Proceedings of the 2011 Winter Simulation Conference
S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds.

SIMULATION VALIDATION USING CAUSAL INFERENCE THEORY WITH MORPHOLOGICAL CONSTRAINTS

William N. Reynolds

Least Squares Software, Inc.
12231 Academy Rd. NE #301-192
Albuquerque, NM 87111, USA

Frank Wimberly

Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

We present an approach for the validation of complex simulation based on the structured elicitation of expert knowledge. Knowledge capture is based on the technique of Morphological Analysis, which is used to capture expert information on causal linkages and constraints in a systems and its simulation representation. This information is combined with Causal Inference Theory arguments to develop assertions about statistical dependency relations that should exist in both system and simulation. Causal Techniques for conducting these tests, which include the elicited constraint information are described. Overviews of Morphological Analysis, Causal Inference Theory and Statistical Testing Approaches are provided in the context of a Bayesian simulation of an example problem.

1 INTRODUCTION: BREADTH-DEPTH SIMULATION VALIDATION

The fundamental idea behind any validation approach is *comparison.* The scientific method is a comparison between ground truth (experiment) and a theoretical model (Popper 2001). In situations where ground truth is unavailable, comparisons can still be made – in social science parlance, this is *triangulation* (Brewer and Hunter 2006, Reynolds et al. 2010).

For complex simulation, a general validation approach has been triangulation against expert knowledge, known as *Face Validation* (DoD 2006, Zacharias, MacMillan, and Van Hemel 2008, Brewer and Hunter 2006, Brade 2004, Sargent 1999). Although relying on expert knowledge remains problematic (Tetlock 2006), it remains a central part of any validation methodology. A more sophisticated version of triangulation against expert knowledge is *model docking* - the comparison of two detailed simulations (Zacharias, MacMillan, and Van Hemel 2008, Axtell et al. 1996).

In Reynolds et al. (2010), one of the authors suggests an intermediate approach between face validation and model docking – *breadth-depth* validation. This approach argues that a central consideration in simulation is *cost* – in hours, data collection, experiment and operator and computer time. Simulation is cheaper than experiment and is consequently a desirable proxy. This too is a central consideration in validation – experimental validation can be very expensive and model docking requires huge effort. Face validation by subject matter experts is, relatively speaking, extremely inexpensive. *Breadth-Depth* is a framework that tries to elucidate these arguments by characterizing the cost of a given descriptive framework, from human expertise to high-resolution simulation.

Reynolds (2010) describes a validation framework based on the structuring of expert knowledge using Morphological Analysis (MA). MA is a modeling framework meant to rapidly and cheaply capture expert knowledge and use it to infer allowed configurations of a complex system (Zwicky 1966, Ritchey 2010). Comparing allowed configurations provides a generalized validation framework that is almost as flexible and inexpensive as expert-based validation. There is good evidence that structuring reasoning processes leads to improvements in cognitive processes (Gawande 2009). It is therefore expected that

structured validation approaches like MA will exhibit improved efficacy compared to unstructured face validation.

In this paper, we provide a variation on breadth-level structuring of expert knowledge. The previous work has focused on using MA to solicit constraints from SMEs to validate complex simulation; however, expert knowledge is often too rich to capture with elementary constraints. The present approach tries to address this problem by enriching the vocabulary that experts have to specify the behavior of a system, while retaining the simplicity and expressive power of breadth approaches like MA.

A powerful framework for expressing relationships in complex systems is the *causal graph* - a mapping which takes elements and causal relationships between those elements into a graph consisting of vertices and edges. In this formulation, vertices, or nodes, represent elements of a system and edges the causal dependencies between those elements (Pearl 2000, Jensen 2000).

Over the last two decades, groups at Carnegie Mellon University (CMU) and University of California at Los Angeles (UCLA) have developed *Causal Inference Theory (CIT)* - a formal description of the statistics implied by the relationships represented by causal graphs. CIT enables statistical tests to be performed on *observational* data to determine causal relationships between variables in a system. The fact that CIT uses observational data is important – it means that many important causal facts can be determined *without* performing experiment on a system. For complex, real-world systems, experiment is often impractical or immoral (e.g., requiring subjects to smoke), so the ability to operate on observational data is crucial. From a validation standpoint, this reduces the complexity of the validation process – causal facts can be determined without complex interventions into a system or simulation. Naturally, more advanced applications of CIT can suggest experiments that can be conducted that reveal an optimal amount of causal information (Pearl 2000; Spirtes, Glymour, and Scheines 2000).

In the present work, we develop a hybrid MA/CIT-based approach for simulation validation. The approach consists of eliciting, from experts, causal relationships and system constraints using MA. Given the elicited causal relationships, CIT provides statistical tests that can be performed on simulation data to confirm or falsify these relationships, which provides a quantitative triangulation of the expert's knowledge against the simulation. MA-elicited constraints modify the standard CIT procedures in a way that we argue will improve the reliability of the tests. These improvements should be realized in testing of either simulation or real-world data.

2 REVIEW OF CIT

Causal inference theory seeks to identify causal relations among random variables via statistical tests on the observed joint distributions of those variables. Colloquially, if we have a large amount of observational data on the states of a system, CIT enables us to figure out some of the causes and effects in a system.

As a simple example of a causal system and tests suggested by CIT consider the water level in a river. The river is fed by melting snow and by a smaller tributary upstream. Rafters travel the river when the water is high. We give a model of this system, along with potential interactions between the factors, in Figure 1.

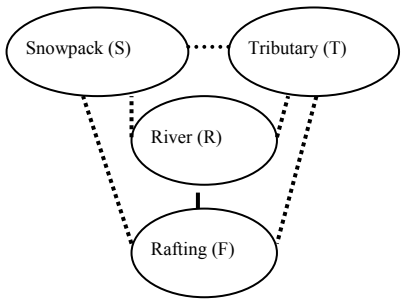

Figure 1: Snowpack-River-Tributary-Rafting (SRTF) model

The nodes in this graph represent our variables: *River (R), Snowpack (S)* and *Tributary (T)* and *Rafting (F)*. The edges capture our notion that these variables could be "related."

Now, by observing the behavior of the snow-river-tributary-rafting (SRTF) system, we would like to test the causal assertions made in the observations above – how does observational data tell us that *S* affects *R* and not vice-versa? This is the question addressed by Causal Inference Theory (Spirtes, Glymour, and Scheines 2000; Pearl 2000).

The fundamental idea of CIT is to identify *statistical independence* relations implied by a causal relationship between variables. It turns out that different causal relations between variables imply different conditional independence relations. By using standard statistical tests to check various combinations of relations (Sokal and Rohlf 1999), investigators may confirm or falsify hypothesized causal relations using observational data. The fundamental argument of this paper is that CIT can be a valuable validation technique for complex simulation: one can capture expert knowledge in the form of causal relations and test that knowledge against observed simulation data.

Consider a fragment of the SRTF system, the relationship between *Snowpack* and *River.* The model is represented as $S \rightarrow R$. This diagram implies that *S* causally affects values of *R.* Intuitively, variations in *S* should lead to variations in *R*. If an investigator had control over either of these variables, experiments could be conducted; however, in the absence of such control, observational data must be used. If only observational data is available, then we must test the statistical statement of our intuition that *R* is *not independent* of *S*.

Formally, the independence of *S* and *R* is expressed using the symbol $\perp\!\!\!\perp$: $S \perp\!\!\!\perp R$. In terms of the joint probability distribution, P(R,S), the relation $S \perp\!\!\!\perp R$ is expressed: $S \perp\!\!\!\perp R \rightarrow P(R,S) = P(R)P(S)$.

The causal relation $S \rightarrow R$ implies that S and R are not independent: $\neg (S \perp\!\!\!\perp R)$. In terms of distributions: $\neg (S \perp\!\!\!\perp R) \rightarrow P(R,S) \mathrel{!}= P(R)P(S)$. More than this cannot be inferred using only observational data – i.e., for two variables, independence captures the edge, but not the arrow.

We now consider a more involved model fragment, a causal "chain," $S \rightarrow R \rightarrow F$. There are three causally related variables: *S*, *R* and *F – R* depends on *S* and *F* depends on *R*. As before, this implies two simple dependence relations: $\neg (S \perp\!\!\!\perp R) \rightarrow P(R,S) \mathrel{!}= P(R)P(S)$ and $\neg (R \perp\!\!\!\perp F) \rightarrow P(R,F) \mathrel{!}= P(R)P(F)$.

However, there is more to this model - intuition tells us that as we vary *S*, this should affect the level of *R* which will in turn affect the value of *F*, therefore *S* and *R* are not independent. However, if we were to only consider *fixed* values of *R*, then the relationship between *S* and *F* should go away – in other words, *S* and *F* are independent given *R*. This is expressed as a *conditional independence relation* $S \perp\!\!\!\perp F \mid R$. In terms of the joint conditional probability distribution, $P(S,F|R)$: $S \perp\!\!\!\perp F \mid R \rightarrow P(S,F|R) = P(S|R)P(F|R)$.

We should note that conditional independence does not imply unconditional independence and vice-versa. The above relation does *not* imply that $S \perp\!\!\!\perp F \rightarrow P(S,F) = P(S)P(F)$ (a relationship that is not actually true in this case). In fact, exploiting the non-equivalence of conditional vs. unconditional dependence lies at the heart of CIT. Consider an airplane with two control yokes. One pilot turns his yoke to the left, but he observes the direction of the plane does not change; he can than infer the state of the other control yoke – it is turned to the right. Knowing the attitude of the plane (the collider node) couples the information about the state of the parent nodes (the control yokes).

We now consider a third model fragment, the *common cause:* $F \leftarrow R \rightarrow H$, here we have two variables, *F* and *H,* (*Trout Fishing,* which we have added to the SRTF model for this illustration), that both depend on *R,* this is known as a *common cause* or a *fork.* Here, the causal edges imply $\neg (R \perp\!\!\!\perp H)$ and $\neg (R \perp\!\!\!\perp F)$. Of course, *F* and *H* will be related, since they both depend on *R*; however, if we consider only *fixed* values of *R,* then the relationship between *H* and *F* should disappear, thus for the shared cause, we have: $F \perp\!\!\!\perp H \mid R \rightarrow P(F,H|R) = P(F|R)P(H|R)$.

The final model fragment is a "causal collider," $S \rightarrow R \leftarrow T$ In this model, *R* is causally dependent on both *S* and *T*. As before, the causal edges imply that $\neg (S \perp\!\!\!\perp R)$ and $\neg (T \perp\!\!\!\perp R)$. However, if we consider fixed values of *R*, then the state of *T* and *S* become related – if *R* = low and *S* = low, then we know that *T* != high. Thus, *R* and *T* are conditionally dependent given R: $\neg (S \perp\!\!\!\perp T \mid R) \rightarrow P(S,T|R) \mathrel{!}= P(S|R)P(T|R)$.

To review, we have the following mapping between graph structures and independence relations. *Edge:* A→ B : ┐(A⫫B). *The Chain:* A→ B → C: ┐(A⫫B), ┐(B⫫C), A ⫫ B | C. *The Shared Cause:* A←B→C: ┐(A⫫B), ┐(B⫫C), A ⫫ C | B. *The Collider:* A→B←C: ┐(A⫫B), ┐(B⫫C), A ⫫ C, ┐(A ⫫ C|B).

This provides a set of statistical "fingerprints" that can be used to check causal relations through statistical tests on observational data. There are few things to observe – one is that the "chain" and the "shared" cause have identical fingerprints – this is an instance of *Observational Equivalence* (also called *Markov Equivalence)* – two graphs that share the same skeleton (the structure left when arrows are removed from the graph) and the same collider sites cannot be distinguished through statistical observation alone (Pearl 2000, 19; Spirtes, Glymour, and Scheines 2000, 300). To discriminate between these cases, one must conduct experiments, or disambiguate them by identifying colliders elsewhere on the graph. Intuitively, if one can go from one structure to another by flipping arrows that do not destroy or create colliders, then the two structures are observational/Markov equivalent. For example, this implies that we can reverse the direction of a chain, and assuming there are not colliders created or destroyed, then the reversed system is observationally indistinguishable from the original.

For simple models, like the one in Figure 1, one can often "read off" the dependence relations from the graph. Formally, the process for "reading off" dependency relations is applying the Causal Markov Condition (CMC) (Spirtes, Glymour, and Scheines 2000, 29); this yields a set of dependency relations which are equivalent and can be reduced to the dependencies generated by d-separation (see below). The CMC simply says that conditioned on its parents, a node is independent of all other nodes that aren't its parents or descendants. The CMC typically yields a complex and redundant set of independence relations – the d-separation procedure yields a minimal set of dependency relations. For complex models, the procedure is not so straightforward; for example, if A and B collide into state C, but they share a common cause, then we can no longer say that they are independent; they are independent given the common cause and are dependent given the common cause and C.

The general procedure for identifying dependence relations from causal graphical models is known as *d-separation* (Pearl 2000; Spirtes, Glymour, and Scheines 2000). Two nodes that are graphically d-separated are conditionally independent in the associated probability distribution. Nodes that are not d-separated are said to be *d-connected.* The d-separation procedure is a systematization of the above informal arguments (Pearl 2000; Spirtes, Glymour, and Scheines 2000). Wimberly (2010) provides a java applet that identifies when two nodes on a graph are d-separated by a third set of nodes.

Returning to the SRTF example, we see that there is only one observationally equivalent structure – inverting any arrow in the diagram would lead to the creation or destruction of a collider, thus tests we conduct on the observational data will confirm or falsify only this model.

To see how these assumptions are used and to see how d-separation facts (graphical) are related to statistical facts consider the case of four variables X1, X2, X3 and X4 which are related by the following conditional independence facts: X1 ⫫ X2, X ⫫X4 |X3 and X2 ⫫ X4 | X3. The task is to find a graph consistent with those statistical facts. We illustrate the procedure in Figure 2.

It is also necessary that data not be aggregated from numerous systems – for example, protein concentration data aggregated from large numbers of cellular systems leads to profound problems of computational complexity when trying to perform automated learning of gene regulatory networks (Wimberly et al. 2010).

Since the graphical objects that represent causal networks are the graphs of Bayesian networks, it would appear that the methods apply only to systems that can be represented with Directed Acyclic Graphs. However there are techniques that permit analysis of systems with feedback. Richardson (1996) has developed the Cyclic Causal Discovery (CCD) Algorithm. That and other Tetrad (Tetrad 2010) algorithms were tested on the problem of inferring genetic regulatory networks from microarray data (Wimberly et al. 2010); gene networks are highly non-linear and have feedback.

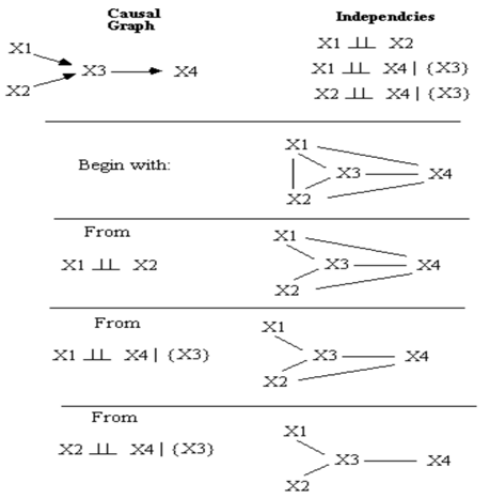

Figure 2: Illustration of graphical model discovery from independence relations

3 MA-BASED ELICITATION OF EXPERT KNOWLEDGE OF SYSTEM CAUSALITIES AND CONSTRAINTS

We now describe how we can integrate elicitation of expert knowledge using MA with CIT-based validation. Pursuing the SRTF model, we could go directly to construction of a causal graph of these variables, however, we will illustrate how an MA approach can be used to support the generation of this graph. We first assign states to the variables in the SME's model: *S: {low, high}; T: {low, high}, R: {low, medium, high}* and *F :{no, yes}*. We construct the following MA table:

Table 1: Morphological Analysis of SRTF Problem

		Snowpack		Tributary		River Level		
		low	high	low	high	low	medium	High
Snowpack	low							
	high							
Tributary	low							
	high							
River Level	low		A		B			
	medium							
	high							
Rafting	no							
	yes					C	D	

As in Reynolds (2010), we have indicated the expert-identified constraints with letters in the table. The expert's reasoning is as follows:

A) *Snowpack = high* implies that the river cannot be *low*, since runoff will cause water to rise.
B) *Tributary = high* implies that the river cannot be *low,* since upstream water will cause the water to rise.
C) *River Level = low* implies that *Rafting Activity* cannot be *yes*, since the rafting company will not conduct tours in low-water conditions.
D) *River Level = medium* implies that *Rafting Activity* cannot be *yes*, as in constraint C.

We can now use this information to initiate construction of a causal graph. Each constraint implies a causal linkage between the two variables involved – for example, constraint A implies that *River Level* is causally related to *Snowpack*. To see this quantitatively, we observe that constraint A implies that the joint probability, *P(River Level = low, Snowpack = high) = 0*.

If *Snowpack* and *River Level* were independent (which is to say, causally unrelated, see below), then *P(River Level = x, Snowpack = y) = P(River Level = x)P(Snowpack = y)* for all states *x* of *River Level* and all states *y* of *Snowpack.* This can only be true if either *P(River Level = low) = 0* and/or *P(Snowpack = high) = 0.* In other words, neither of these events would ever be observed, which, in turn, leads to substantive questions about the modeling framework. Therefore, when the expert places a constraint in a morphological table, and he believes that the relevant states could occur under some circumstances, he is making a statement that the two variables in that block of the table are statistically dependent, and hence causally related. Based on the MA in Table 1, we are led to Figure 2.

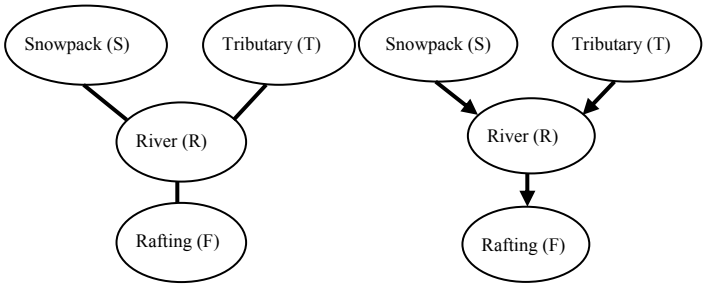

Figure 3: Relational and causal fragments implied by MA in Table 1 and expert knowledge

Note that the MA does not tell us the *causal direction* of the interactions between the variables, nor does an *absence of a linkage* in the fragment derived from the MA imply that there is an *absence of a causal relation* – further SME analysis may indeed indicate additional causal relations not indicated by the MA. However, this graph does tell us some important things: the absence of an edge between *Snowpack* and *Tributary* implies that the level of the snowpack is not "related" with level in the tributary (our experts tell us that it is spring-fed).

Although one could dispense with the MA step, and use standard causal elicitation techniques to generate the graphical model based on SME knowledge, it turns out there are sound technical reasons for using the MA approach. CIT tends to be at its weakest when deterministic causal relations are present – as we will show below, traditional independence tests break down for highly-constrained system. These are precisely the sort of relations implied by the constraints in the MA. As we will see, the presence of these strong constraints will lead to substantive modifications of the statistical procedures we use to test data generated by simulation (and, for that matter, tests on data from observations of the real-world system, if it is available).

A natural question at this point is what are all possible causal models that are implied by the MA-based capture of the expert's knowledge? This is a subtle question – for example, the link between Snowpack and River level could imply three things: Snowpack causes River, River causes Snowpack or a third factor could cause both Snowpack and River; or there is both a common third cause and Snowpack causes River or vice versa. The latent factor could be one included in the model, such as *Tributary*, or a factor not included in the model at all. This difficult question is an appropriate topic for further research; its answer undoubtedly is related to techniques developed for model discovery from data (Spirtes, Glymour, and Scheines 2000, chapter 5). A related question is what statistical tests are implied by the partial "oracular" information being provided by the expert. Again, our approach will remain incomplete at this point – we will assume that we can construct causal models from expert knowledge that provide sufficient certainty that we can identify tests to perform by inspection (in detail, we test for colliders). Systematizing this is a topic for future research.

Starting with the fragment in left of Figure 3, the expert can now refine the causal model, based on his understanding of the system. In our example, this result is given on the right in Figure 3. The expert has claimed that *Snowpack* causes *River*, as does *Tributary*. He has also asserted that *Tributary's* state is independent of *Snowpack,* as is *Rafting*, which is itself independent of Snowpack. This model has no

additional causal relations; in fact, it has stronger statements in that it asserts there are no direct relations between Snowpack, Tributary and Rafting. The direct relations that do exist have had their directions specified. Qualitatively, this model predicts that manipulations of River Level will not affect the Snowpack or Tributary; whereas manipulations in the Tributary or Snowpack levels will affect River Level.

4 CAUSAL INFERENCE THEORY FOR SIMULATION VALIDATION

For models that have many observationally equivalent structures, the question becomes which dependency relations to test? Since the only relations that can be unambiguously tested are colliders, we focus on the collider tests in our diagram. Having identified these, the identified two variable associations will often resolve much of the causal information in the system. From the point of view of MA, the only place where colliders can occur is where different blocks have constraints in the same column or row (of course, these could also be forks or chains; nevertheless, the only way to connect three variables is by sharing constraints in co-row or co-column blocks).

We now provide an explicit example of the mechanics of how to use observational data to check causal relations implied by a particular causal graph. The purpose of this is twofold – first, we demonstrate how causal constraints, such as those elicited from MA, can be incorporated into statistical tests, second, this section provides a brief, accessible overview to the use of statistical independence testing methodologies in CIT.

We are using *observational data* to infer causalities in the system. This means we sample the states of the system over some period of time; we do not perform experiments or control for any of the variables in the system. Questions of how causal information and identified ambiguities and data collection constraints should guide experimental design are questions for future research.

Using the CMU *Tetrad* causal inference tool (Tetrad 2010), we have constructed simulation of the Snowpack/River/Tributary/Rafting (SRTF) system. The model is a Bayesian statistical constrained to disallow states forbidden by the Morphological Analysis from Table 1. Implementing this model in Tetrad, we generated 10,000 observations of the system at regular time intervals. A contingency table of the data is presented in the left of Table 2. Each cell in the six constituent tables represents a state of the system – the numbers in the cells represent the number of times the system was observed to be in that state. The cells that are affected by the MA constraints from Table 1 are marked in yellow – note that the observed frequencies for these states are zero; thus the system passes the first MA-based validation of the model (Reynolds 2010). Marginal sums of the various rows, columns and tables are also given in Table 2. Some of these sums are entirely composed of constrained (yellow) cells, and are consequently constrained themselves. For example, the marginal sum over *S* in the {*F = No, R = Low, T = Low}* column is a sum of two observed frequencies – {*F = No, R = Low, T = High, S = Low}* and {*F = No, R = Low, T = High, S= High}*. Both of these terms are constrained to be zero, so the marginal sum is also constrained to be zero. This is an example of constraint propagation, an effect that will be instrumental when we compute the likelihood that these observations were generated by given causal models.

4.1 Validating Simple Dependency Statements of the Causal Model: The χ^2 and G Tests.

There are a number of two-variable dependency facts implied by the model of Figure 3. For example, *R* should depend on *S, T* and *F*. Each of these dependencies should be tested in order to validate the model. There are also indirect dependencies implied – for example, although there is no causal link between them, *Snowpack* should depend on *Rafting,* due to the mediating influence of *River*.

Testing these dependency facts is a standard exercise from statistics – numerous handbooks detail the procedures, our favorite is Sokal and Rohlf *(*1969). The general test procedure is to assume that the observed data were generated by an underlying statistical model. Expected values are generated from that model, whose parameters are calculated from the observed data. The observed values are then compared with the generated values. A *mismatch* function is calculated, and the likelihood of observing the

mismatch is estimated. If the likelihood of observing the mismatch is below a predetermined acceptance threshold, the model being tested is rejected; otherwise it is accepted. This will either confirm or falsify the underlying dependence relation that is being tested. The constraint information from the MA can be incorporated into this procedure. We will illustrate this procedure in the following examples.

Table 2: Data, Predicted Frequencies, Probabilities and Contingency Tables for Tetrad-Based River Simulation

Contingency Table of Values from Tetrad-Based River Simulation

F	No	Samples	7708				
	T				T		
S	Low	High		S	Low	High	
Low	849	0	849	Low	843	1258	2101
High	0	0	0	High	1268	1291	2559
	849	0			2111	2549	
	R	Low			R	Medium	
	Total	849			Total	4660	

	T		
Snowpack	Shut	Open	
Low	393	585	978
High	626	595	1221
	1019	1180	
	R	High	
	Total	2199	

F	Yes	Samples	2292				
	T				T		
S	Low	High		S	Low	High	
Low	0	0	0	Low	0	0	0
High	0	0	0	High	0	0	0
	0	0			0	0	
	R	Low			R	Medium	
	Total	0			Total	0	

	T		
S	Low	High	
Low	421	628	1049
High	645	598	1243
	1066	1226	
	R	High	
	Total	2292	

Total Samples	10000

Observed Frequencies (R, F Marginalized Out)

	T			
S	Low	High	N(S)	P(S)
Low	2506	2471	4977	0.4977
High	2539	2484	5023	0.5023
N(T)	5045	4955		
P(T)	0.505	0.5	Total (N)	10000

Predicted Probabilities, P(S,D)=P(S)P(T)

	T	
S	Low	
Low	0.2510897	0.2466
High	0.2534104	0.2489

Expected Frequencies N(S,D) = P(S)P(T)N

	T	
S	Low	High
Low	2510.9	2466.1
High	2534.1	2488.9

Contingency Tables of S,R and D.

	R = Low		R = Medium		
	T		T		
S	Low	High	High	Open	N(S|R)
Low	849	0	843	1258	2101
High	0	0	1268	1291	2559
N(T|R)	849	0	2111	2549	
	Total	849	Total	4660	

	R=High		
	T		
S	Low	High	
Low	814	1213	Total Samples
High	1271	1193	10000
N(T|R)	2085	2406	
	Total	4491	

We begin with a "vanilla" test that does not involve constraints. The causal model predicts that *S* and *T* should be statistically independent. We first generate the observed values of *S* and *T* by marginalizing out the other variables. This yields a marginal contingency table, given in Table 2. We test the hypothesis that *S* is independent of *T* - this implies the joint distribution can be factorized: $P(S,T) = P(S)P(T)$.

We compute the distributions $P(S)$ and $P(T)$ from the observational data - these are also given in Table 2: . We multiply these distributions to get the hypothetical joint distribution, and then multiply by the total number of samples (10,000) to get the predicted number of observations given the independence hypothesis, $P(S,T) = P(S)P(T)$.

Table 2 gives the observed and expected frequency tables. Each of the four corresponding cells differs across the two tables. How likely is it to observe these differences in the cells? The traditional procedure is the *Chi-square test*. In this test, we construct a *mismatch* between the observed and expected value in each cell of the two tables by taking the difference between cells, squaring it and normalizing by the expected value. The most well-known mismatch is the *Pearson's chi-square test statistic* or (χ^2), formally:

$$\chi^2 = \sum_{cells} \frac{(O-E)^2}{E} \tag{1}$$

where O is the observed value of a cell and E is the expected value of the cell.

The idea underlying this mismatch function is that if we assume each observed/expected cell difference is a normally distributed random variable, then each term in (1) is the square of a standard (zero mean, unit variance) normal variable. The sum of squares of N independent, standard normal random variables, X_i, is distributed according to the *Chi-Square Distribution with N degrees of freedom,* χ^2_N. The test statistic (1) is assumed to be distributed according to χ^2_N, $\Sigma_i^N X_i^2 \sim \chi^2_N$ (Sokal and Rohlf 1969). By comparing the mismatch function (X) with the distribution, χ^2_N, we can estimate the likelihood of observing the mismatch.

Another widely used mismatch function is G, given by:

$$G = \sum_{cells} O \ln(\frac{O}{E}) \tag{2}$$

G is employed identically to χ^2, with which it often agrees to several decimal places. G is currently preferred by practitioners, especially in cases where there are cells in the observed table where the mismatch is greater than the expected value |O-E| > E (Sokal and Rohlf 1969).

Naively, for S and T, we would expect there to be four degrees of freedom, since there are four cells in the mismatch table. However, this is not correct, since we are assuming that the observed values are drawn from the statistical process *P(S)P(T)*. Thus, we are not asking how likely it is to obtain the observed table from a space of all possible 2x2 contingency tables, but rather from the space of 2x2 contingency tables having 10,000 samples and relative proportions given by *P(S)* and *P(T)*. The four cells in the observed table *cannot* be varied independently. For example, if we were to observe an additional 10 states in the {*S = low, T = High}* cell, we would have to reduce the sum of the other three cells by 10 to maintain a total number of 10,000, and further adjust the values in these cells such that the proportions *P(S)* and *P(D)* were maintained. The number of *degrees of freedom* of the contingency table is given by the total number of cells minus the number of constraints associated with maintaining the total number of samples and the probabilities. For S and T, there are 2 constraints from the probabilities ($\Sigma P(S) = 1$, $\Sigma P(T) = 1$) *and* 1 constraint from the number of samples (N=10,000). Thus, there is 4 – 1 – 1 - 1 = **1** degree of freedom. Effectively, the quantity X (or G) calculated for S and T represents the square of a single standard normal variable. We know the mean of a standard normal variable is 0 and its variance is 1 – so, intuitively, the square of a single variable bounded by one standard deviation should be in the "ballpark" of $(0)^2 = 0$ to $(1)^2=1$. Using Table 2, we compute $\chi^2 = 0.03837$ (*G* is identical to 5 decimal places). Comparing this to the χ^2 distribution yields an 84% chance that we would observe this value for a sample generated by the distribution *P(S)P(T)*. From this we conclude that it is likely that S and T were generated by the process *P(S)P(T)* and consequently they are independent, which is consistent with the causal model.

Do we need to test independence in a table with an MA constraint? We have argued that a constraint implies association. A natural test is to examine whether constrained cells are observed in the data – if they are, there is a clear mismatch between the MA representation and the simulation (Reynolds 2010). Is this test sufficient? For some types of variables it will not - Bishop, Fienberg, and Holland (1973) provide the example of two variables – the scores of winners, S_W, and losers, S_L, in an ensemble of sporting events. An MA of S_L and S_W would imply constraints wherever $S_L > S_W$; however, this does not imply that

$S_L \perp\!\!\!\perp S_W$ in any scientifically interesting sense – we are interested in whether S_L and S_W are associated only *in the non-zero cells of the contingency table.* Such tests are termed by Bishop, Fienberg, and Holland (1973) as tests of *quasi-independence.* The researcher conducting the validation must determine whether MA-elicited constraints do imply causal relations or not. For our STRF model, the constraints are not artifacts of the way we have defined the variables. The fact that *S = high* disallows *R = low* implies a relation between the variables. Thus, the MA tests of the data are sufficient to indicate association between the variables (in this case *S-R, T-R, R-F).*

4.2 Validating Dependency Statements of the Causal Model: Colliders

So far we have described standard statistical correlation tests as a means for validating simulations based on expert knowledge – such approaches have been discussed in the validation literature (Brade 2003, Sargent 1999, DoD 2006). We would like to extend this dependency to analysis to test causal assertions from expert knowledge. As discussed in Section 2, causal facts follow from collider structures in the expert-derived graph – once we have established the validity of colliders, then other causal relations may be determined beyond the simple dependence relations.

In our example SRTF model, there is one collider structure: S→R←T. The independence relations implied by this structure are: S⫫T, ~(S⫫R), ~(T⫫R) and ~(S⫫T|R). We have confirmed the two-variable dependence relations in the previous section, we now test the relation ~(S⫫T|R), which will necessitate considering the MA constraints. We begin by marginalizing out the *F* variable, given in Table 2.

$$P(S,T,R) = P(S|R)P(T|R)P(R) \tag{3}$$

As before, we calculate the parameters of the test distribution from the observed data – see Table 2. Examining this calculation, we see that there is another way for MA constraints to propagate across contingency tables. In the conditional probability tables, the cells *P(S = High| R = Low)* and *P(T = High | R = Low)* are both constrained to 0, since all terms in the sums for *N(S = High| R = Low)* and *N(T = High | R = Low)* were constrained to be zero. In these sum calculations, *all* terms in the sum need to be zero for the constraint to be preserved across marginalization. However, the consequence of constraints that survive marginalization is quite pronounced in the joint tables. Since the joint tables are calculated by taking *products* of the conditional distributions, the zero constraints tend to propagate, since only a single zero operand is needed to render an entire product zero. In the example, the two constraints from the conditional probability distributions lead to three constraints in the joint distribution. In testing causal colliders, which involves joint distributions, constraints are more prevalent and will have an increased impact on the tests (although, since the number of variables increase, so to do the degrees of freedom – it is an open question whether the tendency for constraints to increase is mitigated by the increase in degrees of freedom). Note that the argument made in section 3 that constraints imply lack of independence in the full distributions does not apply to joint distributions, which may be completely zero for a given state of the conditioning variable – see the contingency tables for *{F = Yes}* in Table 2.

Having computed a contingency table of expected frequencies, we are now in position to compute the G or χ^2 mismatches from the expected and the observed frequencies. However, to compute the likelihood of observing these mismatches, we must also determine the degrees of freedom for the tables. There are 12 cells in the three tables, the constraints are:

- 1 constraint for the total number of samples (10,000).
- 3 constraints for the requirement that each column of the conditional distribution *P(S|R)* must sum to 1.
- 3 constraints for the requirement that each column of the conditional distribution *P(T|R)* must sum to 1.
- 1 constraint for the constraint that the distribution *P(R)* must sum to 1

- 3 constraints from the MA.

This leads to 12 – (1 + 3 + 3 + 1 + 3) = 1 degree of freedom. Computing G from the observed frequencies in Table 2 and the expected frequencies in Table 3 gives a value G = 100.05, ($\chi 2$ = 99.74) – this is a 10-standard deviation draw for a single random variable, which is to say vanishingly unlikely. This confirms that S, R and T form a causal collider.

Table 3: Computation of Expected Frequencies for Collider: N(S,T,R)

Observed Frequencies

N(R)	River (R) Low	Med	High
	849	4660	4491

Observed Probabilities

P(R)	River (R) Low	Med	High
	0.0849	0.466	0.4491

Observed Conditional Frequencies N(S|R)

Snowpack	River Low	Med	High
Low	849	2101	2027
High	0	2559	2464
	849	4660	4491

Total 10000

Conditional Probabilities P(S|R)

Snowpack	River Low	Med	High
Low	1	0.4509	0.4513
High	0	0.5491	0.5487
	1	1	1

Observed Conditional Frequencies N(D|R)

Tributary	River Low	Med	High
low	849	2111	2085
high	0	2549	2406

Conditional Probabilities P(D|R)

Tributary	River Low	Med	High
low	1	0.4530	0.4643
high	0	0.5470	0.5357

Predicted Conditional Probabilities P(S,D|R) = P(S|R)P(D|R)

Snow	River = Low, Tributary (T) Low	High
Low	1	0
High	0	0

Snow	River = Med, Tributary (T) Low	High
Low	0.2042	0.2466
High	0.2488	0.3004

Snow	River = High, Tributary (T) Low	High
Low	0.2095	0.2418
High	0.2547	0.2939

Predicted Total Probabilities P(S,R,D) = P(S|R)P(D|R)P(D)

Snow	River = Low, Tributary (T) Low	High
Low	0.085	0
High	0	0

Snow	River = Med, Tributary (T) Low	High
Low	0.0952	0.1149
High	0.1159	0.1400

Snow	River = High, Tributary (T) Low	High
Low	0.0941	0.1086
High	0.1144	0.1320

Predicted Frequencies N(S,R,D) = P(S|R)P(D|R)P(D)N

Snow	River = Low, Tributary (T) Low	High
Low	849	0
High	0	0

Snow	River = Med, Tributary (T) Low	High
Low	951.8	1149.2
High	1159.2	1399.8

Snow	River = High, Tributary (T) Low	High
Low	941.1	1085.9
High	1143.9	1320.1

When calculating G, we have to contend with some divergences in the summands Oln(O/E) – for some cells, both O and E are zero. Since frequencies cannot be negative, we take limits from above, yielding zero. Similar limiting procedures prevent divergences in calculating χ^2. Divergences in test statistics will occur in over-constrained systems, where probability densities go to 1. If a system is so constrained that it becomes deterministic, then statistical tests lose their meaning, and checks to make sure the constraints are not violated, as in (Reynolds 2010) are likely sufficient for validation.

A further point regarding constraints is that real data can have measurement error – for example, data on the river level may have been incorrectly assigned to bin *{Low}* rather than bin *{Medium}*. This can lead to non-zero frequencies in the constrained cells – the values in these cells could be small enough that an investigator does not perceive a constraint violation, and is willing to conduct a causal dependency analysis. For this case, how should the values in these cells be handled? One cannot simply set the expected value of these cells to zero, since that will cause the mismatch functions to diverge. Investigating this question in detail is a task for future work. At this point, the most reasonable procedure seems to be to proceed with the mismatch calculation retaining the observed low-frequency cells, which will lead to a small expected frequency and a well-defined mismatch term. As before, calculate the mismatch and reduce the degrees of freedom by the number of constraints.

5 CONCLUSIONS AND FUTURE WORK

We have presented a novel approach for simulation validation. Based on explicit and systematic elicitation of expert knowledge using structured Morphological Analysis, we use Causal Inference Theory to identify non-obvious statistical dependency relations that must exist in simulation (or real-world) data, to be consistent with experts' mental models. This provides a quantitative methodology for validation/ triangulation of expert knowledge with simulation. We have demonstrated the approach on a small model – an open question is the utility of this approach for very large models requiring very large elicitations of expert knowledge and consequent simulation tests.

There are a number of methodological questions left unanswered by this preliminary work. In particular, we have not explored the implications of validating expert knowledge that is incomplete or uncertain. Similar shortcomings exist for validating systems that violate some of the assumptions of Causal Inference Theory – although it seems reasonable that our approach will work on acyclic fragments of a system that is not generally acyclic, it remains to examine the limitations of such an approach.

Another unaddressed area is using expert elicited information to inform experimental interventions that should be conducted on simulations (or real-world) systems that provide the largest amount of validating information per test. There is a large amount of work that has been done on interventions in the literature (Pearl 2000; Spirtes, Glymour, and Scheines 2000) and we anticipate this will be a fruitful area for further study.

ACKNOWLEDGEMENTS

This material is in part based upon work supported by the Office of Naval Research under Contract No. N03N0001409C0612. We would also like to acknowledge the valuable insights of David Danks, Peter Brooks and Marta Weber. This work is patent pending.

REFERENCES

Axtell, Robert, R. Axelrod, J. M. Epstein, and M. D. Cohen. 1996. "Aligning Simulation Models: A Case Study and Results." *Computational & Mathematical Organization Theory* (1)2:123-131.

Bishop, Y. M. M., S. E. Fienberg, and F. W. Holland. 1975. *Discrete Multivariate Analysis,* Massachusetts Institute of Technology Press.

Brade, D. 2004. "A Generalized Process for the Verification and Validation of Models and Simulation Results." PhD Dissertation, Universität der Bundeswehr, Neubiberg.

Brewer, J., and A. Hunter. 2006. *Foundations of Multimethod Research: Synthesizing Styles*. Thousand Oaks, CA: Sage Publications.

DoD (U.S. Department of Defense) 2006. *VV&A Recommended Practices Guide Build 3.0*. Accessed 8 October 2010. http://vva.msco.mil/.

Gawande, A. 2009. *The Checklist Manifesto.* Metropolitan.

Jensen, F. V. 2000. *Bayesian Networks and Decision Graphs*. Springer.

Pearl, J. 2000. *Causality: Models, Reasoning and Inference.* Cambridge University Press, New York.

Popper, K. 2001. *All Life is Problem Solving*. Routledge.

Reynolds, W. N. 2010. "Breadth-Depth Triangulation for Validation of Modeling and Simulation of Complex Systems." In *IEEE Intelligence and Security Informatics Conference, Workshop on Current Issues in Predictive Approaches to Intelligence and Security Analytics (PAISA-10)*, 190-195. Vancouver, BC, Canada, May 26.

Reynolds, W. N., M. S. Weber, R. M. Farber, C. Corley, A. J. Cowell, and M. Gregory. 2010. "Social Media and Social Reality: Theory, Evidence and Validation." In *IEEE Intelligence and Security Informatics Conference, Workshop on Current Issues in Predictive Approaches to Intelligence and Security Analytics (PAISA-10)*, 221-226. Vancouver, BC, Canada, May 26.

Richardson, T. 1996. "Discovering Cyclic Causal Structure." Report CMU-PHIL-68, Department of Philosophy, Carnegie Mellon University, February. Accessed 15 October 2010. http://www.hss.cmu.edu/philosophy/techreports/68_Richardson.pdf.

Ritchey, T. 2010. *Wicked Problems/Social Messes: Decision Support Modelling with Morphological Analysis*. Swedish Morphological Society.

Sargent, R. G. 1999. "Validation and Verification of Simulation Models." In *Proceedings of the 1999 Winter Simulation Conference*, edited by P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, 39-48. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Spirtes, R., C. Glymour, and R. Scheines. 2000. *Causation, Prediction and Search.* MIT Press.

Sokal, R. R., and F. J. Rohlf. 1969. *Biometry: The Principles and Practice of Statistics in Biological Research,* W.H. Freeman and Company. San Francisco.

Tetlock, R. E. 2006. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Tetrad. 2010. "The TETRAD Project." Accessed 8 October 2010. http://www.phil.cmu.edu/projects/tetrad/.

Wimberly, F., D. Danks, C. Glymour, and T. Chu. 2010. "Problems for Structure Learning: Aggregation and Computational Complexity." In *Computational Methodologies In Gene Regulatory Networks*, edited by S. Das, D. Caragea, W. H. Hsu, and S. M. Welch, 310-333. Hershey, PA: IGI Global Publishing.

Wimberly, F. 2010. "D-Separation Applet" Accessed 8 October 2010. http://www.phil.cmu.edu/~wimberly/dsep/dSep.html.

Zacharias, G. L., J. MacMillan, and S. B. Van Hemel, eds. 2008. *Behavioral Modeling and Simulation: From Individuals to Society*. Committee on Organizational Modeling from Individuals to Societies, National Research Council of the National Academies. Washington, D.C.: National Academies Press.

Zwicky, F. 1966. *Discovery, Invention, Research through the Morphological Approach*. New York: The Macmillan Company.

AUTHOR BIOGRAPHIES

WILLIAM NASH REYNOLDS, President of Least Squares Software, has been a principal researcher and innovator in the field of complexity for twenty years. Over the past five years, Dr. Reynolds has been focused on the role of complex systems in intelligence analysis, directing his research toward complexity-based analytic methodologies of practical value to analysts in the intelligence community. Dr. Reynolds has been focusing on marrying analytic methodology with quantitative tools, focusing on sensing frameworks and simulation validation using Morphological Analysis, and most recently, Causal Inference Theory. Dr. Reynolds graduated in 1986, *magna cum laude,* with degrees in Physics and Comparative Literature at the University of Massachusetts. He followed this with a Ph.D. in Theoretical Physics from the University of California, San Diego in 1993. His e-mail address is bill@leastsquares.com.

FRANCIS C. WIMBERLY has worked in the area of scientific computing for over 35 years. He received a Ph.D. from the University of Pittsburgh where his dissertation addressed problems in finite element methods for structural analysis. After working at Bell Laboratories he joined Carnegie Mellon University and worked in multiple groups over the years including the Robotics Institute, Heinz School of Public Policy and Management, the Pittsburgh Supercomputing Center, and the Philosophy department. He is one of the developers of Tetrad, open source software hosted at CMU. In late 1998 he joined BiosGroup where he worked on agent-based software. He is currently working with his CMU philosophy department colleagues in assessing whether their causal reasoning techniques are applicable to the inverse problem for genetic regulatory networks. His e-mail address is wimberly3@gmail.com.