Interactive Goal Model Analysis Applied - Systematic Procedures versus Ad hoc Analysis

Jennifer Horkoff[1], Eric Yu[2], Arup Ghose[1]

[1] Department of Computer Science, University of Toronto, Canada
[2] Faculty of Information, University of Toronto, Canada
jenhork@cs.utoronto.ca, yu@ischool.utoronto.ca, arup.ghose@utoronto.ca

Abstract. Intentional modeling, capturing the goals of stakeholders, has been proposed as a means of early system elicitation and design for an enterprise, focusing on social and strategic requirements. It is often assumed that more utility can be gained from goal models by applying explicit analysis over models, but little work has been devoted to understand how or why this occurs. In this work we test existing hypotheses concerning interactive goal model analysis via multiple case studies. Previous results have indicated that such analysis increases model iteration, prompts further elicitation, and improves domain knowledge. Results of the new studies do not provide strong evidence to support these claims, showing that such benefits, when they occur, can occur both with systematic and ad-hoc model analysis. However, the results reveal other benefits of systematic analysis, such as a more consistent interpretation of the model, more complete analysis, and the importance of training.

Keywords: Goal Modeling, Model Analysis, Empirical Studies

1 Introduction

Goal modeling has been proposed as a tool for system analysis and design, increasing awareness of the social-driven goals which motivate system design or redesign in an enterprise, aiming to increase the success of the system in practice. Intentional modeling was included as the first sub-model in the EKD (Enterprise Knowledge Development) method aimed to capture and make decisions over aspects of an enterprise [1]. Goal modeling techniques have been elaborated in software requirements engineering frameworks (e.g., [2] [3]). Certain goal modeling frameworks allow users to capture qualitative, imprecise goals (softgoals), difficult to quantify in early analysis, as well as interactions and dependencies between various stakeholders and systems within an enterprise [4].

Studies have shown the benefits of intentional modeling as part of systems analysis, e.g., [5]. Further work has argued that more utility can be gained from goal models by applying systematic analysis over model constructs, guiding users to use the models to answer questions about the domain. A wide variety of analysis procedures have been introduced, e.g. [6] [7]. One class of analysis procedures allows placement of values to intentions in the model reflecting an initial question,

then guides propagation of those values, including ways to deal with conflicting values, ways to backtrack over conflicts, and how to draw overall conclusions. For example, in a model that connects information systems solutions such as CRM (Customer Relationship Management) or predictive analytics to enterprise goals, one might ask "Will adopting CRM improve enterprise agility?" (forward analysis, from means to ends). Conversely, if we want to achieve enterprise agility, "What solutions (among the ones included in the model) should one adopt?" (backward analysis).

Most of the research on analysis procedures focuses on the analytical power and mechanisms of the various analysis procedures, typically demonstrating utility by providing a single example application, often in the context of an industrial project. To our knowledge, little work has been done to study how modelers analyze goal models – to compare ad hoc analysis (without a systematic procedure) with the application of proposed procedures. Without a systematic analysis procedure, the modeler/analyst may be examining the model in an ad hoc manner, possibly mixing forward and backward propagation of values, or assigning values to model intentions without following a predetermined systematic process.

Previous work by the authors provided goal model analysis procedures specifically suited to early stages of enterprise system analysis, supporting qualitative analysis over imprecise concepts, and encouraging iteration over models and elicitation over the domain [8]. Specifically, the work outlined several hypotheses concerning the benefits of qualitative, interactive analysis for agent-goal models:

- **Analysis**: aids in finding non-obvious answers to domain analysis questions,
- **Model Iteration**: prompts improvements in the model,
- **Elicitation**: leads to further elicitation of information in the domain, and
- **Domain Knowledge**: leads to a better understanding of the domain.

An exploratory experiment was conducted to test if these benefits were specific to systematic analysis, or a product of any detailed examination of the model, even if ad hoc. Result did not produce a strong conclusion one way or another, although the experiment suffered from a small number of participants and flaws in the design.

In this work, we designed and administered two types of case studies to further test the hypothesis concerning interactive analysis suggested by previous work. Following our earlier work, we use i* as the goal modeling framework in these studies. Due to the great number of confounding variables, we chose to use case studies as the research method, rather than experiments producing statistically significant data. Specifically, we conducted ten case studies using subjects with some experience in i* modeling. Half of the participants analyzed models using no explicit procedure (ad-hoc analysis) while the other half used implementations of the forward ([8]) and backward ([9]) interactive analysis procedures.

Previous work hypothesized that interactive analysis provokes useful group discussions [8]. In order to gain some insight into analysis by individuals versus analysis in a group, we administered a separate multi-session case study involving a project team designing tool support for modeling "back of the envelope" calculations.

Qualitative and quantitative analysis of results are used to compare treatments in both studies, to gather evidence to support or deny the hypotheses, and to gain an understanding of the benefits of and barriers to systematic goal model analysis, helping to guide the application of goal analysis for systems within an enterprise.

The paper is organized as follows. Section 2 provides background on goal modeling and interactive analysis. Section 3 describes the study design. Section 4 presents results and analysis. Section 5 contains discussion, including threats to validity. Section 6 summarizes related work, while Section 7 provides conclusions.

2 Background: Goal Modeling and Interactive Analysis

2.1 The i* Framework

The i* Framework facilitates exploration of an enterprise emphasizing social aspects by providing a graphical depiction of system actors, intentions, dependencies, responsibilities, and alternatives [4]. An example i* model with a legend is shown in Fig. 1. The social aspect of i* is represented by *actors*, including *agents* and *roles*, and the associations between them. Actors depend upon each other for the accomplishment of *tasks*, the provision of *resources*, the satisfaction of *goals* and *softgoals*. *Softgoals* are goals without clear-cut criteria for satisfaction. Actors have *boundaries* containing the *intentions* of an actor: desired goals and softgoals, tasks to be performed, and resources available. The relationships between intentions inside an actor are depicted with *Decomposition* links, showing the elements which are necessary in order to accomplish a task; *Means-Ends* links, showing the alternative tasks which can accomplish a goal; and *Contribution* links, showing the effects of softgoals, goals, and tasks on softgoals. Positive/negative contributions representing evidence which is sufficient enough to satisfy/deny a softgoal are represented by *Make*/*Break* links, respectively. Contributions with positive/negative evidence that is not sufficient to satisfy/deny a softgoal are represented by *Help*/*Hurt* links.

Analysis labels are used to represent the degree of satisfaction or denial of an intention. Following [2], the *(Partially) Satisfied* label represents the presence of evidence which is *(insufficient)* sufficient to satisfy an intention. *Partially denied* and *denied* have the same definition with respect to negative evidence. *Conflict* indicates the presence of positive and negative evidence of roughly the same strength. *Unknown* represents the presence of evidence with an unknown effect.

2.2 Forward Interactive i* Analysis

The forward analysis procedure starts with an analysis question of the form “How effective is an alternative with respect to goals in the model?” The analysis starts by assigning qualitative evaluation labels to intentions related to the analysis question. These values are propagated along links in the forward direction (i* links are directed) using defined rules. The nature of a *Dependency* indicates that if the element depended upon (*dependee*) is satisfied then the element depended for (*dependum*) and element depending on (*depender*) will be satisfied. *Decomposition* links depict the intentions necessary to accomplish a task, indicating the use of an AND relationship, selecting the "minimum" value amongst intentions in the relation, ordered from satisfied to denied. Similarly, *Means-Ends* links depict the alternative tasks which are able to satisfy a goal, indicating an OR relationship, taking the maximum value.

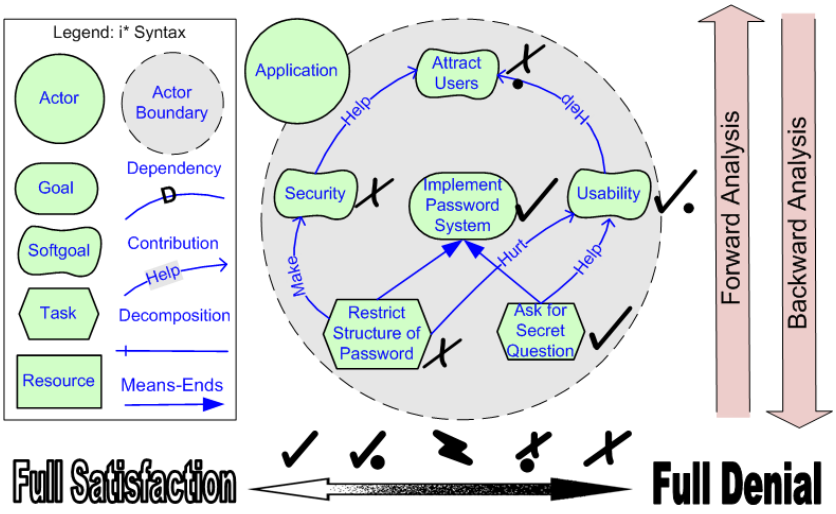

Fig. 1. i* Model Example of a Simple Application with Legend

The procedure adopts *Contribution* link propagation rules from the NFR procedure [1]. Positive values (✓, ✓•) propagated through positive links (Make, Help) produce positive values, weakened with the latter link. Positive values propagated through negative links (Break, Hurt) propagate full or partial negative values (✗, ✗•). Links in i* are symmetric: negative values propagated through positive links produce negative values, and negative values propagated through negative links produce positive values.

The procedure prompts for interactive input when human judgment is needed to combine multiple incoming conflicting or partial values to determine the resulting label for a softgoal. Human judgment may be as simple as promoting partial values to a full value, or may involve combining many sources of conflicting evidence.

Once the procedure has finished interactive propagation, the final analysis values for the intentions of each actor are examined in light of the original question. By looking at the degree of satisfaction or denial of key intentions, an assessment is made as to whether the alternative would work in the domain. More information concerning the procedure can be found in [8].

2.3 Backward Interactive i* Analysis

Backward analysis allows users to ask questions of the form "Is it possible for a set of intentions to be satisfied? If so, how?". The procedure uses the same propagation rules as in the forward procedure, but now propagates evidence both forward and backward. The backward propagation is implemented via a formalization of i* expressed in conjunctive normal form (CNF) and passed to a SAT solver. Human judgment is needed for intentions which have conflicting analysis values assigned. In the backward procedure judgment takes the form "I want Intention X to have the

value V. Give me a combination of values for the contributing intentions that would result in the target value". Users are shown a list of contributing intentions and their associated links, and then are expected to choose a value for each contributing intention. The user can say "No Combination" if no combination of values would produce the target value. The procedure is iterative in that it repeatedly calls the SAT solver until a satisfying assignment is found and no more human judgment is needed. Each iteration involves more human judgment questions. When judgments produce conflicting results, the procedure "backtracks", re-asking the last round of questions involved in the conflict. Currently, when conflicts occur the user is provided with a list of intentions involved in the conflict (derived via a SAT solver which provides a resolution proof). A solution may not be found if the model is conflicting and does not require human judgment, or if the user cannot enter a human judgment which produces a solution. More information can be found in [9].

3 Case Study Design

We designed and administered ten case studies involving individuals and one multi-session case study with a group of participants, all applying interactive analysis over i* models. In the first type of study, our unit of analysis was the individual participants, while in the second it was the group as a whole. As our aim was for interesting qualitative and quantitative findings without statistical significance, changes were made to the procedure under analysis and to the case study designs at various points. We describe the initial and modified study designs in the following. Study design choices and threats to validity are discussed in Section 5.

3.1 Individual Case Studies

Overview. The studies were administered in two rounds. In the first round, six participants were provided i* refresher training and instructions for the study. They were given an introductory sheet describing the model domain, introduced to the three subject models and twelve analysis questions, then given time to answer the questions over the models. In the second round, four participants were given i* refresher training and study instructions, then spent about 25 minutes creating an i* model about life as a student, and then followed an analysis methodology which guided the application of various questions over the model. In both rounds, half of the subjects used the systematic analysis procedure while the other half answered the questions using ad-hoc analysis. The subjects using systematic i* analysis received an additional round of training for the forward and backward procedures (15 minutes). All participants were told that they could make changes to any model at any point, but that they should not feel obligated to do so. The study involved a "think-aloud" protocol, with one of the authors present to observe the progress and answer questions. Participants were encouraged to ask questions about the model if they had them. Results were recorded via audio recording, screen capture and saving versions of the models. All participants used the i* drawing implementation in the OpenOME tool [10]. Every participant was asked a series of follow-up questions concerning

their experience. The total time for each study in both rounds was two hours or less.

Participants. Participants were recruited via a call for participation to students who had learned about i* in one or more system analysis courses, or to students involved in i*-related tool or research projects. Selection was purposive rather than random, we wanted subjects with some knowledge of i* but who were not very familiar with goal model analysis of any form. The resulting participants were students at either the graduate or undergraduate level in Computer Science, Information Systems or Health Informatics. The students had previously created anywhere from one to ten i* models of varying detail, all within the last year. Participants had from none to ten years of industry experience, mostly in technical-related fields. Subjects were paid $40 regardless of the time taken or the results, and results were not made available to anyone who had an influence on course evaluation.

Training. The first two participants of Round 1 were given an i* refresher handout, similar to an expanded version of Section 2.1. The subject using the systematic analysis procedure was given a similar handout describing the forward and backward procedures. After these initial runs of the study, it was apparent that the subject's i* knowledge was not particularly strong. The time devoted to reading the refresher and training documents was not significant. The study was revised such that the facilitator gave a ten-minute i* refresher lesson, and for the participants using systematic analysis, a 10-15 minute instruction session.

Model Domain. In Round 1, subjects were asked to analyze models from the ICSE Greening domain. The models were the result of a study which analyzed the possibilities for "greening" the ICSE'09 conference, conducted via the construction of several medium to large i* models, focusing on the tradeoffs between greening and non-greening goals for the conference chairs [11]. Three models were used from this study, containing between 36 and 79 intentions, 50 and 130 links, and 5 and 15 actors. These were the same models used in the exploratory study described in [8].

The results of the first round of the study performed with six participants showed minimal model changes or elicitation questions, as well as participant difficulties in understanding the models. The decision was made to revise the study and instead allow participants to make their own models over a domain they were familiar with – student life. In the second round, the four participants were provided with some leading questions, (e.g., Who is involved? What do the actors want to achieve?), and then spent 25 minutes creating a model describing their student experiences. P1 to P6 used the ICSE Greening Models, while P7 to P10 used their own student models.

Analysis Questions. In the first round of study, twelve analysis questions over the three models were presented to the participants, four per model, two each aimed at forward and backward analysis. The questions were aimed to represent interesting questions over the domain. For example "If every task of the Sustainability Chair and Local Chair is performed, will goals related to sustainability be sufficiently satisfied?" (forward question) and "Is it possible for both sustainability and successful conference to both be at least partially satisfied? If so, how?" (backward question).

Results from the Group Case study, described in the next section, indicated that it was challenging to motivate modelers to analyze their own models, and that it was sometimes difficult for modelers to come up with interesting analysis questions. As a result, a suggested methodology for model analysis was created using our experiences in evaluating our models in practice. We summarize the methodology in Fig. 2.

1. Alternative Effects (Forward Analysis)
   Identify all leaf intentions in the model (no incoming links)
   a) Implement as much as possible: Evaluate the situation where all leaves are satisfied.
   b) Implement as little as possible: Evaluate the situation where all leaves are denied.
   c) Reasonable Implementation Alternatives: Evaluate likely alternatives in the domain.
2. Achievement Possibilities (Backward Analysis)
   Identify all roots in the model (no outgoing links)
   a) Maximum targets: Evaluate the situation where all roots must be fully satisfied. Is this possible? How?
   b) Minimum targets: Evaluate the lowest permissible values for the roots. Is this possible? How?
   c) Iteration over minimum targets: If a solution was found in b), try gradually increasing the targets in order to find maximum targets which still allow a solution.
3. Domain-Driven Analysis (Mixed)
   a) Use the model to answer interesting domain-driven questions, if possible. If the model cannot answer these questions, can it be expanded to do so, or is it a limitation of the notation?

Fig. 2. Suggested Analysis Methodology

Generally, the first two sections were meant to act as "sanity checks" in the model, checking that it produced sensible answers for a variety of questions, while the second part was intended to support more useful analysis in the domain. Round 2 participants were asked to use this methodology to analyze the student life model they had created. The same methodology was used for all participants, as it did not explicitly reference the forward or backward analysis implementations.

3.2 Group Case Study

A second study was conducted involving a group of four graduate students and a professor who were in the process of designing and implementing a tool (Inflo) to support modeling and discussion of "back of the envelope" calculations. The participants wanted their tool to support informed debate over subjects, such as carbon footprint calculations, containing references to easily understandable models which themselves contain clear references to information sources.

Three two-hour modeling and analysis sessions occurred. Each session had one of the authors present as an i* expert and modeler, and anywhere from two to four of the participant group members. Most of the time in these sessions was devoted to constructing and discussing a large i* model representing the tool, its users, and their goals. During each session, some time was devoted to applying both the forward and backward analysis procedures, letting the participants make decisions over the human judgments posed by the procedures. In this study, the author/facilitator played more of a participatory role, drawing the model and administering the analysis with constant feedback and input from the participants. The first session concluded with a survey concerning the participant's experience with the analysis procedures, while the second and third sessions had audio and/or video recording.

3.3 Data Analysis Methods

The studies produced approximately 24 hours of audio and video, many versions of models, and pages of observer notes. Quantitative data was collected by counting how many and what type of changes to the model were made, (e.g., change a link type, add an intention), and how many domain-related questions were asked for each type of question for each participant (e.g., "What do they mean by collaborate?").

Qualitative data was coded as per the study hypothesis described in the introduction section, allowing for extra fields to capture additional interesting observations. The process of finding results not related to our initial hypotheses was similar to Grounded Theory [12], where qualitative data was grouped according to relevant categories or codes relating to potentially interesting observations or theories. Analysis of further subjects potentially added more evidence to these categories, or produced new categories. What resulted was a list of interesting observations or theories with an associated list of qualitative support classified by participants.

4 Results and Analysis

4.1 Initial Hypotheses

Analysis aids in finding non-obvious answers to domain analysis questions. Results for this hypothesis were mixed. Some participants gave explicit answers to the questions, some referred to analysis labels in the model as answers to the question, while yet others had difficultly producing answers to the questions. One participant was not sure when they were done answering a question. Ideally, participants would be able to interpret the results of the question in the model in the context of the domain; however, only some participants were able to do so. Similarly, participants often had difficulty in translating questions into initial labels in the model.

These results point to a difficulty in mapping the model to the domain, both in starting the analysis and in translating the results back to the world. Presumably, this is a skill which comes with modeling experience. It is interesting to note that these difficulties seemed more prevalent in Round 1 where participants were analyzing large models created by others. It seems that knowledge of i* and the domain may have a significant effect on the ability to apply and interpret analysis.

Model Iteration: prompts improvements in the model. Counts of the number of changes made for each participant are shown in the 3rd and 4th columns of Table 1. Generally, few changes were made with the exception of P1 who redrew much of the model at the start of the study independent of the analysis questions. We omit detailed data on types of changes; however, some specific examples include changing decomposition to contribution links and adding or renaming tasks. Note that a few changes were suggested by the participants but not made, and are not included in the counts. The number of changes was not significant for most participants, and there are more changes made with ad-hoc than systematic analysis. There is no notable difference between participants analyzing their own or others models. For the five participants who made changes, we asked if those changes were helpful, four said yes, while one said it depends on whether changes would be helpful to domain experts.

Table 1. Number of Model Change and Questions Asked for each Participant

		# Model Changes		# Questions Asked		
Treatment	Partic.	Forward Questions	Backward Questions	Forward Questions	Backward Questions	Round
	P1	59	10	10	1	
	P4	0	0	1	0	
	P5	5	13	6	6	1
	P7	2	5	0	0	
Ad-hoc	P9	0	5	0	0	2
	P2	0	0	2	3	
	P3	0	0	2	0	
	P6	0	3	5	1	1
	P8	0	0	2	2	
Systematic	P10	0	0	0	1	2

Elicitation: leads to further elicitation of information in the domain. The number of domain-related questions asked by each participant is shown in the 5th and 6th columns of Table 1. Again, we see no interesting differences between groups.

We can try to understand the results for model changes and elicitation, and why they differ from the results found in previous studies ([8]), by examining the reasoning behind these hypotheses. Previous studies have claimed that it is the interactive nature of the analysis that prompts for changes to the model and drives elicitation. We can expand on this claim, by considering the differences between a goal model representing a domain and the participant's mental model of the domain. An i* model can be considered an incomplete representation of the mental model of its creator. When human judgment is needed in a model, the evaluator is asked to use their mental model of the world to supplement the contents of the physical (explicitly expressed) model. The hypotheses rely on differences between the mental model of the participant and the explicit i* model, especially if they were not the creator of the model. Although such differences could be discovered at any point, they may become particularly apparent when answering human judgment questions.

When these differences are discovered, they may prompt changes to the model, or may cause inquiries concerning the domain. For an example of the former, in the Inflo case study, when asked "Is it possible to make (Inflo) models at least partially trustworthy?" one of the participants decided that validation of a model was not relevant to trustworthy, and the link was removed. The model did not match that participant's mental model of the domain. In other cases, missing elements, inaccurate contributions, or questions concerning the meaning of elements could arise. For example, a human judgment concerning Make conference participation fun made one participant make changes to the model to make the conference more fun and sustainable, renaming task and changing a link from a *hurt* to a *help*.

Because a small number of changes were made to the model, and a modest amount of questions were discovered, we can hypothesize that either the evaluation did not typically reveal differences between the mental model of the evaluator and the

explicit model, or these differences existed, but were not used to modify the model. We can find several examples where the evaluator seemed to ignore the structure of the model and answer human judgment questions using only their mental model. For example, in one case, in the forward judgment for the softgoal "Make conference participation fun" the three contributing intentions all contributed partially denied. The participant decided the value was unknown because "I'm not sure how any of these directly related to fun". It seems this would lead to a conclusion that the model is incomplete or inconsistent with the mental model of the participant, and thus needs to be changed, but no changes were made. In another type of example the participants treated the model and judgment situations as an oracle, deferring to the explicit model, "it's telling me that it's weakly satisfied".

A tentative conclusion is that correcting the model and producing questions relies on more extensive knowledge of the syntax, and may require explicit training in detecting differences between physical and mental models. Further studies could continue to test these hypotheses, in different situations, for example with an experienced modeler or in an industrial setting.

Domain Knowledge: leads to a better understanding of the domain. At the end of every individual study, we asked: do you feel that you have a better understanding of the model and the domain after this exercise? Seven out of ten participants said yes. One participant who did not say yes was commenting on the complexity and learning curve associated with i*, another complained that they were already very familiar with being a student, and didn't learn anything further, and the last said that they learned more about the model, but not about being a student. Selection of complex models and a familiar domain seemed to hinder this potential benefit. Analysis was helpful for both systematic and ad-hoc approaches. Participants provided specific comments concerning evaluation: analysis brings out the flaws in the model, and it was helpful for understanding the effects of goals and relations and in choosing between alternatives.

Promote Discussion in Group Setting: Application of systematic evaluation in a group setting did produce several situations where human judgment caused discussion among participants. For example, the participants discussed whether getting feedback was really necessary in order to make models trustworthy after this contribution appeared in a backward judgment situation for Make models trustworthy. In other examples, the group had discussions about the exact meaning of goals appearing in judgments situations, for example "what is meant by Flexibility?" This revealed that different participants thought it meant slightly different things. To be fair, not all judgment situations provoked discussion; more experience is needed to determine how to maximize this positive effect.

4.2 Additional Findings

In addition to findings supporting or denying our initial hypotheses, our qualitative analysis produced other categories of findings, resulting in new tentative hypotheses.

Model Interpretation Consistency. When examining the differences between ad-hoc and systematic analysis, we can see that some participants using ad-hoc analysis made use of the analysis labels and performed some form of label propagation (2/5),

while others explained the answer to the question over the model without propagating (1/5), while some participants did both in the same study (2/5). The i* training received by all participants contained an explanation of evaluation labels.

Because the i* Framework was defined in such a way as to leave room for interpretation of its symbols and syntax, by creating systematic procedures we extend the definition of the language, making its meaning more precise. It could be argued that the interpretation used by the analysis procedures is not the best/most obvious; however, what is more important is that i* users and evaluators make consistent and similar interpretations of the model. Thus we are interested in whether or not the participants are consistent with each other (and themselves). Collected evidence shows a variety of interpretations of the model expressed via the propagation of evaluation labels, showing that ad-hoc propagation can be inconsistent among evaluators. For example, one participant interpreted the AND decomposition intentions as having to be at least weakly satisfied for the parent to be satisfied (in the procedure they would have to all be satisfied). In the same model, the participant decided that one intention in another AND decomposition was necessary for the satisfaction of the parent, but the other was optional. In several other cases propagation was consistent with the rules of our procedure. Future studies could ask participants to explicitly propagate in order to collect further examples.

Coverage of Model Analysis. Further analysis of the difference between ad-hoc and systematic analysis revealed significant differences in the coverage of analysis across the model. Subjects who used ad-hoc analysis considered the effects of far fewer intentions and actors in the models. For example, one of the participants who did propagation without systematic analysis ignored the links between the actor under analysis and another actor entirely. Several participants when propagating manually forward or backward only propagated one level or one link jump without continuing to consider the affects of other factors in the model. When participants did not propagate at all they often missed the effects of various links or intentions in their verbal analysis. For example, when considering the satisfaction of goals related to Attendee experience without propagation, a participant only looked at contributions from the Sustainability Chair and did not acknowledging positive effects from goals within the Local Chair.

Although use of the explicit analysis procedures increased the coverage of the analysis, it did not ensure complete coverage. Depending on the choices for initial values, the propagation results often did not cover the entire model. Most participants did not see any problems with such incomplete propagation. If propagation is to be complete more often, more training concerning the selection of initial values and the interpretation of analysis results is needed.

Model Completeness and Analysis. Several participants made interesting comments about the relationship between model completeness and the effectiveness of model analysis. In the Inflo case study, the participants felt that analysis was not useful until the model reached a sufficient level of completeness. One individual participant thought that the study should urge people to make a more complete model before analysis. Another participant said that the model would have been much better if there had been more time to work on it, yet this participant finished creating the model before time was up. For this participant, the analysis revealed that model was incomplete. Another participant, when applying analysis, noticed that the model had

no negative links. We can conclude that analysis may be more useful for answering domain questions when the model is complete, but that analyzing over an incomplete model has the potential to reveal its incompleteness.

5 Discussion

5.1 Study Design Selection

Several study design choices were available, the most applicable of which being controlled experiments, action research, or case studies. An experiment would have required the isolation of as many control variables as possible in order to convince the reader that the results in terms of dependent variables (for eg., model changes, questions asked) followed from the manipulation of independent variables (using or not using the procedure, analyzing your own or others models). In the case of goal model analysis, many variables exist which are difficult to control, including: the participants experience with i* and other goal modeling frameworks, their experience with goal model analysis, their experience and openness to modeling in general, their industry experience, and the nature and subject matter of their education. Given that we want to use participants with some i* experience, the second barrier to the application of experiments is finding enough participants to produce statistically significant results. Despite the popularity of i* in research [13], in practice it is not widely used, and a large pool of i* users is not available.

Action research was a further alternative, similar to the types of case studies performed in most work which introduces goal model analysis procedures (for eg. [6] [7]). The forward interactive procedure used in this study has already been applied in one such large-scale study, producing results which led to the formation of the initial hypotheses [8]. Although future studies of this type are useful, we believed it would be advantageous to collect evidence from multiple cases, in an effort to collect a greater quantity of qualitative data. Case studies are useful in that they can provide evidence not only to confirm the existence of hypotheses, but also to explain why such phenomena occur, particularly useful in cases with many confounding variables.

5.2 Threats to Validity

Several threats to the validity of our studies exist.

Construct Validity. We used several measures to test our hypotheses. To test analysis capabilities we looked at how participants were able to use the model to answer questions, whether they could apply some default questions to the models, and whether they could create and analyze their own questions over their own models. However, it was challenging to measure the difficulty participants had in performing these tasks. Often it was hard to determine if the participants were able to take analysis results and use them to draw conclusions over the domain.

To measure model iteration, we counted changes made to the model, or in some cases suggested changes. However, it is difficult to know if these changes are always beneficial. To measure elicitation, we collected questions asked over the model

domain during the study. However, classification of questions versus comments can be subjective, and not all domain questions asked over the model would realistically lead to further elicitation. We used a follow-up question to measure improvements in domain knowledge. However, it is difficult to isolate whether analysis was the source of improved understanding and not simply reading or creating the models.

All other exploratory hypotheses are measured using the collecting of qualitative data. This collection can be subjective, although we battled this subjectivity to some degree by having more than one person involved in the data analysis, and by performing systematic classification of qualitative observations.

Internal Validity. We must show that the design of our study adequately tests the initial hypotheses. The extra analysis training given to participants using explicit analysis may have affected the results, although these participants didn't make any more model changes or ask any more questions. Although the study facilitator tried to encourage honest opinions, the presence of one of the authors in all study sessions may have influenced the results. The think-aloud protocol may have affected participant actions, avoiding actions they could not justify. Some of the participants were not comfortable with the think-aloud protocol, and were quiet, making it hard to understand the motivations behind their actions. It is possible that the choices of model domains influenced results, with the domains being too unfamiliar or familiar.

External Validity. As our study used upper-year undergraduate or graduate students as participants, it is possible that results may not generalize to other groups with less technical background. As our studies used the i* framework and interactive analysis procedures, it is questionable whether the results generalize to other goal modeling frameworks or analysis procedures. We believe that results are applicable to frameworks which have a syntax similar to i* (Tropos, GRL). However, it is unlikely that results will generalize to fully-automated analysis procedures.

Reliability. The study was administered by someone with expert knowledge of i* and i* analysis. If the experiment was repeated with someone with less i* or analysis knowledge, the quality of the training or of questions answered in the study may differ, and so may the results. The researcher in question is the creator of the analysis portions of the OpenOME Tool and the Analysis Methodology in question. Some of the potential bias was avoided by having each participant either use or not use the procedures, avoiding an unintentional promotion of one over the other. Every effort was made to avoid influencing the participants during the study; however, it is difficult to avoid all bias or potential effects in such cases.

6 Related Work

We can find examples of studies applying intentional modeling in repeated case studies or experiments. In [1], Stirna and Persson describe multiple participatory cases to illustrate guidelines for participatory Enterprise Modeling (EM). Related work uses two studies to derive conclusions and recommendations about participatory EM and tool support [14]. An interesting conclusion of this work is that EM modeling requires an EM expert. Our findings concerning the need for more extensive i* and analysis training reflect this finding; however, we believe it is too

restrictive to say that i* and associated analysis should only be used with an expert present. Existing work shows that even i* novices who misuse the notation benefit from its use [15], and our individual participants generally increased their knowledge of the domain through modeling and analysis.

Work in [16] investigates the role of NFR catalogues in creating i* modeling of a software project using a controlled team experiment. Another study tested the effects of patterns on i* modeling using both a case study in practice and an exploratory experiment in a classroom setting [17]. Similar work in [18] evaluated patterns developed in EKD via workshop experiments involving experienced professionals.

7 Conclusions

In this study we applied interactive i* analysis in ten studies with individual participants and one group study with the aim of testing existing hypotheses concerning the benefits of analysis, and discovering new knowledge about interactive goal model analysis. Despite the small participant sample size, the results are interesting, and not as anticipated. The results can be summarized as follows:

- **Analysis**: Both systematic and ad hoc analysis can be useful for answering analysis questions over the domain, although training is needed to apply initial analysis values and interpret the results.
- **Model Iteration**: Both systematic and ad hoc analysis prompted small amounts of iteration over the model.
- **Elicitation**: Both systematic and ad hoc analysis prompted a small number of questions over the domain. The iteration and elicitation effects observed in previous studies may require explicit training in adjusting the model to match the analysts mental model, and using the model to reveal gaps in knowledge.
- **Domain Knowledge**: Both systematic and ad hoc analysis lead to a better understanding of the domain.
- **Model Interpretation Consistency.** Ad hoc analysis will often use interpretations of the model which are inconsistent within one analysis and amongst modelers. Use of systematic analysis promotes a consistent interpretation of the model.
- **Coverage of Model Analysis.** Systematic analysis increases the coverage of intentions and actors considered in answering analysis questions.
- **Model Completeness and Analysis.** A certain level of model completeness may be necessary for effective analysis. In some cases analysis may reveal the incompleteness of the models.

Study results can guide the application of intentional modeling and analysis in an enterprise setting, illustrating the potential benefits of ad-hoc vs. systematic analysis and emphasizing the role of training or the presence of an experienced facilitator.

Acknowledgments. Financial support has been provided by the Natural Sciences and Engineering Research Council of Canada and the Ontario Graduate Scholarship Program.

References

1. Stirna, J., Persson, A., Sandkuhl, K.: Participative Enterprise Modeling: Experiences and Recommendations. In: Krogstie, J, Opdahl, A, Sindre, G. (eds.) CAiSE 2007. LNCS, vol. 4495, pp. 546—560. Springer, Heidelberg (2007)
2. Chung, L. Nixon, B.A., Yu, E., Mylopoulos, J.: Non-Functional Requirements in Software Engineering. Kluwer Academic Publishers, Norwell, MA (2000)
3. Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-Directed Requirements Acquisition. Science of Computer Programming, 20, 3—50 (1993)
4. Yu, E.: Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering. In 3rd IEEE International Symposium on Requirements Engineering (RE'97), pp. 226--235. IEEE Press, New York (1997)
5. Maiden, N. A. M., Jones, S. V., Manning, S., Greenwood, J., Renou, L.: Model-Driven Requirements Engineering: Synchronising Models in an Air Traffic Management Case Study. In: Persson, A., Stirna, J. (eds) CAiSE'04, LNCS, vol. 3084, pp. 367--383, Springer, Heidelberg (2004)
6. Giorgini, P., Mylopoulos, J., Sebastiani, R.: Simple and Minimum-Cost Satisfiability for Goal Models. In: Persson, A., Stirna, J. (eds) CAiSE'04. LNCS, vol. 3084, pp. 20--35. Springer, Heidelberg (2004)
7. Franch, X.: On the Quantitative Analysis of Agent-Oriented Models: In Dubois, E., Pohl, K. (eds) CAiSE'06. LNCS, vol. 4001, pp. 495--509. Springer, Heidelberg (2006)
8. Horkoff, J., Yu, E.: Evaluating Goal Achievement in Enterprise Modeling – An Interactive Procedure and Experiences. In: Persson, A., Stirna, J. (eds.) PoEM'09. LNBIP, vol. 39, pp. 145-160. (2009)
9. Horkoff, J., Yu, E.: Finding Solutions in Goal Models: An Interactive Backward Reasoning Approach. In: 29th Int. Conf. on Conceptual Modeling (ER'10). (to appear)
10. OpenOME, https://se.cs.toronto.edu/trac/ome/wiki
11. Cabot, J. Easterbrook, S., Horkoff, Mazon, J., Lessard, L., Liaskos, S.: Integrating Sustainability in Decision-Making Processes: A Modelling Strategy. In: ICSE 2009 New Ideas and Emerging Results (NIER'09).
12. Seaman, C. B.: Qualitative Methods in Empirical Studies of Software Engineering. IEEE Trans. Softw. Eng. 25, 4, 557-572 (1999)
13. Fourth International i* Workshop (iStar'10). CEUR Workshop Proceedings, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-586/ (2010)
14. Persson, A., Strirna, J.: An Exploratory Study into the Influence of Business Goals on the Practical Use of Enterprise Modeling Methods and Tools. In: Proc. 13th Int. Conf. on Information Sys. Development (ISD 2004), pp. 275—288, Springer, Heidelberg (2002)
15. Elahi, G., Yu, E., Annosi, M. C.: Modeling Knowledge Transfer in a Software Maintenance Organization - An Experience Report and Critical Analysis. In: Stirna, J., Persson, A. (eds.) PoEM'08. LNBIP, vol. 15, pp. 15-29 (2008)
16. Cysneiros, L. M.: Evaluating the Effectiveness of Using Catalogues to Elicit Non-Functional Requirements. In: Proc. Workshop em Engenharia de Requisitos (WER07), pp 107 – 115 (2007)
17. Strohmaier, M., Horkoff, J., Yu, E., Aranda, J., Easterbrook, S.: Can Patterns improve i* Modeling? Two Exploratory Studies. In: Paech, B., Rolland, C. (eds) REFSQ'08. LNCS, vol. 5025, pp. 153--167. Springer, Heidelberg (2008)
18. Rolland, C., Stirna, J., Prekas, N., Loucopoulos, P., Persson, A., Grosz, G.: Evaluating a Pattern Approach as an Aid for the Development of Organisational Knowledge: An Empirical Study. In: Wangler, B., Bergman, L., (eds.) CAISE'00. LNCS, vol. 1789, pp. 176-191. Springer, Heidelberg (2000)