*Proceedings of the 2015 Winter Simulation Conference*
*L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.*

# A COMPUTATIONAL MODEL OF TEAM ASSEMBLY IN EMERGING SCIENTIFIC FIELDS

Alina Lungeanu

Technology and Social Behavior Program
Northwestern University
2240 Campus Drive
Evanston, IL 60208, USA

Sophia Sullivan

Think Big Analytic
156 N. Jefferson St.
Chicago, IL 60661, USA

Uri Wilensky

Departments of Learning Sciences and Computer Science

Northwestern University
2120 Campus Drive
Evanston, IL 60208, USA

Noshir S. Contractor

Departments of Industrial Engineering and Management Sciences, Communication Studies, and Management and Organizations
Northwestern University
2240 Campus Drive
Evanston, IL 60208, USA

## ABSTRACT

This paper examines the assembly of interdisciplinary teams in emerging scientific fields. We develop and validate a hybrid systems dynamics and agent-based computational model using data over a 15 year period from the assembly of teams in the emerging scientific field of Oncofertility. We found that, when a new field emerges, team assembly is influenced by the reputation and seniority of the researchers, prior collaborators, prior collaborators' collaborators, and the prior popularity of an individual as a collaborator by all others. We also found that individuals are more likely to assemble into an Oncofertility team when there is a modicum of overlap across its global ecosystem of teams; the ecosystem is defined as the collection of teams that share members with other teams that share members with the Oncofertility team.

## 1 INTRODUCTION

Interdisciplinary scientific teams are frequently at the root of innovative breakthroughs (Uzzi and Spiro 2005). As a result, understanding the mechanisms behind the assembly of scientific teams has attracted scholarly interest. A first step has been to examine the compositional and relational mechanisms affecting the formation of scientific teams (Guimera et al. 2005; Lungeanu and Contractor 2015; Lungeanu, Huang, and Contractor 2014). Most prior research has treated teams as well-defined entities with a stable set of members who work interdependently toward a common goal (Cohen and Bailey 1997). However, the reality is that most knowledge workers hold membership in multiple teams simultaneously (O'Leary, Mortensen, and Woolley 2011), making membership in an ecosystem consisting of multiple teams with overlapping members the rule rather than the exception. Such ecosystems are dynamic and complex networks of prior collaborations (Poole and Contractor 2011) which enable and constrain the assembly of future scientific teams. Yet the effects of the ecosystem on team assembly have not been explored, perhaps because it entails complex statistical analyses across multiple levels.

In response to this research gap we develop a multi-theoretical multilevel model that incorporates the impact of ecosystem factors on the assembly of interdisciplinary teams. Specifically, we draw upon

theories on the formation of social networks (Contractor, Wasserman, and Faust 2006) and their application to the assembly of teams (Contractor 2013), as well as the more extensive research on groups and teams (Levine and Moreland 1998), to examine factors leading to assembly of interdisciplinary scientific teams.

We implement a hybrid - agent-based and system dynamics - computational model that articulates the multilevel multi-theoretical mechanisms team assembly. We empirically validate our model with data on the assembly of teams in the emerging scientific field of Oncofertility from its inception in 1996 until 2010. Oncofertility is an appropriate context to study assembly mechanisms because (1) teams are often assembled on an ad-hoc basis reflecting the autonomous and individual choices of scientists in the absence of confounding outside influences and (2) the emergence of the field allows us to have a natural starting point to explore how individuals change their motivations in choosing their collaborators as the field grows and begins to attract systemic institutionalized funding.

We begin by reviewing, in Section 2, the theoretical rationale and empirical evidence for factors that affect the assembly of interdisciplinary teams during the emergence of a scientific field. Section 3 provides the rationale for the hybrid computational modeling approach. Section 4 describes the implementation of the hybrid model. Finally, section 5 summarizes our findings and its implications.

## 2 HYPOTHESIZED MECHANISMS FOR TEAM ASSEMBLY

Members of interdisciplinary scientific team need to hold knowledge mutually understood by all parties and know how to coordinate their collaborative tasks (Teasley and Wolinsky 2001). In order to be innovative they also need to incorporate diverse expertise, concepts, methodologies, and theoretical approaches, which produces significant heterogeneity within the team and increases the risk of assembling teams in a suboptimal manner.

To mitigate against these challenges, we propose investigating the assembly of teams based on (i) compositional level mechanisms that focus on the attributes of individuals, (ii) relational level mechanisms that focus on prior relationships among members, and (iii) the ecosystem level mechanisms that focus on the extent to which team members are currently or previously embedded in multiple other teams that have overlapping team membership. These three sets of factors, summarized in Table 1, operate at different levels of analyses and incorporate different theoretical mechanisms prompting the development of a multi-theory, multilevel model of team assembly.

Table 1: Hypothesized theoretical mechanism.

| Mechanism | Description | Citation/studies |
|---|---|---|
| *Compositional mechanisms* | | |
| M1: Seniority | Researchers prefer to collaborate with senior researchers. | (Bozeman and Corley 2004; Lungeanu et al. 2014) |
| M2: H-index | Researchers prefer to collaborate with high performing researchers (h-index) | (Lungeanu et al. 2014) |
| M3: Gender inertia | Researchers' preferences for (or against) gender homophily when choosing new collaborators will remain the same as their preferences in choosing prior collaborators. | (Cummings and Kiesler 2005; Lungeanu and Contractor 2015; Moody 2004) |
| M4: Institution affiliation inertia | Researchers' preferences for (or against) institution homophily when choosing new collaborators will remain the same as their preferences in choosing prior collaborators. | (Cummings and Kiesler 2005; Lungeanu and Contractor 2015; Moody 2004) |
| *Relational mechanisms* | | |
| M5: Prior successful | Researchers are more likely to collaborate with prior successful collaborators in a proportion equal to the | (Guimera et al. 2005; Lungeanu and Contractor 2015; Lungeanu |

| Mechanism | Description | Citation/studies |
|---|---|---|
| collaboration | success of their prior collaboration. | et al. 2014) |
| M6: Friend of a friend | Researchers with prior common collaborators are more likely to collaborate. | (Newman 2001) |
| M7: Preferential attachment | Researchers prefer to collaborate with well-connected researchers. However, well-connected researchers are less likely to accept collaborations with less well-connected researchers. | (Barabási and Albert 1999; Newman 2002) |
| *Ecosystem mechanisms* | | |
| *M8*: Global ecosystem closure | Scientific teams are more likely to be assembled when their scientific ecosystem represents a "coherent intellectual neighborhood." | (O'Leary et al. 2011; Poole and Contractor 2011) |
| M9*:* Local ecosystem brokerage | Scientific teams are more likely to be assembled when their immediate scientific neighborhood is not redundant. | (O'Leary et al. 2011; Poole and Contractor 2011) |

## 3 HYBRID COMPUTATIONAL MODELING OF TEAM ASSEMBLY

To test our hypothesized mechanisms we use a *hybrid agent-based system dynamics simulation* (other hybrid models are reviewed in Lattila, Hilletofth, and Lin 2010). While system dynamics (SD) (Forrester 1961) and agent-based (ABM) simulation models have been used for studying complex systems (North and Macal 2007), these models are complementary paradigms: SD is a top-down approach based on modeling factors (Akkermans 2001; Sterman and Wittenberg 1999), while ABM is a bottom-up approach based on modeling actors (Macy and Willer 2002).

SD is a well-established and commonly used technique that has been applied to explore the impact of factors on dynamics of a system. SD models resources (stocks) and dynamics (flows) within the system as a whole. Stocks are aggregated representations of the system's entities, while flows capture the rates at which entities within the system change state. SD captures feedback and delay processes to model system behavior over time. This approach has the advantage of looking systemically and simultaneously at the impact of multiple positive and negative feedback loops on overall system dynamics. However, many SD models assume homogeneity of the population (i.e. differences in individuals' characteristics and their social network ties to other individuals). Compartmentalized SD models have been used to reflect the different behaviors of subpopulation. However, within the subpopulation, behavior is still assumed to be homogeneous (Hethcote 2000).

ABM, on the other hand, represents system entities as actors (i.e., agents). In ABM, autonomous agents interact with each other and their environment based on using simple behavior rules to act on local information. The agents learn from their interactions and adapt their behavior accordingly (Macal 2010; Macy and Willer 2002). Hence ABMs overcome the limitations of SD models by incorporating heterogeneity in individual attributes and their social networks. However, a limitation of ABMs is that they restrict knowledge of the system to an agent's local point of reference, although the behaviors of all agents together frequently generate emergent patterns that may be unexpected (Holland 1998).

Although SD models and ABMs have been utilized separately to model the assembly of scientific teams and the evolution of new scientific fields (Bettencourt et al. 2008; Guimera et al. 2005; Sun et al. 2013), a hybrid approach enables us to simultaneously examine the effect of heterogeneous individual attributes and relations on the assembly of teams in emerging scientific fields, as well as, in turn, the impact of systemic characteristics of the scientific field on subsequent team assembly. Therefore, a hybrid model has the potential to explore more realistically (and accurately) the assembly of teams in emerging scientific fields.

# 4 IMPLEMENTATION OF COMPUTATIONAL MODEL

## 4.1 The Oncofertility Field and Dataset

The field of Oncofertility represents an appropriate context in which to examine the compositional, relational, and ecosystem mechanisms that affect the assembly of interdisciplinary teams. Oncofertility investigates fertility preservation for young patients with fertility-threatening diseases. This explicitly requires interdisciplinary collaboration among researchers from two different research areas – oncology and fertility. Although the term *'oncofertility'* was first used in 2006, publications in the Oncofertility field can be traced as far back as 1993, thus providing sufficient history to empirically test the model. We first identified all scientific articles that were published in the Oncofertility field using the keywords *oncofertility*, or *cancer* and *ovarian tissue cryopreservation*, or *cancer* and *fertility preservation*. We used the *Web of Science (WoS)* database provided by Thomson Reuters to construct researchers' bibliometric information. Since there were articles that were not indexed in the *WoS* database, we supplemented the dataset with the articles' index in the *PubMed* database. This yielded a total of 553 publications by 1,696 authors between 1996 to 2010.

Demographic and institution information were manually coded. We obtained gender information using text (e.g., text references such as "her work") and image searches on researchers' institutional Web pages. Institution affiliation was extracted from researchers' vitae and publication information. Additional bibliometric data, such as prior co-authorship and citation, were extracted from the *WoS* database.

Our goal was to model the compositional, relational and ecosystem mechanisms that influenced the assembly of teams carrying out scientific collaborations. Compositional mechanisms, which focus on characteristics of the individuals, include *Seniority (M1)*, *H-index (M2), Gender (M3)* and *Institution affiliation (M4) inertia preference.* Relational mechanisms, which focus on relations among the individuals, include *Prior successful collaboration (M5)*, *Friend of a friend (M6)* and *Preferential attachment (M7)*. Ecosystems mechanisms focus on characteristics of the ecosystem in which the potential team would be embedded. *Global ecosystem closure (M8)* was computed as the average clustering coefficient of the relevant team's immediate and distal neighborhood in the Oncofertility ecosystem. The clustering coefficient is defined as the total number of closed triads in the network relative to the number of possible triads, where a closed triad represents a unit of clustering with all people connected to each other. In other words, clustering means that researchers tend to collaborate with the collaborators of their collaborators. Relevant team's neighborhood was considered as the total number of researchers and collaborations links (teams) three-steps away from the focal team. Finally, *Local ecosystem brokerage (M9)* was computed as the relevant team's inverse clustering coefficient in the immediate Oncofertility ecosystem. We followed a process of initiating and accepting/rejecting invitations to collaborate, steps that act as generative mechanisms for team assembly.

## 4.2 Model Description

As mentioned previously, we use a hybrid simulation approach to implement our mechanisms and we use empirical data to estimate the extent to which our hypothesized mechanisms are in fact influencing team assembly. The evolution of a new scientific field shares similarities to both the spread of a disease and to the adoption of a new innovation, which have traditionally used SIR (Susceptible, Infected, Recovered) models as modeling techniques. Bettencourt et al. (2008) use an SIR model to represent the evolution of several academic fields, which we adapt for our model. Our model has three populations: *Unaware* agents that have not heard of the oncofertility field, *Aware* agents that have heard of the field but have not yet published, and *Active* agents that have already published in the field.

### 4.2.1 Hybrid Agent-Based System Dynamics Model Implementation

The development of the computational model relies heavily on empirical data and is focused on team-level properties such as number of teams and distribution of team sizes. In order to simulate team assembly, we assume that the distribution of team sizes and an individual's participation in a certain set of team sizes in a given year are *constant* and *equal to what was empirically observed*. In order to preserve the distribution of team sizes, we do not require that those who are already Active have to publish every year. Rather, if they do become Active in a year, then they are assigned a team size distribution of the teams in which that individual participated in that specific year. For example, if a particular Active agent is invited to participate in a team of size 3, then the actor is assigned the entire list of team sizes in which an individual from the empirical data participated. Consequently, $[2000; A_i: 3,6]$ would indicate that an individual $A_i$ collaborated on two papers in the year 2000, one with three authors (including herself), and the other with five other collaborators. In the next year, she may or may not collaborate again and, if she was, would be assigned a new set of co-authorship teams.

*Model initialization.* When the model is initialized, each agent is assigned an age, H-index, gender and institution that match those of a person from the empirical data. The agent is also assigned the prior collaboration network of that person with weighted ties that represent the number of times that agent collaborated with other agents outside of the oncofertility field. Next, each pair of agents receives a success score that represents the number of citations received by the papers the two agents co-authored. All agents except one, which is set as Aware, are initially set as Unaware. The model starts in 1996 and loads the distribution of collaboration teams for that year: one team of two, one team of three, and one team of five. As a result, the 1996 team distribution is [1996; 2,3,5] and the Aware agent receives a randomly picked team size (either 2, or 3, or 5).

*Team assembly*. First, a random Aware or Active agent decides to form a team. An Active agent is assigned a team size distribution for that year, and will decide on forming a team with size based on that distribution. For an Aware agent, a team size is randomly picked from the remaining sizes for that year, and a distribution is assigned to the agent that contains a team of that size. Preference in team formation is given to agents who are already Active and have other teams in which they must participate in that year.

Next, the agent starts to build the team. The founding agent will always be the first to invite other agents to join. Preference is given to Active agents who must be on a team of that size in that year. In order to decide whether to invite another agent to join the team, each agent calculates a score for the other agent based on the theoretical mechanisms defined above. The invited agents compute a utility function score to decide whether to accept the collaboration invite. This decision is based on all properties of the current team members, not just on the properties of the inviting agent. If the invited agent accepts, s/he is added to the team, and another round of invitations begins. If the invited agent does not accept, and was Unaware, the agent then becomes Aware but does not join the team.

Appendix A contains the analytical representation of the scores for each hypothesized mechanisms. Each score is normalized, so its value is between 0 and 1, and each mechanism has an equal influence on the decision to assemble the team. So, the total score is:

$$P(Team_i) = \sum_{k=1}^{9} M_k \Big/ \sum_{k=1}^{9} \max\,(M_k) \tag{1}$$

where $Team_i$ is the team $i$ to be assembled, and $M_k$ represents the score for the mechanism $k$. The final score is then compared to a random uniform number (i.e., a floating point number uniformly distributed between 0 and 1) to see whether $Team_i$ will be formed.

Each team member invites other prior collaborators to join sequentially, or if they have none, another random agent. This process continues until all the spots on the team have been filled. The team members then add all the other members of the team to their oncofertility collaboration network.

*Moving to the next year*. Once all the teams for the particular year have been formed, the simulation moves to the next year. While the oncofertility collaboration network is determined by the model, the

prior non-oncofertility collaboration is obtained from the empirical data and it is loaded in the model each year. For example, when the model starts the simulation for year 2000, the collaboration network outside the oncofertility field prior 2000 is loaded into the model. Then the team assembly process begins again. The simulation ends when all the teams from the last year, 2010, have been formed.

#### 4.2.2 Exogenous Event Affecting the Emergence of a Scientific Discipline

The computational model includes an exogenous event that influences the field evolution. In 2007, the National Institutes of Health (NIH) provided a $6.5 million dollar grant to fund the Oncofertility Consortium to promote the field and make it more visible to researchers and practitioners (https://www.woodrufflab.org/ul1-oncofertility-consortium). Thus, before 2007, scientists that were Active agents in the field were more likely to promote it, while after 2007, Aware agents also began to promote the field in order to have access to institutional funds. Therefore, we used the number of prior Actives to determine when the NIH grant was made and the following equation to describe the analytical representation of this event:

$$dU/dt = -\alpha_1 * (A + W/2)/N * (1 - NIH) - \alpha_2 * (A + W)/N * (NIH) \quad (2)$$

where U represents the number of people unaware about the field (inactive) at time $t$, A represents the number of actives at time $t$, W represents the number of aware at time $t$, and N represents the size of the population modeled. $\alpha_1$ represents the contact rate before NIH intervention, and $\alpha_2$ represents the contact rate after NIH intervention. This equation is called every time a new team is assembled in order to capture the current number of Unaware, Aware, and Active agents.

### 4.3 Model Validation

#### 4.3.1 Parameter Estimation

The model was implemented in the Netlogo ABM platform (Wilensky 1999) using the process described above. Once the model was built, the parameters were fitted to the empirical data to assess the relative importance of the hypothesized team assembly mechanisms. The parameters were fit using the BehaviorSearch tool (Stonedahl and Wilensky 2010). BehaviorSearch is a powerful and robust tool that calibrates models implemented in NetLogo (Thiele, Kurth, and Grimm 2014). Calibration simply describes the process of manipulating a model to get closer to a desired behavior. In this case, the desired behavior is matching the simulated teams to the empirical teams as closely as possible. The objective function used was to minimize the mean squared error between the average clustering coefficients of the simulated teams and the empirical teams.

To investigate our space of parameters, we used the standard genetic algorithm (GA) search method, with ‘‘GrayBinaryChromosome” representation. GAs offer a flexible meta-heuristic search mechanism which has been successful in combinatorial optimization and search problems. Gray codes have generally been found to give better performance for search representations. The optimization function was measured as the minimum objective function over 100 simulations. Each simulation contained 1000 model runs with 100 replications of each previous best model obtained. Two separate analyses were performed to empirically fit the model. First, the BehaviorSearch minimization was run across all years. This analysis yields one set of parameters. Additionally, the BehaviorSearch minimization was run for each year separately. This yields a distinct set of parameters for each year.

#### 4.3.2 Results and Analysis

All variables were weighted to fall between 0 and 1. Additionally, all parameters were specified to range between 0 and 1 because of the positive relationship hypothesized the model. This analytical strategy allows us to compare directly the effect sizes of all parameters specified. The magnitude of the parameter

is a measure of the effect size for each mechanism and describes how important each factor is relative to the others. Important effects are defined as effects that have larger effect sizes relative to the other factors assessed. Table 2 shows the value of parameters when the analysis was run for the entire dataset.

Table 2: Model parameters estimated across entire period from the empirical data.

| Mechanism | Parameter estimate |
|---|---|
| *Compositional mechanisms* | |
| M1: Seniority | 0.79 |
| M2: H-index | 0.99 |
| M3: Gender inertia | 0.12 |
| M4: Institution affiliation inertia | 0.19 |
| *Relational mechanisms* | |
| M5: Prior successful collaboration | 0.44 |
| M6: Friend of a friend | 0.88 |
| M7: Preferential attachment | 0.64 |
| *Ecosystem mechanisms* | |
| M8: Global ecosystem closure | 0.45 |
| M9: Local ecosystem brokerage | 0.12 |

The computational model shows that of the four compositional mechanisms, H-index is the most important factor, followed by seniority. Gender and institution affiliation preferences are much less important factors when deciding collaboration relationships. Out of the three relational mechanisms, friend of a friend is the most important factor, followed by preferential attachment and prior successful collaboration factors. Out of the two ecosystem mechanisms, global ecosystem closure is the most important factor. Overall, these results suggest that, when a new field emerges, team assembly is influenced at the compositional level by the reputation and seniority of the researchers. At the relational level, team assembly is influenced by choosing prior collaborators, collaborators' collaborators or the prior popularity of an individual as a collaborator by all others. At the ecosystem level, individuals are more likely to assemble into a focal team when there is a modicum of overlap across the global ecosystem of teams comprising members who are on teams that share members with other teams that share members with the focal team. However, the local ecosystem brokerage, the lack of overlap among teams in which members of the focal team participated, was among the least important factors. One would expect that the team having access to novel information from their participation in teams that do not overlap would make it more likely to assemble in an emerging scientific field. Instead the results suggest that assembling into teams with others who are on teams that do not share collaborators might be considered risky, at least at the beginning of a new field.

The results of the computational model presented above reflect an aggregate snapshot across all 15 years from 1996 to 2010 and we should not expect that these results are stable over time. Therefore, we examined whether the factors affecting team assembly vary in strength and/or influence during the life-cycle of the emerging field. Table 3 presents the results of this post-hoc analysis. Given the small team network at the beginning of the field (1996 – 2000), our analysis focuses on the period 2001 – 2010.

Table 3: Model parameters estimated across each year from the empirical data.

| | Parameter estimate by year | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
| *Compositional mechanisms* | | | | | | | | | | |
| M1: Seniority | 0.67 | 0.30 | 0.15 | 0.15 | 0.18 | 0.50 | 0.72 | 0.17 | 0.18 | 0.94 |

| | **Parameter estimate by year** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
| M2: H-index | 0.51 | 0.22 | 0.49 | 0.13 | 0.27 | 0.71 | 0.89 | 0.47 | 0.18 | 0.84 |
| M3: Gender inertia | 0.12 | 0.69 | 0.95 | 0.96 | 0.89 | 0.81 | 0.50 | 0.74 | 0.82 | 0.34 |
| M4: Institutional affiliation inertia | 0.94 | 0.89 | 0.91 | 0.89 | 0.74 | 0.89 | 0.94 | 0.87 | 0.72 | 0.14 |
| *Relational mechanisms* | | | | | | | | | | |
| M4: Prior successful collaboration | 0.41 | 0.82 | 0.75 | 0.96 | 0.94 | 0.78 | 0.52 | 0.50 | 0.52 | 0.49 |
| M5: Friend of a friend | 0.30 | 0.74 | 0.60 | 0.64 | 0.93 | 0.34 | 0.50 | 0.85 | 0.59 | 0.86 |
| M6: Preferential attachment | 0.30 | 0.11 | 0.95 | 0.64 | 0.81 | 0.87 | 0.57 | 0.79 | 0.54 | 0.24 |
| *Ecosystem mechanisms* | | | | | | | | | | |
| M7: Global ecosystem closure | 0.67 | 0.64 | 0.64 | 0.69 | 0.44 | 0.43 | 0.40 | 0.25 | 0.22 | 0.26 |
| M8: Local ecosystem brokerage | 0.30 | 0.42 | 0.26 | 0.12 | 0.10 | 0.12 | 0.12 | 0.14 | 0.15 | 0.15 |

The impact of the four compositional mechanisms on team assembly vary greatly with three of the factors (i.e., seniority, H-index, and institution affiliation inertia) recording a peak in year 2007. Prior successful collaboration is stable with a small decline starting in 2007. The preferential attachment factor is very low at the beginning of the field, increases gradually, and then decreases after 2006-2007. Both ecosystem mechanisms decrease over time. It is noteworthy that the changes that appear around 2006-2007 coincide with NIH funding the creation of the Oncofertility Consortium. This illustrates the impact of external events such as funding on the motivations of individuals to assemble into teams.

### 4.3.3 Model validation

Finally, we tested the significance of the parameters obtained from the computational model. First, we generated individual attributes using the same mean and standard deviation from the empirical data. Then, we generated the prior non-oncofertility network using the same degree distribution from the empirical data. Then, we ran the computational model keeping the parameters constant. We repeated this process 100 times for each year. The simulations were run in R using the RNetLogo package (Thiele 2014). Finally, a one sample *t-test* was performed to determine whether the set of errors estimated with the proposed computational model are less than those estimated using the null model. Results show that the computational model performs better than the null model for all years ($p < 0.01$).

## 5 CONCLUSION

We developed a hybrid agent-based and system dynamics computational model to understand what factors affect the assembly of interdisciplinary teams in an emerging scientific field. Our research makes three important contributions. First, we propose a multilevel framework that incorporates compositional, relational, and ecosystem mechanisms to study the assembly of teams. Second, we implemented and validated a hybrid agent-based and system dynamics computational model to examine team assembly using data from the emerging scientific field of Oncofertility. Future research could examine the applicability of our methods and conclusions to other new or to mature scientific fields. Finally, the changes in motivations for team assembly observed in 2007 illustrate the impact of funding decisions by agencies such as the NIH.

## ACKNOWLEDGMENTS

The preparation of this manuscript was supported by funding from the US Army Research Laboratory (W911NF-09-2-0053), the Army Research Office (W911NF-14-10686), and the National Institutes of Health (UL1DE019587, UL1RR025741, UL1RR024146-06S2, 1U01GM112623). The views, opinions,

and/or findings contained in this manuscript are those of the authors, and should not be construed as an official Department of the Army or National Institutes of Health position, policy, or decision, unless so designated by other documents..

## APPENDIX A

When agent $A_i$ decides whether to invite agent $A_j$ to join the team or not, $A_i$ calculates an invite score $Score_k(A_i \rightarrow A_j)$, where $k$ is the theoretical mechanism. Next, agent $A_j$ decides whether to accept the invite to join the team or not. Agent $A_j$ calculates an acceptance score of the team, $Score_k(A_j \rightarrow Team_i)$, based on the characteristics of the agents already members of $Team_i$. The mechanism score is $M_i = w_i * Score_i$, where $M_i$ represents the final score for mechanism $i$, $w_i$ represents the parameter estimate for mechanism $i$ , and $Score_i$ represents the invite / acceptance score. Tables 4 and 5 present the analytical representation of invite and acceptance actions for each mechanism.

Table 4: Analytical representation of the invite score.

| | **Invite score** |
|---|---|
| M1 | $Score_1(A_i \rightarrow A_j) = age(A_j) / \underset{k \in network}{\text{MAX}} (\text{age}(A_k))$<br>where age $(A_i)$ represents the seniority of agent $A_j$ |
| M2 | $Score_2(A_i \rightarrow A_j) = hindex(A_j) / \underset{k \in network}{\text{MAX}} (\text{hindex}(A_k))$<br>where hindex $(A_i)$ represents the H-index of agent $A_j$ |
| M3 | $Score_3(A_i \rightarrow A_j) = \mathbf{1}_{[G_i = G_j]} * P_{Gi} + \mathbf{1}_{[G_i \neq G_j]} * (1 - P_{Gi})$<br>where $\mathbf{1}_{[G_i = G_j]}$ and $\mathbf{1}_{[G_i \neq G_{j2}]}$ are indicator functions that have a value of 1 if the genders of $A_i$ and $A_j$ are the same, and 1 if the genders of $A_i$ and $A_j$ are different respectively, and zero otherwise, and $P_{Gi}$ is the proportion of prior collaborators of agent $A_i$ with the same gender as $A_i$. |
| M4 | $Score_4(A_i \rightarrow A_j) = \mathbf{1}_{[I_i = I_j]} * P_{Ii} + \mathbf{1}_{[I_i \neq I_j]} * (1 - P_{Ii})$<br>where $\mathbf{1}_{[I_i = I_j]}$ and $\mathbf{1}_{[I_i \neq I_{j2}]}$ are indicator functions that have a value of 1 if the institutions of $A_i$ and $A_j$ are the same, and 1 if the institutions of $A_i$ and $A_j$ are different respectively, and zero otherwise, and $P_{Ij}$ is the proportion of prior collaborators of agent $A_j$ with the same institution as $A_j$ |
| M5 | $Score_5(A_i \rightarrow A_j) = success(A_i - A_j) / \underset{k \in network}{\text{MAX}} (success(A_i - A_k))$<br>where $success(A_i - A_j)$ represents the number of citations of all papers co-authored by $A_i$ and $A_j$ |
| M6 | $Score_6(A_i \rightarrow A_j) = \sum_{k \in network} (\exists tie_{A_i - A_k} \wedge \exists tie_{A_j - A_k}) / \deg(A_i)$<br>where deg $(A_i)$ represents the degree centrality of agent $A_j$ in the entire network |
| M7 | $Score_7(A_i \rightarrow A_j) = \deg(A_j) / \underset{k \in oncofertility\ network}{\text{MAX}} (\deg(A_k))$<br>where deg $(A_j)$ represents the degree centrality of agent $A_j$ in the simulated oncofertility network |
| M8 | $Score_8(Team_i) = AvgCC_{Team_i} - ERcc$<br>where $AvgCC_{Team_i}$ represents the average clustering coefficient of $Team_i$ neighborhood, and $ERcc$ represents the average clustering coefficient of a random simulated network with the same size as $Team_i$ neighborhood |
| M9 | $Score_9(Team_i) = AvgCC_{Team_i} - CC_{Team_i}$<br>where $AvgCC_{Team_i}$ represents the average clustering coefficient of $Team_i$ neighborhood, and $CC_{Team_i}$represents the clustering coefficient of $Team_i$ |

Table 5: Analytical representation of the acceptance score.

| | **Acceptance score** |
|---|---|
| M1 | $Score_1(A_j \rightarrow Team_i) = \max_{k \in Team_i} Score_1(A_j \rightarrow A_k)$<br>where $Score_1$ represents the invite Score 1 from $A_j$ to $A_k$ |
| M2 | $Score_2(A_j \rightarrow Team_i) = \max_{k \in Team_i} Score_2(A_j \rightarrow A_k)$<br>where $Score_2$ represents the invite Score 2 from $A_j$ to $A_k$ |
| M3 | $Score_3(A_j \rightarrow Team_i)$<br>$= \begin{cases} \frac{\sum_{k \in Team_i} \mathbf{1}_{[G_k = G_j]}}{\lvert Team_i \rvert} - P_{Gj} + min(P_{Gj}, 1 - P_{Gj}), if \ \frac{\sum_{k \in Team_i} \mathbf{1}_{[G_k = G_j]}}{\lvert Team_i \rvert} < P_{Gj} - min(P_{Gj}, 1 - P_{Gj} \\ P_{Gj} + min(P_{Gj}, 1 - P_{Gj}) - \frac{\sum_{k \in Team_i} \mathbf{1}_{[G_k = G_j]}}{\lvert Team_i \rvert}, if \ \frac{\sum_{k \in Team_i} \mathbf{1}_{[G_k = G_j]}}{\lvert Team_i \rvert} > P_{Gj} + min(P_{Gj}, 1 - P_{Gj} \\ 0, \quad otherwise \end{cases}$<br>where $P_{Gj}$ is the proportion of prior collaborators of agent $A_j$ with the same gender as $A_j$ |
| M4 | $Score_4(A_j \rightarrow Team_i)$<br>$= \begin{cases} \frac{\sum_{k \in Team_i} \mathbf{1}_{[I_k = I_j]}}{\lvert Team_i \rvert} - P_{Ij} + min(P_{Ij}, 1 - P_{Ij}), if \ \frac{\sum_{k \in Team_i} \mathbf{1}_{[I_k = I_j]}}{\lvert Team_i \rvert} < P_{Ij} - min(P_{Ij}, 1 - P_{Ij}) \\ P_{Ij} + min(P_{Ij}, 1 - P_{Ij}) - \frac{\sum_{k \in Team_i} \mathbf{1}_{[I_k = I_j]}}{\lvert Team_i \rvert}, if \ \frac{\sum_{k \in Team_i} \mathbf{1}_{[I_k = I_j]}}{\lvert Team_i \rvert} > P_{Ij} + min(P_{Ij}, 1 - P_{Ij}) \\ 0, \quad otherwise \end{cases}$<br>where $P_{Ij}$ is the proportion of prior collaborators of agent $A_j$ with the same institution as $A_j$ |
| M5 | $Score_5(A_j \rightarrow Team_i) = \max_{k \in Team_i} Score_5(A_j \rightarrow A_k)$<br>where $Score_5$ represents the invite Score 5 from $A_j$ to $A_k$ |
| M6 | $Score_6(A_j \rightarrow Team_i) = \max_{k \in Team_i} Score_6(A_j \rightarrow A_k)$<br>where $Score_6$ represents the invite Score 6 from $A_j$ to $A_k$ |
| M7 | $Score_7(A_j \rightarrow Team_i) =$<br>$0_{\left[\max_{k \in Team_i}\left(\frac{\deg(A_k)}{\max_{f \in field}(\deg(A_f))}\right) > .5 \ \& \ \frac{\deg(A_j)}{\max_{f \in field}(\deg(A_f))} > .5\right]} * \max_{k \in Team_i} Score_7(A_j \rightarrow A_k) +$<br>$0_{\left[\max_{k \in Team_i}\left(\frac{\deg(A_k)}{\max_{f \in field}(\deg(A_f))}\right) < .5 \ \& \ \frac{\deg(A_j)}{\max_{f \in field}(\deg(A_f))} < .5\right]} * \max_{k \in Team_i} Score_7(A_j \rightarrow A_k) +$<br>$1_{\left[\max_{k \in Team_i}\left(\frac{\deg(A_k)}{\max_{f \in field}(\deg(A_f))}\right) > .5 \ \& \ \frac{\deg(A_j)}{\max_{f \in field}(\deg(A_f))} < .5\right]} * \max_{k \in Team_i} Score_7(A_j \rightarrow A_k) +$<br>$-1_{\left[\max_{k \in Team_i}\left(\frac{\deg(A_k)}{\max_{f \in field}(\deg(A_f))}\right) < .5 \ \& \ \frac{\deg(A_j)}{\max_{f \in field}(\deg(A_f))} > .5\right]} * \max_{k \in Team_i} Score_7(A_j \rightarrow A_k)$<br>where deg $(A_j)$ represents the degree centrality of agent $A_j$ in the simulated oncofertility network |
| M8 | $Score_8(Team_i) = AvgCC_{Team_i} - ERcc$<br>where $AvgCC_{Team_i}$ represents the average clustering coefficient of $Team_i$ neighborhood, and $ERcc$ represents the average clustering coefficient of a random simulated network with the same size as $Team_i$ neighborhood |

| | |
|---|---|
| M9 | $Score_9(Team_i) = AvgCC_{Team_i} - CC_{Team_i}$<br>where $AvgCC_{Team_i}$ represents the average clustering coefficient of $Team_i$ neighborhood, and $CC_{Team_i}$represents the clustering coefficient of $Team_i$ |

**REFERENCES**


Akkermans, H. 2001. "Emergent Supply Networks: System Dynamics Simulation of Adaptive Supply Agents." In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, edited by Institute of Electrical and Electronics Engineers, Inc.

Barabási, A.-L., and R. Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439):509-512.

Bettencourt, L. M. A., D. I. Kaiser, J. Kaur, C. Castillo-Chávez, and D. E. Wojick. 2008. "Population Modeling of the Emergence and Development of Scientific Fields." *Scientometrics* 75 (3):495-518.

Bozeman, B., and E. Corley. 2004. "Scientists' Collaboration Strategies: Implications for Scientific and Technical Human Capital." *Research Policy* 33 (4):599-616.

Cohen, S. G., and D. E. Bailey. 1997. "What Makes Teams Work: Group Effectiveness Research from the Shop Floor to the Executive Suite." *Journal of management* 23 (3):239-290.

Contractor, N. S. 2013. "Some Assembly Required: Leveraging Web Science to Understand and Enable Team Assembly." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (1987).

Contractor, N. S., S. Wasserman, and K. Faust. 2006. "Testing Multi-Theoretical Multilevel Hypotheses About Organizational Networks: An Analytic Framework and Empirical Example." *Academy of Management Review* 31:681-703.

Cummings, J. N., and S. Kiesler. 2005. "Collaborative Research across Disciplinary and Organizational Boundaries." *Social Studies of Science* 35 (5):703-722.

Forrester, J. W. 1961. *Industrial Dynamics*. Cambridge, MA: MIT Press.

Guimera, R., B. Uzzi, J. Spiro, and L. A. N. Amaral. 2005. "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance." *Science* 308 (5722):697-702.

Hethcote, H. W. 2000. "The Mathematics of Infectious Diseases." *SIAM review* 42 (4):599-653.

Holland, J. H. 1998. *Emergence: From Chaos to Order*. Oxford: Oxford University Press.

Lattila, L., P. Hilletofth, and B. Lin. 2010. "Hybrid Simulation Models - When, Why, How?" *Expert Systems with Applications* 37 (12):7969-7975.

Levine, J. M., and R. L. Moreland. 1998. "Small Groups." In *The Handbook of Social Psychology*, edited by D. T. Gilbert, S. T. Fiske, and G. Lindzey, 415-469. Boston & New York: McGraw-Hill

Lungeanu, A., and N. S. Contractor. 2015. "The Effects of Diversity and Network Ties on Innovations: The Emergence of a New Scientific Field." *American Behavioral Scientist* 59 (5):548-564.

Lungeanu, A., Y. Huang, and N. S. Contractor. 2014. "Understanding the Assembly of Interdisciplinary Teams and Its Impact on Performance." *Journal of informetrics* 8 (1):59-70.

Macal, C. M. 2010. "To Agent-Based Simulation from System Dynamics." In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 371-382. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Macy, M. W., and R. Willer. 2002. "From Factors to Actors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology* 28:143-166.

Moody, J. 2004. "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999." *American Sociological Review* 69 (2):213-238.

Newman, M. E. J. 2001. "The Structure of Scientific Collaboration Networks." *Proceedings of the National Academy of Sciences* 98 (2):404.

Newman, M. E. J. 2002. "Assortative Mixing in Networks." *Physical Review Letters* 89 (20):208701.

North, M. J., and C. M. Macal. 2007. *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation*. Oxford, UK: Oxford University Press.

O'Leary, M. B., M. Mortensen, and A. W. Woolley. 2011. "Multiple Team Membership: A Theoretical Model of Its Effects on Productivity and Learning for Individuals and Teams." *Academy of Management Review* 36 (3):461-478.

Poole, M. S., and N. S. Contractor. 2011. "Conceptualizing the Multiteam System as an Ecosystem of Networked Groups." In *Multiteam Systems: An Organizational Form for Dynamic and Complex Environments*, edited by S. J. Zaccaro, M. A. Marks, and L. A. DeChurch. New York, NY: Routledge Academic.

Sterman, J. D., and J. Wittenberg. 1999. "Path Dependence, Competition, and Succession in the Dynamics of Scientific Revolution." *Organization Science* 10 (3):322-341.

Stonedahl, F., and U. Wilensky. 2010. BehaviorSearch [Computer Software], Center for Connected Learning and Computer Based Modeling, Northwestern University, Evanston, IL.

Sun, X., J. Kaur, S. Milojević, A. Flammini, and F. Menczer. 2013. "Social Dynamics of Science." *Scientific reports* 3.

Teasley, S., and S. Wolinsky. 2001. "Scientific Collaborations at a Distance." *Science* 292 (5525):2254.

Thiele, J. 2014. "R Marries NetLogo: Introduction to the RNetLogo Package." *Journal of Statistical* 58 (2):1-41.

Thiele, J. C., W. Kurth, and V. Grimm. 2014. "Facilitating Parameter Estimation and Sensitivity Analysis of Agent-Based Models: A Cookbook Using NetLogo and 'R'." *Journal of Artificial Societies and Social Simulation* 17 (3).

Uzzi, B., and J. Spiro. 2005. "Collaboration and Creativity: The Small World Problem." *American Journal of Sociology* 111 (2):447-504.

Wilensky, U. 1999. NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

## AUTHOR BIOGRAPHIES

**ALINA LUNGEANU** is a Ph.D. candidate in Technology and Social Behavior at Northwestern University. Her research examines the assembly of scientific teams and their effect on the emergence and evolution of new scientific fields. Her email address is alina.lungeanu@u.northwestern.edu.

**SOPHIA SULLIVAN** is a data scientist at Think Big, A Teradata Company. She holds a Ph.D. in Industrial Engineering and Management Sciences from Northwestern University. Her e-mail address is sophiasull@gmail.com.

**URI WILENSKY** is Professor of Learning Sciences, Computer Science, and Complex Systems at Northwestern University. He is the founder and current director of the Center for Connected Learning and Computer-Based Modeling and a co-founder of the Northwestern Institute on Complex Systems. He is the author of the NetLogo language. His email address is uri@northwestern.edu.

**NOSHIR CONTRACTOR** is the Jane S. & William J. White Professor of Behavioral Sciences in the Departments of Industrial Engineering and Management Sciences, Communication Studies, and Management and Organizations at Northwestern University. His research investigates the emergence and outcomes of social and knowledge networks. His email address is nosh@northwestern.edu.