

EXPLORATORY SIMULATION OF COLLECTIVE INNOVATIVE BEHAVIOR IN GLOBAL PARTICIPATORY SCIENCE COMMUNITIES

Guangyu Zou

Auburn University
3101 Shelby Center
Auburn, AL 36849, USA

Levent Yilmaz

Auburn University
3116 Shelby Center
Auburn, AL 36849, USA

ABSTRACT

Better understanding of how and why networks of open innovation and global participatory science communities form and evolve, and how they can be governed or influenced toward sustainable innovation and productive states are critical questions. To this end, a simulation-based exploratory study is conducted to better understand the conditions that confer increased rates of innovation in such socio-technical systems. Three types of open science communities are identified and simulated using agent simulation as a method of inquiry. Simulation results show that centrality, as a measure of degree of connectedness, exhibits positive influence for innovation output in exploratory and service communities up to a point. Also, utility-oriented communities have social network structures with low density and high centrality, suggesting high potential for innovation.

1 INTRODUCTION

A diverse set of factors are driving the globalization of science. Factors that catalyze these trends include the emergence of new information architectures and laboratories that enable the emergence of a new mode of collaboration among scientists distributed over the globe to share and co-develop knowledge over the cyber-infrastructure. The practice of science is more open and global, as the access to knowledge, as well as its production is becoming increasingly transparent. Service oriented science (Foster 2005) and e-Science initiatives lead to scientific communities, where shared domain knowledge is no longer solely documented in the journal articles or patents, but is also documented in software, simulations, and databases that represent an evolving collective knowledge-base that is governed and maintained by community members. Just like open source software communities, "SourceForge for science" style in scientific production and collaboration provide the requisite infrastructure that encompass community membership services, catalogs, storage services, and workflow orchestration service. We denote this style that revolves around open and distributed communities of scientific practice as *Global Participatory Science* (GPS).

GPS is a socio-technical practice. Scientists and engineers progress through an orientation phase, during which they become familiar with the norms and governance mechanisms of the community of practice. Recently a number of virtual scientific laboratories emerged and continue to successfully bring together scientists over the globe to collaborate to not only share and aggregate data, but also create new knowledge (Smith and Ashburner 2007, NanoHUB 2009). For instance, OBO Foundry (Smith and Ashburner 2007) is collaboration among a group of communities that are active in developing ontologies to standardize data acquisition and use in the health sciences community. As laboratories over the cyberinfrastructure become sophisticated in terms of capabilities that support remote access, collaboration, and cooperative activity management, virtual organizations as open science socio-technical systems are becoming prominent and increasingly central to science and engineering projects. To harness their full potential and the promise they offer in learning, discovery, and innovation, it is critical to understand what structural and behavioral conditions confer sustained innovation and facilitate enhancing scientific and engineering production in such communities.

Using agent simulation, we examine three types of communities (i.e., exploratory, service, utility) and observe emergent socio-technical networks to determine relations between emergent network structures and innovation output. Specifically, we are interested in exploring the following questions:

what is the impact of alternative scientific community cultures and leadership styles on the sustainability and innovation potential of GPS?

Network metrics such as centrality, density, and clustering coefficient are used to determine if they are related to innovation output under alternative community governance mechanisms. Simulation results indicate that centrality, as a measure of integration, positively influences innovation up to a point. That is communities that generate high degrees of innovation output tends to have more focal nodes in their network structure than communities that have low or moderate degrees of innovation output. Social networks in utility-oriented community exhibit low density and high centrality, which suggest high potential for innovation. Also, utility-oriented communities exhibit large degree of variation in terms of average knowledge fitness defined in terms of knowledge utility and embeddedness.

2 BACKGROUND

2.1 Open Science Communities

The loss of richness in communication, diversity in membership, and increased mobility in open science communities pose unique challenges. Such challenges involved in sustaining scientific laboratories are examined in (Bos, Zimmerman, Olson, Yew, Yerkie, Dahl, and Olson 2003). Traditions of scientific independence, difficulties in sharing knowledge, and formal organizational barriers are presented among the challenges that influence the sustainability of such communities. On the other hand, recently a number of virtual scientific laboratories emerged and continue to successfully bring together scientists over the globe to collaborate to not only share and aggregate data, but also create new knowledge. The following are among such growing and active open science laboratories and innovation networks.

- OBO Foundry- Open Biomedical Ontologies
- NanoHUB - Simulation, Education, Technology for Nano Technology
- NEES Grid - Network for Earthquake Engineering Cyberinfrastructure
- CABIG - Cancer Biomedical Informatics Grid

OBO Foundry (Smith and Ashburner 2007) comprises over 60 ontologies, and its role as an ontology information resource is supported by the NIH Roadmap National Center for Biomedical Ontology (NCBO). OBO Foundry serves as a hub for a network of communities to foster interoperability and to align and integrate their efforts. Some of the participating communities such as Open Biomedical Investigations (OBI), are international, collaborative efforts to build artifacts for annotation of Biomedical Investigations. NanoHub (NanoHUB 2009) is a network comprised of scientists, engineers, and educators that share and exchange resources related to Nano technology. NanoHub is based on an NSF-funded initiative that aims to establish a virtual organization and network for computational nanotechnology. Shared resources include simulations, learning modules such as course materials, publications, presentations, and tools. Similarly NEESGrid, which is an active consortium of earthquake engineering centers, institutes, researchers, and practitioners in United States and around the world and is the result of a collaboration effort led by the National Center for Supercomputing Applications (NCSA).

2.2 Computational Models of Science and Innovation

Earlier studies pertaining to the application of computational models to scientific discovery processes focus on simulating cognitive processes and re-enact discoveries (Klahr and Simon 1999). Specifically, computational modeling of concept formation is viewed as central to discovery and has a long history (Hovland and Hunt 1960).

Although significant research has been conducted on social aspects of scientific communities, simulation modeling of such communities is rare. One notable example is (Gilbert 1997), where the citation patterns and growth of knowledge are simulated to exhibit empirical regularities observed in scientific communities. Yet, this study does not aim to consider social processes pertaining to enculturation and innovation. On the other hand, the simulation study presented in (Edmonds 2007) views scientific discovery as a social process. However, unlike the model presented in this article, its underlying generative process does not take interactions between agents (i.e., scientists) into account. In the context of innovation, the use of simulation of collective invention and innovation diffusion (Cowan and Jonard 2004) revealed the significance of social network structure in knowledge creation and diffusion. Building on these earlier studies, the model introduced herein (1) explicitly specifies the underlying generative mechanisms related social dynamics of knowledge creation and (2) examines the implications of these mechanisms within the context of emergent virtual forms of scientific communities.

3 A COMPUTATIONAL MODEL OF GLOBAL PARTICIPATORY INNOVATION COMMUNITIES

Following the basic tenets of systems model of creativity (Csikszentmihalyi 1999), Figure 1 presents components of the conceptual model. *Individuals* are scientists that operate on knowledge structures that are generated by *creation*, *combination*, and *elaboration* operations (Ward, Smith, and Vaid 1997).

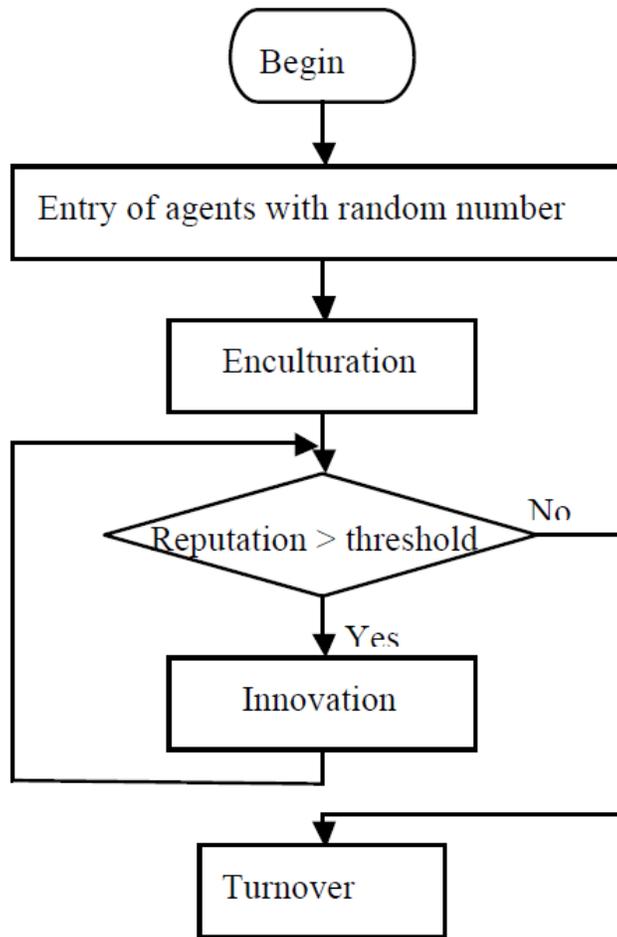


Figure 2: Life Cycle of an Agent

3.2 Innovation via Kene Generation

Evolutionary models of innovation suggest conceptual combination, elaboration, metaphor, and analogy as the major processes that operate on concept structures to produce outcomes that are novel but nevertheless rooted in existing knowledge. The innovation process in our model mimics concept creation, combination, and elaboration operators to generate kene structures, where each kene is represented by two vectors of binary bits. Vectors represent x and y dimensions that are used to spatially locate kenes in the knowledge space.

The kene creation operator generates a random vector in each dimension and measures its Hamming Distance from the reference kene that is located in the center of the domain. Kene combination uses two or more kenes and selects subsets from each kene that are shuffled into a new kene that combines traits from selected source kenes. Kene elaboration selects a reference kene and elaborates by randomly changing a number of bits to create a new kene that incrementally updates the source kene. The location of a new kene is computed based on its Hamming Distance from the source or reference kene on each dimension. In the case of combination, each source kene pulls the new kene to its location in the knowledge space through a weighted formula, where the weight is decided in terms of relative number bits that a source kene provides to create the new one.

3.3 Principle of Kene Selection

Kene elaboration and combination operations require selecting kenes on which new knowledge depends. This is similar to citation mechanism in traditional scientific communities. In open science communities, this can be interpreted as the mechanism for selecting open source knowledge on which to build new services. Kenes with higher fitness are cited (used) more frequently than those with lower fitness. In order to implement this strategy, which is known as epistemic

contagion (Crane 1972) we define the probability of selecting a kene using roulette wheel selection (Holland 1995). Assume that the fitness of kene i is f_i and the total number of kenes is N . Then the probability that kene i will be chosen is

$$p_i = \frac{f_i}{\sum_{j=1}^N f_j} \quad (5)$$

3.4 Kene Evaluation

Contributed kenes are selected by evaluators to decide whether they should be retained for inclusion in the domain. The fitness of a kene is based on two factors: its quality (novelty) and relational fitness. Relational fitness measures the extent to which the new kene is rooted in the existing knowledge and used by others. If the total fitness of a kene is less than a particular threshold, it will be discarded. On the contrary the kene with a fitness greater than the threshold will be retained in the domain to be reused by other individuals. The novelty of a kene is assigned randomly, while relational fitness is computed as follows:

$$R = w_{out}N_{out} + w_{in}N_{in}, \quad (6)$$

where $w_{out} + w_{in} = 1$. N_{out} and N_{in} depict the out and in degree of the kene, denoting the number of kenes that it uses and is used by, respectively. Overall fitness F of the kene is defined as a function of relational fitness R and quality Q :

$$F = aR + (1 - a)Q, \quad (7)$$

where a indicates sensitivity of the evaluator to relational fitness (usefulness) as opposed to novelty.

3.5 Turnover

If the motivation of an agent is less than the exit threshold, it leaves the community or transfers to another community in which it has more potential to increase its motivation and make contributions. As discussed in section 5, exploration, utility, and service-oriented communities have varying degrees of turnover rates. Service-oriented community has the maximum threshold, utility-oriented has a moderate value and exploration-oriented community has the minimum threshold. The higher the threshold for turnover is, the easier an agent will leave the community.

3.6 Generation of the Socio-technical Network

The socio-technical network of a community is generated through the kene generation and evaluation processes described above. A socio-technical network is comprised of nodes that denote kenes and members as well as links that specify relations among them.

- **Relation among kenes:** The relationship between kenes indicates that a kene is elaborated or combined with others.
- **Relation among individuals:** The producer of new a kene is related to owner of the kene that is used to derive the new kene. This relation captures the influence between individuals.
- **Relation between kenes and individuals:** That an individual is related to the kene indicates the individual is the producer (i.e., generator) of the kene.

Figure 3 depicts snapshot of a generated socio-technical network leveraging these relations. Green nodes are kenes and red nodes are individuals. Purple lines represent the relationship who creates which kenes. Blue lines represent the citation relationship between kenes. The influence relationship among agents is denoted by white line.

4 Metrics and Indicators for Measuring Creative Output

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

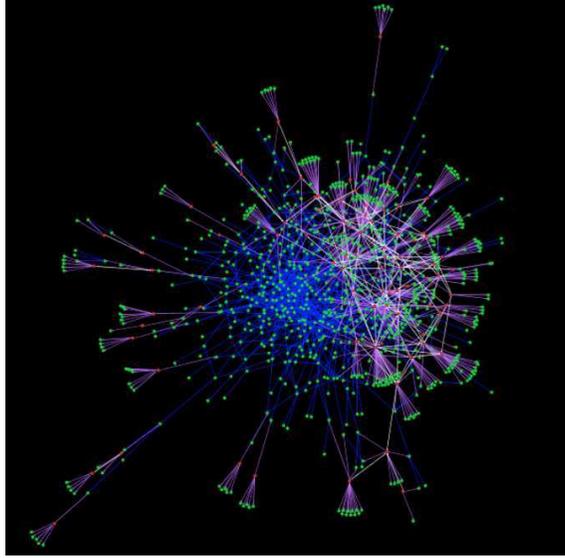


Figure 3: Socio-technical Network of a Simulated Community

4.1 Innovation Metrics

To measure creative output and effectiveness of diffusion kenes, we artificially divide the community into subdomains. Each subdomain is associated with its own focus areas and interacts with other subdomains through boundary processes through kene combination that enables trait transfer. The metrics are (1) average kene fitness, (2) domain impact factor, and (3) diffusion.

4.1.1 Kene Count and Average Kene Fitness

A basic measure of innovation output is the number of kenes accepted by the community. A more sophisticated measure takes into account not only the count, but also the degree usefulness of the generated knowledge. If we assume that the total number of accepted kenes is N , then the equation to calculate the average kene fitness akf is as follows.

$$akf = \frac{1}{N} \sum_{j=1}^N F_j, \quad (8)$$

where F_j is the fitness of kene j , which includes the novelty and relational fitness components specified in equation (7).

4.1.2 Domain Impact Factor

Impact Factor (IF) is frequently used as a metric for the importance of a sub-domain to its field. Impact Factor has the advantage over raw citation count that it is situated in time and accounts for changes in sub-domain importance over time (Mann, Mimno, and McCallum 2006). DIF for subdomain D_s is defined in terms of the number of links from the kenes in the overall domain D to kenes in D_s at time $t-1$ and $t-2$.

$$dif(s,t) = \frac{citations(D, D_s^{t-1}, D_s^{t-2})}{|D_s^{t-1}| + |D_s^{t-2}|} \quad (9)$$

4.1.3 Diffusion

Impact factor cannot accurately reveal the importance of a specific kene. One of the drawbacks of Impact Factor is whether a particular kene has broad or narrow impact. Does a highly cited kene dominate a prolific sub-field or does it have broad appeal and utility across many fields? We use sub-domain based impact measures that reveal more than

citation count alone. The metric for evaluating broad-based impact is Diffusion (Mann, Mimno, and McCallum 2006), which is defined for a given sub-domain s :

$$diffusion(s) = H(P_s) = \sum_{s'} P_s(s') \log P_s(s'), \quad (10)$$

$$\text{where } P_s(s') = \frac{citations(D_s, D_s)}{|D_s|}$$

4.2 Network Metrics

Structural properties of networks, as they relate to creative output, can help measure the degree of integrated differentiation (Stankiewicz 1992). As a general measure of the degree of social interaction, we use density, centrality, clustering coefficient, and percentage of strong ties so as to elucidate the relationship between structure and innovativeness.

5 TYPES OF OPEN SCIENCE COMMUNITIES

To examine creative output of an open scientific community based on different evaluation strategies, we considered three types of open source societies: *Exploration-oriented*, *Utility-oriented* and *Service-oriented* (Yilmaz 2009). Table 1 below defines the characteristics of these communities.

Table 1: Community Types

Community Type	Characters	
Exploration-oriented	Recruitment Selectivity	High
	Growth rate	Low
	Turn over	Low
	Decision-making Style	Centralized
Utility-oriented	Recruitment Selectivity	Moderate
	Growth rate	High
	Turn over	Moderate
	Decision-making Style	Emergent selection
Service-oriented	Recruitment Selectivity	Low
	Growth rate	Moderate
	Turn over	High
	Decision-making Style	Council

Recruitment selectivity indicates the threshold that determines whether an agent begins to contribute or leave the community after the process of enculturation. If the enculturation level is greater than the threshold, the agent begins to contribute. Otherwise the agent leaves the community. Growth rate indicates the number of new individuals entering the community at each time interval. Turnover specifies the threshold that determines whether or not an agent will leave the community. If the motivation level is less than the threshold, the agent will leave the community. Otherwise, the agent stays in the community and continue to make contributions to the domain. Decision making style involves the process that determines whether to accept or reject a contribution. Centralized decision making, which is often observed in exploratory communities, is performed by a single evaluator that represents the leader of the project. In service-oriented communities, core members of the group are allowed to vote. Majority of votes is considered to make the decision to exclude or retain the contribution. Emergent selection delays evaluation for a specific time interval, at the end of which the number of individuals who adopted (e.g., cited) the contribution is taken into account to make a decision.

6 Simulation Results and Evaluation

The model is implemented using the Recursive Porous Agent Simulation Toolkit (RePast) that is a free and open source agent-based modeling and simulation environment.

6.1 Preliminary Experiments

To examine (1) the impact of alternative community structures on the creative output, (2) the role of social network structures in conferring innovativeness in community forms of organization, and (3) factors that help distinguish innovative communities from the rest under each type of community, we conducted a series of preliminary experiments.

First, each type of community is simulated 100 times to collect innovation and network metrics to comparatively observe their knowledge creation potential as well as the structure of social networks generated during the process.

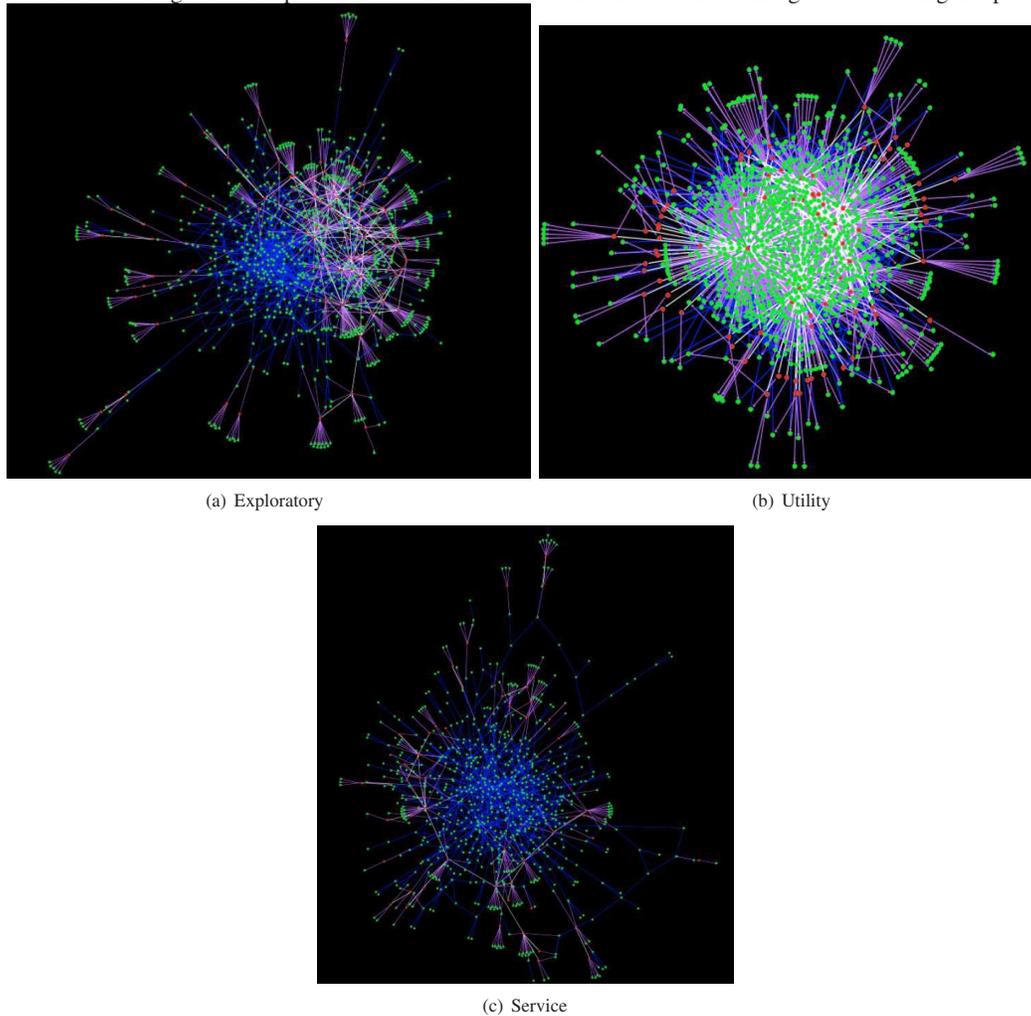


Figure 4: Generated Socio-technical Networks under Alternative Community Parameters

Recall that in the final state of the simulated communities shown in Figure 4(a,b,c), green nodes are kenés and red nodes are individuals. Purple lines represent the relationship between individuals and kenés. Blue lines represent the citation relationship between kenés. The influence relationship among agents is denoted by white line. Qualitatively, one can observe a strongly connected relational structure between the kenés in both the service and exploratory projects. This is expected, because kenés that are tied to core knowledge are more likely to be accepted in centralized and council style decision making. Utility communities are not as successful as service and exploratory communities in developing a strongly integrated domain knowledge. However, they are more effective in creating multiple strongly connected sub-communities at the periphery. This is indicative of their potential in generating integrated differentiation within the community.

6.2 Innovation and Network Structure in OSCs

Comparatively, on the average, exploratory and service communities generate more accepted contributions than utility communities. This is shown in Figure 5. Having small number of accepted contributions in utility communities does not necessarily mean low innovativeness. In utility communities, due to emergent selection style of decision making, only those contributions that are highly beneficial to others are accepted. This is evident in part (b) of the figure, as utility communities exhibit higher level of average fitness, which takes into account the degree of utility and impact of the produced knowledge.

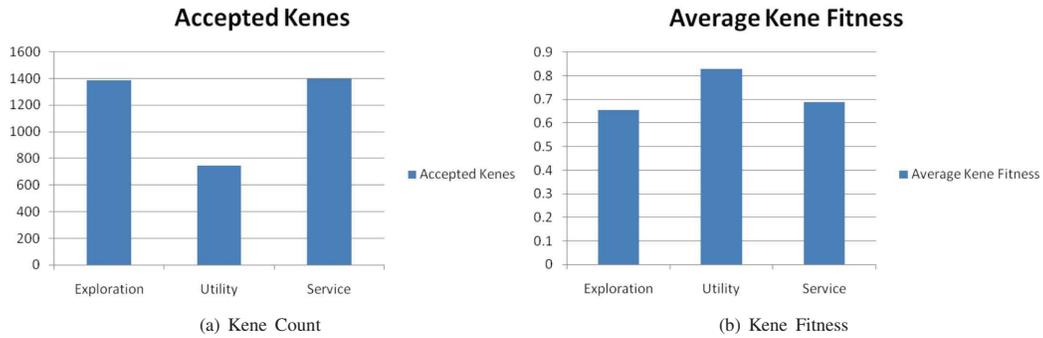


Figure 5: Knowledge Output

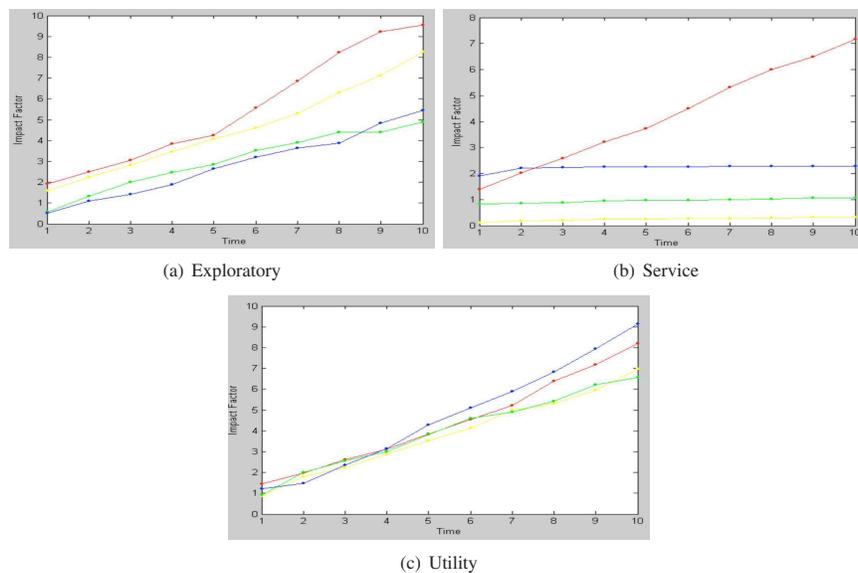


Figure 6: Comparative Analysis of Domain Impact Factor

To examine the intensity of boundary processes among artificially created subdomains, we measure the domain impact factor for each subdomain under exploratory, utility, and service-oriented communities. Each different line in Figure 6 depicts the accumulation of impact over time for the corresponding subdomain. The results suggest that utility and exploratory communities exhibit sustained growth in transfer of kenes across sub-communities, while service-oriented communities impede transfer. In utility-oriented communities variation in knowledge transfer intensity among subdomains is small compared to exploratory communities.

The comparison of alternative communities with respect to structural network metrics is shown in Figure 7. In contrast to exploratory and service-oriented communities, the knowledge creation processes in utility communities result in low density/high centrality that is characteristic of many innovation networks (Dhanaraj and Parkhe 2006). On the other hand, the centralized decision making strategy used in exploratory communities results in cohesive and strongly connected communities. The high-level of centrality, in conjunction with low density, observed in utility communities indicate not only differentiation, but also integration. This is because members tend to be organized around multiple focal individuals that are loosely coupled to each other.

From the perspective of cliquishness, exploration and service communities tend to be more cliquish as compared to utility communities. Although cliquishness when occurs with average smallest short path lead to high rate of knowledge growth, large values of clustering coefficient result in strongly connected communities that are not sufficiently differentiated. Finally, in utility communities we observe repeated interactions among the members of a utility community. This

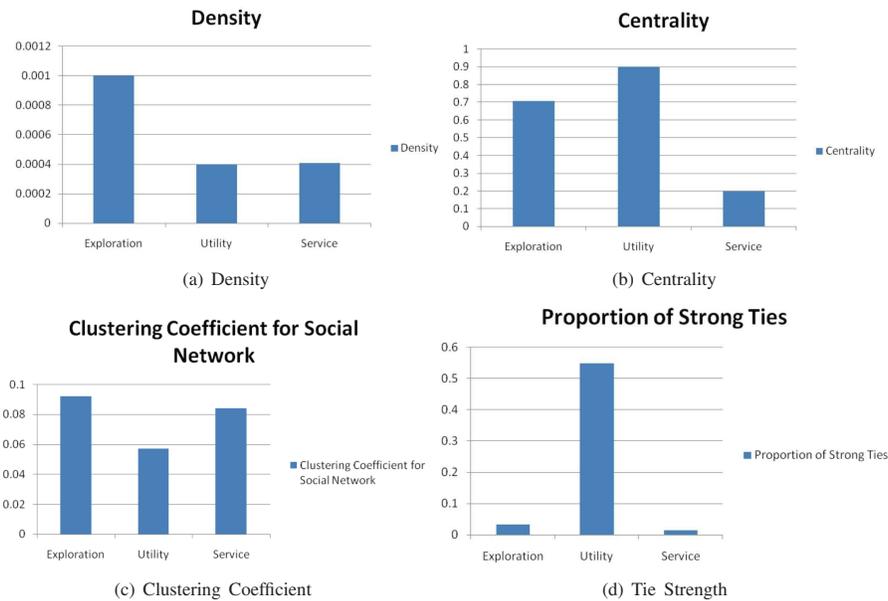


Figure 7: Comparative Analysis of Social Network Structures

observation is reasonable because small number of individuals that are organized around a focal member is more likely to interact with each other.

6.3 What Differentiates Innovative Communities

The purpose of the second experiment is to explore the characteristics that help distinguish innovative communities from the rest. Therefore, we examine relationship between average kene fitness and social network metrics in each type of community. To analyze the relationship large number of communities of each type are simulated. The experiment consists of two steps: the first step is to categorize the data into three groups (i.e., low, moderate and high innovation) based on the average kene fitness; the second step is to compute the average network metrics for each group.

For each type of community, Figure 8 plots the change in density, centrality, clustering coefficient, and strength of ties in the social network. Although the results are not conclusive, for exploratory communities, density and clustering coefficient do not help differentiate innovation potential, but centrality slightly increases as average kene fitness increases. This may indicate the role of centrality on the quality of kenes in exploration-oriented community. Additionally, we observe slight decrease in the proportion of strong ties as average kene fitness increases. For utility communities, clustering coefficient is not significantly different across different innovation performance categories. However, density and proportion of strong ties, and especially centrality, increase with the increase in average kene fitness. Centrality seems to be a stronger indicator for innovation potential in utility communities as compared to exploratory and service communities.

7 CONCLUDING REMARKS

We examined the impact of community governance mechanisms on collective creativity in innovation communities such as virtual scientific organizations. The view advocated in the study is that such innovation communities exhibit the characteristics of self-organizing complex adaptive systems and the tools in complexity science can bring useful insight into the study of creativity and innovation. To this end, we developed a model for representing the structure and dynamics of open science communities. The model is conceptually grounded and credible in terms of its congruence to conceptual elements of such communities. The dynamics of the model is patterned after the systems model of creativity.

Based on the simulation results, social networks generated by knowledge creation processes under the utility-oriented community style exhibit high centrality and low density, both of which are indicative of innovative communities. Also, tendency to organize around focal members of the community significantly effects the quality of kenes in all three kinds of scientific communities. At the same time, we observe greatest variation of average kene fitness in utility-oriented communities.

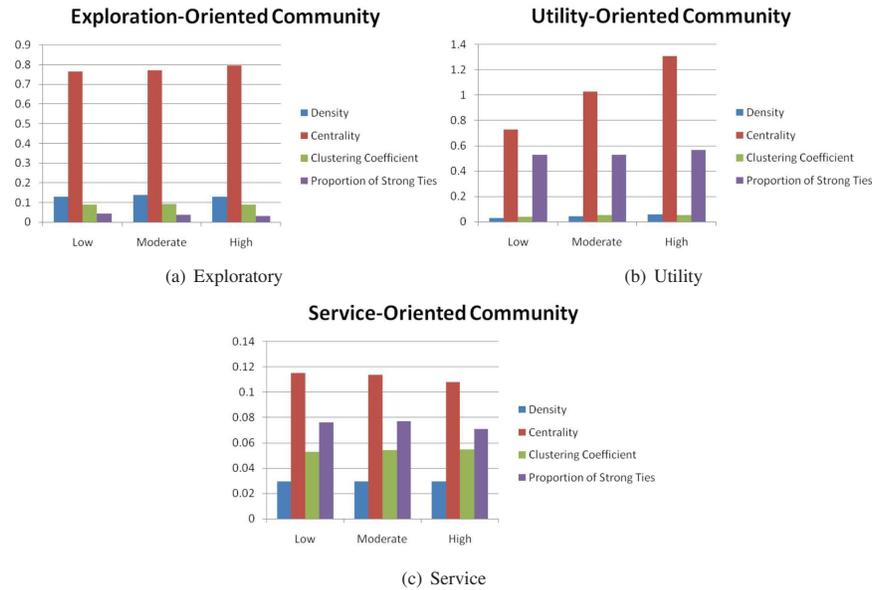


Figure 8: Network Structure vs. Innovativeness

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation Grant NSF-SBE-0830261. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government.

REFERENCES

- Bos, N., A. Zimmerman, J. Olson, J. Yew, J. Yerkie, E. Dahl, and G. Olson. 2003, May. From shared databases to communities. *Technical Report: Oxford Technical Institute*.
- Cowan, R., and N. Jonard. 2004. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics & the Diffusion of Knowledge* 28:1557–1575.
- Crane, D. 1972. *Invisible colleges: Diffusion of scientific knowledge in scientific communities*. Chicago, IL: The University of Chicago Press.
- Csikszentmihalyi, M. 1999. Implications of a systems perspective for the study of creativity. *Handbook of Creativity*:313–338.
- Dhanaraj, C., and A. Parkhe. 2006. Orchestrating innovation networks. *Academy of Management Review* 31 (3): 659–669.
- Edmonds, B. 2007. Artificial science - a simulation to study the social processes of science. *Social Simulation: Technologies, Advances and New Discoveries*:61–67.
- Foster, I. 2005, May. Service-oriented science. *Science* 308:814–817.
- Gilbert, N. 1997. A simulation of the structure of academic science. *Sociological Research Online* 2 (2): <http://www.socresonline.org.uk/socresonline/2/2/3.html>.
- Harrison, J. R., and G. R. Carroll. 2006. *Culture and demography in organizations*. Princeton, NJ: Princeton University Press.
- Holland, J. H. 1995. *Hidden order: How adaptation builds complexity*. Perseus Books.
- Hovland, C., and E. Hunt. 1960. The computer simulation of concept attainment. *Behavioral Science* 5:265–267.
- Klahr, D., and H. A. Simon. 1999. Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin* 125 (5): 524–543.
- Mann, G. S., D. Mimno, and A. McCallum. 2006. Bibliometric impact measures leveraging topic analysis. *Proceedings of the 2006 JCDL Conference*:25–34.
- NanoHUB 2009, May. Network for computational nanotechnology. <http://nanohub.org/>.
- Smith, B., and M. Ashburner. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25 (11): 1251–1255.

- Stankiewicz, R. 1992, Springer-Verlag. Technology as an autonomous socio-cognitive system. *Dynamics of Science-Based Innovation*.
- Ward, T. B., S. M. Smith, and J. Vaid. 1997. Conceptual structures and processes in creative thought. *Creative Thought: An Investigation of Conceptual Structures and Processes*:1–30.
- Yilmaz, L. 2009. An agent simulation study on conflict, community climate, and innovation in open source communities. *International Journal of Open Source Software and Processes* 1 (4): 1–25.

AUTHOR BIOGRAPHIES

GUANGYU ZOU is a Ph.D. candidate at the Computer Science and Software Engineering Department at Auburn University. He received his M.S. degree from Industrial and Systems Engineering at Auburn University. His research interests are in modeling and simulation, agent-based modeling, and complex adaptive systems. His email address is gzz0001@auburn.edu.

LEVENT YILMAZ is Associate Professor of Computer Science and Software Engineering and holds a joint appointment with the Industrial and Systems Engineering at Auburn University. He received his B.S. degree in Computer Engineering and Information Sciences from Bilkent University and M.S. and Ph.D. degrees from Virginia Tech. His research interest are in Modeling & Simulation, Agent-directed Simulation, and Complex Adaptive Systems with a focus in (1) advancing the theory and methodology of modeling and simulation via novel formalisms and their use in computational models of scientific discovery and theory formation and robust decision support and (2) models of socio-technical, cognitive, and cultural systems (e.g., science of science and innovation policy). He is a member of the Board of Directors of SCS and is the Editor-in-Chief of *Simulation: Transactions of the Society for Modeling & Simulation International*. He is member of ACM, IEEE Computer Society, Society for Computer Simulation International (SCS), and Upsilon Pi Epsilon. His email and web addresses are yilmaz@auburn.edu and <http://www.eng.auburn.edu/users/yilmale>.