# Effects of network topology on the OpenAnswer's Bayesian model of peer assessment

Maria De Marsico[1], Luca Moschella[1], Andrea Sterbini[1], Marco Temperini[2]

[1] Computer Science Dept., Sapienza University, Rome, Italy

{demarsico, sterbini}@di.uniroma1.it

[2] Computer, Control, and Management Eng. Dept., Sapienza University, Rome, Italy

marte@dis.uniroma1.it

Abstract. The paper investigates if and how the topology of the peer-assessment network can affect the performance of the Bayesian model adopted in OpenAnswer. Performance is evaluated in terms of the comparison of predicted grades with actual teacher's grades. The global network is built by interconnecting smaller subnetworks, one for each student, where intra-subnetwork nodes represent student's characteristics, and peer assessment assignments make up inter-subnetwork connections and determine evidence propagation. A possible subset of teacher graded answers is dynamically determined by suitable selection and stop rules. The research questions addressed are: RQ1) "does the topology (diameter) of the network negatively influence the precision of predicted grades?"; in the affirmative case, RQ2) "are we able to reduce the negative effects of high-diameter networks through an appropriate choice of the subset of students to be corrected by the teacher?" We show that RQ1) OpenAnswer is less effective on higher diameter topologies, RQ2) this can be avoided if the subset of corrected students is chosen considering the network topology.

Keywords. Peer assessment, Open answers, Bayesian networks, Bayesian model of peer assessment, Network topology.

## 1 Introduction

In Bloom's taxonomy of educational objectives [1] learners need wider and deeper comprehension of topics when passing from pure knowledge (just remembering), to comprehension, application, analysis, evaluation and finally synthesis, as higher metacognitive skills. Peer assessment is a possible tool to help students exercise/enforce these abilities [6]. Open answers to questions allow challenging assessment methods, e.g., exercises, free text answers to questions, etc., which are more effective than multiple-choice tests [5], but also harder to handle for teachers. OpenAnswer [7][8][9] (OA) allows (semi-)automated grading of open answers through peer assessment. During an OA session, each student is assigned a number of peers' answers to grade. To enforce the reliability of the grading results, the sys-

tem provides the (not mandatory) possibility for teacher grading of a subset of answers, chosen step by step according to some select-next-or-stop strategy.
In OA, each student cognitive/metacognitive state is modeled by a fragment of a global Bayesian Network (BN). Assignments of peer answers to grade make up the interconnections among the subnetworks and determine the topology of the global one. Assessments fed by peers (and possibly by the teacher) are propagated within the BN. The system allows providing the students not only marks, but also an estimate of their knowledge and ability to judge, that spurs metacognitive awareness.
Earlier works [4][7][8][9] analyzed several factors affecting the accuracy of predicted grades. Present research questions are: RQ1) "does the topology of the network (in particular its diameter) negatively influence the precision of predicted grades?" If the response to RQ1 is positive, RQ2) "are we able to reduce the negative effects of high-diameter networks through an appropriate choice of the subset of students to be corrected by the teacher?" To answer RQ1 we modified OA: 1) to produce topologic indicators (e.g. diameter of the peer assessment graph, coverage percentage of corrected students plus their immediate neighbors, average distance between inferred and corrected students); 2) to choose the set of corrected students through topological strategies. The available datasets were used to generate also graphs with higher diameter than the original ones. As hypothesized, OA is less effective on higher diameter topologies. While this represents a general result regarding Bayesian models, the response to RQ2 indicates how this can be avoided in the specific case of peer assessment, if the students to be corrected are chosen by considering the network topology. The results provide an educational-specific operational strategy for using OA in a concrete setting, and for designing a suitable peer assessment network for each single session. The topology-based strategies perform even better than the formerly identified best one, as the experiments section shows.

## 2 Related work

Peer-assessment entails a higher cognitive level activity [1]. It pursues different goals [10], especially to allow the learner to appreciate the personal cognitive state and progress. A comprehensive study of peer assessment in a prototype application is in [2]. OA evaluates open answers through peer-assessment, by modeling students and assessment by Bayesian Networks. A different machine learning approach to student modeling is in [3], where Bayesian Networks are used within an Intelligent Tutoring System.

## 3 The model underlying OpenAnswer

The OA system models peer-assessment as a Bayesian network composed of interconnected individual sub-networks. Each such subnetwork represents a student, and is made of three discrete nodes/variables, representing respectively: K - student's knowledge about the topic; C - the correctness of student's answer being evaluated; J - student's ability to judge/assess the answer of a peer; one variable G for each grade

given to a peer (G variables represent the interconnections among subnetworks). K, J and C are updated by information propagation. The final values of C variables represent the estimated answers correctness (grades).

Each variable above has a 6-valued discrete domain ranging from A (best) to F (fail). A-E corresponds to 10-6 (sufficient marks one by one), and F is from 5 below.

For all student's marks of a peer's answer, a corresponding Grade variable (G) and value is injected as evidence into the network, and propagates its effects depending on both the current value of J of the grading student, and on current estimation of C of the answer corrected. Variables C and J are assumed depend from K with conditional probabilities P(C | K) and P(J | K). The C dependence is because writing an essay cannot be easily guessed as it happens in multiple-choice quizzes. As for J, the inspiration is from Bloom's taxonomy of cognitive levels [1] assuming that judging a peer's answer can be considered as a more difficult task than knowing the topic and answering it. The distribution of values for G is conditioned by J and C with distribution P(G | J, C).

When the teacher corrects the essays, OA suggests the next answer to grade and notifies her when no further correction is needed. The possible alternatives for stop condition can be found in [4]. In this work we use a fixed condition, i.e., reaching 30% of answers. The next answer to grade is chosen to maximize the information gain achieved by its teacher correction. Possible criteria are in [4]. Here only the best achieving of past experiments is compared with the new topological selection criteria. Such rule is *maxEntropy*: where the next answer to grade is the one with the highest entropy, i.e., the one the system knows less about.

## 4 Methodology

To show (RQ1) if the propagation of information through the BN drops in quality the more it moves far from the set of corrected nodes, we need networks with higher diameter (the maximum of minimal distances between any two nodes). In our datasets we have two groups of real assessments (datasets I and M, respectively with 2 and 6 peer-assessment sessions) where each student graded the 3 next peers in order from the group (modulo the size of the group). This produced "ring-shaped" networks with a diameter proportional to the number of nodes and inversely proportional to the number of corrected peers (Fig. 1, left).

To get even higher diameters we can either cut the ring (Fig. 1, bottom, "broken" network) or reduce the number of assessments by each student (Fig. 1, right, "2-peers" network), or both (Fig. 1, bottom right). The figure shows also the teacher grades for each student (node color, darker=better) and the grades given by each peer (edge color, darker=better). When a new teacher grade is asserted for a student C variable (new evidence), two information propagations happen in two directionsfrom the corrected student towards her "judging" peers, and towards her "judges" peers.. Because the information flows both directions the edges are considered as undirected (propagation has same weight in both directions).

Fig. 1. Peer-assessment from dataset I - left: 3 peers - right: 2 peers – top: ring, bottom: broken - node color intensity: teacher's grade - edge color intensity: peer's grades, (darker=better)

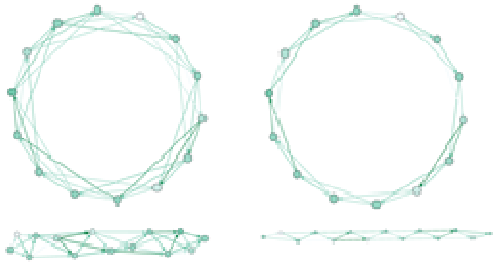

Table 1 shows for each assessment in datasets I and M the diameter for the graphs depending on the transformations: 3 peers or 2 peers, and ring shaped or broken.

The aim of the experiments is to show that the distance traversed by the information in the model has an adverse effect. This requires: 1) topological indicators that describe the network with respect to both static properties (diameter) and to dynamic ones (e.g. the current average distance between corrected students and inferred ones); 2) a new group of selection strategies which takes into account the static and dynamic topological measures of the network when selecting the next student to be corrected.

To this aim three new "good" strategies for the OA greedy selection algorithm are introduced here, to counteract the (expected) negative outcomes of higher distances: maxCoverage: chooses the next student so to maximize the network coverage (the number of corrected students union their immediate neighbors); maxAvgDistCorrected: chooses the next student so to maximize the average distance among corrected students (to distribute them better through the network); minAvgDistInferredCorrected: chooses the next student so to minimize the average distance among the inferred students and their nearest corrected peer (to reduce the average distance the information should traverse). For further verification, three corresponding "bad" strategies are tested that try to keep higher the distance among inferred and corrected students: minCoverage, minAvgDistCorrected, maxAvgDistInferredCorrected.

The available ground truth (complete teacher grades) allows us to simulate different settings and strategies for teacher's correction. The experiments presented here adopt and compare the above strategies to select the next answer to grade, together with the former *maxEntropy*, and stop teacher's grading when 30% of the students have been corrected. The remaining grades are inferred. To compare the prediction performances we examine the percentage of exact inferred grades (OK/INFERRED), and the average distance in the graph between inferred students and their nearest corrected peer (AVG_PEER_DISTANCE). The simulations have been run with these different sets of parameters: Selection strategy: *maxEntropy* or the above topology based ones; Initialization of the P(K) distribution: *flat* or *TgradeDist* ; Number of peers corrected by each student: 3 or 2; Shape of the network: ring or broken.

Table 1. Diameter of original and modified peer-assessments for datasets I and M

| | | NUM. PEERS | 2 peers | | Original (3 peers) | |
|---|---|---|---|---|---|---|
| | | TYPE | ring | broken | Ring | broken |
| DATASET | ASSESSMENT ID | NUM. STUDENTS | DIAMETER | | | |
| I<br>High School, physics | 3 | 14 | 4 | 7 | 3 | 5 |
| | 4 | 12 | 3 | 6 | 2 | 4 |
| M<br>University, C programming | 3 | 13 | 3 | 6 | 2 | 4 |
| | 4 | 13 | 3 | 6 | 2 | 4 |
| | 6 | 11 | 3 | 5 | 2 | 4 |
| | 7 | 11 | 3 | 5 | 2 | 4 |
| | 8 | 9 | 2 | 4 | 2 | 3 |
| | 9 | 11 | 3 | 5 | 2 | 4 |

## 5 Experimental results

Table 2 shows the OK/INFERRED percentage and the maximum AVG_PEER_DISTANCE at the end of the correction. Because of space limits we show only one of the topology-based "good" and "bad" strategies. As a first observation, the maximum AVG_PEER_DISTANCE is very low (1) for the "good" topology-based selection strategies, and that also maxEntropy shows a low value for this outcome (near 1.4). Conversely, the "bad" topology-based strategies show higher AVG_PEER_DISTANCE, as expected (in particular, 2 peers-based networks and broken networks show the highest distances). Yet, the max AVG_PEER_DISTANCE when 30% of students have been corrected is not too high (max 2.9).

When we examine the OK/INFERRED results we see that the "good" topology-based selection strategy outperform the "bad" one and the maxEntropy strategy. In this we affirmatively answer to both our research questions RQ1 and RQ2, regarding the accuracy of correctly inferred marks: network diameter seems to have a negative effect on prediction accuracy, but this can be addressed by suitable topology-oriented strategies for selecting the answers to grade by the teacher.

Other observations can be drawn from the table. The P(K) initialization affects the outcome with better results for TgradeDist , i.e. OA, as expected, works better with some global knowledge about the class. Cutting the ring to increase the diameter ("broken" shape) reduces the OA performances for almost all selection strategies. Reducing the number of peers from 3 to 2 reduces the performances as expected. More investigation is due on the "perfect" number of corrected peers per student.

## 6 Conclusions

Higher-diameter networks induced by assignments of peer grading tasks to students, reduce the prediction precision of OpenAnswer. However, an appropriate choice of the selection strategy for teacher graded answers can counteract this negative effect and perform even better than the earlier best selection strategy, *maxEntropy*.

Table 2. OK/INFERRED vs NUM. PEERS, STRATEGY, RING/BROKEN, P(K) initialization averaged over all assessments in the I and M datasets (green = best values, red = worst values)

| | | | Average of OK/INFERRED | | Max of Avg Peer Distance | |
|---|---|---|---|---|---|---|
| | | SHAPE | ring | broken | ring | broken |
| NUM. PEERS | P(K) init. | STRATEGY | | | | |
| 2 | flat | maxEntropy | 33% | 35% | 1.5 | 1.5 |
| | | maxCoperture | 38% | 38% | 1 | 1 |
| | | minCoperture | 29% | 29% | 1.9 | 2.9 |
| | TgradeDist | maxEntropy | 38% | 39% | 1.5 | 1.5 |
| | | maxCoperture | 47% | 49% | 1 | 1 |
| | | minCoperture | 34% | 39% | 1.9 | 2.9 |
| 3 | flat | maxEntropy | 32% | 33% | 1.4 | 1.4 |
| | | maxCoperture | 37% | 36% | 1 | 1 |
| | | minCoperture | 34% | 32% | 1.5 | 2 |
| | TgradeDist | maxEntropy | 39% | 38% | 1.5 | 1.5 |
| | | maxCoperture | 45% | 42% | 1 | 1 |
| | | minCoperture | 43% | 35% | 1.5 | 2 |

# 7 Bibliography


1. Bloom, B.S., Engelhart, M.D., Furst,. E.J., Hill, W.H., Krathwohl, D.R., 1956. Taxonomy of educational objectives: The classification of educational goals. Handbook I. McKay.
2. Chung, H., Graf, S., Robert Lai, K., Kinshuk, 2011. Enrichment of Peer Assessment with Agent Negotiation. IEEE TLT Learning Technologies, 4(1), pp.35-46.
3. Conati, C., Gartner, A. , Vanlehn, K., 2002. Using Bayesian Networks to Manage Uncertainty in Student Modeling. User Modeling and User-Adapted Interaction 12, pp. 371-417.
4. De Marsico, M., Sterbini, A., Temperini, M., 2015. Towards a quantitative evaluation of the relationship between the domain knowledge and the ability to assess peer work. Proc. ITHET 2015 (pp. 1-6). IEEE.
5. Palmer, K., Richardson, P., 2003. On-line assessment and free-response input-a pedagogic and technical model for squaring the circle. In Proc. 7th CAA Conf. (pp. 289-300).
6. Sadler, P. M., E. Good, P. M., 2006. The Impact of Self- and Peer-Grading on Student Learning. Ed. Ass., 11(1).
7. Sterbini, A., Temperini, M., 2012. Dealing with open-answer questions in a peer-assessment environment. Proc. ICWL 2012. LNCS, vol. 7558, pp. 240–248. Springer.
8. Sterbini, A., Temperini, M., 2013a. OpenAnswer, a framework to support teacher's management of open answers through peer assessment. Proc. FIE 2013.
9. Sterbini, A., Temperini, M., 2013b. Analysis of OpenAnswers via mediated peer-assessment. Proc. 17th IEEE Int Conf. on System Theory, Control and Computing (ICSTCC 2013).
10. Topping, K., 1998. Peer assessment between students in colleges and universities, Rev. of Ed. Research, 68, pp. 249–276.