AN APPROACH TO IDENTIFYING MEANINGFUL ACTION PATTERNS IN STUDENT-TUTOR INTERACTIONS

Anna L. Rowe and Nancy J. Cooke
Rice University *New Mexico State University*
Houston, Texas *Las Cruces, New Mexico*

Part of the success of computerized intelligent tutoring systems will be associated with their ability to assess and diagnose students' knowledge in order to direct pedagogical interventions. What is needed is a methodology for identifying general relationships between on-line action patterns and patterns of knowledge derived off-line. Such a methodology would allow an assessment and diagnosis of knowledge, based only on student actions. The focus of this initial research is the development of a means of identifying meaningful action patterns in student-tutor interactions. Actions executed by subjects on a set of verbal troubleshooting tests (Nichols et al., 1989) were summarized using the Pathfinder network scaling procedure (Schvaneveldt, 1990). The results obtained from this work indicate that meaningful patterns of actions can be identified using the Pathfinder procedure. The network patterns are meaningful in the sense that they can differentiate high and low performers as defined by a previous scoring method. In addition, the networks reveal differences between high and low performers suggestive of targets for intervention.

INTRODUCTION

As tasks become more cognitively complex and demand more specialized skill, training issues become increasingly critical. Recent research (e.g., Nichols, Pokorny, Jones, Gott, & Alley, 1989) suggests that computerized intelligent tutoring systems may successfully supplement traditional training in complex tasks such as troubleshooting ill-defined problems. Intelligent tutoring systems enable individuals to spend time learning a skill in a one-on-one environment in which a computer takes on the role of a human tutor. One goal of intelligent tutoring systems is to incorporate more individualized instruction through a detailed on-line assessment of student knowledge, diagnosis of student strengths and weaknesses based on the assessment, and intervention targeted at those strengths and weaknesses. However, the best approach to assessment and diagnosis of student knowledge remains to be delineated.

Most methods typically used to assess student knowledge involve the gathering of data (e.g., verbal reports) not typically collected during interactions with a tutor, thereby interrupting the tutoring process for data collection. The single exception to this limitation is the case in which assessment and diagnosis are based on ideal or incorrect actions. Indeed, most of the intelligent tutor approaches to assessment and diagnosis have relied on this method (e.g., Anderson, Boyle, & Reiser 1985; Burton, 1982; Pokorny & Gott, 1992; Stevens, Collins, & Goldin, 1979). However, action data are impoverished relative to the much richer data obtained from methods such as verbal reports and structural analyses. These richer methods go beyond the student's actions to reveal the cognitive processes underlying those actions. Thus, what is needed is a method for mapping student actions onto a rich representation of student knowledge. Identifying general relationships between on-line action patterns and patterns of knowledge derived off-line would allow later predictions to be made about knowledge, based only on student actions. The focus of this initial effort is the development of a means of identifying meaningful action patterns in student-tutor interactions.

METHOD

Actions executed by subjects on a set of verbal troubleshooting tests (Nichols et al., 1989) were used in the present analyses. Nichols et al. examined the effects of an intelligent tutoring system called SHERLOCK, a tutor designed to give technicians opportunities troubleshooting faults in the "test station". The test station is a device typically used to diagnose aircraft equipment malfunctions; however, on occasion, the test station itself malfunctions, and technicians must then troubleshoot it. Nichols et al. (1989) assessed SHERLOCK by comparing the performance of Air Force technicians who received both on the job training (OJT) and SHERLOCK training to the performance of technicians who received only OJT. A brief description of the methodology used by Nichols et al. follows.

Subjects

The subjects were 37 manual avionics shop technicians identified by supervisors as being at a beginning or intermediate skill level and available for the study duration. The 30 subjects who completed a specific set of three verbal troubleshooting problems were used in the present analyses.

Materials and Procedure

All subjects participated in a training period in which they were randomly assigned to receive either OJT or OJT and SHERLOCK. A set of four pre- and posttest measures were administered before and after this training period, respectively. In the present study only data resulting from one of these measures were analyzed: pre- and posttest versions of a verbal troubleshooting test, an individually administered structured problem solving test. The test begins with the examiner describing a fault that had occurred. The subject then attempts to isolate the fault and repair the equipment through a series of recursive action-result steps, expressed verbally. In each step, the subject specifies an action he or she would take, and the examiner responds by informing the subject of the action's effect on the equipment (i.e., the test or check results in a pass or a fail).

Only data from three troubleshooting problems were used in the present analyses (pretest 1, pretest 2, and posttest 1). The complete analysis described below was conducted on data from the latter two problems because these were comparable in terms of type and difficulty. The first problem was primarily analyzed to determine the optimal coding scheme.

RESULTS AND DISCUSSION

Overview of Analysis

The action data were summarized using the Pathfinder network scaling procedure, a descriptive statistical technique that represents pairwise proximities in a graphical form (Schvaneveldt, 1990). Three steps were taken in the analysis process. First, a coding scheme for students' actions was developed using the data from the pretest 1 problem. The scheme was then applied to and modified slightly for the remaining two problems, referred to herein as pretest and posttest. The purpose of the scheme was to be able to classify discrete actions into meaningful action units that could be represented as nodes in a Pathfinder network. Second, conditional transition probabilities for all pairs of actions (in both directions) were calculated for individual subjects and for four groups of subjects by dividing the frequency with which specific action transitions occurred by the frequency with which the first item in the sequence occurred. Finally, the matrices of transition probabilities were submitted to the Pathfinder network scaling technique. The four group solutions are presented in Figures 1 through 4.

Assessment

A major question to be asked of this approach is whether Pathfinder action networks can distinguish high and low performers for the purposes of assessment. In this study the subjects' score for each problem derived using Pokorny and Gott's (1992) scoring worksheet is assumed to be the "true score" indication of their performance on that problem. Therefore, to answer the above question we examined the correlation between the score derived from the scoring worksheet procedure and an assessment measure derived from Pathfinder networks. To assess students via Pathfinder an ideal or expert network is necessary. Thus, for each problem an expert network (based on the six highest performers, see Figures 1 and 3) was compared to the network representations of the remaining 24 nonexpert individuals, using the C measure (Goldsmith & Davenport, 1990), a quantitative index of network similarity. The correlations between troubleshooting scores and this network similarity measure for the nonexperts in each problem are presented in Table 1. Two additional assessment measures were also calculated and included in the analysis: 1) a measure derived from a correlation between an individual's action frequencies and action frequencies associated with the aggregate expert protocol, and 2) a measure based on the number of actions executed by each subject.

Examination of Table 1 indicates that the Pathfinder similarity measure is predictive of troubleshooting scores for the pretest but not for the posttest. However, the action frequency measure is predictive of the score for both the pre- and the posttest. Other significant correlations indicate that the Pathfinder similarity measure and action frequency are highly correlated, as would be expected given that both measures take shared actions into account. However, the Pathfinder measure focuses on action sequences rather than action frequency. Indeed, at least for the pretest, both measures seem to independently account for a portion of the variance. Specifically, the correlation between the troubleshooting score and the action frequency measure remains significant when the Pathfinder similarity measure is partialed out (r (21) = .53, p < .01). Also, the correlation between the troubleshooting score and the Pathfinder similarity measure is marginally significant when the action frequency measure is partialed out (r (21) = .39, p < .07). These results suggest that the types of actions subjects perform and the frequency with which they perform them are predictive of both the pre- and posttest scores. In addition, action sequence is predictive of the pretest scores. There was a much wider range of actions performed by the low performers in the posttest compared to the pretest which may have overwhelmed any predictive power of sequential variation.

Finally, the assessment measures were also compared in terms of their ability to discriminate subjects in the experimental and control groups. The mean scores of experimental and control subjects for the pre- and posttest are presented in Table 2. As should be expected, there were no pretest differences between experimental and control groups. Interestingly, the only significant difference between these two groups at posttest is for the Pathfinder similarity measure (t (22) = 2.07, p < .05). Subjects in the experimental condition had networks that were more similar to the ideal network than did subjects in the control condition. The lack of a significant score difference between the two groups is most likely due to the restriction of range that occurred by eliminating the six highest performers on the posttest. The fact that Pathfinder accounts for experimental vs. control differences, whereas the action frequency measure did not suggests that subjects who were trained on SHERLOCK learned more expert-like action sequences than those who were not.

Diagnosis and Intervention

In the following analysis we evaluated the aggregate networks of the high and low performers to identify strengths and weaknesses that could potentially be used for diagnosis and intervention. Note that an identical analysis could be performed at the individual network level.

The Pathfinder networks for the high and low performers differed both quantitatively and qualitatively. General quantitative differences between the two groups can be seen in terms of the number of nodes present in the networks of the high and low performers. The high performers' networks had fewer action nodes (pretest=23, posttest=21) than the low performers' networks (pretest=28, posttest=45), especially at posttest (see Figures 1 through 4). In other words, the high performers as a group executed fewer distinct actions than the low performers, indicating that high performers seem to agree

Table 1. Intercorrelation matrix of the four assessment measures. (VT score = verbal troubleshooting score; PF sim = similarity of Pathfinder network with expert network; Act Freq = correlation of action frequencies with expert action frequencies; No. Act = number of actions)

Table 1a

Pretest	1	2	3	4
1. VT score	1.0	.57**	.65**	.38
2. PF sim		1.0	.47**	.38
3. Act Freq			1.0	.30
4. No. Act				1.0

Table 1b

Posttest	1	2	3	4
1. VT score	1.0	.26	.76**	-.35
2. PF sim		1.0	.55**	.22
3. Act Freq			1.0	-.17
4. No. Act				1.0

* p < .05; ** p < .01

Table 2. Mean assessment measures of experimental and control groups on pre- and posttests (VT score = verbal troubleshooting score; PF sim = similarity of Pathfinder network with expert network; Act Freq = correlation of action frequencies with expert action frequencies; No. Act = number of actions)

	Pretest Mean	Posttest Mean
VT score		
Experimental	42	68
Control	47	59
PF sim		
Experimental	.05	.07
Control	.05	.04
Act Freq		
Experimental	.38	.41
Control	.33	.24
No. Act		
Experimental	12.6	15.8
Control	11.5	16.6

on the relevant actions in comparison to low performers. Although the low performers at posttest executed over twice as many distinct actions as the high performers, they shared all but one of the high performers' actions (shared nodes at pretest=17, at posttest=20). These results suggest that the low performers as a group have knowledge about a wide variety of actions, yet they do not seem to understand *when* these actions apply. Interestingly, the subjects in the experimental group executed fewer distinct actions (35) than those in the control group (48). Thus, SHERLOCK may be effective in teaching students the conditions under which various actions apply.

Closer examination of the differences between the action nodes in the networks offers insight for diagnosis and intervention. The six pretest nodes in the high performers' network that were not contained in the low performers' network consisted of signal flow and data flow tests. In addition, at pretest, low performers executed 11 actions high performers did not, seven of which were data flow and signal flow tests. At posttest, half of the additional actions executed by low performers were data flow and signal flow tests, and half were swaps. Thus, these errors of omission and commission indicate that intervention in these particular cases should be targeted at learning the appropriate data flow and signal flow actions. A more detailed target may be derived by a focus on individual nodes.

The networks of the high and low performers also differed in some more global, qualitative ways. First, the high performers (both tests) appeared to follow a rule about the general sequence of actions which were taken: 1) general checks outside of the test package, including visual checks, equipment checks, and swaps, 2) signal flow tests inside the test station, 3) data flow tests of components inside the test station, and 4) swapping. Low performers, on the other hand, did not closely follow this rule and instead committed violations in this general sequence. This trend was observed both for pre- and posttest networks.

Second, the low performers exhibited what may be termed a meaningless action sequence at both pre- and posttest, whereas high performers did not. For example, after completing a signal flow or data flow check which indicated that the component was functional, some low performers chose to swap the component anyway. The high performers did not exhibit meaningless action sequences such as these.

CONCLUSIONS

The results obtained from this work indicate that meaningful patterns of actions can be identified using the Pathfinder network scaling procedure. The network patterns are meaningful in the sense that they can differentiate high and low performers as defined by a previous scoring method. In addition, the networks reveal differences between high and low performers suggestive of targets for intervention (e.g., data flow and signal flow tests). Finally, this bottom-up approach to knowledge elicitation resulted in general action patterns that may not have been verbalized in a typical knowledge elicitation interview (i.e., the general sequence of high performers: checks outside, signal flow tests inside, data flow tests, swaps).

These results are even more promising when the source of the ideal or expert network used to make these comparisons is considered. Specifically, high performers were used here as the ideal. An even better ideal would probably result from the use of recognized subject matter experts. In addition, the use of subjects with more expertise would widen the range of performance, which would likely result in enhanced assessment and diagnostic capabilities.

ACKNOWLEDGMENTS

This work was accomplished under the Air Force Office of Scientific Research Summer Research Program for the Job Structures Branch of the Manpower and Personnel Research Division, Human Resource Directorate of the Air Force Armstrong Laboratory. We would like to thank Drs. Sherrie Gott, Ellen Hall, and Bob Pokorny for their assistance during the course of this project.

REFERENCES

Anderson, J.R., Boyle, C.F., & Reiser, B.J. (1985). Intelligent tutoring systems. *Science*, *228*, 456-462.

Burton, R.R. (1982). Diagnosing bugs in a simple procedural skill. In D. Sleeman and J.S. Brown (Eds.), *Intelligent Tutoring Systems* (pp. 157-183). London: Academic Press.

Goldsmith, T.E., & Davenport, D.M. (1990). Assessing structural similarity of graphs. In R. Schvaneveldt (Ed.), *Pathfinder Associative Networks: Studies in Knowledge Organization* (pp. 241-253). Norwood, NJ: Ablex.

Nichols, P.D., Pokorny, R., Jones, G., Gott, S.P., & Alley, W.E. (1989). *Effective Instructional Processes Within an Intelligent Tutoring System.* Air Force Technical Report. Brooks AFB, TX: Air Force Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division

Pokorny, R., & Gott, S. (1992, in preparation). *The Evaluation of a Real-World Instructional System: Using Technical Experts as Raters.* Air Force Technical Report. Brooks AFB, TX: Air Force Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division

Schvaneveldt, R. (Ed.), (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization* . Norwood, NJ: Ablex.

Stevens, A., Collins, A., & Goldin, S.E. (1979). Misconceptions in student's understanding. *International Journal of Man-Machine Studies*, *11*, 145-156.

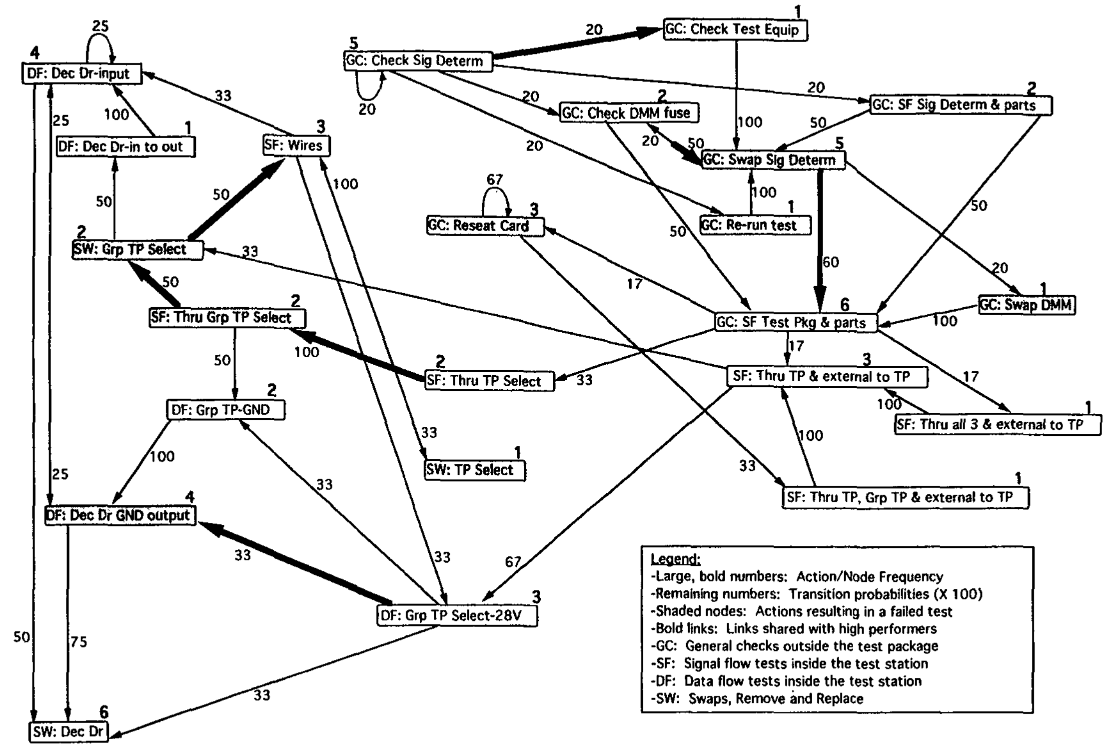

Figure 1. Pretest network based on aggregate transition probabilities of the six high performers.

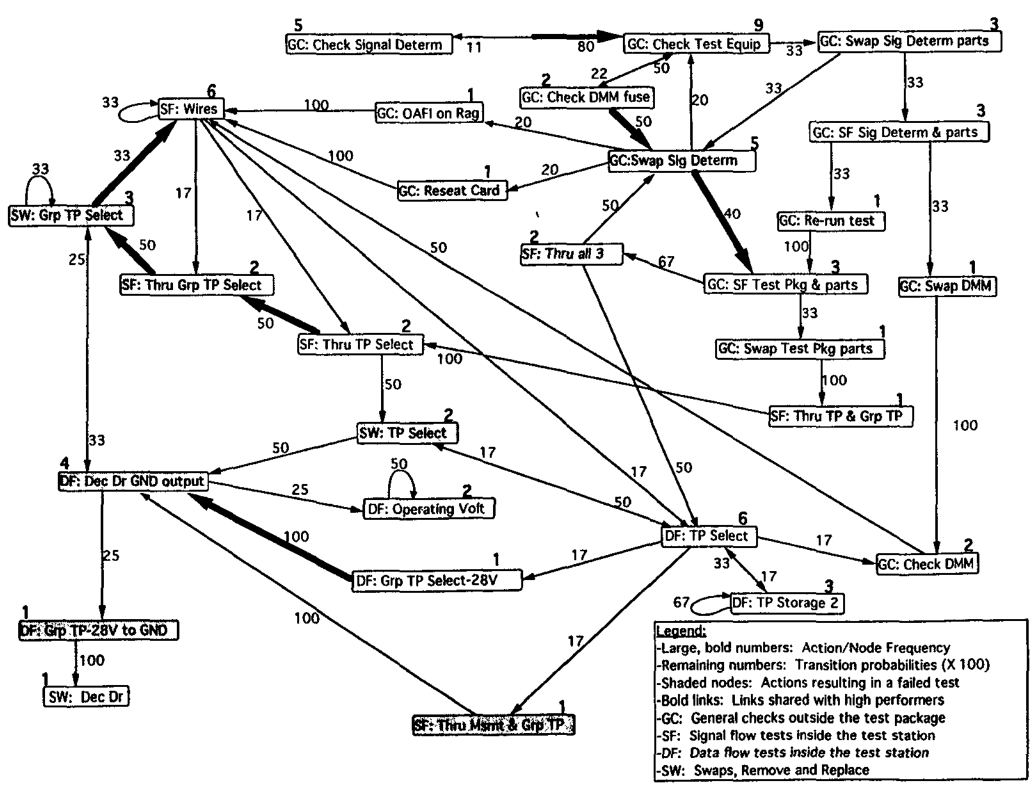

Figure 2. Pretest network based on aggregate transition probabilities of the eight low performers.

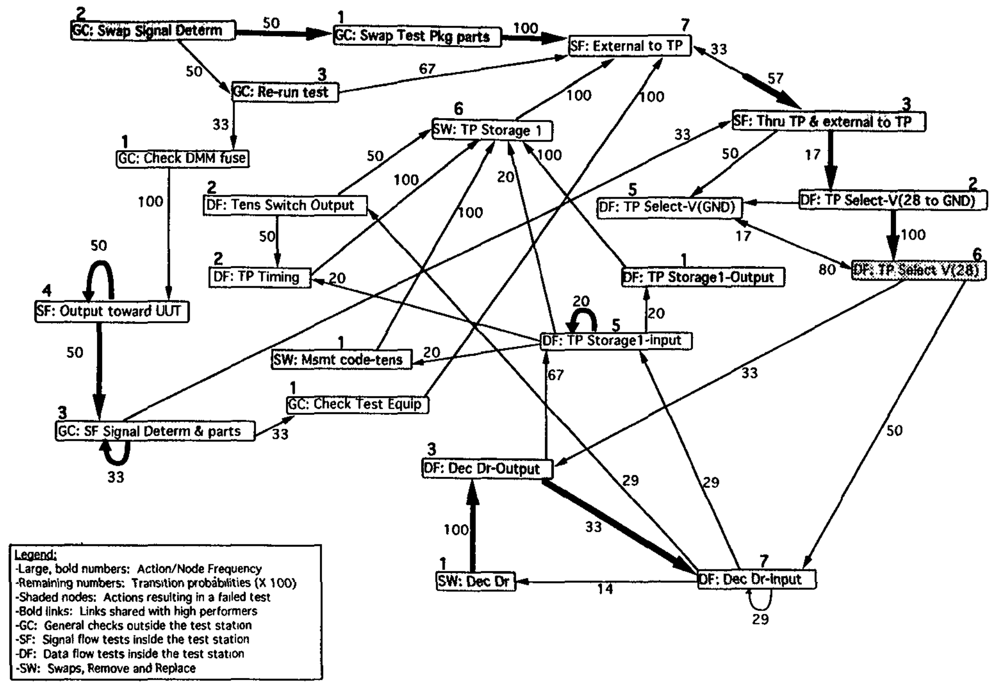

Figure 3. Posttest network based on aggregate transition probabilities of the six high performers.

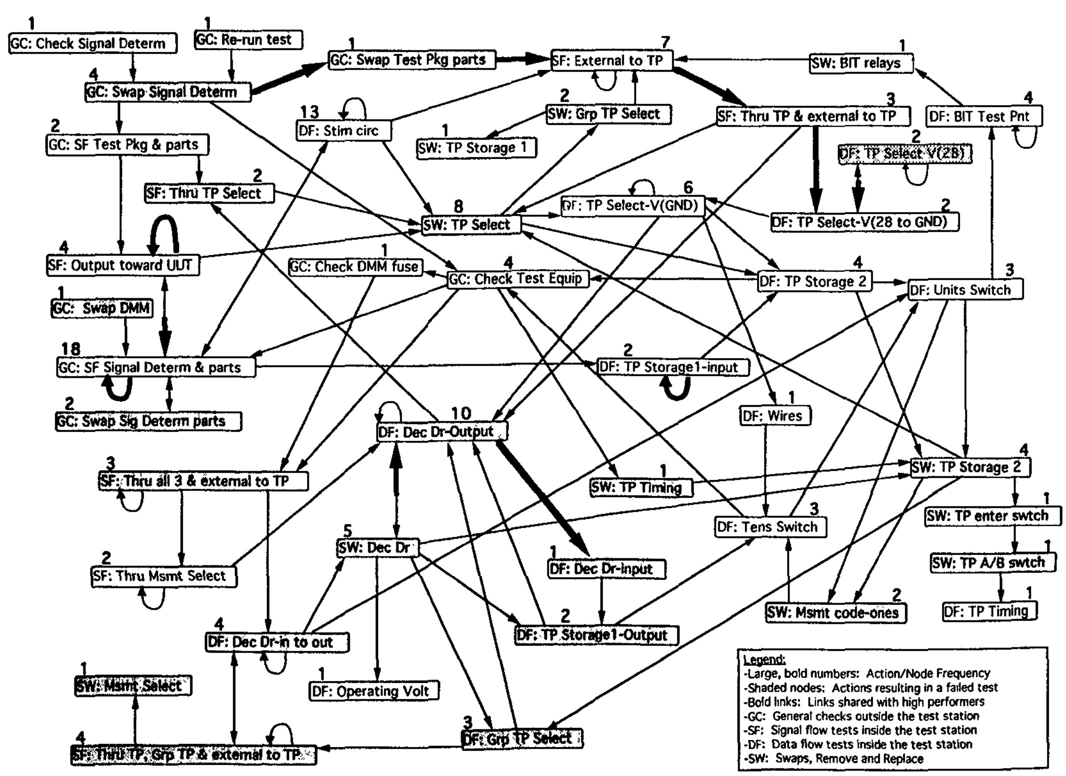

Figure 4. Posttest network based on aggregate transition probabilities of the eight low performers (link weights are omitted due to graph complexity).