Exploring user data from a game-like math tutor: a case study in causal modeling

D. RAI AND J. E. BECK
Worcester Polytechnic Institute, USA

We have used causal modeling to understand data from a game-like math tutor, *Monkey's Revenge*. We collected student data of various types such as their attitude and enjoyment via surveys, performance within tutor via logging, and learning as measured by a pre/post test. Although the data are observational, we want to understand the causal relationships between the variables we have collected. We contrast the causal modeling approach to the results we achieve with traditional approaches such as correlation matrix and multiple regression. Relative to traditional approaches, we found that causal modeling did a better job at detecting and representing spurious association, and direct and indirect effects. We found that the causal model, particularly one augmented with domain knowledge about likely causal relationships, resulted in much more plausible and interpretable model. We present a case study for blending exploratory results from causal modeling with randomized controlled studies to validate hypotheses.

Key Words and Phrases: Causal modeling, confounders, structural equation modeling, case study

1. INTRODUCTION

Making causal inferences based on non experimental statistical data has been a controversial topic [Freedman, 1987, Rogosa, 1987, Denis, 2006]. While randomized controlled trials are the standard approach to take care of intervening third variables, causal modeling is an alternative method of making causal inferences [Pearl, 2009, Sprites et al., 2001] based on observational data making certain causal assumptions. We are using the causal modeling approach to analyze and explore the data from a game-like math tutor, *Monkey's Revenge.*

We created four different versions of Monkey's Revenge with varying degree of game-like properties. A total of 297 middle school (12-14 year olds) students from four Northeastern schools in the United States participated in the study. We logged their tutor activities and asked 16 survey questions using 5 point Likert scale from "strongly disagree" to "strongly agree." We then used factor analysis to reduce the variables into six categories: ***likeMath, mathSelfConcept, pedagogical preference*** (prefer Computers to books; find real world examples helpful for learning Math.)***, tutorHelpful, tutorConfusing, likeTutor.*** From students' log data, we calculated variables ***%correct*** (ratio of correct problems to total problems); ***avgAttemptTime*** (average time spent on each attempt) and ***avgHints*** (average number of hints asked on each question). We also collected ***preTestScore*** and ***prePostGain*** and used a variable ***game-like*** as an ordinal parameter (taking on values 1 through 4). Along with using causal modeling to explore and analyze our data, we analyze the causal modeling approach itself. We compare it against the standard statistical approaches of correlation and multiple regression.

2. CAUSAL MODELING, CORRELATION MATRIX

Based on the data we collected, we used TETRAD, a free causal modeling software package [Glymour et al., 2004] with the PC search algorithm to generate a causal graph (fig 1). We also generated a graph based on the correlation matrix, computed by finding the correlations between the variables (fig 2).

Authors' addresses: {D. Rai, J. E. Beck} Department of Computer Science, Worcester Polytechnic Institute, MA, USA. E-mail: {dovan, josephbeck}@wpi.edu

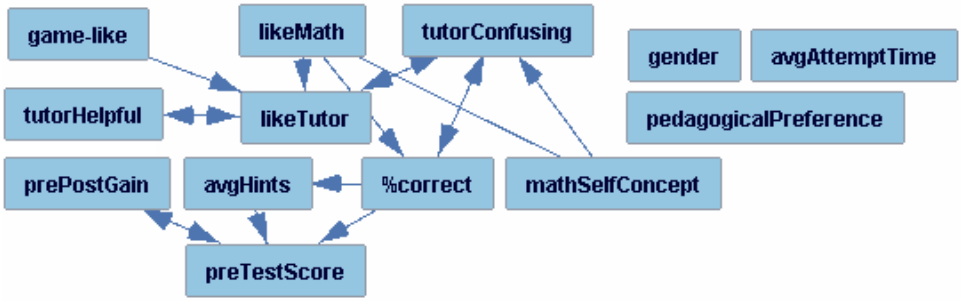

Figure 1 Causal model from PC algorithm without domain knowledge

True negatives (indirect and spurious associations): Correlation is not causation as there might be possible confounders causing the spurious association, and causal modeling controls for all third variables regarding them as possible confounders. From the correlation matrix, we see that *likeTutor* and *%correct* are correlated which would suggest that students who like the tutor performed better. This result could be interpreted as evidence for student engagement, since students who liked the tutor are presumably more engaged while using it. But the causal model (Fig 1) infers that this is a spurious association confounded by *likeMath*. Students who like math tend to like tutor more and to have better performance. Once we control for *likeMath*, there is no relation between *likeTutor* and *%correct.* Thus, the scientific interpretation of the data changes.

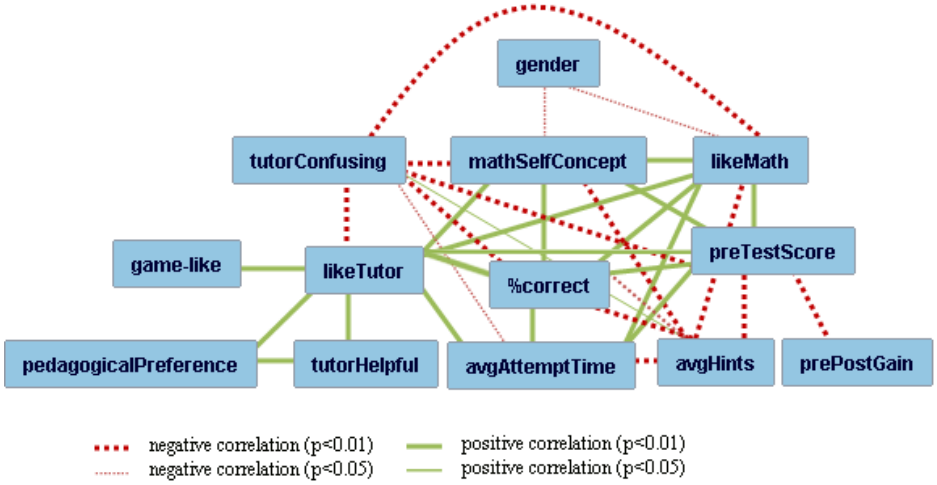

Figure 2 Graph based on correlation matrix

Still, the causal model is limited to assertions about the observed variables as there might be other confounders which we have not observed. Causal modeling makes also distinction between direct and indirect association. For example, *likeMath* and *avgHints* are negatively correlated, which suggests that the students who like math ask for fewer hints. But once we control for *%correct*, that correlation is gone (see Fig 1), suggesting that the students who like math ask for fewer hints only because they already know the correct responses and so do not need as much help—there is no direct effect.

False negatives (reduced statistical power and multicollinearity): Controlling on third variables reduces statistical power and we might get false negatives if we have few data. Multicollinearity is an extreme case when the independent variables are correlated among themselves. For example: *avgAttemptTime* is correlated with both *%correct* (0.3**) and *preTestScore*(0.3**). But since, *%correct* and *preTestScore* are highly correlated among themselves (0.6**), *avgAttemptTime* is conditionally independent to both of them. We can see that *avgAttemptTime* is an isolated node in figure 1; in contrast, the correlation graph (Figure 2) indicates *avgAttemptTime* is related to both *preTestScore* and *%correct*.

3. CAUSAL STRUCTURE, PATH ORIENTATION AND DOMAIN KNOWLEDGE

In the causal model, some edges have plausible orientations (e.g: *likeMath* →*likeTutor*← *game-like)*. Using the information that *likeTutor* is correlated with both *likeMath* and *game-like,* but *likeMath* and *game-like* are independent between themselves, search algorithm correctly identifies that it is not *likeTutor* influencing *likeMath* and *game-like* but the other way round. However, we see that there are other edges which are incorrectly oriented such as *%correct*→*preTestScore*; student performance on the tutor cannot have influenced a pretest that occurred before students began using the tutor.

Correlation underdetermines causality as covariance in statistical data is rarely sufficient to disambiguate causality. Therefore, even after we use search algorithms to find some structure, there are a number of "Markov equivalent" structures. In TETRAD, we can add domain knowledge in the form of knowledge tiers which represent the casual hierarchy. Causal links are only permitted to later tiers, and cannot go back to previous tiers. We used the following knowledge tier based on our knowledge of assumed causal hierarchy and temporal precedence.

i. Gender,
ii. Game-like, mathSelfConcept
iii. likeMath, Pedagogical preference
iv. preTestScore
v. %correct, avgAttemptTime, avgHints, tutorConfusing, tutorHelpful
vi. likeTutor
vii. prePostGain

We see from Figure 1 and Figure 3 that adding domain knowledge not only fixes the path orientations (*preTestScore*→*%correct*), but have changed the whole causal structure adding some new causal links (*gender*→*mathSelfConcept, pedagogicalPreference*→*tutorHelpful, correct*→*avgAttemptTime*).

At first, it may appear that knowledge of causal hierarchy only helps to orient the edges specifying which one is cause and which one is effect. However, besides distinguishing variables as potential causes and effects, the domain knowledge also restricts the set of variables to be considered as confounders and mediators. Note that many data analyses have such causal hierarchies implicit in the analysis and conclusions.

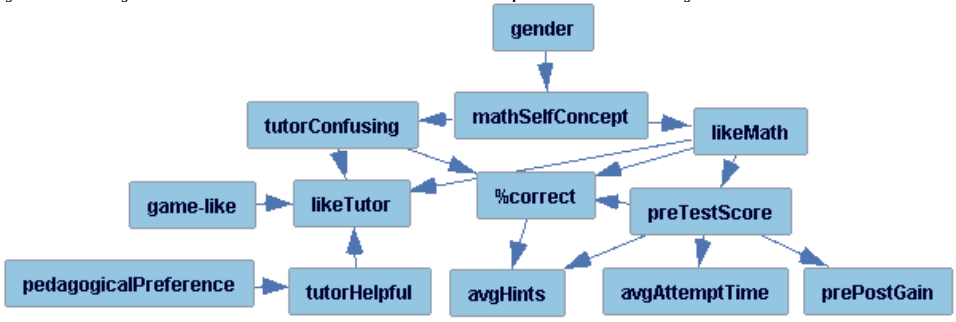

Figure 3 Causal model with domain knowledge

Sometimes, we do not know about the causal hierarchy of the variables we are trying to analyze and may not know which is the cause and which is the effect, but having information of the causal hierarchy of third variables, such as whether they are a potential confounder or a potential mediator, can help infer if there is any causal path between the variables of interest. We can illustrate this with a concrete example in education. Suppose we have observed that engagement and learning are correlated, but want to understand the causal relation between them. Imagine there are two other variables, prior knowledge, a potential confounder, and performance, a potential mediator. Consider two scenarios: if partialling out prior knowledge removes the correlation, then we know there is no causal relationship between engagement and learning, and the causal structure is engagement←prior knowledge→learning. On the other hand, if partialing out performance removes the correlation between engagement and learning, then there is still an *indirect* causal effect between the two, either engagement→performance→learning, or learning→performance→engagement.

Interestingly, adding domain knowledge can also address the problem of multicollinearity. Since we have set *preTestScore* on higher causal tier than *%correct*, *%correct* cannot be a possible confounder or mediator and therefore, the partial correlation (*preTestScore, avgAttemptTime* | *%correct*) is not calculated and the correlation between *preTestScore* to *avgAttemptTime* is maintained.

4. CAUSAL MODELING AND MULTIPLE REGRESSION

Causal modeling is a sophisticated extension to multiple regression which employs a series of multiple regression. Multiple regression only looks at direct effect but fails at identifying **indirect effects**. While multiple regression can be equally robust when it comes to predictive accuracy, causal modeling provides a better representation and framework to *understand* interrelationships of variables. Since causal modeling allows multiple layers of associations of variables, it adds affordance to insert **domain knowledge** in the form of a causal hierarchy. On top of the statistical assumptions used by statistical methods such as regression, causal modeling adds **causal assumptions** such as faithfulness and causal sufficiency [Sprites et al., 2001]. Stronger assumptions add more analytical power but also higher chances of inaccuracy. It is up to researcher to select these assumptions based on their data and domain. We have accepted the causal assumptions made by TETRAD since they seem reasonable for our data and purpose.

5. CONFIRMATORY, EXPLORATORY AND GRAPHICAL TOOL

Causal modeling can be used to test goodness of fit of a model constructed *a priori* from theory. We did not start with such a model, but the causal model has supported some of our prior hypotheses (*pedagogicalPreference→tutorHelpful→likeTutor*) and (*mathSelfConcept→tutorConfusing→likeTutor*).

Using causal modeling as an exploratory tool to generate new theories is controversial as the possibility of unobserved confounders and under determination of causality from correlation pose serious limitation to generate new valid conclusions. But, conditional independencies in data and domain knowledge can offer some new inferences which can be helpful in guiding us towards further analyses and examination. For example, in our causal model, we found that *likeMath* has both direct (*likeMath→%correct*) and indirect (*likeMath→preTestScore→%correct*) effect on *%correct.* Based on this, we are considering two possible causal models as shown in figure 4. Model I suggests that *pretestScore* does not capture all of the variance in prior knowledge of the student, as represented by the latent node “Prior knowledge.” So,

students who like math and have high prior knowledge may have a low pre test score but they have high performance nonetheless. Model II on the other hand suggests that students who like math both have higher prior knowledge and are more engaged, and have therefore higher performance. In other words, likeMath affects both prior knowledge and engagement.

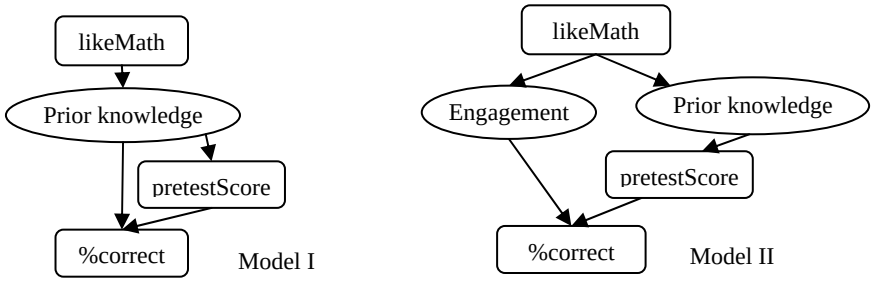

Figure 4 Two possible causal models linking *likeMath* and *%correct*

We were not able to make any conclusive findings with the causal model, but it has at least made interesting inferences and raised questions that are very important for us. It has directed towards the possibilities that we would like to make further examination and possibly run some controlled randomized trials.

Even if researchers are skeptical of the domain knowledge we have brought to bear and are dubious of the causal modeling assumptions, it is still possible to consider Figure 1 without the assumption that the edges represent causality. This graph would be a compact representation of the partial correlation relationships among the variables

6. CONCLUSIONS

We have used a causal modeling approach to explore the data from a game-like math tutor. Based on statistical independence within data and our domain knowledge, TETRAD's causal inference algorithm generated a causal model which not only confirmed some of our prior hypotheses about data but also made some interesting new findings. Causal modeling does a good job of identifying not only cause and effect but also confounders so that we can find spurious associations and mediators so that we know both direct and indirect effects. But those inferences cannot be claimed as accurate since causal modeling cannot identify spurious association caused by unobserved confounders and there are multiple Markov equivalent causal models that can be generated from the same data. Still, causal modeling is the best approach we have found, particularly when compared with common statistical techniques such as correlation and multiple regression to generate most plausible inferences from observational educational data sets.

REFERENCES

DENIS, D. J. LEGERSKI, J. 2006. Causal Modeling and the Origins of Path Analysis. *Theory & Science*, Vol. 7, No. 2

FREEDMAN, D. A. 1987. As Others See Us: A Case Study in Path Analysis. *Journal of Educational Statistics*, (12:2), pp. 101-128.

GLYMOUR, C., MADIGAN, D., PREGIBON, D., SMYTH, P. Statistical Themes and Lessons for Data Mining. *Data Mining and Knowledge Discovery*, 2004. p. 11-24

GLYMOUR,C., SCHEINES,R. 2004. Causal modeling with the TETRAD program.Synthese.37-64

PEARL J. 2009. Causality.2nd Edition. Cambridge University Press.

ROGOSA, D. 1987. Causal models donot support scientific conclusions: A comment in support of Freedman. *Journal of educational statistics*. Vol. 12, No. 2, pp.185-195

SPRITES, P. GLYMOUR, C. SCHEINES, R. 2001. Causation, Prediction, and Search. 2nd Edition. MIT press