes is based on how dissimilar are the nodes that are connected. Interestingly, it indicates a binomial

Figure 5.14: Maturity Distribution

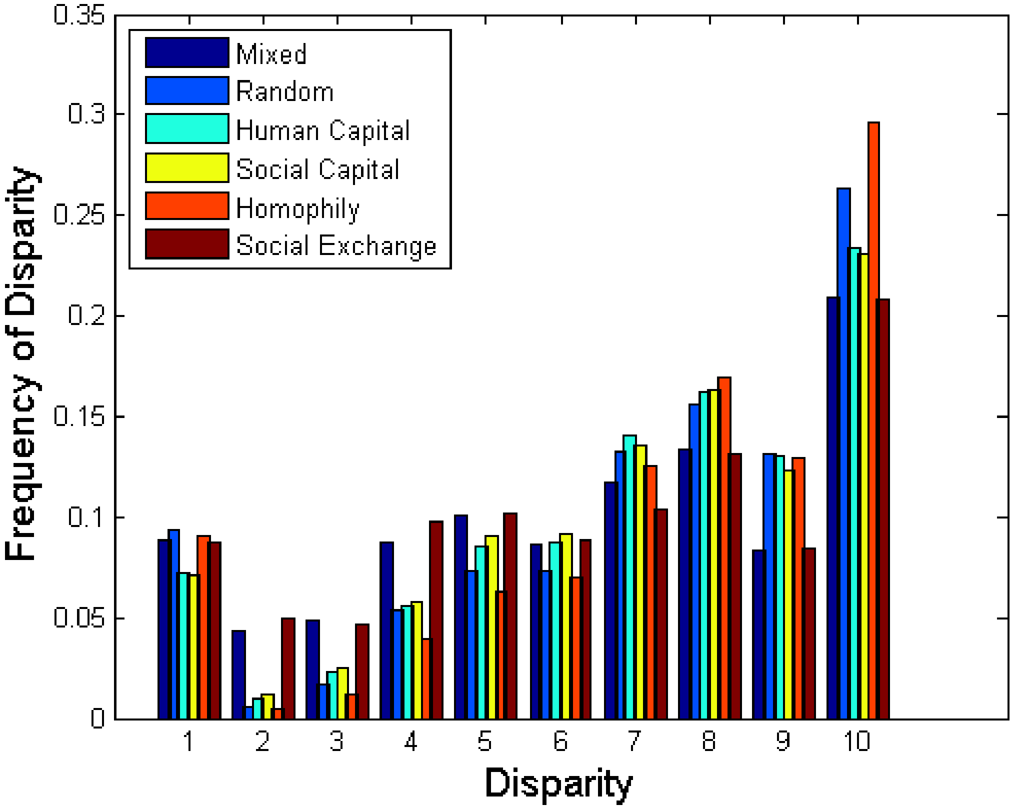


distribution. Since each bit on binary arrays can be perceived as bernoulli trials, the number of matches or dissimilar bits between two binary arrays produces binomial distribution in long run, which is quite similar to normal distribution.

Figure 5.17 and Figure 5.18 represent the average number of active scientists and artifacts. It is shown that *mixed communications* and *social capital* mechanisms result in relatively higher level of activity than *random connections* and *homophily* mechanisms. It is not possible to distinguish *human capital* and *social exchange* mechanisms from the others.

Figure 5.19 illustrates small-world phenomenon information. The calculation method, which basically divides clustering coefficient by the average path length is

Figure 5.15: Disparity Distribution Among Nodes



X-axis is the bin number. Each bin has width of 0.1. Y-axis is the proportion of the pairs of nodes which has disparity levels fall on respective bin

adopted (Uzzi and Spiro, 2005). Uzzi and Spiro (2005) state that small-world phenomenon is also indicative of creativity, which spurs innovation until certain extent. Random, human capital, and social capital have higher level clustering and smaller average path length that means the network is dense and closely knitted allowing diffusion of ideas and collective productivity.

Table 5.1 summarizes the results for different communication preferences. For better illustration, relative values of each metric are indicated in three levels: low, medium, and high. Bold characters are used to discern the values that promote innovation potential.

Figure 5.16: Disparity Distribution Among Links

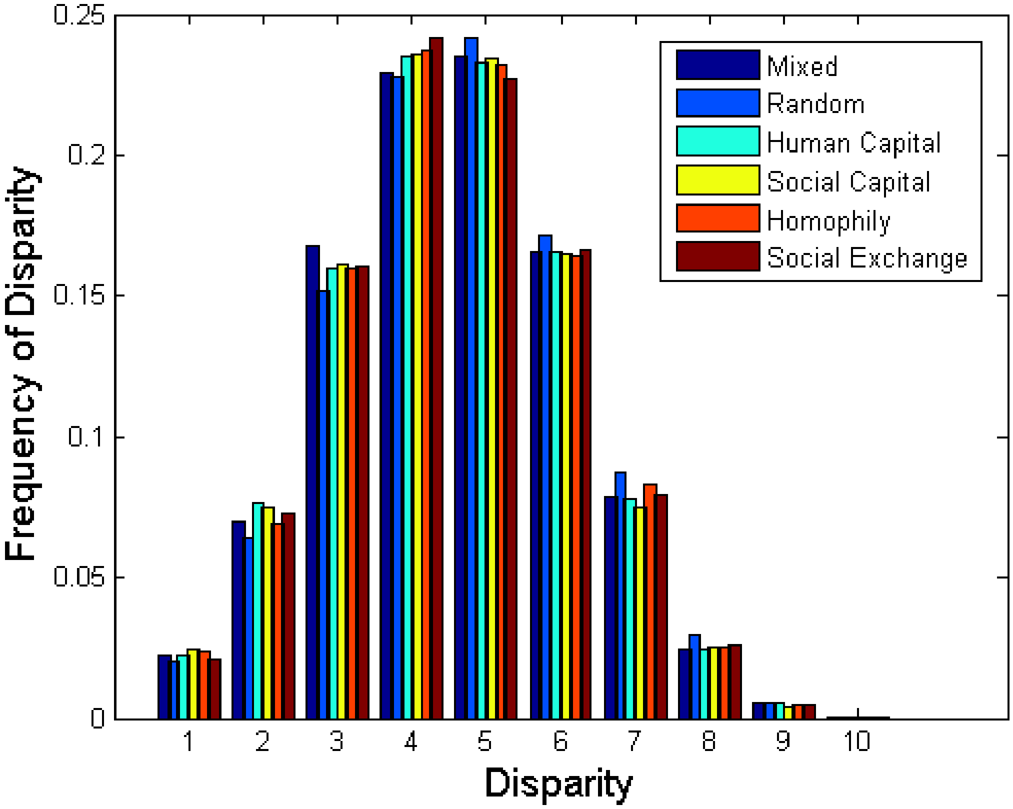


Figure 5.17: Active Scientists

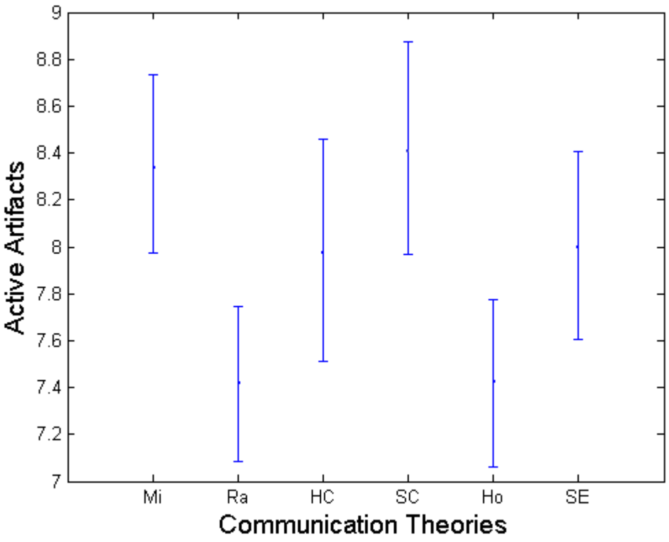


Figure 5.18: Active Artifacts

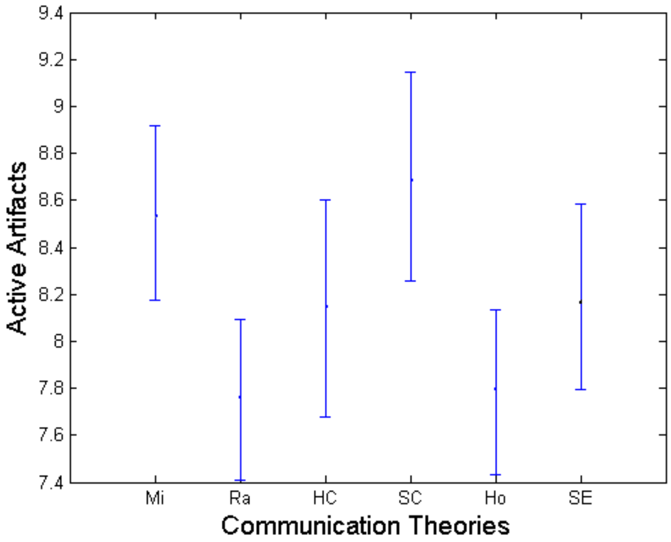
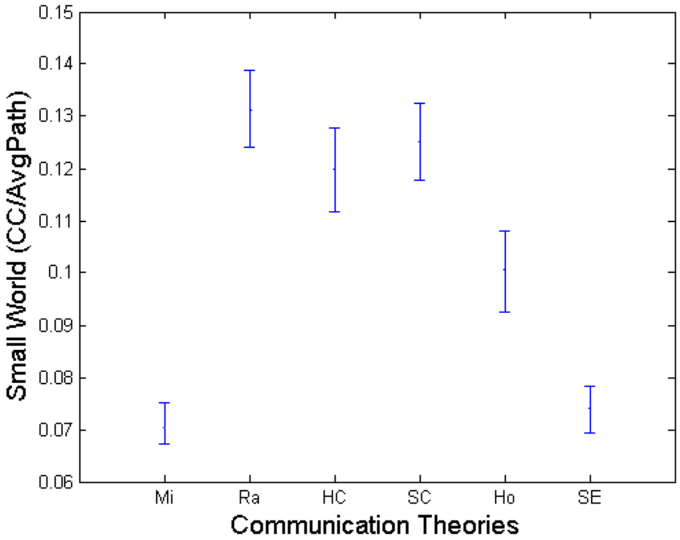


Figure 5.19: Small-world Phenomenon



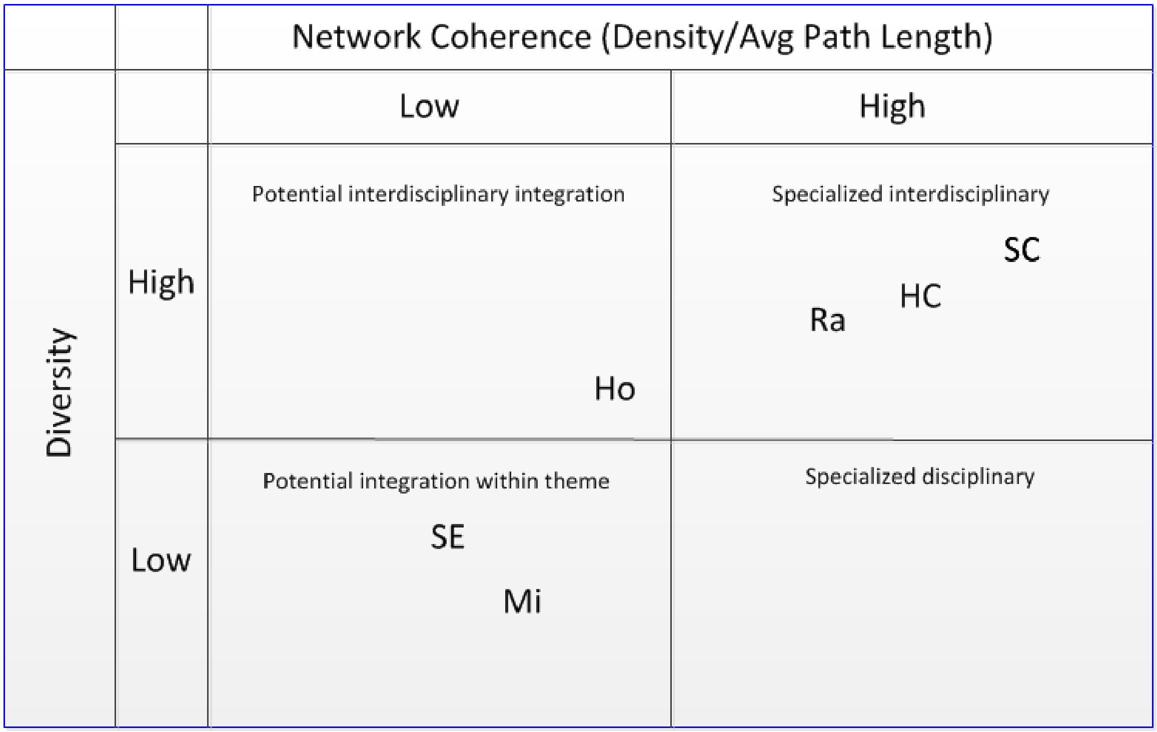
Interdisciplinarity is illustrated by Figure 5.20 below. Interdisciplinarity is also a desirable feature that fosters innovation. The heuristic is adopted from the study by Rafols and Meyer (2010).

In conclusion, *social capital* theory supports all indicators of innovativeness more than the other candidate theories. *Social capital* theory promotes connections among

Table 5.1: Summary of the Sensitivity Analysis

Criteria	Mi	Ra	HC	SC	Ho	SE
Density	Low	**High**	**High**	**High**	Med	Low
Degree Centrality	Low	**High**	**High**	**High**	Med	Low
Clustering Coefficient	Low	**High**	**High**	**High**	Med	Low
Avg Path	High	**Low**	**Low**	**Low**	Med	High
Core/Periphery	**Low**	High	Med	Med	Med	**Low**
DiversityL	Low	Med	Med	**High**	Med	Low
DiversityN	Low	**High**	**High**	**High**	Med	Low
Activity	**High**	Low	Med	**High**	Low	Med
Expertise	Med	Med	**High**	**High**	Med	Med
Maturity	Med	Med	**High**	**High**	Med	Med
Small World	Low	**High**	**High**	**High**	Med	Low

Figure 5.20: Node Diversity vs Network Coherence


Network Coherence (Density/Avg Path Length)
Low
High
Diversity
High
Low
Potential interdisciplinary integration
Ho
Specialized interdisciplinary
SC
HC
Ra
Potential integration within theme
SE
Mi
Specialized disciplinary

central scientists and periphery scientists. Apparently, highly central members are also more likely to be more active and accumulate more expertise. More expert scientists in the population increase the complexity of the artifacts and lead to more mature artifacts in the environment, which means more knowledge creation. Highly central scientists also broadcast the knowledge to periphery members and spur the diversity among the network members. *High centrality-fewer structural holes* hypothesis is supported more by *social capital* theory.

Additionally, core/periphery structures are desirable to promote diffusion of innovative ideas. Mixed connections and social exchange favor the balance of popularity among the scientists, which also results in lower degree centrality. Therefore, they have small number of core members while the number of periphery members are high and they usually have moderate level of expertise.

The implemented simulation model can be used to conduct more sensitivity analysis under further environmental conditions and the different communication preferences can be tested along with different parameter set-ups (more altruistic, more difficult, more elasticity etc.). In the following chapter, a search algorithm (genetic algorithm) is introduced and more robust communication landscapes are explored in the scenario space. The goal is to capture more robust parameter set-ups with an evolutionary algorithm that enables intelligent search and avoids the exhaustive parameter sweep.

Chapter 6

ROBUSTNESS IN GLOBAL PARTICIPATORY SCIENCE

Developing innovation coordination mechanisms that are robust and resilient under environmental uncertainty is critical for sustained innovation. For that reason, *robustness* is valuable to explore as opposed to only searching for an optimal behavior in terms of a fitness function in an unchanging environment. The motivation to explore the scenario space is to gain deeper insight and generalize from the observed behavior. Exploration vs. exploitation is a well-known trade-off (March, 1991). An important issue is to decide when to stop exploring and when to start exploiting the parameter space to discover robust system configurations. The aim in this chapter is to exploit the possible scenario space to discover robust landscapes in terms of an evolutionary search algorithm. To measure robustness, the fitness function is defined in terms of the degree of variability among innovation potential metrics.

6.1 Exploratory Modeling

There is insufficient knowledge and a high level of uncertainty for the modeled target system (Global Participatory Science). The estimates on initial and boundary conditions or the nonlinearities in the models can cause even small levels of initial uncertainties to generate remarkable levels of uncertainties in the results. The critical question to be addressed is: “What is the appropriate method for using the model considering its limitations?” In this chapter, an *exploratory modeling* approach is adopted, which is defined as a series of computational experiments to explore the implications of mechanisms (e.g., communication mechanisms) and parameter changes (Bankes, 1993).

An exploratory modeling can involve a search for key configurations of the system (Bankes, 1992). Initially, boundaries of the plausible scenario space (possible parameter ranges) and an ensemble of plausible scenarios are generated. The process of selecting which ensemble of the plausible scenarios to run depends on the question of interest. Through the search, an output metric is observed. In this research, the output metric is defined as an indicator of robustness (related to the purpose of the study). Moreover, a search strategy is needed in exploratory modeling. In this work, a genetic algorithm (GA) is implemented to evolve the regions of parameter values where more robust (less variable) social network structures are observed.

The ability to discern patterns from output metrics depends on being able to define a topology (set of model configurations) in the ensemble such that similar parameter ranges have similar outcomes (Bankes, 1993). In this research, the search is guided by a heuristic (specifically a GA) and cannot guarantee an absolute optimal scenario. Thus, the evolution of the ensemble of scenarios at different generations are recorded.

Furthermore, the ensemble is interactively tested against different metamorphic relations to bound the plausible scenario space (by interventions). This search can be called human-mediated interactive robustness analysis. The ensemble of scenarios is revised over time resulting in an evolving scenario space. In the following section, the use of genetic algorithm in exploring plausible scenario space is discussed.

6.1.1 The Use of Genetic Algorithms

In genetic algorithms, through randomness associated with selection and crossover, alternatives with desired outputs mate to generate new ensemble members. A scenario with better output metric (with respect to a fitness function) has a higher probability to be selected. Over iterated generations, an increasingly desired behavior is expected to accumulate (Atmar, 1994).

The selection decision of GA is motivated by simulation optimization studies. Simulation optimization can be defined as the process of finding the most effective and optimal input values among all possibilities without explicitly evaluating each possibility (Carson, 1997). However, there exist other techniques that are used in simulation optimization: "Gradient-based search, response surface method, stochastic approximation, and Ranking and Selection etc." (Carson, 1997; April et al., 2003). Most available tools use evolutionary algorithms to optimize the inputs of a given model (Fu and Glover, 2005). Artificial neural networks (ANN) simulation is also a well-published method used for training a model with a given set of data (Sexton et al., 1999). Fu and Glover (2005) state that there is no clear answer for why the use of evolutionary algorithms are dominant, but they point out the benefits, such as the ability to explore the entire state space and robust properties in practice.

Genetic algorithms are used in various simulation studies as parameter optimization tools. For example, in GENOSIM, authors manipulate the values of control parameters in a traffic micro-simulation and globally search for an optimal set of values that minimize the gap between real field data and the simulation output data (Ma and Abdulhai, 2002). Similarly, Zou (2012) explores the parameter space with a genetic algorithm that minimizes the discrepancy between the real world community data (also gathered from OBO) and various social network metrics. In a dynamical system like a social network, Zou (2012) builds the fitness function based on the point estimators at the termination state of the simulation. It adds credibility to the models by stating that the model is capable of creating snapshots of some real world networks at certain states, but it can not assert that the simulation arrive to observed states following the same transient events.

Bäck and Schwefel (1993) compare different evolutionary algorithms and promote the use of GA by stressing its ability to assign a nonzero selection probability to each individual (called as preservative selection or proportional selection). Additionally, Paul and Chanev (1998) address the appropriateness of GA for simulation optimization. Unlike this research, Paul and Chanev (1998) simulate an existing steady-state Steelworks model. For each scenario, they run only a single replication for a long period of time, which they believe is long enough to gather sufficient statistics at the steady-state of the system.

Nazzal et al. (2011) state that when the goal is to optimize a stochastic system with high variability, the performance of GA can be inadequate. The authors address that it is important to consider variance when evaluating the alternative scenarios produced by GA over generations. Hence, Nazzal et al. (2011) propose a methodology that incorporates an indifference-zone (IZ) ranking and selection procedure under common random numbers (CRN). The methodology also aims to reduce the required number of replications for each scenario. In contrast, Pierreval and Tautou (1997) note that when a simulation model is considered, as a common practice in GA selection operation, scenarios are compared based on the differences between mean values of an output metric (or fitness value). If the proportional selection operation is used, the comparison tests (e.g., sensitivity analysis) between scenarios are less important and the scenarios can evolve based on the mean values of the output metric (Pierreval and Tautou, 1997).

In general, genetic algorithms are used to determine the discrete input values (Liepins and Hilliard, 1989). Genetic algorithms mimic the evolutionary process of biological systems to create new generations guiding the search towards optimal solution (Swisher et al., 2000). Genetic algorithms cannot guarantee optimality, but the intention is to improve the ensemble, and if possible derive robust scenarios under environmental uncertainty. As discussed in Chapter 4, for the sake of generalization, the aim of this research is not to find the optimal parameter set that mimics the real world data, but rather to explore the scenario space to measure the robustness of different communication theories under different conditions and if possible, to discover diverse scenarios that are more robust than the others. In order to limit the search space or understand the response surface of the model, metamorphic testing is introduced as a candidate process. In the following section, metamorphic testing is delineated and characterized.

6.1.2 Metamorphic Testing

Metamorphic testing approach is introduced by Chen et al. (1998) to address the problem of testing programs with no oracle. Metamorphic testing is a testing method that is based on the expected properties of the application or model. The properties, called metamorphic relations (MRs), are basically functions that define the relationships between program, model, or function input and the expected changes in output. Specifically, those relationships can provide means to define V&V test cases.

Suppose a case in which x is a test input and produces output $f(x)$. The metamorphic properties of the function f can be used to develop a transformation function, which when applied to the test input produces x'. This then enables prediction of the expected output $f(x')$ based on the known $f(x)$. If the outcome $f(x')$ is consistent with the expectation, it is not necessarily correct. However, violation of the metamorphic property indicates that one (or both) of the outputs, $f(x)$ or $f(x')$, is wrong. So, though it may not be possible to know with a single test whether an output is correct, it is determined if the output is incorrect. Metamorphic testing has been used at the function (Guderlei and Mayer, 2007), application (Xie et al., 2011), and simulation (Ding et al., 2011; Pullum and Ozmen, 2012) levels.

As an example to explain metamorphic relations, consider a function that calculates the standard deviation of a set of numbers. For some transformations of the input set, we expect no change in the result, e.g., if the order of the members of the input set is permuted or if each member of the input set is multiplied by -1. Other transformations of the input set will predictably alter the output, e.g., multiplying each member of the input set by 2 will result in a standard deviation twice that of the original input set.

To date, there has been only a single published effort to apply metamorphic testing to agent based models (Murphy et al., 2011). Murphy et al. (2011) investigate the use of metamorphic testing on a simulation tool of a hospital in healthcare domain. ABMs are often used as exploratory tool for discovery of unknown regularities. Prior studies on metamorphic relations test if the implementation is correct, while in this work, metamorphic relations are intended to be used to identify the boundaries of the model response surface. Instead of classifying the model as wrong when the model behaves differently than the expectation, the corresponding response area can be flagged to be eliminated from the analysis. Metamorphic relations can be updated iteratively to make sure that the analysis is conducted on the scenario space of interest. In this research, it is only used to test expected behavior and derive new expected behaviors for future use.

6.2 Design Decisions Relative to the GA

In this section, the components of the genetic algorithm module and the implementation of the algorithm are explained. It is worth mentioning that the main intention is not to create a novel meta-heuristic, but to synthesize the existing methods for better search within a computationally feasible time frame. In genetic algorithms, variation is achieved via various operators (e.g., combination, mutation, crossover) and the selective pressure is based on the fitness function. The relationship between context of this study and the biological systems are listed below:

- A scientific community of scientists and its traits (phenotype) can be interpreted as a member of the population.

- The population consists of different scenarios/communities.
- In the algorithm, different communication theories are metaphorically perceived as different species of the population.
- Each community has a scalar measure that identifies its fitness, which depends on the purpose of the study. In this research, fitness is defined in terms of the aggregation of variabilities that are observed in different metrics.
- Gene is the genotype. It is a vector that retains the parameter values for each community.

Regarding the simulation optimization studies, a widely observed approach in the literature is to run each scenario for a sufficient number of replications (using CRN) and to compare mean values of the outputs while conducting proportional selection. When the desired termination condition is met, the GA provides a single best scenario (Tompkins and Azadivar, 1995; Bäck and Schwefel, 1993; Faccenda and Tenga, 1992). However, GAs are random algorithms and if they are asked to solve exactly the same problem twice, they are likely to come up with two different scenarios (if exact optimum is not found).[1] If the goal is to generate a set of scenarios (diverse portfolio of solutions), one method is to re-initiate the GA multiple times. Likewise, Zeigler et al. (1997) provide a high performance environment for modeling large-scale systems at high resolution to enable parallel GA runs. Parallel GA modules are initiated to identify possible parameter configurations using CRN for each GA module. Corresponding to the goal of exploratory modeling, these parameter configurations can serve as a basis for drawing general conclusions about the system of interest. Zeigler et al. (1997) conclude that parameter estimations of simulation-based studies for large-scale models must await new generation computers. Even though it has

[1] http://www.burns-stat.com/documents/tutorials/an-introduction-to-genetic-algorithms/ - As of 4.07.2013

been more than a decade since this paper was published, today' s desktop computers have only reached to Gflops of computational power, on average. Along with the advancements in hardware, simulation studies have become more popular and high resolution-more complex simulation models are being developed that still need High Performance Computing (HPC).

Similarly in the analysis of exploratory modeling, the challenge is the problem of deciding on the limited number of experiments that can be run practically (consuming reasonable amount of computational resources) to best inform the question of interest. The sampling strategy (the number of replications) involves human judgment. Bankes (1993) states that:

> Consequently, the result of an exploratory analysis will typically not be a mathematically rigorous answer, but rather an imperfect image of the complete ensemble that improves gradually as more cases are run. Given a fixed analytic budget (in dollars, people, or time), the analysis must provide the most useful results possible based on what is known about problem on hand.

Given the complexity of the developed socio-communication model in this research, an exhaustive parameter sweep across all plausible scenarios is not possible. The sampling strategy (30 replications) is dedicated to produce a reasonable amount of help in converging to robust configurations from a limited number of computational experiments.

In the search, diversity is aspired to be maintained in the population. Diverse configurations of parameters are explored and it is desired to produce likelihood (in proportional selection) to potentially robust scenarios. In order to add diversity to the ensemble and avoid premature convergence, randomness is perceived as a non-parametric environmental condition that can not be controlled. At each generation, the set of random numbers is changed. This approach is motivated by two studies: (1) the search method suggested in (Dibble, 2006) and (2) the explicit separation of environmental conditions and model parameters (Mitchell and Yilmaz, 2009). Dibble (2006) explores the parameter space running a small number of replications (2-3) per scenario, then suggests a search across the sets of random number seeds to test worst case combinations of stochastic events. Mitchell and Yilmaz (2009) also run each scenario for a small number of replications and observe the adaptation of the converged scenario space against non-parameterized environmental conditions modeled by a separate simulator that emulates the environment. Random or purposeful changes in the environment is fed back into the GA algorithm through explicitly separated environmental parameters. This approach assures that the competing GA solutions are in synch with the evolving environmental conditions under which they are competing. However, considering the complexity of the model, the design of the robustness metric, and the implementation of the GA, this approach is likely to generate variable results rather than convergence to a single best scenario. Therefore, narrowing the focus on determining the best scenarios is postponed to future analysis. A similar technique is delineated by Dibble (2006):

> Using supervisory genetic algorithms to discover highly effective treatments or to search for exceptional or surprising simulation outcomes has the potential to profoundly enhance our ability to make the most effective use of limited computational and analytical resources. It permits us to discover and test incisive empirical insights, effective normative designs or interventions, and surprising heuristic insights. Once such treatments or outcomes have been identified by the genetic algorithm, subsequent ordinary batches of simulations can be carefully targeted in order to evaluate the accuracy, uncertainty, risk, and inference power of results obtained from any well-specified agent-based simulation model.

When the decision maker (analyst) decides that the GA exploration is complete, the task is to select a portfolio of scenarios that can be a basis on policy making decisions. The key parameter configurations can be identified by observing the fittest scenarios over generations. In this research, the portfolio is selected from the fittest scenarios to which the GA converges. The decision is qualitative and aims to identify different level sets for parameters. In like manner, Bankes (2002) states that a portfolio of models or level sets for parameter configurations that behave in a reasonable range provide more information than does a single optimal configuration.

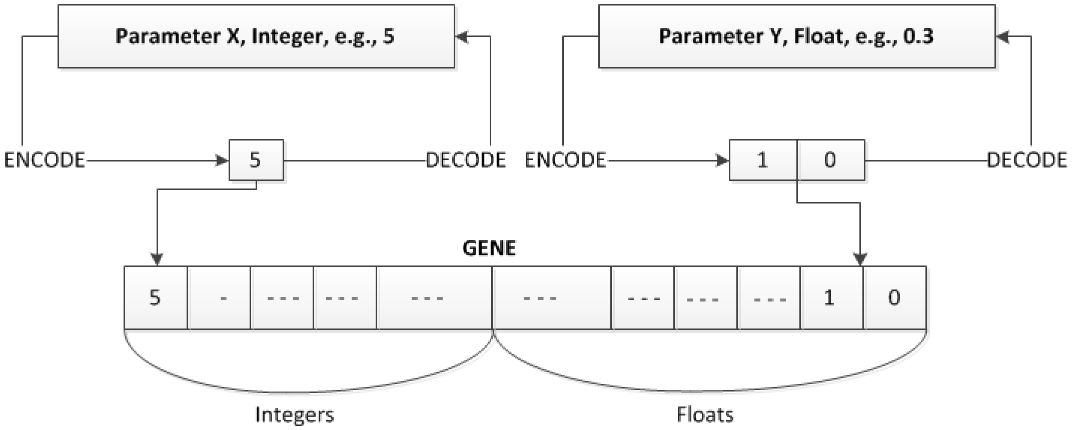
Dibble (2006) indicates that the greater economy in searching for key parameter values can release computational resources that can be dedicated to simulate each candidate key configuration for a sufficient number of replications to test if there are statistically significant differences among them. Also, existing techniques of decision support may need static recommendations such as providing a single scenario as an answer (Bankes et al., 2002). Therefore, in this research for illustration purposes, further batches of runs are conducted to evaluate the identified portfolio (of scenarios) in terms of uncertainty in the robustness results. In the following sections, components of the GA are described in detail.

6.2.1 Encoding and Decoding of the Parameter Space

Encoding refers to the mechanism of mapping the parameters to genes, so that the evolution of the gene in the parameter space can be realized. Genes can be represented as binary strings or real numbers. Allowing continuous values in bits result in an exhaustive search, so it is essential to decide on what values the bits can take and how many bits are needed to represent the plausible scenario space. In this research, if the parameter value is a floating number, the precision is set up as increments to have fixed numbers of values that are feasible. Then the combination of binary bits are used to represent those parameter values.

Decoding is the reverse process of encoding. Decoding partitions the gene into its parts so that the corresponding parameter values can be used in simulation runs. Therefore, the fitness function value can be calculated for every set of replications for different genes/scenarios. Figure 6.1 illustrates two kinds of parameter values and how they are represented in a gene.

Figure 6.1: Encoding and Decoding of the Parameters.



The initial search is conducted on the scenario space that is bounded by the experience of the modeler of this research. The gene consists of 22 bits that describe the parameter set-up. The parameter values are identified in two classes: (1) integers and (2) floats. The Table 6.1 lists the bits and value ranges to interpret in the decoding process.

Table 6.1: The Bit Values in Initial Genome

Bits	Parameter	Code	Values/Range
1	Communication Preference	Integer	[0,5]
2	Maximum Time Expectation	Integer	[2,10]
3	Foraging Mechanism	Integer	[1,2]
4	Migration Threshold	Integer	[2,5]
5	Maximum Scope	Integer	[1,10]
6-7	Maximum Altruism	{00,01,10,11}	{0.1, 0.3, 0.5, 0.9}
8-10	Minimum Tension	{000,001,...,111}	{0.1,0.2,...,0.8}
11-13	Minimum Burden	{000,001,...,111}	{0.1,0.2,...,0.8}
14	Mutation Rate	{0,1}	{0.01,0.05}
15-16	Probability to Recover	{00,01,10,11}	{0.1,0.2,0.3,0.4}
17-18	Probability to Leave	{00,01,10,11}	{0.1,0.2,0.3,0.4}
19-20	Artifact Creation Probability	{00,01,10,11}	{0.05,0.1,0.2,0.3}
21-22	Arrival Rate	{00,01,10,11}	{0.05,0.1,0.2,0.3}

6.2.2 Activity Flow of the GA Module

Figure 6.2 represents the activity-flow specification of the genetic algorithm module. The population is randomly initialized. For the bits that have an integer value, the value is drawn uniformly from the list of possible values. If the bits are represented as binary numbers, the values are uniformly assigned to each bit. As a note, all possible combinations of the bits represent a number, so all permutations are valid scenarios.

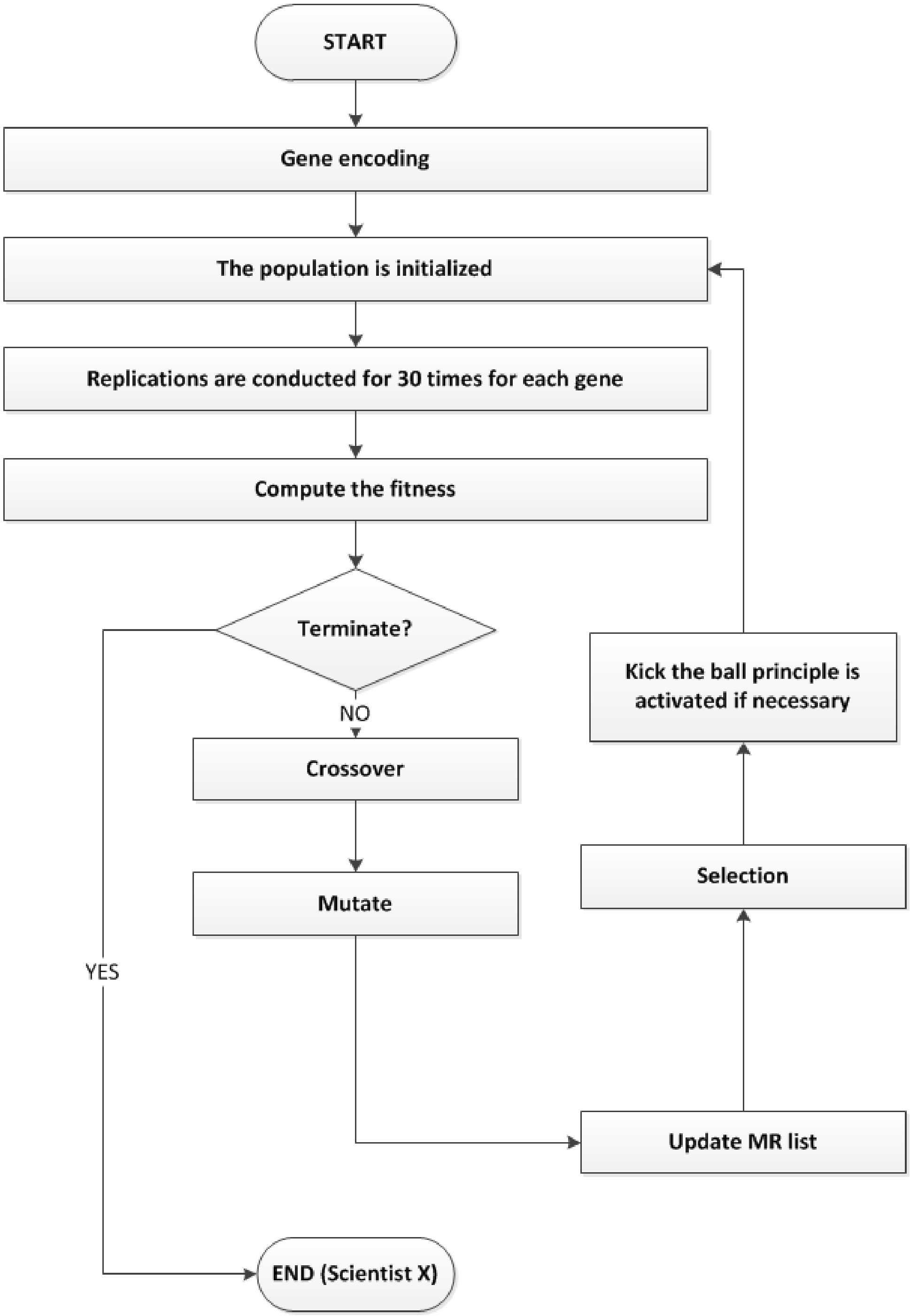
6.2.3 Metamorphic Relations

Based on the analysis that are conducted in previous chapters and the observations on the behavior of the model, initial metamorphic relations are identified. By identifying the metamorphic relations, it is aimed to understand the response surface of the model. Since the model is a dynamical system, bounding the search space in a way consistent with the goal of the research is critical. Even though individual values for each parameter are valid, the combined effect of different parameters can steer the search toward undesired regions. Especially, the parameters that are incorporated in the *collective action* formula can have this impact. Below, some initial metamorphic relations are listed to test the effectiveness of the method.

- Initially *cognitive burden*, *tension*, and *altruism* are uniformly distributed between 0 and minimum or maximum values. Tension starts at 1 and gradually decreases with new contributions. Considering the *collective action* formula, in order to have an active initial population, an MR about the expected values of selected parameters is identified as follows:

$$\frac{minCognitiveBurden}{2} < \frac{maxAltruism}{2} + 0.25 \tag{6.1}$$

Figure 6.2: Activity Flow Diagram of the Genetic Algorithm


START
Gene encoding
The population is initialized
Replications are conducted for 30 times for each gene
Compute the fitness
Terminate?
NO
YES
Crossover
Mutate
Update MR list
Selection
Kick the ball principle is activated if necessary
END (Scientist X)

where the values are divided by 2 to find the expected value. 0.25 is the expected value of the initial benefits. In these cases, the outputs are expected to be highly variable, because $cost$ is higher than the initial $benefits$. Therefore, it is harder for a scientist to become active, resulting in fizzling activity.

- If *arrival rate* and *artifact creation rate* are at low, while *probability to recover* is high, the scenario might lead to low level of activity and a small population size, resulting in fizzling activity. In those scenarios, the activity level is highly dependent on the initial conditions.

These metamorphic relations are initially identified for test purposes, and the evolution of the ensemble is monitored against those conditions. In the initial generations, rather than bounding the plausible scenario space by excluding the regions that violate the MRs, the results are recorded. The results are used to verify if the violations of identified MRs are observed and under what conditions they are observed, so that, in future replications, refined MRs can be derived from the outputs and if desired, the scenario space can be bounded.

6.2.4 Fitness Function

In genetic algorithms, likewise in biology, there is a selection process by which the fittest parameter value configurations are retained in the population. The fitness is quantitatively represented as a function that includes the fusion of five output values. As aforementioned, identifying possible robust landscapes and to determine how the communication theories behave under diverse range of scenarios is important. The output values that are under consideration are related to innovation potential metrics discussed in Chapter 5. The Core/Periphery ratio is excluded from the analysis due to the high level of variation under the majority of the scenario space. Including Core/Periphery in the calculation of the fitness function would cause bias due to the number of connections that is assigned to identify core members in the method of calculation. Further information can be found in Appendix A about the activity flow of the calculation method for Core/Periphery ratio. The metrics that are used in the fitness function are:

- Density

- Degree Centrality
- Clustering Coefficient
- Average Path Length
- Diversity Among Nodes

Functional robustness definition by Krakauer (2006) is interpreted to define the robustness metric. This kind of robustness can be achieved by invariance of the output metrics. Thus, the fitness function is described as minimization of aggregated variance measures (for each relative social network metric). There are two goals of the implemented fitness function. First goal is to minimize the variability among various metrics to discover more robust scenario space. The second goal is to measure average variability comparing different communication preferences. The fitness function to minimize is defined by Equation 6.2 below:

$$f_j = \frac{\sum_{i=1}^{N} \frac{MAE_i}{Mean_i}}{N} \tag{6.2}$$

where j is a gene (i.e., scenario), N is the number of output metrics under consideration, and MAE_i is the mean absolute error for the i_{th} element of the output vector. In the analysis, MAE is used since the deviation among the data-points for each time-tick in the time-series data is measured and it is aimed to diminish the effects of outliers on the fitness function. Different output metrics may have MAE values at different scales, so MAE is divided by mean values of each metric to normalize.

As a note, MAE and mean values are calculated based on the last N time ticks of time series data for each metric. The formula used in calculating the mean of density is represented in Equation 6.3.

$$Mean\,of\,Density = \frac{\sum_{j=1}^{N} \frac{\sum_{i=1}^{R} Density_{i,j}}{R}}{N} \tag{6.3}$$

where i is the total number of replications per scenario and N is the number of time ticks that are considered at the end of each time series data. The same logic is used to calculate means and MAE' s for each metric.

6.2.5 Selection

The selection operator is built to select the parent scenarios, which are used to reproduce offspring population to replace the actual population in the following generation. The selection operator is applied after each scenario (the population) is run for 30 replications. After the runs are completed, the mean fitness function value for each scenario is calculated. Subsequently, the probability for each scenario to be selected as a mate is determined by the following equations.

$$p_j = \frac{\frac{1}{f_j}}{\sum_{i=1}^{N} \frac{1}{f_j}} \tag{6.4}$$

$$P_j = \sum_{i=1}^{j} p_j \tag{6.5}$$

where p_j is the probability for gene j to be selected, P_j is the cumulative probability that is used in roulette wheel algorithm, and f_j is the fitness of gene j. N is the total number of genes representing the addition of fittest members of each communication theory and the whole population.

The fittest population member of each communication theory is shuffled in the selection process twice since parameter set-ups might behave differently under different communication mechanisms. In order to keep track of the fittest population members for each communication theory and to avoid dominance of the parameter set-ups that behave better under certain theories, selection is done among the fittest genes for each theory and the existing population. This mechanism adds scaling to the proportional selection algorithm; however as a disadvantage, it may cause *genetic* *drift* to be altered, evolving the population members to be identical. That is why two crossover operators are introduced in the next section. The selection of gene j is based on the following condition:

$$P_j \geq R > P_{j-1} \tag{6.6}$$

where R is the random number that is drawn between 0 and 1. If the condition is satisfied, then gene j is selected to mate. This process is replicated until the operator finds the second mate, which must be different than the first mate. As described, proportional selection algorithm is adopted among the other techniques such as "Rank Based Selection," "Tournament Selection," and "Truncation Selection" (Dréo et al., 2005).

6.2.6 Crossover and Mutation

The reproduction of genes is realized by the crossover and mutation operators. A total of $N-1$ new offsprings are reproduced after $N-1$ couples are determined by tournament selection. Among each couple, the gene with better fitness value is determined and the offspring is created with identical genome to the selected gene. Then one point crossover is applied twice. One for selecting a random bit among binary bits and equalizing that bit to the mate' s value. The other one is for equalizing a randomly selected bit among integer bits (bits 2,3,4,5) to the average of both parties' bit values. The process is stochastic, so crossover of the same distinct parents can generate different offsprings. This process is repeated until $N-1$ offsprings are reproduced.

If a randomly generated number is less than or equal to a certain probability, i.e., 1%, then a randomly selected binary bit is flipped. Mutation is a unary operator like crossover operators and realized on a single bit. First bit is not changed through crossover and mutation operators, because first bit represents communication preferences and it is intended to stay the same among all generations. As aforementioned, metaphorically, communication ttheories can be perceived as different species, while other evolving parameters are traits that are passed between species.

6.2.7 Culling

As the last step, the population members except the fittest member in terms of the fitness function value are replaced by the offsprings. It is generational replacement with elitist approach. *Elitism* consists of preserving at least one of the individuals with the best fitness from one generation to another (Dréo et al., 2005). The intention in this research is to avoid getting away from the optimum/sub-optimum areas easily by giving more chance to the best gene for reproduction. Followed the culling, the next generation runs are started. *Steady-state replacement* is not adopted. Because, due to the computational complexity and to promote *diversity*, the GA needs to excessively disturb the scenario space rather than gradual evolution.

6.3 Analysis

During the initial analysis, the ensemble consists of 30 scenarios/population members (5 scenarios for each communication mechanism). Before the first intervention, the ensemble is evolved for ten generations. Then metamorphic relations and the fittest scenarios are evaluated. It is observed that *maximum altruism* dominated the other variables, and genes converged to similar scenarios, which always have the highest value of *maximum altruism* (0.9).

In the early generations, some scenarios are observed to violate initial MRs, but they result in highly variable landscapes, so they are disappeared from the ensemble at later generations. Also, in some scenarios, it is identified that, *collective action* formula might be redundant for the majority of scientists. In those scenarios, high level of activity is expected to be observed. So, another metamorphic relation is added to the list, which is represented below:

$$\frac{1 + \frac{MinTension}{2}}{2} > \frac{maxAltruism}{2} \tag{6.7}$$

This MR is included to mitigate the dominant effect of *maximum altruism* on the activeness. Additionally, *maximum altruism* value is set to 0.5 to avoid the region the second metamorphic relation has violated.

Moreover, it is observed that both the foraging mechanism and the mutation rate stabilized. (Foraging = 2 and Mutation = 0.01). Those parameters that do not have an impact on the fitness function are eliminated. As a last step, two more parameters are introduced (*foraging extension* and *expertise to recover*) on the bits that are idle after the elimination. Then the fittest scenario is kept, *kick the ball* principle is applied by varying the parameter values of the other population members, and new generations are run. In the following sections, the results are delineated.

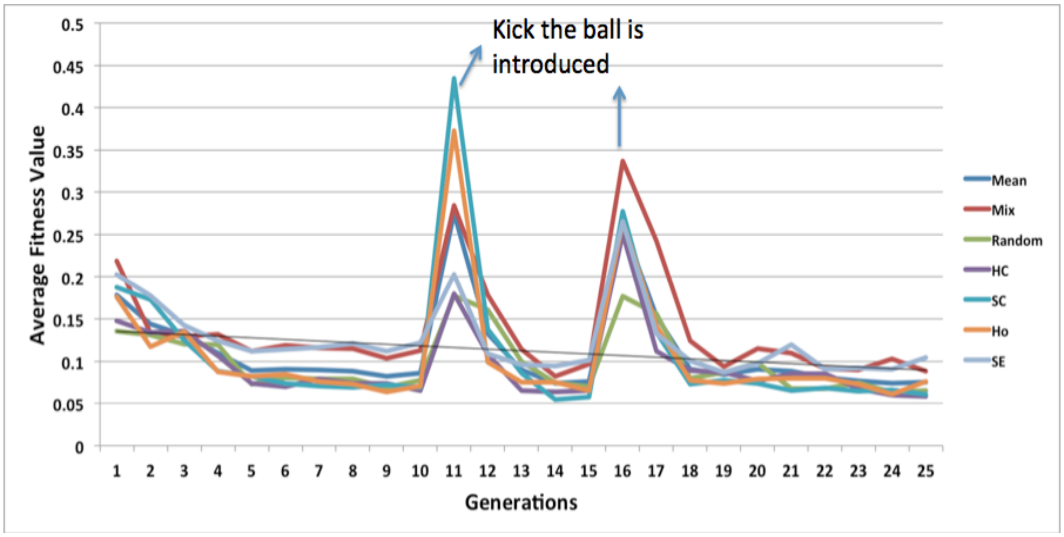
6.3.1 Results

In Figure 6.3, the evolution of the average fitness value for each communication mechanism along with the mean fitness in whole population are represented.

It is observed that the GA module improves the results over generations, which verifies the implemented algorithm. As a note, peaks are observed at generation 11 and generation 16 as a result of implemented *kick the ball* principle.

As a first observable, mean robustness for each communication mechanism is presented. This representation is inspired by the use of ensembles in the machine learning domain (Dietterich, 2000). Likewise comparing the average behavior of model ensembles, the average behavior of the sub-populations (for each communication mechanism) are compared. The goal is to measure how each communication mechanism behaves under various scenarios in Figure 6.4 is illustrated.

Figure 6.3: Average Fitness Values over Generations

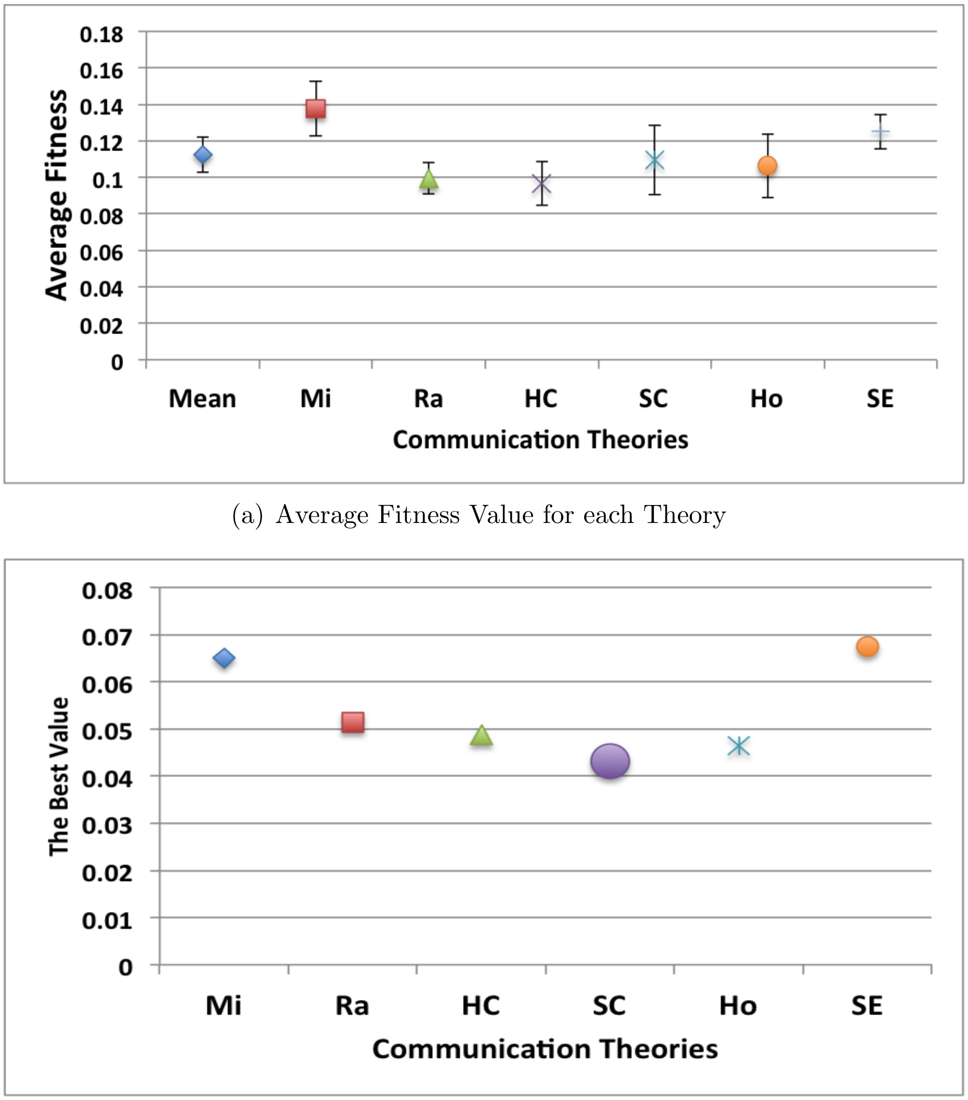


The minimum fitness value that is observed over generations is a scenario with *social capital* theory. However, considering the average performance, *human capital* theory behaves more robust than the others. When the 95% confidence intervals are presented in Figure 6.4, due to the small sample size, it is not possible to prefer one theory over another in terms of robustness. Further runs are recommended.

Additionally, accepting that the ensemble improves in terms of robustness over generations, examining the distinct parameter values can lead to draw conclusions on the converged scenario space and the impacts of variability. It is a scenario space that GA module discovers. The ensemble could converge to a scenario or diverse scenarios that provide basis for further exploration. In the analysis, the fittest scenarios converged to an identical parameter set-up with different communication mechanisms. Figure 6.5 represents the evolution of the fittest member in the ensemble for each integer parameter.

When the *maximum scope* is increased, the scientists have more global information, which gives them the ability to identify the artifacts to study on more effectively. Interestingly, the scope does not evolve to the maximum value and stabilizes around

Figure 6.4: Minimum and Average Fitness Values for each Theory



(a) Average Fitness Value for each Theory

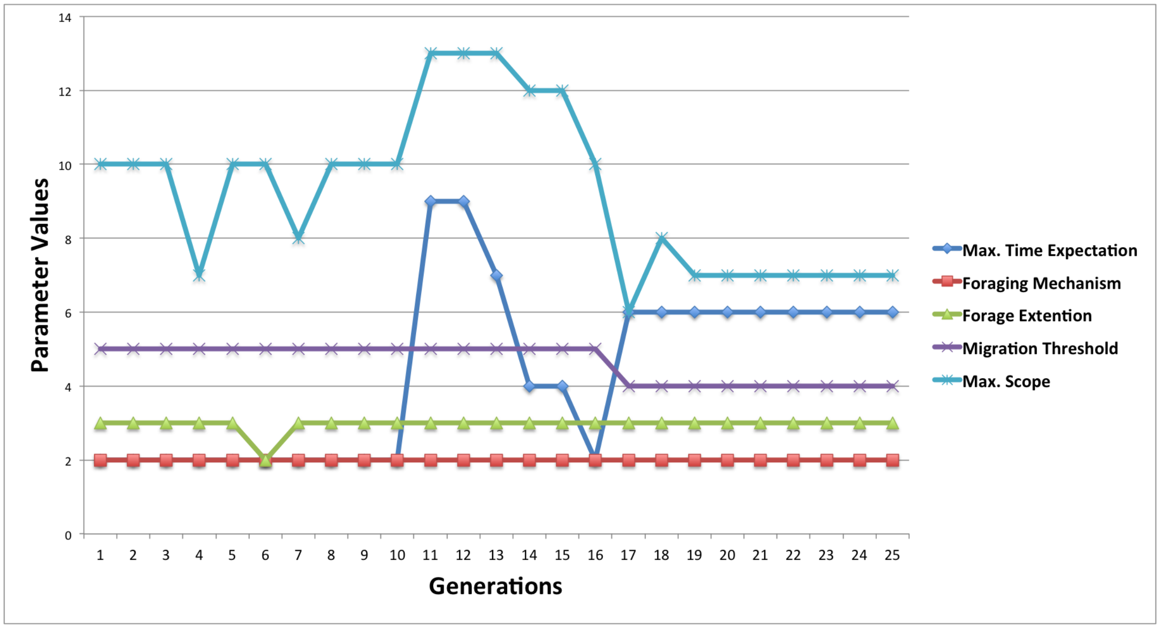
(b) Minimum Fitness Value for each Theory

Mi: Mixed Communication, Ra: Random Communication, HC: Human Capital, SC: Social Capital, Ho: Homophily, SE: Social Exchange

the 7 cells. This observation can be indicative of the presence of a level of diminishing return of the scope in relation to the robustness of the system.

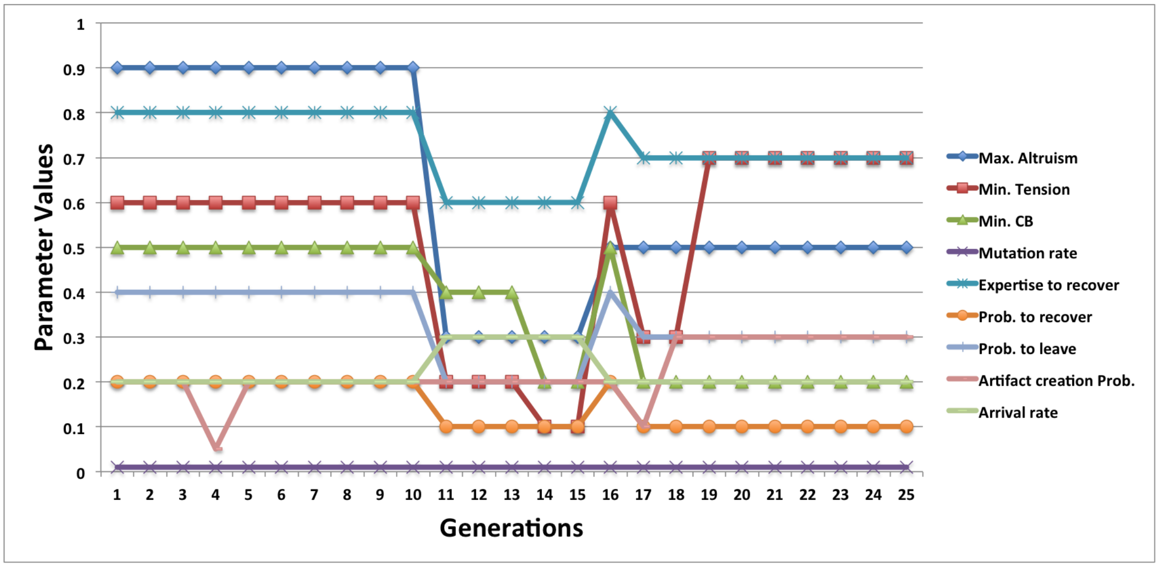
Migration threshold is the number of times that a scientist forages before considering to leave the environment to recover. The greater the threshold, the less likely the scientists are to depart from the network. This is the reason why a higher level

Figure 6.5: Integer Parameters of the Fittest Scenario over Generations



of migration threshold creates more robust networks. However, further exploration is needed to determine if there is a diminishing return as observed for the *scope.* The same comments can be made for the *forage extension*, which defines how broader *scope* scientists perceive in case of foraging. Figure 6.6 represents the evolution trends among the parameter values represented in the genome (floating numbers).

Figure 6.6: Floating Number Parameters of the Fittest Scenarios over Generations



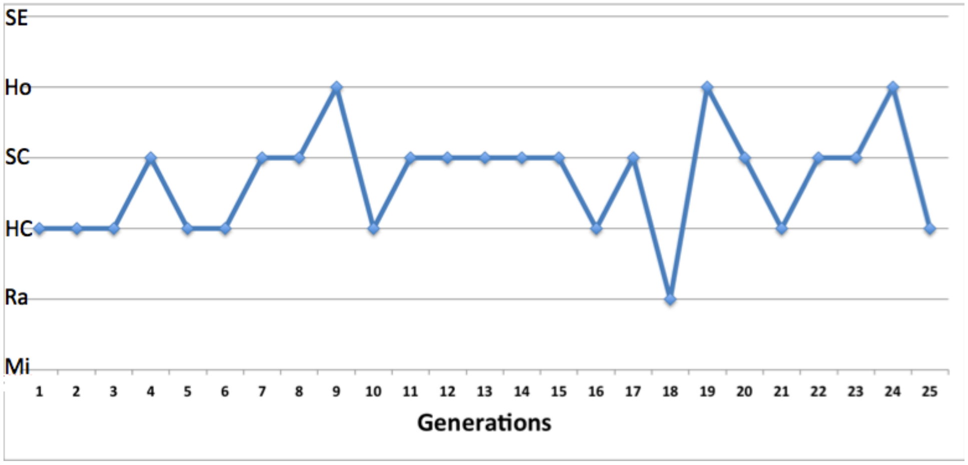
Expertise to recover sets the expertise level above which, scientists consider to depart from the environment. The greater the value is, the less likely the scientists are to leave the environment. Earlier in the analysis, it was expected that greater values would create more robust environments. However, experiments indicate otherwise. Letting scientists dissolve sooner does not improve network centrality. That is, when scientists dissolve from the network, the perturbation to the social network is smaller, resulting in more robust environments. Hence it is converged to value of 0.7.

Maximum altruism value is stabilized at 0.9 for the first 10 generations. After the value is limited to 0.5, it stabilized around 0.5. Higher values of *altruism* create more active and crowded social networks, which are more robust than smaller social networks. Interestingly, when *altruism* values are higher (first 10 generations delineate it better), *minimum tension* and *minimum cognitive burden* move in the same direction. This observation suggests that moderate level of difference between the parameter values that are dialectic forces in the collective action formula can lead to more robust landscapes.

Arrival rate is converged to the value of 0.2. Increased arrival rate result in more crowded social networks, that can be more robust against perturbations. However, it is stabilized at lower values than the upper limit. *Probability to recover* (related to the turnover rate) decreases as expected, because it decreases the probability of dissolution from the network. *Probability to leave* is related to mobility of the members. Lower level of mobility can cause scientists to stick with the project they reside on, missing other opportunities, while higher levels can cause distraction, abandoning promising opportunities before attracting more attention. Thus, scientists may end up wandering in the environment most of the time and connect to less number of people. This causes the network to be smaller, therefore less robust to perturbations.

Figure 6.7 represents the communication mechanism of the fittest scenarios over time. It is observed that under the same parameter configurations, different communication mechanisms outperform the others at different generations, which can be caused by environmental uncertainty (different random number seeds) or communication mechanisms.

Figure 6.7: Communication Mechanism of the Fittest Scenarios over Generations



Mi: Mixed Communication, Ra: Random Communication, HC: Human Capital, SC: Social Capital, Ho: Homophily, SE: Social Exchange

Table 6.2: Decoded Values of the Identified Portfolio of Scenarios

Parameter	Value
Communication Preference	{0,1,2,3,4,5}
Maximum Time Expectation	4
Mutation Rate	0.01
Foraging Mechanism	Basic
Expertise to Recover	0.7
Migration Threshold	4
Maximum Scope	7
Maximum Altruism	0.5
Minimum Tension	0.7
Minimum Burden	0.2
Forage Extension	3
Probability to Recover	0.1
Probability to Leave	0.3
Artifact Creation Probability	0.3
Arrival Rate	0.2

The portfolio of scenarios is represented in Table 6.2. The parameters that are not presented in the table are the same as the base-line model. The portfolio is determined by observing the scenarios that output the fittest results at the last generations of the search. The decision of selecting the scenarios that are added in the portfolio is qualitative. Depending on the purpose of the study, the decision can be made by just observing the fittest scenarios over time or selecting the scenarios that give results between a range of values. In this study, further batches of runs are conducted to provide an example for the use of exploratory modeling. In order to narrow the focus while enabling the best use of the computational resources, only fittest scenarios of last generations are included in the portfolio.

In conclusion, communication theory parameter is flipped a couple of times in the portfolio. When the overall robustness of the population is observed, non of the theories are significantly better than the others. The analysis needs more number of replications and further exploration. Therefore, in order to test the communication mechanisms at the portfolio scenarios and to measure innovation potential, 200 runs for each theory are conducted. The aim is to assess fitness values of different communication mechanisms and examine whether there is a tradeoff between *robustness* and *innovation potential*. Table 6.3 presents the fitness values for each communication mechanism.

Table 6.3: Fitness Values - The Most Robust Parameter Scenario

Theory	**Fitness**
Mixed Communications	0.072
Random Communication	0.062
Human Capital	0.060
Social Capital	**0.057**
Homophily	0.067
Social Exchange	0.076

For further analysis, confidence intervals are needed. However, fitness function is based on time-series data of all replications and there is no fitness value for a single replication. So, these 200 runs are divided into batches of 10 replications (20 batches) and the analysis of variance is conducted among these batches. Table 6.4 summarizes the number of batches that is guessed to conduct analysis of variance with desired half-width. Figure 6.8 represents 95% confidence intervals for each mechanism. As a note, the interpretation of these confidence intervals are biased on the number of replications for each batch and the assigned half-width. As a result, it is not possible to prefer one theory over another. Further tests can be conducted including more replications for each batch or lower level of half-width (which will require more number of batches) to be able to distinguish the performance of communication mechanisms. It is interpreted from the analysis that the fitness function design (accounting MAE of time-series data) is also effective on variable and similar outputs of some communication mechanisms.

Table 6.4: Analysis of Variance for 20 batches of 10 replications

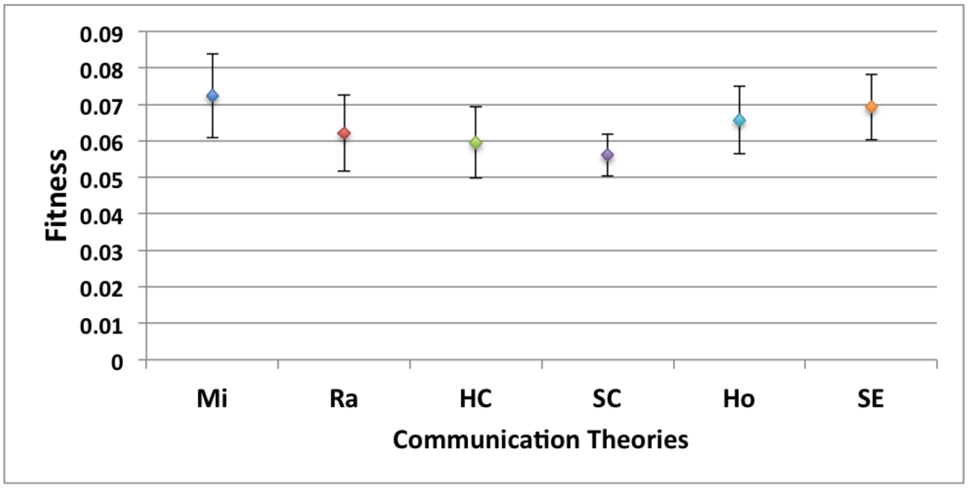
Theory	**Mean**	**Standard Deviation**	**n**	**95% CI**
Mi	0.072	0.012	11.517	0.011
Ra	0.062	0.011	12.771	0.010
HC	0.060	0.010	12.104	0.010
SC	0.056	0.006	4.794	0.006
Ho	0.066	0.009	9.141	0.009
SE	0.069	0.009	7.687	0.009

($t_{19,1-0.025} = 2.093$, half-width = 10% of the mean)

Table 6.5: Mean of Various Metrics - The Most Robust Scenario

Theory	**Density**	**DC**	**CC**	**AvgPath**	**CP**	**DiversityN**	**DiversityL**	**SW**
Mi	0.192	0.255	0.397	2.434	1.493	0.492	0.286	0.163
Ra	0.322	0.33	0.322	1.805	4.600	0.607	0.371	0.178
HC	0.287	0.319	0.287	1.884	3.664	0.559	0.399	0.152
SC	0.281	0.325	0.281	1.896	3.370	0.571	0.390	0.148
Ho	0.254	0.318	0.254	1.933	3.710	0.519	0.422	0.131
SE	0.199	0.257	0.199	2.400	1.585	0.495	0.285	0.083

Figure 6.8: 95% Confidence Intervals for Different Communication Mechanisms



Mi: Mixed Communication, Ra: Random Communication, HC: Human Capital, SC: Social Capital, Ho: Homophily, SE: Social Exchange

Table 6.6: Standard Deviation of Various Metrics - The Most Robust Scenario

Theory	**Density**	**DC**	**CC**	**AvgPath**	**CP**	**DiversityN**	**DiversityL**	**SW**
Mi	0.030	0.021	0.040	0.146	0.400	0.064	0.046	0.274
Ra	0.051	0.021	0.056	0.087	1.594	0.044	0.032	0.644
HC	0.038	0.017	0.048	0.075	0.792	0.042	0.023	0.640
SC	0.035	0.021	0.038	0.070	0.796	0.042	0.022	0.543
Ho	0.040	0.018	0.049	0.082	1.194	0.045	0.019	0.598
SE	0.026	0.020	0.037	0.161	0.408	0.061	0.051	0.230

Comparing the results (Table 6.5 and Table 6.6) of the most robust landscape with the analysis of Chapter 5, more robust networks generate more connections. While degree centrality is observed at similar levels, clustering coefficient is observed at higher levels than the base-line scenario, suggesting that the networks have densely connected cliques, or the network behaves like a single densely connected clique. Since average path length is observed at lower numbers and core periphery ratio is high, it can be concluded that the networks consist of a single highly dense cluster, which is a small world itself. Small-worldliness, lower average path length, and higher density are indicative of a more innovation potential. However, Core/Periphery ratios are increased that indicates the number of periphery members are not many. Diversity among links suffers from highly dense clusters. Additionally, diversity among nodes is spurred creating connections between scientists from different interest levels.

6.4 Limitations and Conclusions

The exploration software in this research provides a process to navigate effectively through the plausible scenario space and identify key configurations that can lead to construct lines of reasoning in terms of robustness and innovation. If more insight is desired, further sensitivity analysis can provide information on whether these scenarios are significantly different or similar. Additionally, the identified portfolio of scenarios are constructed as a basis for further exploration of the genetic algorithm by bounding the plausible scenario space, or for high-resolution exploration by re-determining the possible parameter ranges to search on. The developed human-mediated exploration software is realized to provide a search process and a portfolio of robust scenarios rather than a product (optimal answer).

An important limitation of this research is the amount of computational resources that could be dedicated to exploration. Thirty replications for each scenario are dedicated to provide sufficient level of convergence. Since the research does not aim optimization, strong convergence can also be problematic. Aside from the discussed approach of re-initiating the GA for multiple times (Zeigler et al., 1997), an alternative implementation would be to run each scenario for 2 or 3 replications and dedicate the remaining computational resources to further exploration (Dibble, 2006). However, standard GA (as implemented in this research) have strong convergence tendencies (Burke et al., 2002). So, in order to interpret whether this alternative method would provide more diversity as expected, the GA operators should be varied and the performance should be evaluated.

In this chapter, four particular operations are implemented to promote diversity in the population: (1) the set of random number seeds are changed at each generation,

(2) by the interventions, *kick the ball* principle is introduced (randomly varying the population), (3) during interventions, *metamorphic relations* are tested to bound the plausible scenario space, and (4) the population is divided in sub-populations of communication mechanisms (which are preserved over time). Alternative approaches to maintaining the diversity in the population are listed such as *crowding models*, *assortative mating*, *dividing the population in sub-populations*, and *fitness sharing* (Smith et al., 1993; Burke et al., 2002). Another limitation of the study is related to the implemented GA design. In particular, experimentation with alternative GA designs can be performed to understand the features that can make the exploration software faster and more effective. The results can be compared to determine the effectiveness of the operations that are applied to promote diversity in this research.

To conclude, in the most robust landscape, high level of activity and knowledge creation generate highly dense, clustered small-world networks that increase the robustness while the diversity and core-periphery structures are mitigated. Further runs with more number of replications are suggested to compare communication mechanisms and their robustness performance. If desired, different fitness functions can be tested based on the standard deviation of average behavior among replications or coefficient variation (not MAE) etc. Also, the portfolio can be extended by including the scenarios (i.e. fittest scenario of generation 14) that perform close enough to the scenarios in the portfolio for further sensitivity analysis. In the following chapter, the summary of the research is described along with suggestions for future research.

Chapter 7
CONCLUSIONS AND FUTURE RESEARCH

The *Science of Science and Innovation Policy* program of the National Science Foundation (NSF) is particularly interested in development of computational models that explore different aspects of knowledge creation. The aim is to identify the theories and the mechanisms that mimic the dynamics in knowledge creation and then to conduct exploratory analysis under various conditions to develop better understanding of innovation and creativity. Policymakers might eventually aim to develop open-scientific environments and maintain cyber-infrastructures that provide a landscape for scientific collaboration.

In his well-known work, Nielsen (2011) coins the term *designed serendipity*. He anticipates that the world is approaching to the second *open science* revolution (first one was the publication system around 300 years ago), that will alter the way people publish and the publication system itself by creating new norms (new agora) (Gibbons et al., 1997) and tools that people conduct research on. IBM[1] and MIT Collective Intelligence Lab[2] have already been working on online collaboration tools that will foster innovation and creativity. Nielsen (2011) states that human society is in the era to figure out how to design these tools in a way to promote serendipity coming out of the system. That is why he calls the phenomenon *designed serendipity.* **The results presented in this dissertation provide insight about which components have impacts on collaboration in Global Participatory Science (GPS) and which communication behaviors can be promoted to improve innovation potential**

[1] http://www.research.ibm.com/labs/watson/index.shtml - As of 4.07.2013
[2] http://cci.mit.edu - As of 4.07.2013

and capacity. In the long run, these findings can be used for orchestrating the collaboration in GPS and designing online cyber-infrastructures.

In this study, *self-organizing complex adaptive system* viewpoint is adopted. The CAS domain provides a trans-disciplinary research framework, allowing a wide-variety of disciplines to benefit from. Initially, self-organization principles and theoretical foundations that can explain the behaviors of the scientists in GPS are examined. **It is observed that *collective action* theory is a plausible theory on which to ground the models of *collaborative environments***. Collective action dynamics is modeled in terms of self and mutual interest variables based on the findings of Olson (1974). The theory is used to explain different phenomena in late ′80s and early ′90s including computational models, however the theory has vanished in the last decades (Oliver and Marwell, 2001). Recently, the theory is brought up to explain the open science phenomenon.[3] **To the best of our knowledge, to date, there is no computational model that incorporates *collective action* theory to explain scientific knowledge generation.**

Self-organization mechanisms that explain the dynamics in GPS are implemented over the collective action model. Unlike prior studies on OSSD communities that use the *stigmergy* mechanism (Cui et al., 2009), *preferential attachment* mechanism is used to represent the artifact/project selection process. *Information foraging* theory is interpreted along with the domain knowledge, and it is implemented as a mechanism that explains the migration of scientists among the projects. Learning and influence mechanisms are implemented by monotonic increase functions, representing the cumulative knowledge creation. Since GPS is an open system with new arrivals and departures of the scientists, population dynamics are also introduced. All these mechanisms are implemented with *positive and negative feedback* dynamics based on the local information that is available to scientists. *Bounded*

[3] http://michaelnielsen.org/blog/the-logic-of-collective-action/ - As of 4.07.2013

rationality assumption is an essential principle perceived in all human decision processes. The mechanisms that are related with those decisions incorporate *roulette wheel* algorithms to represent proportional selection rather than optimal selection. The main principle is: "People satisfice, rather than optimize." **All mechanisms that are implemented in this research are grounded on theory and empirical findings.**

One of the limitations at the first phase of this research is the use of the V-shaped function that approximates *tension* over time. A U-shaped function can be introduced to the model in future studies. Different learning techniques can be employed to build up expertise of scientists and the complexity of the artifacts. Additionally, agents can perceive different levels of appropriate information in effectively finding the artifact opportunities for contribution. Different information types can be introduced to the *preferential attachment* mechanism based on the objective of the study.

Following the representation of the intrinsic motivation dynamics of scientists, the conceptualization of the base-model is concluded. The base-model development was an important stage for this research, since it formally defines the parameters and effectively formulates the known dynamics and spatio-temporal characteristics of GPS. **Verification and validation studies are conducted during and after the implementation with performance tests as described in Chapter 4. Additionally, OBO data is used to conduct tests for validation of emergent patterns. In the future, if additional social network data become available, the model can be calibrated to represent specific communities for further validation.**

The second milestone, was the identification of more innovative communication traits among scientists. Communication preferences are stated as important factors that have an impact on the evolution of GPS and

social networks. To conduct sensitivity analysis, the candidate communication theories that are related with the problem domain are identified. Then those theories are implemented over the base-model. In Chapter 5, the interpretation of each theory and their relative implementations are described. Under the base-model conditions that the validation studies are conducted on, innovation potential of the identified communication traits are explored. Innovation potential is represented in terms of different hypothesis: (1) *High centrality-fewer structural holes*, (2) *Density*, (3) Diversity, and (4) Interdisciplinarity. **Most studies calculate the diversity in terms of individual disparities in the population, while in this research, another diversity metric that identifies discrepancy among the atomic social networks of individuals is introduced. It was effective to identify disparity among the nodes in the social network, so it is promising to be used in different studies.**

All metrics are represented to policymakers by a summary table (matrix of levels that different metrics converge), so that the behavioral patterns of different theories can be distinguished. Moreover, if it is desirable to conduct further runs, the outputs can be interpreted by the policymakers to aid their multi-criteria decision making process. While a theory might seem more innovative under a set of conditions, another theory can outperform that theory under different set of conditions. Interestingly, *social capital* theory dominated the other theories during the analysis in Chapter 5.

***Social capital* theory supports innovativeness relatively more than the other candidate theories. In this research, it is revealed that if the information about the social degrees of scientists are broadcasted on online tools and if link formation between highly central members and periphery members are fostered, then networks generate more innovation potential.** Specifically, the *social capital* theory promotes connections among central scientists and periphery scientists, enabling innovation diffusion among the scientists. Highly central members are also more likely to have more contributions and gain more expertise. As the expertise levels of the scientists increase, complexity of the artifacts also benefit. This leads to increased maturity in the environment, resulting in sustained knowledge creation. *High centrality-fewer structural holes* hypothesis is supported more by the *social capital* theory, because the theory promotes highly central members, who are connected to the periphery members from different clusters. In terms of link diversity, *social capital* gives opportunity to form connections between different cliques, causing diverse atomic individual networks. This mechanism also decreases the observed average path length in the network.

Core/periphery structures promote diffusion of innovative ideas. It is observed that *mixed connections* and *social exchange* favor the balance among the scientists. They generate a small number of core members, while the number of periphery populations is large. Regarding the balance of expertise, they usually result in moderate level of expertise in the population.

This research provides the base-model as a computational laboratory that is developed with object oriented programming language (Java) principles using the Repast framework, so that it is extensible and portable across platforms. Repast is selected as the development framework, because it is porous; that is, it provides a framework to model agents that are not necessarily atomic but can be distinct by design. Repast is also called *recursive* and hence allows designing nested combinations of agents and spaces. Therefore, the model can be extended by adding higher level contexts or agents, including communities as different layers and these different layers can still communicate.

Robustness, as a systemic characteristic, is an important objective for policymakers. The scenario space is needed to be explored by testing different communication preferences under different parameter set-ups. For that reason, *exploratory modeling* method is adopted and **a GA module is implemented to conduct intelligent**

search on the scenario space while improving the scenarios in terms of the robustness metric. Robustness is defined by aggregating the mean absolute errors of different output metrics. The less variable the outputs are, the more robust are the scenarios. Furthermore, there was a need to build a feedback mechanism between generations to identify the domain specific or model specific regions of the scenario space that behave highly variable or give trivial results. ***Metamorphic relations*** **(MRs) are introduced as a feedback mechanism. It is addressed as a promising method in identifying the expected behaviors of the model under different conditions as well as bounding the search space.** In this research, the feedback mechanism between MRs and GA module is realized manually, however there are studies that implement automatic creation of metamorphic testing (Gotlieb and Botella, 2003). Future research includes the evaluation of metamorphic relations to automatically generate further relations that can be used. **Experimental results indicate that the implemented GA module that is coupled with MRs can be applied in different domains for different problem sets.**

The search for robust landscapes in this research is limited due to the computational complexity and the lack of computational resources. In order to draw more significant results, the number of replications can be re-determined. The simulations are run on an Apple laptop, that has 8 GB of memory and 2.3 Ghz i5 CPU with 5400 Rpm hard-drive. Even though the kick the ball principle is applied, the runs converged to the same region after each randomization. Different techniques to promote *diversity* can be used in the GA design as a future work. Also, more runs can be conducted after the boundaries of the parameter values are relaxed and continuous values are allowed in the parameter space. Therefore, diminishing returns of various parameters can be identified. **For the objective of this research, providing the exploration software and exploring the general behaviors of the system**

that is bounded by the MRs and expert opinion (opinion of the author) is considered sufficient.

The further experiments could not reveal any communication theory that globally outperforms the others in terms of robustness. In general, the parameter values that generate more arrivals and cause more populated social networks (by mitigating the effect of dissolution) are observed to create more robust landscapes, since they are more resilient against perturbations by the removal of central members. Interestingly, the tendency to balance certain values is observed. For example, the expertise threshold after which scientists might leave the environment, is set at a medium level. Experimental observations suggest that more expert scientists are likely to become central in the network over time and in the case of a removal, they are more likely to perturb the network more. So, medium level values result in more robust landscapes. Additionally, the gap between *maximum altruism* value and the combined effect of *minimum cognitive burden* and *minimum tension* values is oriented to be at medium levels. **A tradeoff is addressed between robustness and innovation potential. It is stated that more robust scenarios have dense and more clustered networks with highly central members promoting trust and sharing of ideas. However, the diversity among links and core/periphery structures that are related with structural holes are hindered.**

The findings of this research can assist policymakers during the design phase of online collaboration tools. The behaviors observed under each theoretical mechanism can be interpreted and related information can be broadcasted transparently in terms of the network structure. For example, the *social capital* theory, that is found to promote higher levels of innovation potential, states that scientists are more likely to connect to the ones who are connected to more people. Consequently, scientists aim to increase the social resources they have. In order to support *social capital* theory, in the tool design, the degree information of scientists and their social reachability can be presented publicly to guide the scientists. Besides, policymakers can develop mechanisms (such as reputation index) to incentivize the connections between central members and periphery members. Hence central members would incline to connect to less central scientists.

Moreover, the evolution of the social network can be observed over time. If the network evolves to undesired stages, other communication styles can be promoted. For example, if the network transtions into a state, which has great number of central members and highly specialized cliques, then *mixed communications* or *social exchange* theory can be supported. Scientists from different backgrounds and expertise levels can be injected into different projects to balance the network. This would promote core/periphery structures, which are known to promote innovation potential and capacity.

Interdisciplinarity is an aspired characteristic highly promoted in the modern scientific culture. **Network coherence matrix can be used by policymakers to identify what properties they want and by which theories they can guide the system toward that particular property.** While specialized interdisciplinary networks are more innovative, more balanced networks with potential interdisciplinary integration can also be desired by policymakers to enable effective management. Such networks can be generated and maintained by promoting *homophily* theory (connection between scientists from similar domains and interests).

Mixed communications mechanism assigns a communication preference to a scientist with equal probability. As a future venue of research, until what extent each communication style should be promoted in the environment can be investigated. Subsequently, the granularity of individual information that enables each communication preference and the way that information is shared can be diversified in the model. Further, incentive mechanisms to encourage different communication preferences can be designed to promote innovation potential and capacity. Besides the implemented communication mechanisms in this research, other mechanisms can also be modeled based on different communication theories (i.e., Cognitive Consistency, Balance, and Network Exchange Theory).

Bibliography

Albert, R. and A. Barabási (2002). Statistical mechanics of complex networks. *Reviews of modern physics 74*(1), 47.

April, J., F. Glover, J. P. Kelly, and M. Laguna (2003, October). Practical introduction to simulation optimization. In *Proceedings of the 2003 Winter Simulation Conference*, pp. 71–78.

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces That Shape Our Decisions* (1 ed.). HarperCollins.

Arndt, J. (1985). On making marketing science more scientific: role of orientations, paradigms, metaphors, and puzzle solving. *The Journal of Marketing 49*(3), 11–23.

Atmar, W. (1994). Notes on the simulation of evolution. *IEEE Transactions on Neural Networks 5*(1), 130–147.

Auer, S. and H. Braun-Thürmann (2011). Towards Bottom-Up, Stakeholder-Driven Research Funding–Open Science and Open Peer Review. Technical report.

Axelrod, R. (1997a). Advancing the art of simulation in the social sciences. *Complexity 3*(2), 16–22.

Axelrod, R. (1997b). The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution 41*(2), 203–226.

Axelrod, R. (2006). Building new political actors. In *Generative social science: Studies in agent-based computational modeling*, pp. 121–144. Princeton Univ Press.

Axelrod, R. and M. Cohen (2001). *Harnessing complexity: Organizational implications of a scientific frontier.* Basic Books.

Axtell, R. and J. Epstein (2006). Coordination in transient social networks: an agent-based computational model of the timing of retirement. In *Generative social science: Studies in agent-based computational modeling.* Princeton Univ Press.

Bäck, T. and H.-P. Schwefel (1993, March). An Overview of Evolutionary Algorithms for Parameter Optimization. *Evolutionary Computation 1*(1), 1–23.

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk (2009, June). Diversity and Complexity in DNA Recognition by Transcription Factors. *Science 324*(5935), 1720–1723.

Balci, O. (1990). Guidelines for successful simulation studies. In *Winter Simulation Conference Proceedings*, pp. 25–32.

Bankes, S. C. (1992). Exploratory Modeling and the Use of Simulation for Policy Analysis. RAND Corporation.

Bankes, S. C. (1993, May). Exploratory Modeling for Policy Analysis. *Operations Research 41*(3), 435–449.

Bankes, S. C. (2002). Tools and techniques for developing policies for complex and uncertain systems. *Proceedings of the National Academy of Sciences of the United States of America 99*(Suppl 3), 7263–7266.

Bankes, S. C., R. Lempert, and S. Popper (2002, November). Making Computational Social Science Effective: Epistemology, Methodology, and Technology. *Social Science Computer Review 20*(4), 377–388.

Barabasi, A. (2002). *Linked: The New science of networks.* Perseus Books.

Becker, G. (1978). *The economic approach to human behavior.* The University of Chicago Press.

Belton, V. and T. Stewart (2002). *Multiple criteria decision analysis: an integrated approach.* Springer.

Benhamou, F. and S. Peltier (2010, November). Application of the Stirling Model to Assess Diversity using UIS Cinema Data. *UNESCO Institute for Statistics*, 1–73.

Bernstein, C., A. Kacelnik, and J. Krebs (1988). Individual decisions and the distribution of predators in a patchy environment. *The Journal of Animal Ecology 57*(3), 1007–1026.

Blank, A. and S. Solomon (2000). 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*(1-2), 279–288.

Blau, P. (1977). *Inequality and heterogeneity: A primitive theory of social structure.* Free Press New York.

Boguna, M., R. Pastor-Satorras, A. Diaz-Guilera, and A. Arenas (2004, November). Models of social networks based on social distance attachment. *Phys. Rev. E 70*(5), 056122.

Bolland, J. (1988). Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social networks 10*(3), 233–253.

Bonacich, P. (1990). Communication dilemmas in social networks: An experimental study. *American Sociological Review 55*(3), 448–459.

Booth, D., H. Haas, F. Mccabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard (2004). Web Services Architecture. Technical report.

Borgatti, S. and M. Everett (2000). Models of core/periphery structures. *Social networks 21*(4), 375–395.

Boyd, J., W. Fitzgerald, and R. Beck (2006). Computing core/periphery structures and permutation tests for social relations data. *Social networks 28*(2), 165–178.

Bratley, P., B. L Fox, and L. E Schrage (1987). *A guide to simulation.* Springer.

Burke, E., S. Gustafson, and G. Kendall (2002). A survey and analysis of diversity measures in genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 716–723. sn.

Burt, R. (1982). *Toward a structural theory of action: Network models of stratification, perception, and action.* New York: Academic Press.

Burt, R. (1995). *Structural holes: The social structure of competition.* Harvard University Press.

Burton, R. (2003). Computational laboratories for organization science: Questions, validity and docking. *Computational and Mathematical Organization Theory 9*(2), 91–108.

Camazine, S. (2003). *Self-organization in biological systems.* Princeton Univ Pr.

Carayol, N. and J. Dalle (2007). Sequential problem choice and the reward system in Open Science. *Structural Change and Economic Dynamics 18*(2), 167–191.

Carley, K. M., N. Y. Kamneva, and J. Reminga (2004). Response surface methodology. Technical report.

Carlson, J. and J. Doyle (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences 99*(Suppl 1), 2538.

Carson, Y. (1997). Simulation optimization: methods and applications. In *Proceedings of the 29th conference on Winter simulation*.

Caruana, R., A. Niculescu-Mizil, G. Crew, and A. Ksikes (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 18. ACM.

Cawley, G. and N. Talbot (2010, August). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res. 99*, 2079–2107.

Charnov, E. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology 9*, 129–136.

Chen, T. Y., S. C. Cheung, and S. M. Yiu (1998). Metamorphic testing: A new approach for generating next test cases. Technical report, Department of Computer Science, Hong Kong University of Science and Technology.

Cowan, R. and N. Jonard (2004). Network structure and the diffusion of knowledge. *Journal of economic Dynamics and Control 28*(8), 1557–1575.

Cui, X., J. Beaver, J. Treadwell, T. Potok, and L. Pullum (2009). A Stigmergy Approach for Open Source Software Developer Community Simulation. *Computational Science and Engineering, 2009. CSE'09. International Conference on 4*, 602–606.

David, P. (1998). Common agency contracting and the emergence of "open science" institutions. *The American Economic Review 88*(2), 15–21.

De Nooy, W., A. Mrvar, and V. Batagelj (2005). Exploratory social network analysis with Pajek. *Network 40*(3), 362.

Dhanaraj, C. and A. Parkhe (2006). Orchestrating innovation networks. *Academy of Management Review 31*(3), 659.

Diaz-Guilera, A., , S. Lozano, and A. Arenas (2009). Propagation of Innovations in Complex Patterns of Interaction. In *Innovation networks: new approaches in modelling and analyzing*, pp. 271–286. Springer Verlag.

Dibble, C. (2006). Computational Laboratories for Spatial Agent-Based Models. In *Handbook of Computational Economics*, pp. 1511–1548. Handbook of Computational Economics.

Dietterich, T. (2000). Ensemble methods in machine learning. *Multiple classifier systems - Lecture Notes in Computer Science 1857*, 1–15.

Ding, J., T. Wu, D. Wu, J. Q. Lu, and X. H. Hu (2011). Metamorphic testing of a Monte Carlo modeling program. In *Proceedings of the 6th International Workshop on Automation of Software Test*, pp. 1–7. ACM.

Dréo, J., A. Petrowski, P. Siarry, and E. Taillard (2005). *Metaheuristics for hard optimization: methods and case studies.* Springer.

Dron, J. and T. Anderson (2009). On the Design of Collective Applications. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, pp. 368–374. IEEE.

Epstein, J., D. Cummings, and S. Chakravarty (2006). Toward a containment strategy for smallpox bioterror: an individual-based computational approach. In *Generative social science: Studies in agent-based computational modeling.* Princeton Univ Press.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling.* Princeton Univ Press.

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation 11*(4), 12.

Faccenda, J. F. and R. F. Tenga (1992). A combined simulation/optimization approach to process plant design. In *the 24th conference*, New York, New York, USA, pp. 1256–1261. ACM Press.

Feynman, R. (1998). *The Meaning of It All.* Penguin.

Finholt, T. and G. Olson (1997). From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science 8*(1), 28.

Flack, J., D. Krakauer, and F. De Waal (2005). Robustness mechanisms in primate societies: a perturbation study. *Proceedings of the Royal Society B: Biological Sciences 272*(1568), 1091.

Foster, I. (2005). Service-oriented science. *Science 308*(5723), 814.

Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social networks 1*(3), 215–239.

Fu, M. and F. Glover (2005). Simulation optimization: a review, new developments, and applications. In *Proceedings of 37th conference on Winter simulation.*

Gibbons, M., L. C, and N. H (1997). *The new production of knowledge: the dynamics of science and research in contemporary societies.* Sage.

Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research Online 2.*

Gilbert, N. (2006). Putting the Social into Social Simulation. In *Keynote address to the First World Social Simulation Conference, Kyoto.*

Goldstein, J. (2011). Attractors and nonlinear dynamical systems. *Deeper Learning*, 1–17.

Gotlieb, A. and B. Botella (2003). Automated metamorphic testing. In *Computer Software and Applications Conference, 2003. COMPSAC 2003. Proceedings. 27th Annual International*, pp. 34–40.

Grunwald, P. (2005). A tutorial introduction to the minimum description length principle. Technical report, Centrum voor Wiskunde en Informatica.

Guderlei, R. and J. Mayer (2007, October). Statistical Metamorphic Testing Testing Programs with Random Output by Means of Statistical Hypothesis Tests and Metamorphic Testing. In *Quality Software, 2007. QSIC '07. Seventh International Conference on*, pp. 404–409.

Hardin, R. (1982). *Collective action.* Resources for the Future.

Hawkins, D. M. (2004, January). The Problem of Overfitting. *Journal of Chemical Information and Modeling 44*(1), 1–12.

Holland, J. (1996). *Hidden order: How adaptation builds complexity.* Basic Books.

Hollingshead, A. B., J. Fulk, and P. Monge (2002). Fostering intranet knowledge sharing: An integration of transactive memory and public goods approaches. *Distributed work*, 335–355.

Jensen, C. and W. Scacchi (2010). Governance in Open Source Software Development Projects: A Comparative Multi-Level Case Study Analysis. In *The 6th International Conference on Open Source Systems: IFIP Working Group 2.13*, Notre Dame, IN, USA. Springer Boston: Springer Boston.

Kendall, B. E. (2001). Nonlinear dynamics and chaos. *eLS*.

Klügl, F. (2008). A validation methodology for agent-based simulations. In *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 39–43.

Krakauer, D. (2006). Robustness in Biological Systems: a provisional taxonomy. *Complex systems science in biomedicine*, 183–205.

Krebs, V. and J. Holley (2002). Building sustainable communities through network building. Technical report.

Kuhn, T. (1996). *The structure of scientific revolutions*. University of Chicago press.

Laine, T. (2006). *Agent-based model selection framework for complex adaptive systems*. ProQuest.

Latour, B. (1998). From the world of science to the world of research? *Science 280*(5361), 208.

Lattemann, C. and S. Stieglitz (2005). Framework for Governance in Open Source Communities. In *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*, pp. 192a–192a.

Liepins, G. E. and M. R. Hilliard (1989). Genetic algorithms: Foundations and applications. *Annals of operations research 21*(1), 31–57.

Lynne, H. and N. Gilbert (2009). Social circles: A simple structure for agent-based social network models. *Journal of Artificial Societies and Social Simulation 12*(2), 3.

Ma, T. and B. Abdulhai (2002). Genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters. *Transportation Research Record: Journal of the Transportation Research Board 1800*(-1), 6–15.

March, J. (1991). Exploration and exploitation in organizational learning. *Organization Science 2*(1), 71–87.

McCormack, J. (2007). Artificial ecosystems for creative discovery. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pp. 307.

Merton, R. (1979). *The sociology of science: Theoretical and empirical investigations.* University of Chicago Press.

Milbergs, E. and N. Vonortas (2006, January). Innovation Metrics: Measurement to Insight. Technical report, IBM Corporation.

Mitchell, B. and L. Yilmaz (2009). Symbiotic adaptive multisimulation: An autonomic simulation framework for real-time decision support under uncertainty. *ACM Trans. Model. Comput. Simul. 19*(1), 2:1–2:31.

Monge, P. and N. Contractor (2003). *Theories of communication networks.* Oxford University Press, USA.

Moritz, M., M. Morais, L. A. Summerel, J. Carlson, and J. Doyle (2005). Wildfires, complexity, and highly optimized tolerance. *Proceedings of the National Academy of Sciences of the United States of America 102*(50), 17912.

Muffatto, M. and M. Faldani (2003). Open Source as a complex adaptive system. *Emergence 5*(3), 83–100.

Mukherjee, A. and S. Stern (2009). Disclosure or secrecy? The dynamics of open science. *International Journal of Industrial Organization 27*(3), 449–462.

Murphy, C., M. S. Raunak, A. King, S. Chen, C. Imbriano, G. Kaiser, I. Lee, O. Sokolsky, L. Clarke, and L. Osterweil (2011). On effective testing of health care simulation software. In *Proceedings of the 3rd Workshop on Software Engineering in Health Care*, pp. 40–47. ACM.

Naveh, I. and R. Sun (2006). A cognitively based simulation of academic science. *Computational and Mathematical Organization Theory 12*(4), 313–337.

Nazzal, D., M. Mollaghasemi, H. Hedlund, and A. Bozorgi (2011, July). Using genetic algorithms and an indifference-zone ranking and selection procedure under common random numbers for simulation optimisation. *Journal of Simulation 6*(1), 56–66.

Newman, M. (2010). *Networks: An Introduction.* Oxford University Press.

Newman, M. and D. Watts (2006). *The structure and dynamics of networks.* Princeton Univ Pr.

Nielsen, M. (2010). The Logic of Collective Action. Technical report.

Nielsen, M. (2011). *Reinventing discovery: the new era of networked science.* Princeton University Press.

Nowotny, H., P. Scott, and M. Gibbons (2001). *Re-thinking science: knowledge and the public in an age of uncertainty.* Polity Press.

Oliver, P. and G. Marwell (2001). Whatever happened to critical mass theory? A retrospective and assessment. *Sociological Theory 19*(3), 292–311.

Olson, M. (1974). *The logic of collective action: Public goods and the theory of groups.* Harvard University Press.

O'Mahony, S. and F. Ferraro (2007). The emergence of governance in an open source community. *Academy of Management Journal 50*(5), 1079–1106.

Ostrom, E. and C. Hess (2007). *Understanding knowledge as a commons: from theory to practice.* MIT Press.

Page, S. (2010). *Diversity and complexity.* Princeton Univ Pr.

Paul, R. J. and T. S. Chanev (1998). Simulation optimisation using a genetic algorithm. *Simulation Practice and Theory 6*(6), 601–611.

Pavard, B., J. Dugdale, N. Saoud, S. Darcy, and P. Salembier (2006). Design of robust socio-technical systems. In *Second Symposium on Resilience Engineering Proceedings, Juan-les-Pins, France, November*, pp. 8–10. Citeseer.

Pierreval, H. and L. Tautou (1997). Using evolutionary algorithms and simulation for the optimization of manufacturing systems - Springer. *IIE Transactions 29*(3), 181–189.

Pirolli, P. (2007, April). *Information foraging theory: Adaptive interaction with information.* Oxford University Press, USA.

Pirolli, P. and S. Card (1999). Information foraging. *Psychological review 106*(4), 643.

Powell, W. W., K. W. Koput, and L. Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly 41*(1), 116–145.

Preece, J. and B. Shneiderman (2009). The Reader-to-Leader Framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction 1*(1), 13–32.

Pullum, L. L. and O. Ozmen (2012, December). Early Results from Metamorphic Testing of Epidemiological Models. In *Workshop on Verification and Validation of Epidemiological Models in ASE International Conference on Biomedical Computing*, Washington DC.

Pyka, A. (2009). *Innovation networks: new approaches in modelling and analyzing.* Springer Verlag.

Rafols, I. and M. Meyer (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics 82*(2), 263–287.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *The Journal of Machine Learning Research 3*, 1371–1382.

Robinson, S., R. Brooks, K. Kotiadis, and D. Van Der Zee (2010). *Conceptual modeling for discrete-event simulation*. CRC Press, Inc. Boca Raton, FL, USA.

Rowley, T. (1997). Moving beyond dyadic ties: A network theory of stakeholder influences. *The academy of management review 22*(4), 887–910.

Sargent, R. (2005). Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation conference*, pp. 130–143.

Saviotti, P. (2009). Knowledge Networks: Structure and Dynamics. In *Innovation networks: new approaches in modelling and analyzing*. Springer Verlag.

Scacchi, W. and C. Jensen (2008). Governance in Open Source Software Development Projects: Towards a Model for Network-Centric Edge Organizations. In *13th International Command and Control Research and Technology Symposium*, Bellevue, WA.

Schank, T. and D. Wagner (2005). Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications 9*, 265–275.

Sexton, R. S., R. E. Dorsey, and J. D. Johnson (1999). Optimization of neural networks: A comparative analysis of the genetic algorithm and simulated annealing. *European Journal of Operational Research 114*(3), 589–601.

Sheard, S. and A. Mostashari (2008). A Framework for System Resilience Discussions. In *Proc Eighteenth Annu Int Symp INCOSE*.

Shlesinger, M. F. (2007, September). Complex Adaptive Systems: An Introduction to Computational Models of Social Life. *Journal of Statistical Physics 129*(2), 409–410.

Shrager, J. and P. Langley (1990). Computational approaches to scientific discovery. *Computational Models of Scientific Discovery and Theory Formation. Morgan Kaufmann*, 1–26.

Silverman, B. and G. K. Bharathy (2011). Modeling and Simulation Fundamentals Theoretical Underpinnings and Practical Domains.

Smith, A. and A. Stirling (2008). Social-ecological resilience and socio-technical transitions: critical issues for sustainability governance. Technical report.

Smith, R. E., S. Forrest, and A. S. Perelson (1993). Searching for diverse, cooperative populations with genetic algorithms. *Evolutionary Computation 1*(2), 127–149.

Standish, R. K. (2008, May). Concept and Definition of Complexity. *ARXIV eprint*.

Stirling, A. (2007, February). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface 4*(15), 707–719.

Swisher, J. R., P. D. Hyden, S. H. Jacobson, and L. W. Schruben (2000). A survey of simulation optimization techniques and procedures. In *Simulation Conference, 2000. Proceedings. Winter*, pp. 119–128. IEEE.

Thagard, P. (1989, April). Explanatory coherence. *Behavioral and Brain Sciences 12*, 435–502.

The FANTOM Consortium (2005, September). The Transcriptional Landscape of the Mammalian Genome. *Science 309*(5740), 1559–1563.

Tompkins, G. and F. Azadivar (1995). Genetic algorithms in optimizing simulated systems. In *the 27th conference*, New York, New York, USA, pp. 757–762. ACM Press.

Udehn, L. (1993). Twenty-five years with the logic of collective action. *Acta Sociologica 36*(3), 239.

Uzzi, B. and J. Spiro (2005). Collaboration and Creativity: The Small World Problem. *ajs 111*(2), 447–504.

Van Aardt, A. (2004). Open Source Software development as a Complex Adaptive System: Survival of the fittest? In *Paper delivered at The 17th Annual Conference of the NACCQ. Christchurch, New Zealand, 6th-9th July.* Citeseer.

Wagner, C. (2008). *The new invisible college: Science for development.* Brookings Inst Pr.

Walker, B., C. Holling, S. Carpenter, and A. Kinzig (2004). Resilience, Adaptability and Transformability in Social–ecological Systems. *Ecology and society 9*(2), 5.

Wasserman, S. (1994a). Social Network Analysis. *Sociology The Journal Of The British Sociological Association 22*(1), 109–127.

Wasserman, S. (1994b). *Social network analysis: Methods and applications.* Cambridge university press.

Watts, D. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology 105*, 493–527.

Wood, M. (2009). The pros and cons of using pros and cons for multi-criteria evaluation and decision making.

Wu, S., Y. Bi, X. Zeng, and L. Han (2009, July). Assigning appropriate weights for the linear combination data fusion method in information retrieval. *Information Processing and Management 45*(4), 413–426.

Wynn, D. (2003). Organizational structure of open source projects: A life cycle approach. In *7th Annual Conference of the Southern Association for Information Systems, Georgia*.

Xie, X., J. W. K. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen (2011). Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software 84*(4), 544–558.

Yam, B. (2005). *Making Things Work: Solving Complex Problems in a Complex World*. NECSI Knowledge Press.

Yang, A. and Y. Shan (2008). *Intelligent complex adaptive systems*. IGI Publishing Hershey, PA, USA.

Yilmaz, L. (2006). Validation and verification of social processes within agent-based computational organization models. *Computational and Mathematical Organization Theory 12*(4), 283–312.

Yilmaz, L. (2008a). Innovation systems are self-organizing complex adaptive systems. In *Association for the Advancement of Artificial Intelligence*.

Yilmaz, L. (2008b). Project Proposal - NSF-SBE-0830261.

Yilmaz, L. (2009). On the synergy of conflict and collective creativity in open innovation socio-technical ecologies. In *Proceedings of the 2008 Winter Simulation Conference*.

Yilmaz, L. and A. Hunt (2010, January). Computational Discovery - Project Description.

Zeigler, B. P., Y. Moon, D. Kim, and G. Ball (1997, February). The DEVS environment for high-performance modeling and simulation. *Computational Science Engineering, IEEE 4*(3), 61–71.

Zou, G. (2012). *ColorScape: A Creative Artificial Ecosystem Model of Communication and Collective Creativity in Global Participatory Science*. Ph. D. thesis, Auburn University.

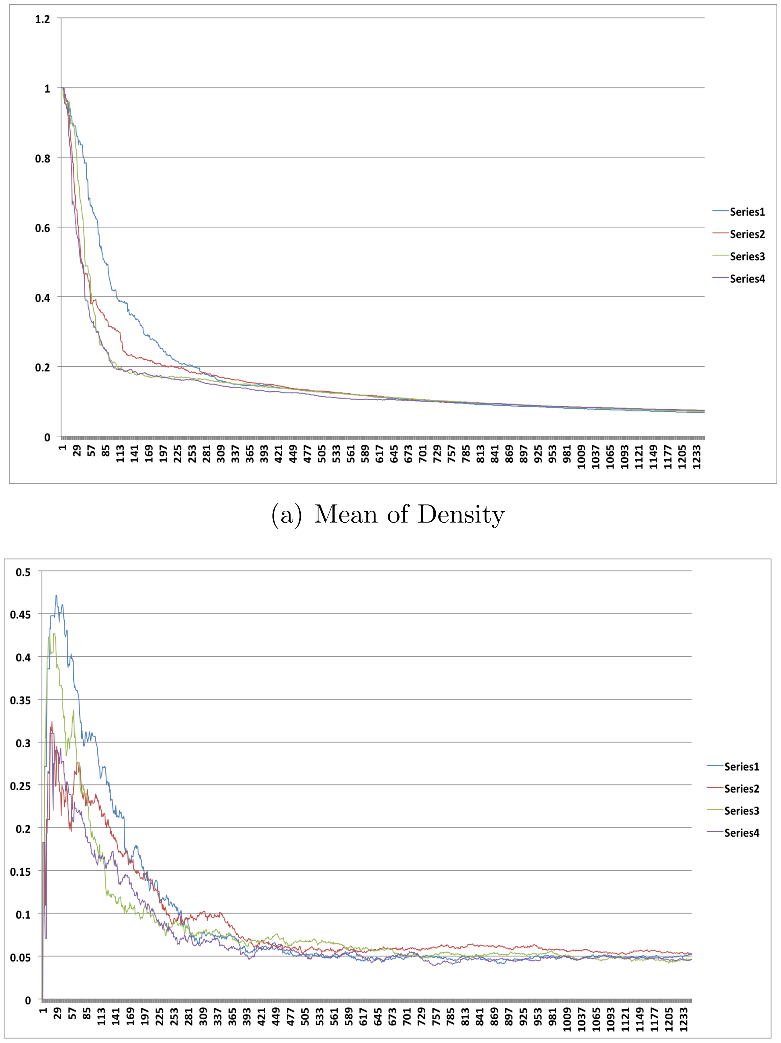
Zou, G. and L. Yilmaz (2011). Dynamics of knowledge creation in global participatory science communities: open innovation communities from a network perspective. *Computational and Mathematical Organization Theory 17*(1), 35–58.

Appendix A

A.1 Termination State Analysis

The snapshots of time-series plots for different performance metrics are illustrated below. The means and standard deviations for each metric are observed to support terminating state decision. The detailed information about the calculation process of *diversity* metrics is provided in Chapter 5. The first set of snapshots are taken for four different scenarios, which are assumed to represent the different patterns encountered in the analysis (only one representative of qualitatively similar patterns are plotted for readability). During the runs, scientists are socially connected when they contribute on the same artifact at the same time-tick (OBO assumption). Also, *recovered* scientists are not removed from the social network.

Figure A.1: Density Over Time - OBO Scenarios

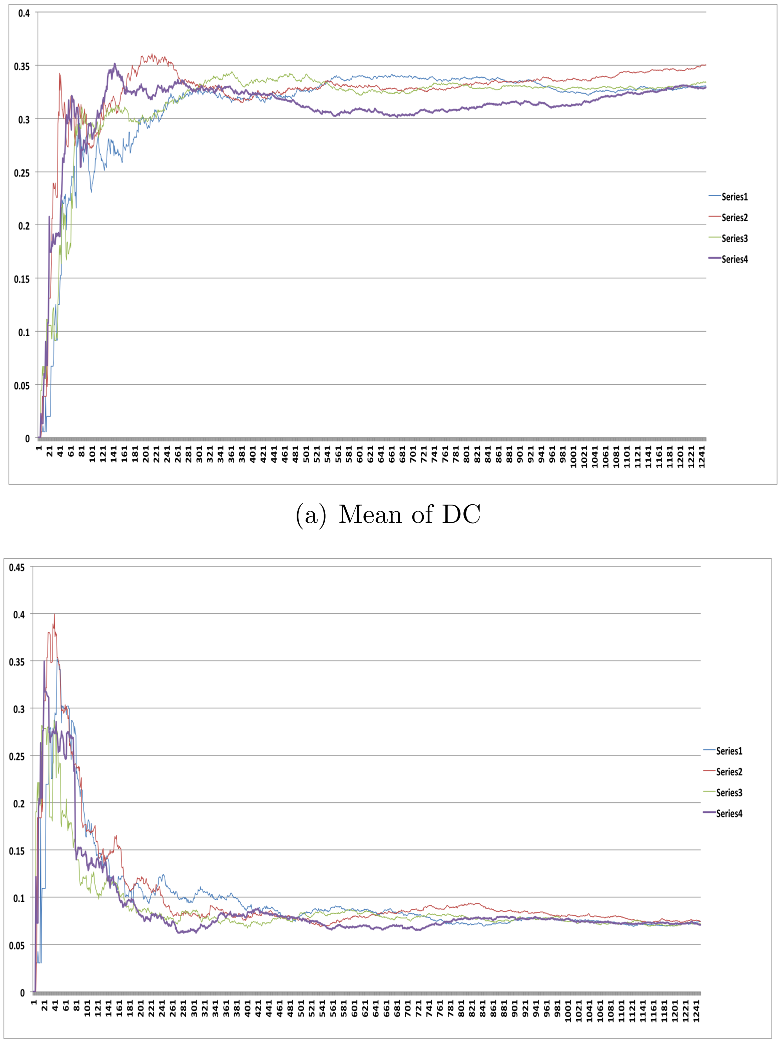


(a) Mean of Density

(b) Standard Deviation of Density

Series-1 is Opt. Foraging/Low initial population, Series-2 is Opt. Foraging/Moderate initial population, Series-3 is Basic Foraging/Low initial population, Series-4 is Basic Foraging/High initial population

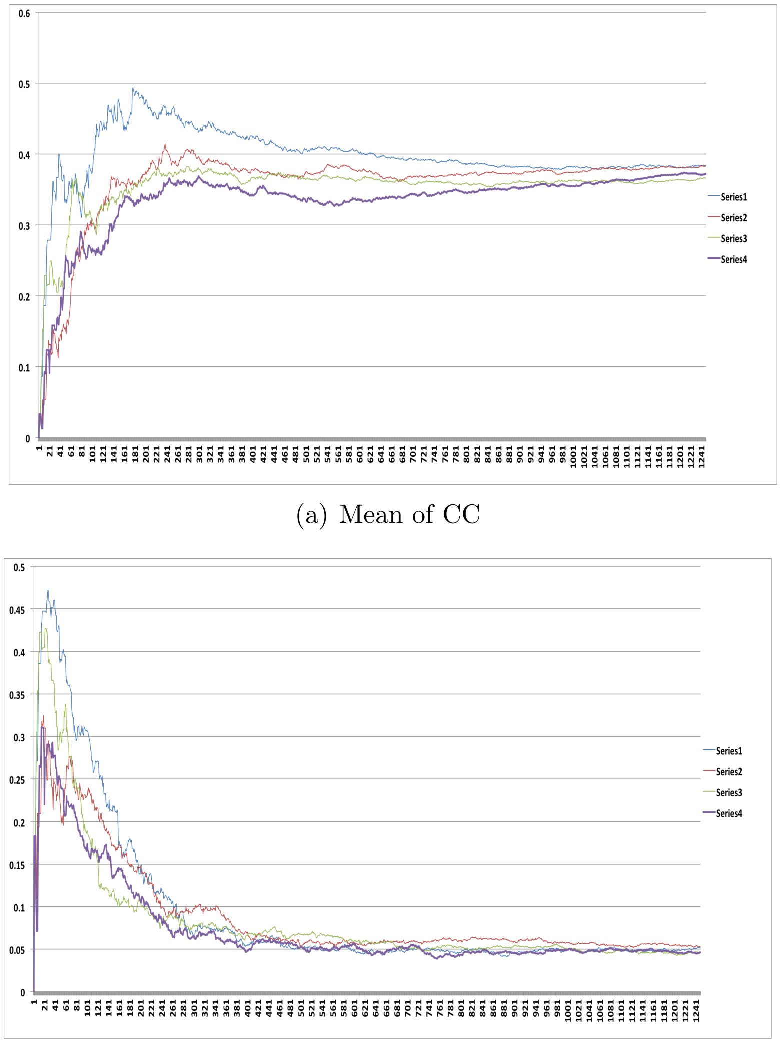
Figure A.2: Degree Centrality Over Time - OBO Scenarios



(a) Mean of DC

(b) Standard Deviation of DC

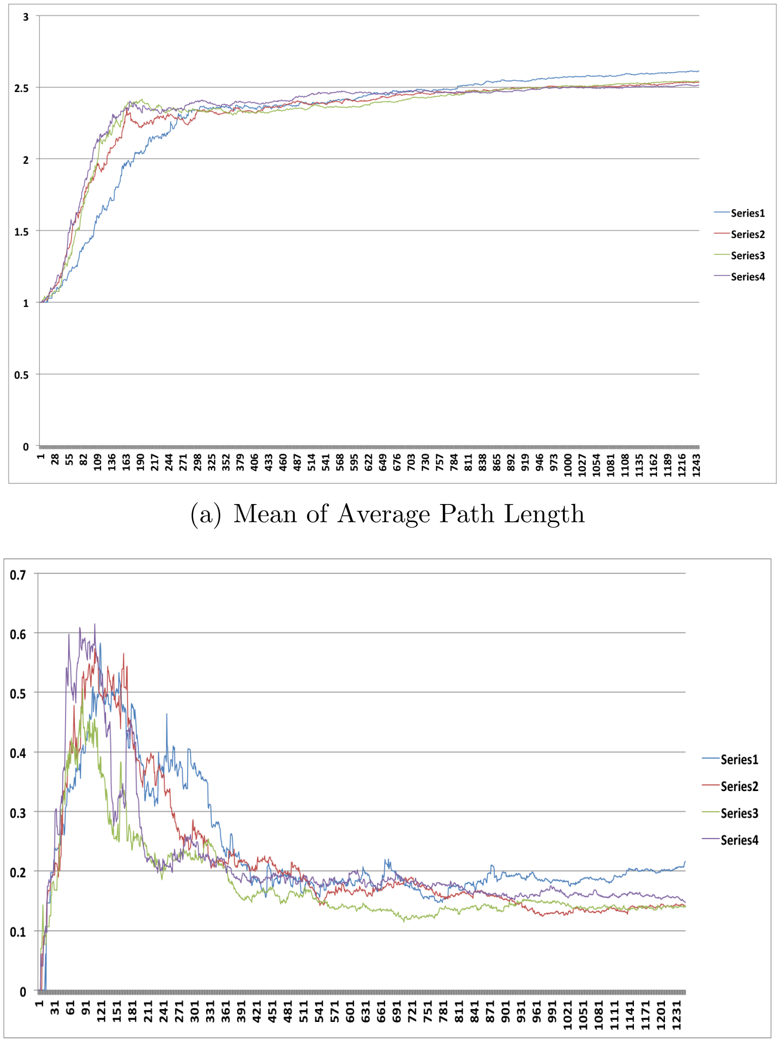
Figure A.3: Clustering Coefficient Over Time - OBO Scenarios



(a) Mean of CC

(b) Standard Deviation of CC

Figure A.4: Average Path Length Over Time - OBO Scenarios

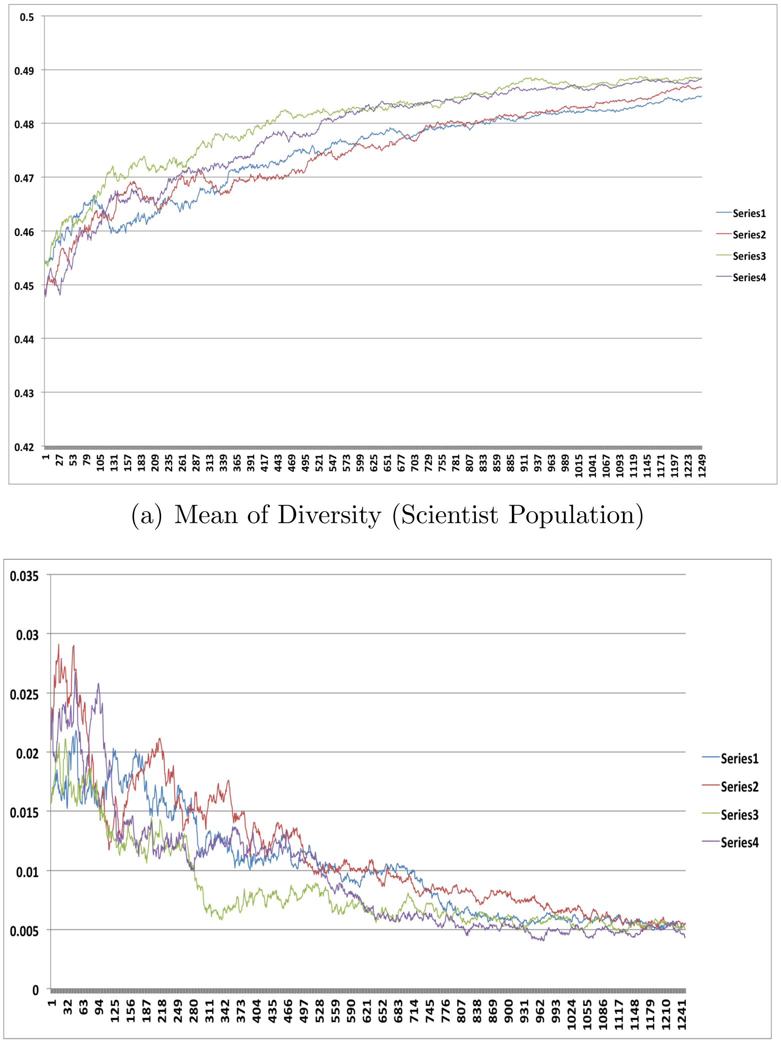


(a) Mean of Average Path Length

(b) Standard Deviation of Average Path Length

The scenarios at which scientists are connected randomly are illustrated in the following plots, and their mechanisms are described in Chapter 5. In this case, *Recovered* scientists are removed from the social network. That is why in some scenarios, after 500 time-ticks, the network starts to dissolve (as central members leave the environment). Compared to OBO scenarios, relatively high variability is observed in social network metrics. Four scenarios out of 40 scenarios are selected, which present different patterns.

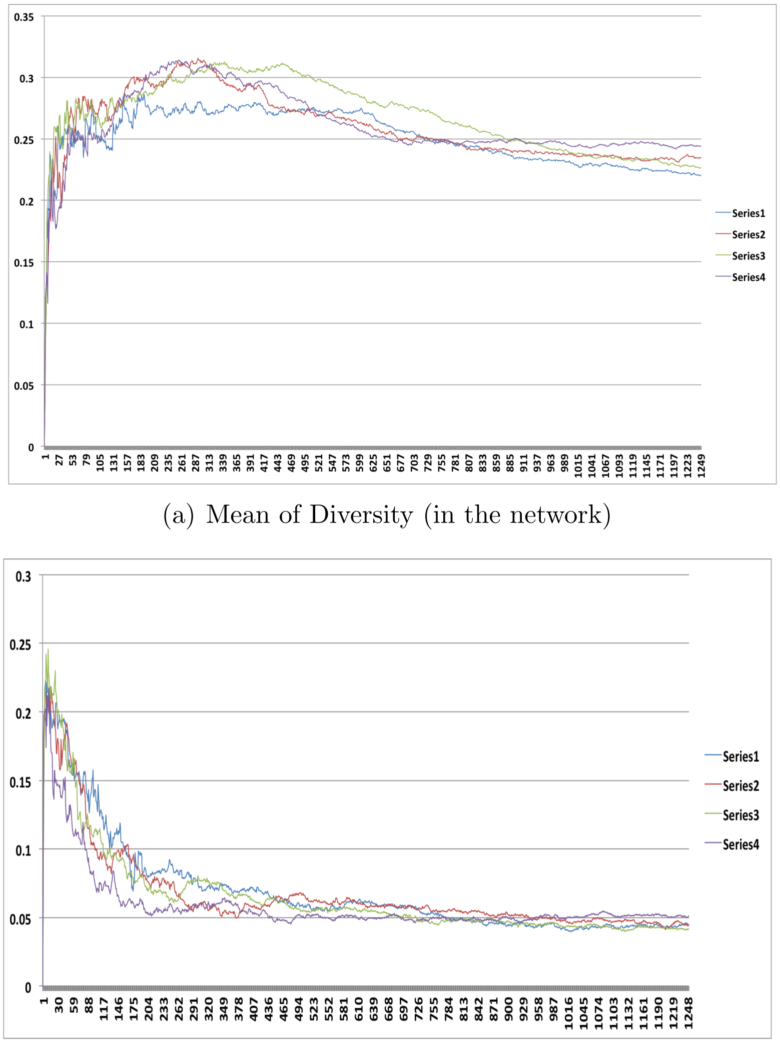
Figure A.5: Diversity (Scientist Population) Over Time - OBO Scenarios



(a) Mean of Diversity (Scientist Population)

(b) Standard Deviation of Diversity (Scientist Population)

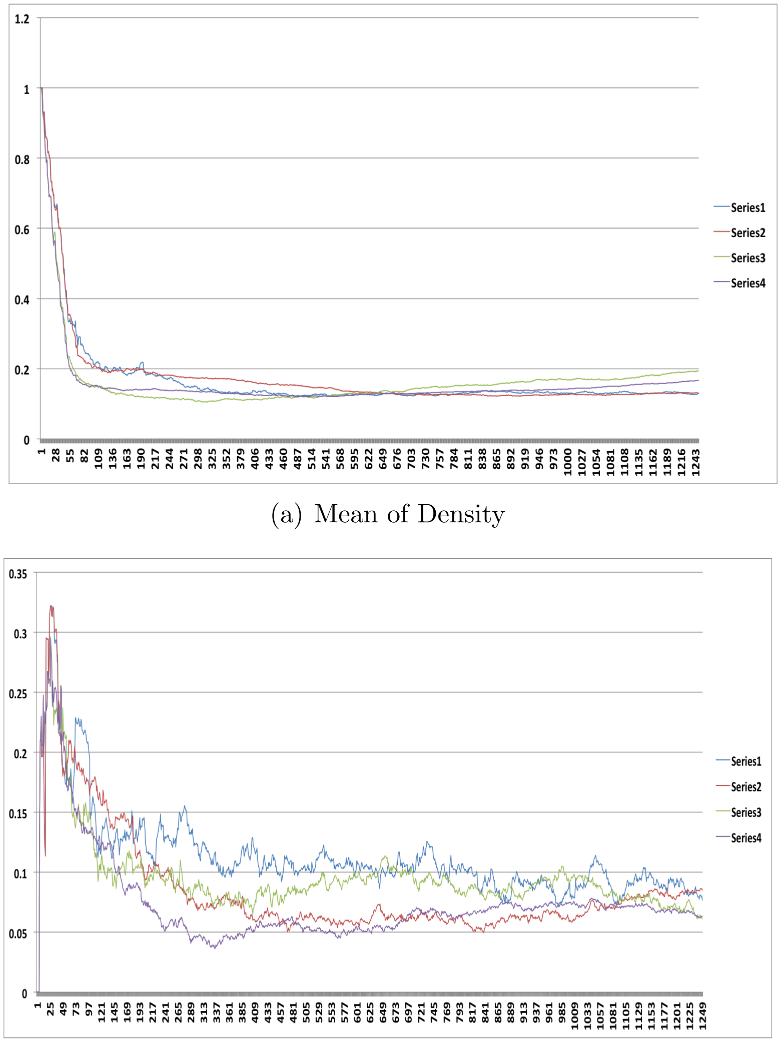
Figure A.6: Diversity (in the network) Over Time - OBO Scenarios



(a) Mean of Diversity (in the network)

(b) Standard Deviation of Diversity (in the network)

Figure A.7: Density Over Time - Random Connection

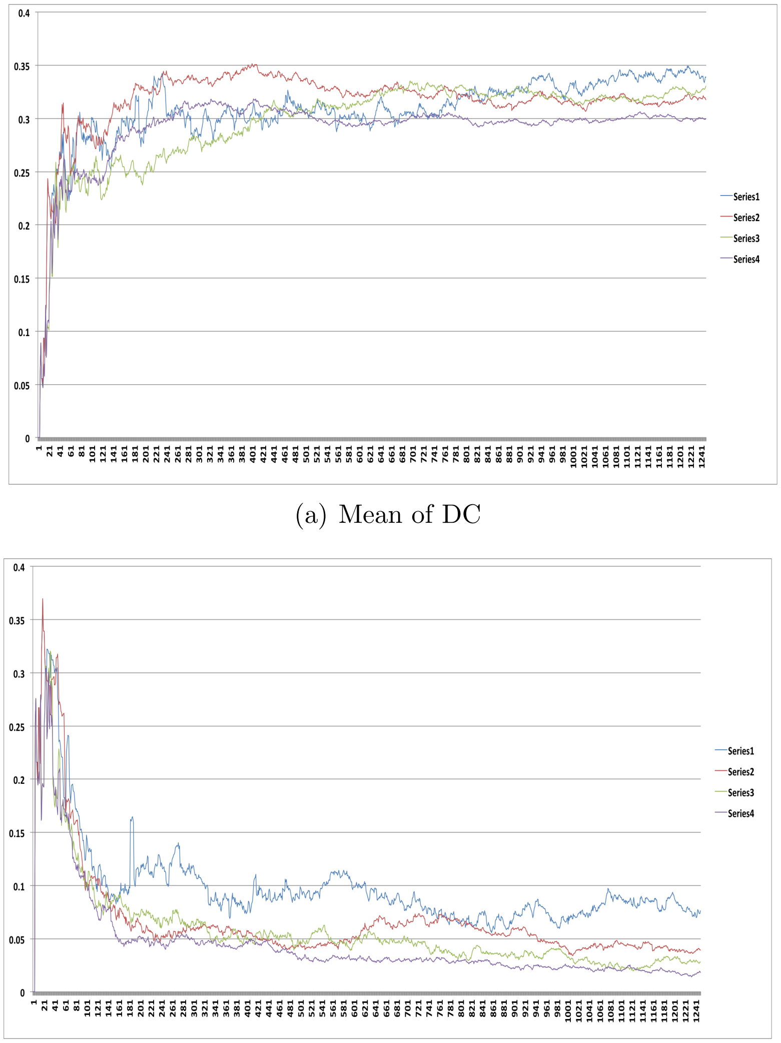


(a) Mean of Density

(b) Standard Deviation of Density

Series-1 is Low Arrival Rate/High Turnover rate, Series-2 is Low Arrival Rate/Low Turnover rate, Series-3 is High Arrival Rate/High Turnover rate, and Series-4 is High Arrival Rate/Low Turnover rate

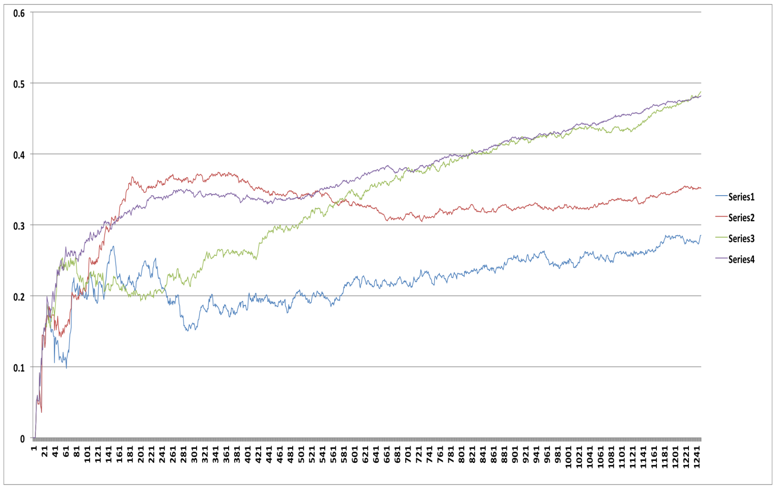
Figure A.8: Degree Centrality Over Time - Random Connection



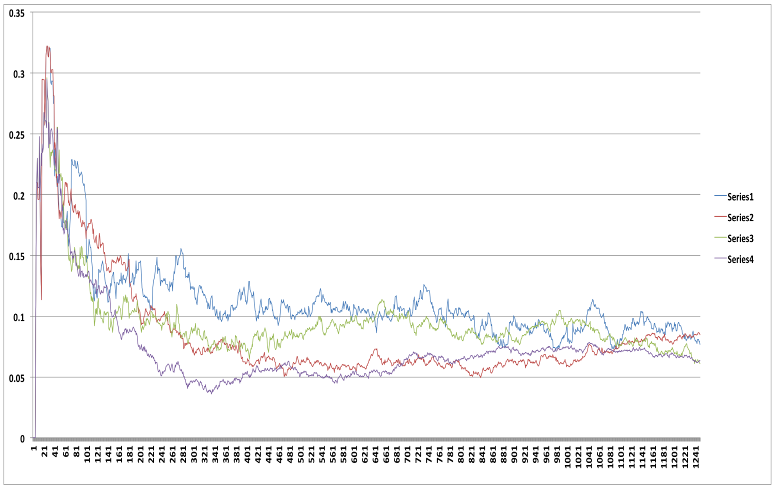
(a) Mean of DC

(b) Standard Deviation of DC

Figure A.9: Clustering Coefficient Over Time - Random Connection

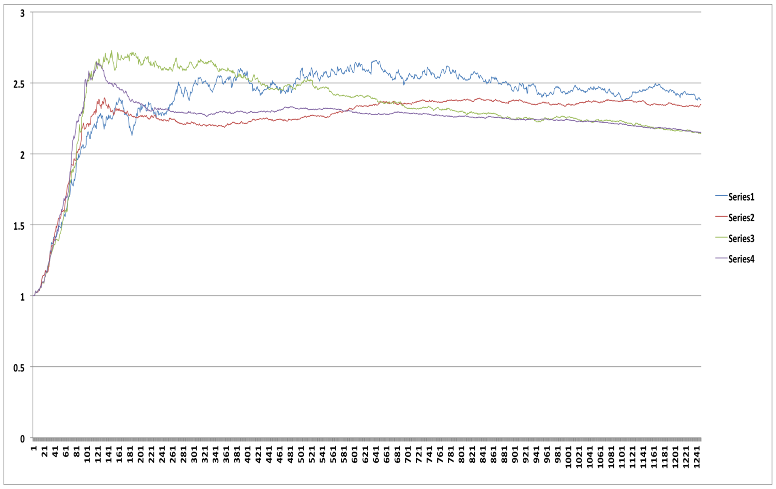


(a) Mean of CC

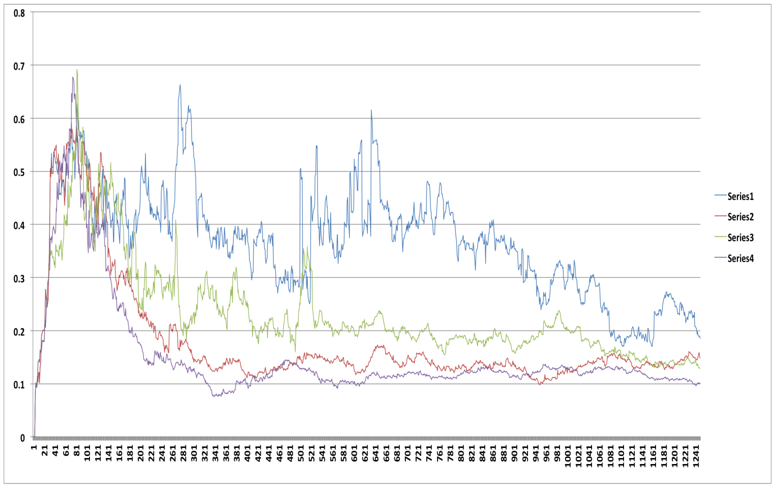


(b) Standard Deviation of CC

Figure A.10: Average Path Length Over Time - Random Connection

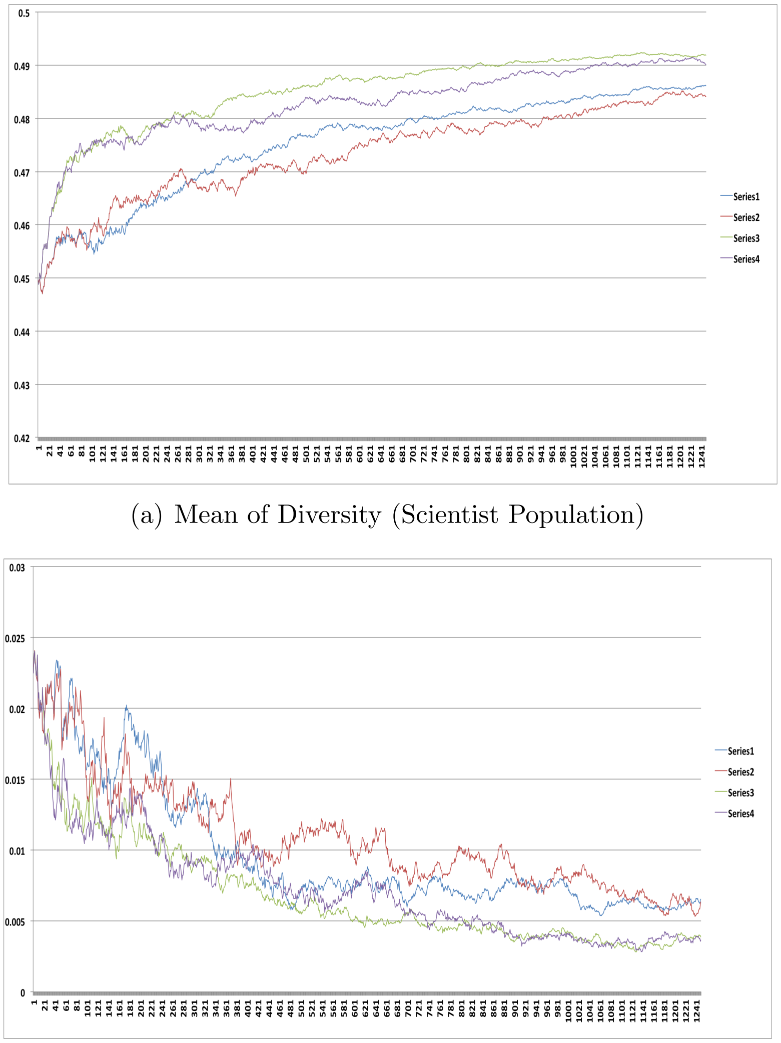


(a) Mean of Average Path Length



(b) Standard Deviation of Average Path Length

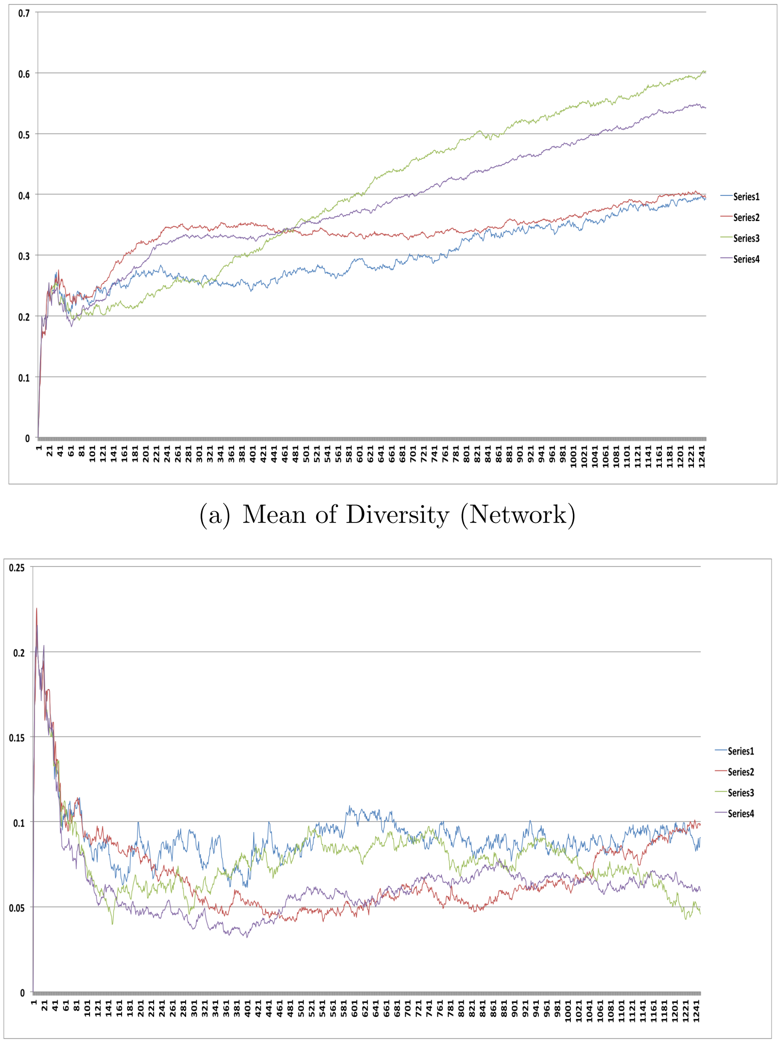
Figure A.11: Diversity (Scientist Population) Over Time - Random Connection



(a) Mean of Diversity (Scientist Population)

(b) Standard Deviation of Diversity (Scientist Population)

Figure A.12: Diversity (in the network) Over Time - Random Connection



(a) Mean of Diversity (Network)

(b) Standard Deviation of Diversity (in the network)

A.2 Response Surface Analysis

Table A.1 represents the variables identified for response surface analysis and their respective values.

Table A.1: Parameter Values for RSM

Parameter	**Scenarios**
Communication Type(Theory)	[1,...,5]
Expertise Level to Recover	[0.7,0.9]
Probability to Leave	[0.1,0.2]
Minimum time Expectation	[1,5]
Arrival Rate	[0.1,0.2]
Migration Threshold	[3,5]
Recover Rate	[0.1,0.2,0.5]
Forage Extension	[2,3,5]
Minimum Tension	[0.1,0.5,1]
Minimum Burden	[0.1,0.5,1]
Maximum Altruism	[0.1,0.5,1]

The following Table A.2 and Table A.3 represent response surface analysis results.

Table A.2: Summary of Response Surface Analysis - 1

Inputs	CC			DC			Avg_Path			CP			Density		
	β	**t**	**p**	β	**t**	**p**	β	**t**	**p**	β	**t**	**p**	β	**t**	**p**
(Constant)	-0.076	-4.572	0.001	0.446	23.292	0.001	0.5	26.711	0.001	1.675	14.087	0.001	0.471	660.381	0
maxAltruism	1.004	43.01	0.001	0.934	24.698	0.001	0.921	22.558	0.001	-0.666	-10.153	0.001	-0.811	-54.058	0.001
MinBurden_ComType	0.051	1.852	**0.065**	0.107	2.451	0.015	0.155	3.331	0.001	-0.234	-3.116	0.002	0	0	0
migrationThreshold	0.169	15.616	0.001	0	0	0	0.124	4.721	0.001	-0.11	-3.742	0.001	-0.091	-9.421	0.001
Recover_ComType	0.045	1.543	**0.124**	0	0	0	0	0	0	0	0	0	0	0	0
expertiseToRecover	0.111	10.446	0.001	-0.049	-4.071	0.005	-0.29	-15.706	0.001	0.164	5.663	0.001	0.041	4.228	0.001
minTension	-0.167	-10.243	0.001	-0.187	-4.385	0.001	-0.153	-5.381	0.001	0.317	7.112	0.001	0.142	15.407	0.001
minBurden	**-0.387**	-14.351	0.001	**-0.517**	-12.119	0.001	**-0.433**	-9.332	0.001	0.421	5.679	0.001	**0.377**	40.984	0.001
Altruism_ComType	-0.233	-8.532	0.001	-0.254	-5.863	0.001	-0.311	-6.693	0.001	0.359	4.782	0.001	-0.03	-1.967	**0.05**
arrivalRate	0.162	11.456	0.001	-0.088	-4.446	0.001	-0.253	-13.256	0.001	0.09	2.6	0.01	0.259	24.37	0.001
recoverRate	-0.203	-8.345	0.001	-0.054	-3.244	0.002	-0.113	-6.216	0.001	0.113	3.969	0.001	0.086	9.223	0.001
ArrivalRate_ComType	-0.175	-6.405	0.001	0	0	0	0	0	0	0.128	2.105	0.036	0	0	0
MinBurden_MinTension	0.088	4.281	0.001	0.1	3.075	0.003	0.081	2.236	0.026	**-0.653**	-11.569	0.001	0	0	0
probToLeave	-0.034	-4.014	0.006	-0.083	-4.162	0.001	0	0	0	0	0	0	0.024	2.233	0.026
MinTension_ComType	0	0	0	0.048	1.094	**0.275**	0	0	0	0	0	0	0	0	0
ForageExt_ComType	0	0	0	0.069	1.078	**0.282**	0	0	0	0	0	0	0	0	0
Foraging_ComType	0	0	0	-0.311	-4.603	0.001	0	0	0	0	0	0	0	0	0
forageExtension	0	0	0	-0.097	-2.462	0.015	0	0	0	0	0	0	0	0	0
minTimeExpectation	0	0	0	-0.05	-2.498	0.013	0	0	0	0	0	0	0	0	0
Migration_ComType	0	0	0	0.073	2.204	0.028	-0.081	-3.059	0.04	0	0	0	0	0	0

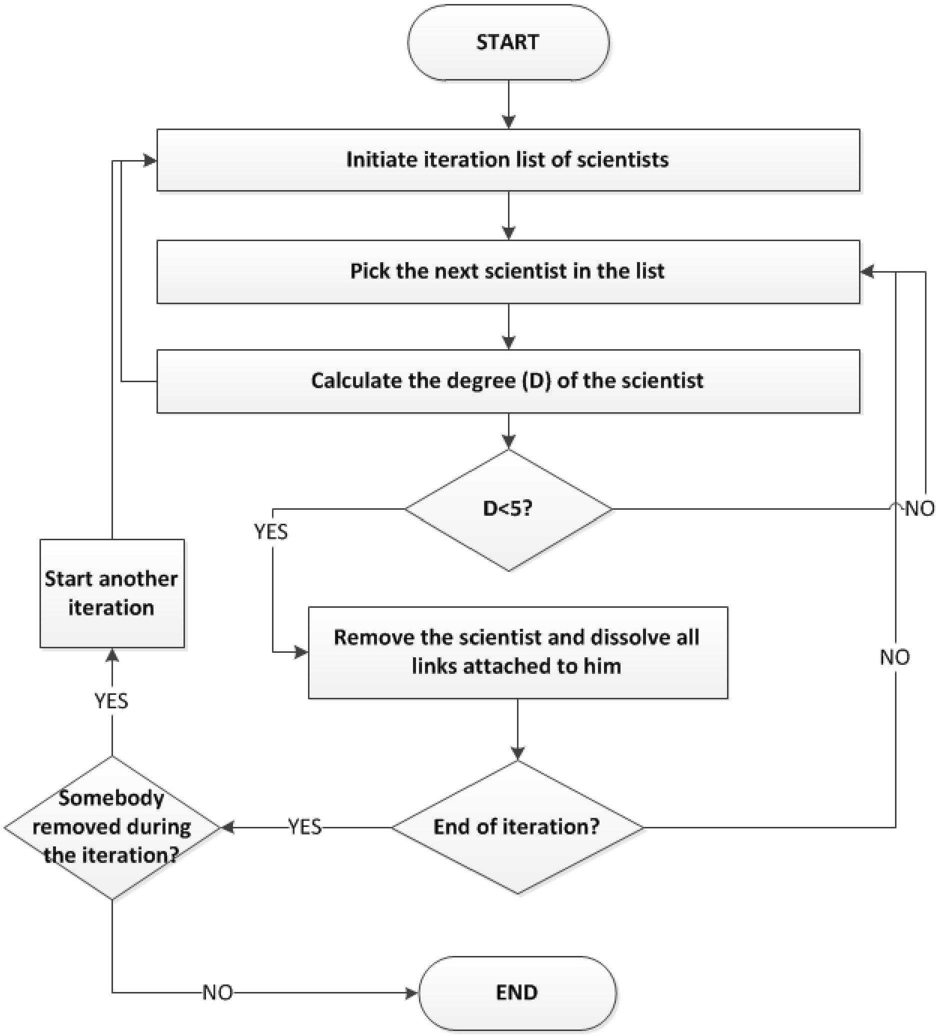
Table A.3: Summary of Response Surface Analysis - 2

Inputs	DiversityS			DiversityA			DiversityN			DiversityL		
	β	**t**	**p**	β	**t**	**p**	β	**t**	**p**	β	**t**	**p**
Constant	0.492	572.778	0	0.471	660.381	0	0.673	27.172	0.001	0.282	14.211	0.001
maxAltruism	-0.745	-43.713	0.001	-0.811	-54.058	0.001	0.755	18.794	0.001	1.042	31.421	0.001
MinBurden_ComType	0	0	0	0	0	0	0	0	0	0.102	2.662	0.008
migrationThreshold	-0.058	-5.297	0.001	-0.091	-9.421	0.001	0	0	0	0.134	9.154	0.001
Recover_ComType	0	0	0	0	0	0	0	0	0	0	0	0
expertiseToRecover	-0.274	-25.525	0.001	0.041	4.228	0.001	-0.063	-3.009	0.003	0	0	0
minTension	0.145	13.818	0.001	0.142	15.407	0.001	-0.155	-4.866	0.001	-0.164	-7.314	0.001
minBurden	**0.429**	41.026	0.001	**0.377**	40.984	0.001	**-0.485**	-15.279	0.001	**-0.487**	-13.005	0.001
Altruism_ComType	-0.051	-2.964	0.004	-0.03	-1.967	**0.05**	-0.174	-3.934	0.001	-0.293	-7.702	0.001
arrivalRate	0.191	17.279	0.001	0.259	24.37	0.001	-0.193	-7.972	0.001	0	0	0
recoverRate	0.095	8.975	0.001	0.086	9.223	0.001	0	0	0	-0.157	-11.002	0.001
ArrivalRate_ComType	0	0	0	0	0	0	0	0	0	0	0	0
MinBurden_MinTension	0	0	0	0	0	0	0.13	3.229	0.002	0.157	5.534	0.001
probToLeave	0	0	0	0.024	2.233	0.026	-0.088	-3.61	0.001	-0.061	-4.125	0.001
MinTension_ComType	0	0	0	0	0	0	0	0	0	0	0	0
ForageExt_ComType	0	0	0	0	0	0	-0.085	-3.006	0.003	0	0	0
Foraging_ComType	0	0	0	0	0	0	0	0	0	-0.09	-3.004	0.003
forageExtension	0	0	0	0	0	0	0	0	0	0	0	0
minTimeExpectation	0	0	0	0	0	0	-0.088	-3.617	0.001	0	0	0
Migration_ComType	0	0	0	0	0	0	0	0	0	0	0	0

A.3 Core/Periphery Calculation Method

In Figure A.13, the activity diagram of calculation method of the core/periphery metric is represented.

Figure A.13: Core/Periphery Activity Diagram


START
Initiate iteration list of scientists
Pick the next scientist in the list
Calculate the degree (D) of the scientist
D<5?
YES
NO
Remove the scientist and dissolve all links attached to him
NO
End of iteration?
YES
Somebody removed during the iteration?
YES
Start another iteration
NO
END