Going Beyond the Data: Empirical Validation Leading to Grounded Theory

Craig Schreiber
Kathleen Carley
Institute for Software Research, International
Center for the Computational Analysis of Social and Organizational Systems
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

The purpose of this study is two-fold. First, a validation study on Construct-TM is conducted to show that modeling the actual and cognitive knowledge networks of a group can produce agent interactions within the model that correlate significantly with the communication network obtained from empirical data. Second, empirically grounded theory is produced by combining empirical data with simulation experiments run on empirically validated models.

Contact:
Craig Schreiber
Institute for Software Research, International
Center for the Computational Analysis of Social and Organizational Systems
Carnegie Mellon University
Wean Hall 1325
Pittsburgh, Pennsylvania 15213
Tel: (412) 268-5866
Fax: (412) 268-2338
Email: craigs@andrew.cmu.edu

Keywords: Model Validation, Agent-Based Simulation, Computational Organization Theory, Communication, Knowledge Management, Knowledge Transfer, Transactive Memory, Databases, Information Technology

Acknowledgement: This work was supported in part by the NSF KDI IIS-9980109 and the NSF IGERT in CASOS-9972762. Additional support was provided by CASOS - the center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University (http://www.casos.ece.cmu.edu).

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 National Science Foundation or the U.S. government.

Going Beyond the Data: Empirical Validation Leading to Grounded Theory

Craig Schreiber and Kathleen Carley

Introduction

The purpose of this study is two-fold. First, a validation study on Construct-TM is conducted to show that modeling the actual and cognitive knowledge networks of a group can produce agent interactions within the model that correlate significantly with the communication network obtained from empirical data. Second, empirically grounded theory is produced by combining empirical data with simulation experiments run on empirically validated models.

Group data

Data was collected from five small groups, each in a different company. The main business function of each of the companies is as follows: Company A – regional office/headquarters of a professional association, Company B – aeronautics, Company C – consulting firm, Company D – aeronautics and Company E – university. The group data was collected as part of the KDI project and using the IKNOW questionnaire. The groups, which ranged in size from 9 to 13 members, all received the same questionnaire and each used a database for knowledge storage and retrieval. The specifics such as people, tasks and knowledge areas varied by group but the questions on the specifics were the same.

Besides database use and group size, the groups differed greatly and are considered independent of each other. The results for each group are separate analyses and any comparisons are for analytic clarity. Each of the groups require different knowledge areas (expertise) for the performance of tasks and the tasks goals are also different. For instance, in the aeronautics groups B and D, one group is performing administrative functions and the other group is performing research and development functions. The number of major tasks that each group is responsible for varies with a range of 3 to 8 tasks. Also, a measure of task complexity was obtained by multiplying the knowledge areas and number of tasks. This measure reflects variance as well – the range is from 15 to 40.

Model

The model used was Construct-TM (Carley, 1990, 1991, 1995, 2002) (Carley and Schreiber, 2002). Construct-TM is a multi-agent model of group interaction whereas the agents communicate, learn, and make decisions in a continuous cycle. The Construct-TM model is used because it is able to represent actual knowledge, cognitive knowledge (transactive memory) and multiple interaction styles.

The representation of transactive memory is meaningful because referential data is a technological form of transactive memory. It is assumed that if an agent can use their own transactive memory to seek out another agent for knowledge then they will do so and not use the technological form. Otherwise, an agent without the ability to use their own transactive memory will always turn to the database and the usage and value of referential data could potentially be overstated.

Agents can interact using one of three interaction styles: relative similarity (homophily), knowledge seeking or a mixture of both relative similarity and knowledge seeking. The relative similarity interaction style calculates the probability of interaction based on the similarity of the agents. For example, agents with similar characteristics, such as knowledge, perceptions and demographics, will have high probabilities of interaction. Agents who are dissimilar on these characteristics will have low probabilities of interaction. The knowledge seeking style calculates interaction based on knowledge asymmetry so that agents with differing characteristics will have high probabilities of interaction and vice versa. The mixture style uses both the relative similarity and knowledge seeking styles based on a probability for how often each is used. The analysis of the separate interaction styles allows for the determination of which style is most predictive of actual group communication patterns.

Each agent has both actual and transactive knowledge and when interaction occurs either type of knowledge can be communicated. When actual knowledge is communicated the receiving agent simply learns the bit of knowledge. When transactive knowledge is communicated it is a referral from the transactive memory of agent X telling agent Y to go to agent Z for the knowledge. It is then up to agent Y to follow up with agent Z. The transactive memory of agents is not always accurate or complete and changes over time as the agents communicate and learn.

Communication Frequency Matrix

The member by member communication frequency matrix is built by compiling the responses from a survey question of how frequently each member communicated with every other member in the group. These are selfreported frequencies so the matrix is not symmetrical. The frequencies range from 0 to 6 where 0 is never and 6 is once per day.

The Knowledge Bit Strings

The actual knowledge is a representation of what each member knows by knowledge area. The data from the survey rated each members expertise per knowledge area along a four point scale (0 = none, 1 = beginner, 2 = intermediate, 3 = expert). There are four different bit strings created for each member per group. A knowledge area is represented by bits on a binary string with a 0 meaning they did not know that piece of knowledge and a 1 meaning they did know that piece of knowledge. For string 1, three bits per knowledge area are used. So if a member was rated as having no knowledge of that area they receive 0's in all three bits. If a member was rated as either a beginner, intermediate or expert then they receive one, two or three 1's respectively. The length of this string is the number of knowledge areas times three and represents the distribution of actual knowledge for each member of the group. String 2 adds demographic data collected from the survey to the distribution of knowledge string 1. The demographic variables used are job type, gender, education and tenure. Strings 3 and 4 use nine bits instead of three to represent each knowledge area. Again, all 0's represented a member with no knowledge but a beginner, intermediate and expert are represented by two, four and six 1's respectively in the nine bits. String 3, like string 1, is only the distribution of actual knowledge. String 4 is string 3 plus the demographic data.

The cognitive knowledge is a representation of the transactive memory for every member in the group. This cognitive representation is the perception of what each member thinks every other member of the group knows about each knowledge area. The data from the survey rated each members transactive memory along a four point scale (0 = none, 1 = beginner, 2 = intermediate, 3 = expert). The cognitive knowledge is represented in the same binary four string scheme as actual knowledge.

The rational of the four different strings/sets is to test which representations will offer the highest correlation with the communication frequency patterns of the groups. String 1 tests if the knowledge representation alone correlates with the communication frequency patterns. String 2 tests the addition of demographic data to string 1. Strings 3 and 4 are like the strings 1 and 2 respectively but testing the sensitivity of the knowledge representation. Expanding the size of the knowledge representation in these two strings will determine if a more detailed knowledge representation correlates higher.[1]

Probability of Interaction Matrices

The Construct-TM probability of interaction matrix is calculated using the actual and cognitive knowledge parameter inputs as well as the interaction style choices of relative similarity, knowledge seeking or a mixture of both. The mixture choice is based on a probability of how often relative similarity or knowledge seeking is used. Previous research has shown that compared to a random assignment of groups, people dyadically interact using homophily (relative similarity) on average 60% of the time (McPherson, Smith-Lovin, 1987). The mixture choice is set according to this research so agents will interact using relative similarity 60% of the time and knowledge seeking 40% of the time. The output probabilities for the probability of interaction matrices are between 0 and 1 inclusive and the matrices are not symmetrical due to the use of perceptions of knowledge and the relative interaction styles.

Validation Design

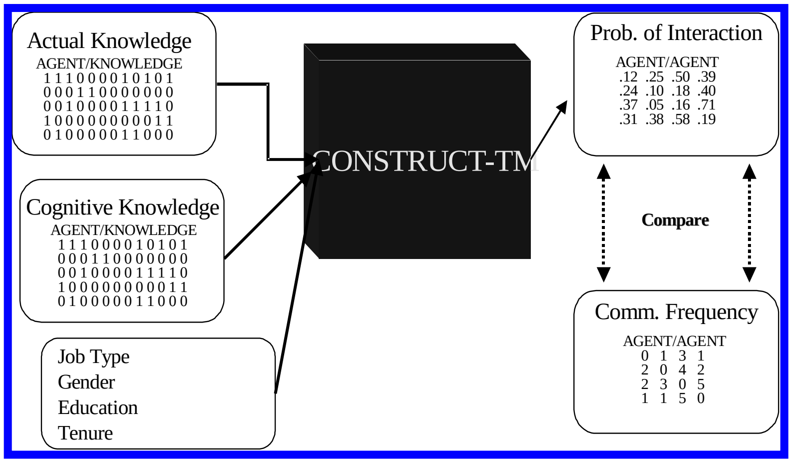
The validation uses a 5x4x3 experimental design, see Table 1. The survey responses are used as parameter input into the model and compares the agent probability of interaction matrix output from Construct-TM to the communication frequency matrix obtained from the survey, see Figure 1. There are sixty probability of interaction matrices output from the model, twelve per group. QAP correlation is used to compare the matrices. QAP correlation is a nonlinear, non-parametric method for comparing relational data.

Table 1 Experimental Design for Validation Study

Variable	Description	Values
Group	Knowledge representations for each group	5 Groups
Knowledge representations	Actual and cognitive knowledge parameters + demographics	4 Strings
Interaction style	Type of interaction	3 Styles

[1] A bit string with just demographic data was also tested but this string showed very low correlations.

Figure 1



Validation Results

Table 2 displays the results of the correlation analysis. Groups B, C and E show significant correlation results across conditions. Groups A and D do not correlate significantly with the Construct-TM model.

Group B has the most significant correlations of any of the groups. The mixed interaction style explains the communication frequencies for group B the best so members of this group are interacting using both relative similarity and knowledge seeking. Knowledge sharing in this company is fostered and this is reflected here as members of the group are seeking expertise. Demographic data adds very little to the correlations so these variables are not important to communication within this group. This is not surprising. Since members are seeking knowledge, demographics are less important as an interaction protocol. The more detailed knowledge strings are able to capture the knowledge seeking style of interaction much better than the less detailed knowledge strings. This result holds for all companies whether or not the correlations are significant.

Table 2 Real World Group Correlations

		String 1	String 2	String 3	String 4
Group A	Similarity	0.178*	0.084	0.124	0.082
	Mixed	0.122	0.06	0.113	0.084
	Seeking	-0.005	0.036	0.041	0.069
Group B	Similarity	**0.442****	**0.44****	**0.377****	**0.392****
	Mixed	**0.457****	**0.458****	**0.447****	**0.453****
	Seeking	**0.18****	0.2*	**0.428****	**0.422****
Group C	Similarity	**0.427****	**0.459****	**0.455****	**0.444****
	Mixed	**0.316****	**0.363****	**0.309****	**0.34****
	Seeking	0.013	0.052	0.18*	0.178*
Group D	Similarity	0.167	0.107	0.133	0.068
	Mixed	0.107	0.108	0.098	0.103
	Seeking	0.116	0.081	0.132	0.116
Group E	Similarity	**0.242****	0.177*	**0.278****	0.221*
	Mixed	**0.223****	0.094	**0.236****	0.16
	Seeking	0.085	-0.163	0.157*	0.046

* - p < .05
** - p < .01

Group C's results show that the members of this group interact based on relative similarity more than the other interaction styles. Out of all the groups analyzed, group C's members worked the most independently. In fact, members of this group indicated that they rely very little on the expertise of other members. Also, the representation of demographic data does add some value to the correlations so collection and use of this data is worthwhile. Because the members rely so little on the expertise of others it makes sense that relative similarity is the predominant interaction style and that demographic variables have a play in interactions.

Group E's interaction is also based on relative similarity. There are not as many significant correlations for group E as there are for groups B and C. Knowledge seeking does not figure into this scenario much at all and demographics decrease the significant correlations considerably. Group E is a university and considering the particular culture of this university these results are consistent. The group members do mostly independent work and this group had the lowest communication frequency ratings from the empirical data. There is not much interaction across the lines of academic disciplines. Therefore, it is no surprise that knowledge seeking is minimal as compared to other groups. When interaction does occur demographics are not important at all. The diversity of the university culture breaks down many demographic barriers that exist in most other situations. The lesser amount of significant correlations for group E are explained by the highly independent work, the lack of knowledge seeking and the low reliance on similarity for interaction.

Groups A and D did not have significant correlations and task interdependency is the reason. Members of groups A and D rated task interdependencies very high but members of groups B, C and E rated task interdependencies low to none. The problem is that Construct-TM does not represent task interdependencies. Therefore, the communication frequency patterns associated with such interdependencies cannot be approximated. The model lacks correlation with the empirical data in these cases.

Another possible explanation in addition to task complexity is a lack of proper knowledge representation. The knowledge representations, including demographics, may not be detailed enough. The complexity of groups A and D may require a very detailed knowledge categorization in order to detect and approximate the communication among members.

An Empirically Validated What-if Analysis: Going Beyond the Data to Produce Theory

A what-if analysis is conducted on groups B, C and E because of the significant correlation that Construct-TM has with the communication frequency networks of these groups. The purpose of this virtual experiment is to determine the impact that database use and data type has on performance.

Hollingshead and Contractor (2002) describe how groups can be represented as communication networks, actual knowledge networks or cognitive knowledge networks. Most studies use only one or two of the representations. This study uses all three. The actual knowledge networks and cognitive knowledge networks of each group are input into the model. The predominant mode of interaction as identified in the validation study is used so the highest significantly correlated communication network is represented. By representing groups as communication networks, actual knowledge networks and cognitive knowledge networks, the Construct-TM model affords complex analysis. The results are monte carlo with 100 runs per cell.

What-if Analysis Results

Group B's performance improves with the use of a database. The predominant mode of interaction for group B is the mixed interaction style and both task and referential data help although the impact of referential data is slight. The slight impact of a referential database makes sense given the knowledge sharing culture of the group which leads to the development of a good transactive memory system.

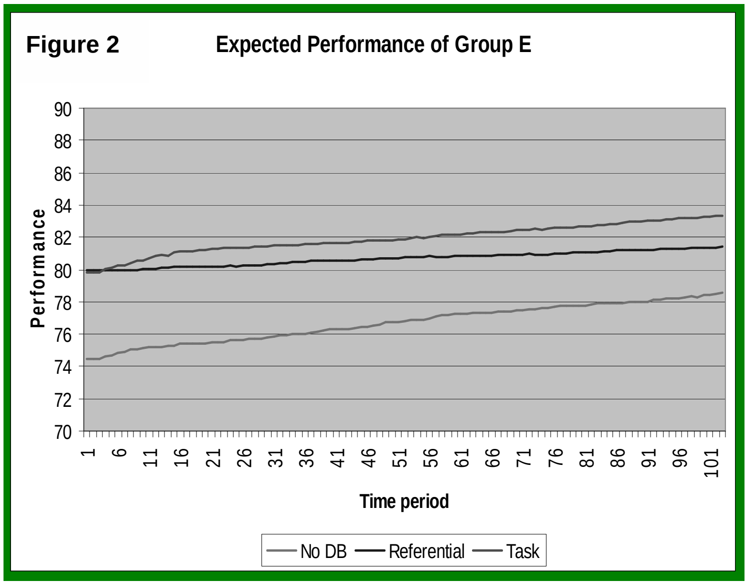


Figure 2 Expected Performance of Group E

Group C's performance improves with the use of a database. The predominant mode of interaction for group C is relative similarity. There is not much difference between the use of task and referential data but the use of referential data has more of an impact on group C than it did on group B. Since the members of group C perform tasks independently they have less transactive knowledge. The effect of referential data on performance makes sense because of this lack of accurate transactive knowledge.

Group E's performance improves with the use of a database, see Figure 2. The predominant mode of interaction for group E is relative similarity. Both task and referential data significantly increase performance. Of all the groups in this study group E gained the most improvement on performance while using referential data and this does make intuitive sense knowing the independent nature of the tasks for the group and the low knowledge seeking behavior as shown in the validation study. Due to the independent completion of work, this group had the most to gain from the use of referential data connecting members across expertise boundaries. Low knowledge seeking behavior and independent work habits is evidenced by this group having the lowest communication frequencies of any group. The ability of referential data to be a catalyst for interaction coupled with the low member interactions show that this group had a real potential for improvement.

The interesting phenomena is referential data improving performance for all the groups but especially for the groups interacting on a relative similarity basis. This was not expected due to previous studies (Carley and

Schreiber, 2002). What seems to be going on here is that groups who interact predominantly on relative similarity are limited in knowledge transfer opportunities. Such homopholous interaction restricts not only knowledge transfer but the development of transactive memory. Referential data, because it is a technological form of transactive memory, compensates for the lack of human transactive memory and directs members to others who have the required task knowledge - thus knowledge transfer can occur. The resulting theory is that referential data improves the knowledge transfer and performance for groups that interact on a basis of relative similarity.

Conclusion

The first purpose of this study is to do a validation of the Construct-TM model. By using the actual and cognitive knowledge of a group as input, Construct-TM did validate against empirical communication network data. The significant correlation shows that actual and cognitive knowledge are predictors of communication patterns. But the use of Construct-TM to represent an actual group is contingent on several factors. First, the task interdependencies must be understood. As the model stands now, it cannot represent these interdependencies. So, if there are high task interdependencies then the model is not a fit. Otherwise, when task interdependencies are low to medium the model does very well at representing the group in terms of communication networks. Second, the predominant mode of interaction needs to be identified. This identification through correlation provides a basis for proceeding with virtual experiments to provide understanding and predictions. By using the predominant mode of interaction we can have confidence that the results of such experiments are reflective of the group under study. Third, if the predominant mode of interaction is knowledge seeking then effort should be taken to represent knowledge in as much detail as possible. The results of this study show that high-level knowledge indicators provide weak correlation at best for this interaction mode. The change in correlation from high to low level indicators is so vast that the cost of any effort is justified. Lastly, demographic data may or may not provide explanatory value. Preliminary correlation analysis can easily indicate whether effort should be taken to include these variables as parameter inputs into the model.

The second purpose is to demonstrate that simulation combined with empirical model validation can go beyond the data and produce empirically grounded theory. By first empirically validating the model, we can use what-if analyses to look at issues that would be impossible to examine empirically. This study produced a theory of referential data knowledge transfer. Groups that interact homopholously can benefit from the use of a referential database because referential data crosses over homogeneous and cognitive barriers. When this occurs, referential databases can have a considerable positive impact on group performance.

More validation studies need to be undertaken. Validating a computational model is a difficult and tedious task, especially with the collection of social network data such as the frequency of communication matrices used in this study. But the difficulty and tediousness pays off as validation is tremendously beneficial. A successful validation allows us to have confidence in the analyses and predictions obtained from using the model. Besides such confidence, validation offers an opportunity to learn the strengths of the model and to understand the boundaries beyond which the model begins to break down. This study has succeeded on all three of these fronts. The field of computational modeling is lacking an emphasis on validation studies. An increase in validation studies will further the improvement of our models and our field.

References

Carley, K. M. (1990). Group Stability: A Socio-Cognitive Approach. In Lawler E., Markovsky B., Ridgeway C., and Walker H. (Eds.) *Advances in Group Processes: Theory & Research.* Vol. VII. (pp. 1-44). Greenwhich, CN:JAI Press.

Carley, K. M. (1991). "A Theory of Group Stability." *American Sociological Review*, 56(3): 331-354.

Carley, K. M. (1995). Communication Technologies & Their Effect on Cultural Homogeneity, Consensus, & the Diffusion of New Ideas. *Sociological Perspectives*, 38(4): 547-571.

Carley, K. M. (2002). Smart Agents and Organizations of the Future. In L. Lievrouw & S. Livingstone (eds), *The Handbook of New Media*, (pp. 206-220). Thousand Oaks, CA: Sage.

Carley K. M. and Schreiber, C. (2002). *Information Technology and Knowledge Distribution in C^3I Teams*, 2002 Command and Control Research and Technology Symposium, Monterey, CA.

Hollingshead, A. and Contractor, N. (2002). New media and organizing at the group level. In L. Lievrouw & S. Livingstone (eds), *The Handbook of New Media*, (pp. 221-235). Thousand Oaks, CA: Sage.

McPherson, J. M. and Smith-Lovin, L. (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review*, 52, 370-379.