Pattern Recognition Letters
journal homepage: www.elsevier.com

Trust-Based Community Assessment

Patricia Gutierrez[a,**], Nardine Osman[a], Carme Roig[b], Carles Sierra[a]

[a]*IIIA-CSIC, Campus de la UAB, Barcelona, Spain*
[b]*INS Torras i Bages, L'Hospitalet de Llobregat, Spain*

ABSTRACT

In this paper we present Community Assessment (COMAS), a trust-based assessment service that helps compute group opinion from the perspective of a specific community member. We apply COMAS in the context of communities of learners, and we compute the group opinion from the perspective of the teacher. Specifically, our model relies on *teacher assessments*, aggregations of *student assessments* and *trust measures* derived from student assessments to suggest marks to assignments that have not been assessed by the teacher. The proposed model intends to support intelligent online learning applications by 1) encouraging students to assess one another, and 2) benefiting from students' assessments. We believe the task of assessing massive numbers of students is of special interest to online learning communities, such as Massive Open Online Courses (MOOCs). Experimental results were conducted on a real classroom datasets as well as simulated data that considers different social network topologies (where we say students assess some assignments of socially connected students). Results show that our method 1) is sound, i.e. the error of the suggested assessments decreases for increasing numbers of teacher assessments; and 2) scales for large numbers of students.

1. Introduction

Self and peer assessments have clear pedagogical advantages (Lu and Zhang, 2012; Topping, 1998; Stepanyan et al., 2009; Hannon, 2009; Jenkins, 2009). Students increase their responsibility and autonomy, get a deeper understanding of the subject, become more active in the learning process, reflect on their role in group learning, and improve their judgment skills. Online learning communities encourage different types of peer-to-peer interactions along the learning process. These interactions permit students to get more feedback, to be more motivated to improve, and to compare their own work with other students accomplishments. Teachers, on the other hand, benefit from these interactions as they get a clearer perception of the students engagement and learning process. Encouraging online interactions, group discussion, feedback and user engagement is an essential part of any intelligent tutoring system or social platform (Bickmore et al., 2013; Yee-King et al., 2013).

In this paper we describe and analyze Community Assessment (COMAS), a trust-based assessment service embedded in our lesson planning tool PeerFlow[1] (de Jonge et al., 2014). The teacher is expected to mark a small set of assignments. Similarly, students are expected to assess an even smaller number of their students' assignments. For all assignments not assessed by the teacher, COMAS suggests assessments based on aggregating students' assessments. This aggregation takes into consideration the similarity between the students' assessments and those of the teacher. COMAS's behaviour is such that the more assessments by the teacher the better the suggested assessments are.

COMAS has been evaluated over 2 groups of 13 years old students of English attending a state school near Barcelona. Different assignments were assigned to students which had to be assessed following a set of evaluation criteria (instructional rubrics).

This work has direct application in communities of learners where the totality of assessments to be made by a teacher is simply not feasible because of the sheer size of the community, as it is so commonly the case in MOOCs. More generally, this work can also be applied to online communities where rat-

**Corresponding author: Tel.: (+34) 93 580 9570 ; fax: (+34) 93 580 9661;
e-mail: patricia@iiia.csic.es (Patricia Gutierrez)

[1] http://peerflow.iiia.csic.es

ings of objects are needed by an assessor but cannot be accomplished because of the large quantity of items that need to be assessed. As a result, the assessor can rely on peer assessments, by giving more weight to peer assessments whose providers had a closer assessment profile to the assessor in question (that is, those whose past assessments were similar to the past assessments of the assessor in question). For instance, think of a senior program committee member in a large peer review process who needs to decide what are the final marks of reviewed papers, or a user in an e-commerce scenario where the user needs to build up the opinion about products evaluated by others.

Our inspiration comes from a use case explored in the EU-funded PRAISE project (www.iiia.csic.es/praise). This project enables online virtual communities of students with common musical interests to come together and share their music practice so the process of learning becomes social. Teachers define *lesson plans* as pedagogical workflows of activities, such as uploading recorded songs, consulting automatic performance analysis tools, engaging in peer feedback (which is considered an essential part of the learning process), or performing a reflexive pedagogy analysis. Once a lesson plan is defined, students can navigate through the activities, upload assignments, practice their assignments, assess each other, and so on. The tools developed within PRAISE allow teachers to monitor what students have done and to assess some of them. COMAS then helps teachers in the assessment process by relying on the assessments of the students to decrease the assessment load of teachers.

The contributions of this work are: 1) presenting a novel algorithm that suggests assessments for assignments in the education domain based on teacher and students' assessments; 2) experimentally showing that the algorithm works well in a real setting; and 3) experimentally showing that the algorithm scales for large numbers of students.

This paper opens with a section on related work (Section 2). We then describe in detail the Community Assessment model (Section 3). We experimentally evaluate the accuracy of our method on a real classroom dataset, as well as on simulated data considering different topologies of student interactions (Section 4). Finally, we conclude with a discussion of the results and future work (Section 5).

2. Related Work

Previous works have proposed different methods of peer assessment as part of the learning process with the added advantage of helping teachers in the sometimes daunting task of marking large quantities of students.

Piech et al. (2013) propose a method to estimate student reliability and correct student biases. They assume the existence of a true score for every assignment, which is unobserved and to be estimated. Every grader is associated with a bias, which reflects the grader's tendency to inflate or deflate his or her assessments with respect to the true score. Also, graders are associated with a reliability, which reflects how close the grader's assessments tend to land near the corresponding true score, after having them corrected for bias. Authors infer the value of these unobserved variables using known approximated inference methods such as Gibbs sampling. The models proposed are therefore probabilistic and they are compared to the grade estimation algorithm used on Coursera's platform, which does not take into account individual biases and reliabilities.

de Alfaro and Shavlovsky (2013) propose the CrowdGrader framework, which defines a crowdsourcing algorithm for peer evaluation. The authors claim that, when performing evaluations, relying on a single evaluator is often impractical and can be perceived as unfair. The method then combines students' assessments into one suggested assessment for each assignment, relying on a reputation system. The reputation of each student (or their accuracy degree) is measured by comparing the student's assessments with the assessments of their fellow students for that same assignment. In other words, the reputation of a student describes how far are his assessments from those of his fellow students. The suggested assessment is calculated by aggregating all student assessments weighted by the reputation of the students providing them. The algorithm executes a fixed number of iterations using the consensus grade to estimate the reputation (or accuracy degree) of students, and in turn uses the updated student's reputation to compute more precise suggested assessments.

PeerRank (Walsh, 2014) is based on the idea that the grade of an agent is constructed from the grades it receives from other agents, and the grade an agent gives to another agent its weighted by the grading agent's own grade. Thus, the grade of an agent is calculated as a weighted average of the grades of the agents evaluating the agent, and the grades of the evaluators are themselves weighted averages of the grades of other agents evaluating them. The method is defined by a fixed point equation, similar to the PageRank method where web-pages are ranked according to the ranks of the web-pages that link to them.

Wu et al. (2015) investigates consensus building between a group of experts in a trust network. New trust relationships are derived from the trust network and the trust scores of such relationships are calculated using an averaging operator that aggregates trust/distrust values from multiple trust paths in the network. The trust score is used to distinguish the most trusted expert from the group and, ultimately, to drive the aggregation of the individual opinions in order to arrive at a group consensual decision making solution. This work also includes a visual consensus model to identify discordant opinions, produce recommendations to those experts that are furthest from the group, and show future consensus status if experts are to follow the recommendations.

Collaborative Filtering (Shardanand and Maes, 1995) is a social information filtering algorithm that recommends content to users based on their previous preferences or ratings, exploiting the similarities between the tastes of different users when recommending items. The basic idea is as follows:

1. The system maintains a user profile, which is a record of the user's ratings over specific items.
2. It computes a similarity measure between users' profiles.
3. It recommends items to users with a rating that is a weighted average of the ratings on that item given by other

users. The weights are the similarity measures between the profiles of users rating the item and the profile of the user receiving the recommendation.

In this paper, and differently from previous works, we define the *reliability of a student as a distance between the student's assessments and the teacher's assessments over the same assignments*. To compute such a reliability measure, we build a trust network conformed of direct and indirect trust values among community members. Direct trust values are derived from common assessments while indirect trust is based in the notion of transitivity. Our target is to accurately estimate unknown assessments from the teacher's point of view, based on the students assessments and reliability.

In the experimental evaluation of our system, we compare with CF but not with others because CF is the only one that biases the final computation towards the opinion of a particular member of the community. Furthermore, CF has been widely adopted by the industry. Recommendation services, as the ones provided by Amazon, Youtube or Last.fm, are typical examples of services based on the CF algorithm.

3. Community Assessment

In this section we introduce COMAS.

3.1. Notation and preliminaries

We say an online course has a teacher τ, a set of students $\mathcal{S}$, and a set of assignments $\mathcal{A}$ that need to be marked by the teacher and/or students with respect to a given set of criteria C.

The suggested assessment state S is then defined as the tuple:

$$S = \langle R, \mathcal{A}, C, \mathcal{L} \rangle$$

$R = \{\tau\} \cup \mathcal{S}$ defines the set of possible referees (or markers), where a referee could either be the teacher τ or some student $s \in \mathcal{S}$. $\mathcal{A}$ is the set of submitted assignments that need to be marked. $C = \langle c_1, \ldots, c_n \rangle$ is the set of criteria that assignments are marked upon. $\mathcal{L}$ is the set of marks (or assessments) made by referees, such that $\mathcal{L} : R \times A \rightarrow [0, \lambda]^n$ (we assume marks to be real numbers between 0 and some maximum value λ). In other words, we define a single assessment as: $\mu_\alpha^\rho = \langle m_1, \ldots, m_n \rangle$, where $\alpha \in \mathcal{A}$, $\rho \in R$, and $m_i \in [0, \lambda]$ describes the mark provided by the referee ρ on criteria c_i.

Similarity between marks. We define a similarity function $sim : [0, \lambda]^n \times [0, \lambda]^n \rightarrow [0, 1]$ to determine how close two assessments μ_α^ρ and μ_α^η are. We calculate the similarity between assessments $\mu_\alpha^\rho = \{m_1, \ldots, m_n\}$ and $\mu_\alpha^\eta = \{m'_1, \ldots, m'_n\}$ as follows:

$$sim(\mu_\alpha^\rho, \mu_\alpha^\eta) = 1 - \frac{\sum_{i=1}^{n} |m_i - m'_i|}{\sum_{i=1}^{n} \lambda} \quad (1)$$

This measure satisfies the basic properties of a fuzzy similarity (Godo and Rodríguez, 2008). Other similarity measures could be used.

3.2. Trust relations between referees

Teachers need to decide how much can they trust the assessments made by students. We define this trust measure based on the following two intuitions. Our first intuition states that if the teacher and the student have both assessed the same assignment, then the similarity of their marks can give a hint of how close the judgments of the student and the teacher are. Similarly, we can define the similarity of judgments of any two students by looking into the common assignments evaluated by both of them. However, cases may arise where there are simply no assignments evaluated by both the teacher and some particular student. In such a case, one may think of simply neglecting (or not taking into account) that student's opinion (or mark) as the teacher would not know how much to trust that student's mark. Our second intuition, however, proposes an alternative approach for such cases, where we approximate that unknown trust between the teacher and the student by looking into the chains of trust between the teacher and the student through other students. In the following, we define these two intuitions through two different types of trust relations:

- *Direct trust*: This is the trust between referees $\rho, \eta \in R$ that have at least one assignment assessed in common. The trust value is the average of their assessments' similarity over the assignments assessed in common. Let the set $A_{\rho,\eta}$ be the set of all assignments that have been assessed by both referees. That is, $A_{\rho,\eta} = \{\alpha \mid \mu_\alpha^\rho \in \mathcal{L} \text{ and } \mu_\alpha^\eta \in \mathcal{L}\}$. Then,

$$T_D(\rho, \eta) = \frac{\sum_{\alpha \in A_{\rho,\eta}} sim(\mu_\alpha^\rho, \mu_\alpha^\eta)}{|A_{\rho,\eta}|} \quad (2)$$

  One may also think of defining direct trust as the conjunction of the similarities for all common assignments as:

$$T_D(\rho, \eta) = \bigwedge_{\alpha \in A_{\rho,\eta}} sim(\mu_\alpha^\rho, \mu_\alpha^\eta) \quad (3)$$

  However, this would not be practical, as a significant difference in just one assessment of those assessed by two referees would make their mutual trust very low.

- *Indirect trust*: This is the trust between referees $\rho, \eta \in R$ that do not have any commonly assessed assignment. We compute this trust as a transitive measure, considering chains of referees for which there are pair-wise direct trust values. We define a trust chain between ρ and η as a sequence of referees $\langle \rho_1, ..., \rho_i, \rho_{i+1}, \ldots, \rho_{m_j} \rangle$ where $\rho_i \in R$, $\rho_1 = \rho$ and $\rho_{m_j} = \eta$ and $T_D(\rho_i, \rho_{i+1})$ is defined for all pairs (ρ_i, ρ_{i+1}) with $i \in [1, m_j - 1]$. Considering $Q(\rho, \eta)$ as the set of all trust chains between ρ and η, indirect trust is defined as follows:

$$T_I(\rho, \eta) = \max_{q \in Q(\rho,\eta)} \prod_{i \in [1, m_j - 1]} T_D(\rho_i, \rho_{i+1}) \quad (4)$$

Hence, indirect trust is based on the notion of transitivity.[2] Ideally, we would like to not overrate the trust of a teacher on a student, that is, we would like that $T_D(a,b) \geq T_I(a,b)$ in all cases. Guaranteeing this in all cases is impossible, but we can decrease the number of overtrusted students by selecting an operator that gives low values to T_I. In particular, we prefer to use the product $\prod$ operator, because this is the t-norm that gives the smallest possible values. Other operators could be used, for instance the *min* function.

Trust Graph. Direct and indirect trust values are represented in a graph:

$$G = \langle R, E, w \rangle$$

where the set of nodes R is the set of referees in S, $E \subseteq R \times R$ are edges between referees with direct or indirect trust relations, and $w : E \to [0,1]$ provides the trust value. We note by $D \subset E$ the set of edges that link referees with direct trust. That is, $D = \{e \in E \mid T_D(e) \neq \emptyset\}$. An similarly, $I \subset E$ for indirect trust, $I = \{e \in E \mid T_I(e) \neq \emptyset\} \setminus D$. Weights in w are defined as follows:

$$w(e) = \begin{cases} T_D(e) & \text{, if } e \in D \\ T_I(e) & \text{, if } e \in I \end{cases} \tag{5}$$

3.3. Trust-based community assessments

To suggest assessments, we propose to aggregate the assessments of all referees on a given assignment, taking into consideration how much trusted is each referee from the point of view of the teacher (i.e. taking into consideration the trust of the teacher on the referee in marking assignments). The computation of the final suggested assessment relies therefore on the trust graph built from past interactions.

Algorithm 1 implements the Community Assessment service (COMAS). We keep the notation (ρ, η) to refer to the edge connecting nodes ρ and η in the trust graph and $Q(\rho, \eta)$ to refer the set of trust chains between ρ and η.

The first thing the algorithm does is to build a trust graph from $\mathcal{L}$. For indirect links, we use Dijkstra's algorithm to calculate the path that maximizes the cost between the teacher and students (where the cost is the product of trust values along the path). For clarity in the pseudocode, we assume that the trust graph is calculated from scratch considering all assessments in $\mathcal{L}$. For efficiency purposes, the trust graph can be stored and updated incrementally every time a new assessment is introduced.

Once trust values are calculated/updated, final assessments are computed as follows. If the teacher marks an assignment, then the teacher's mark is considered as the final mark. Otherwise, a weighted average (μ_α) of the marks of students is calculated for this assignment, where the weight of each student is the trust of the teacher on that student. To give more importance to the opinion of highly trusted students and diminish the impact of assessments made by low trusted students, we power the weight to a factor $\omega \geq 1$. Other forms of aggregation could be considered to calculate μ_α, for instance a student assessment may be discarded if it is very far from the rest of assessments, or if the referee's trust falls below a certain threshold.

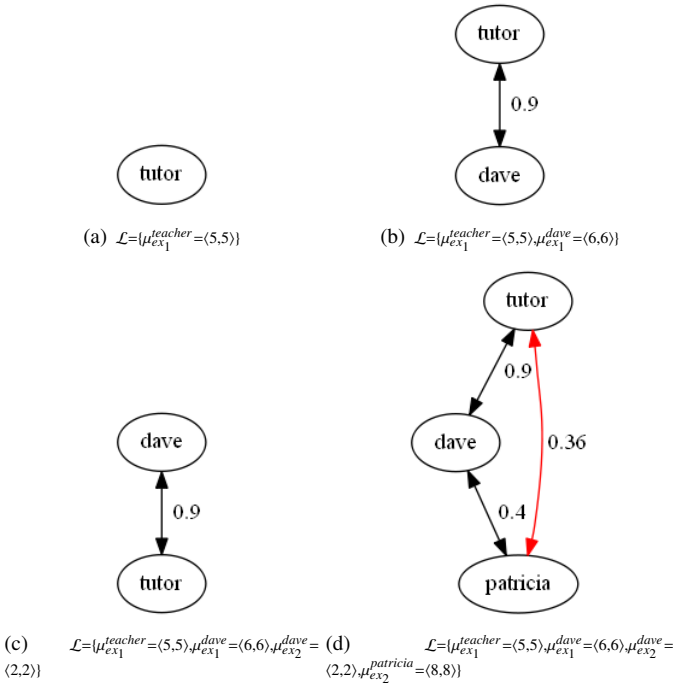
Algorithm 1: CommunityAssessment($S = \langle R, \mathcal{A}, C, \mathcal{L} \rangle$)

▷ Initial trust between referees is zero
$D = I = \emptyset$;
for $\rho, \eta \in R$ *and* $\rho \neq \eta$ **do**
  $w(\rho, \eta) = 0$;
end
▷ Calculate direct trust between students
for $\rho, \eta \in R$ *and* $\rho \neq \eta$ **do**
  $A_{\rho,\eta} = \{\beta \mid \mu^\rho_\beta \in \mathcal{L} \text{ and } \mu^\eta_\beta \in \mathcal{L}\}$;
  if $|A_{\rho,\eta}| > 0$ **then**
    $D = D \cup (\rho, \eta)$;
    $w(\rho, \eta) = T_D(\rho, \eta)$;
  end
end
▷ Calculate indirect trust between teacher & students
for $\rho \in R$ **do**
  if $(\tau, \rho) \notin D$ *and* $Q(\tau, \rho) \neq \emptyset$ **then**
    $I = I \cup (\rho, \eta)$;
    $w(\rho, \eta) = T_I(\tau, \eta)$;
  end
end
▷ Calculate suggested assessments
$assessments = \{\}$;
for $\alpha \in A$ **do**
  if $\mu^\tau_\alpha \in \mathcal{L}$ **then**
    ▷ Teacher assessments are preserved
    $assessments = assessments \cup \mu^\tau_\alpha$
  else
    ▷ Generate suggested assessments
    $R' = \{\rho \mid \mu^\rho_\alpha \in \mathcal{L}\}$;
    if $|R'| > 0$ **then**
      $\mu_\alpha = \dfrac{\sum_{\rho \in R'} \mu^\rho_\alpha * w(\tau, \rho)^\omega}{\sum_{\rho \in R'} w(\tau, \rho)^\omega}$;
      $assessments = assessments \cup \mu_\alpha$;
    end
  end
end
return $assessments$;

Figure 1 shows examples of four trust graphs built from four assessment histories that corresponds to a chronological sequence of assessments. The criteria C in this example are *speed* and *maturity* (taken from the musical domain) and the maximum mark value is $\lambda = 10$. Black edges represent direct links in D and red edges represent indirect links in I. For simplicity we only represent those referees that have made assessments in $\mathcal{L}$.

In Figure 1(a) there is one node representing the teacher who has made the only assessment over the assignment ex_1 and there are no links to other nodes as no one else has assessed anything. In (b) student Dave assesses the same exercise as the teacher and thus a link is created between them. The trust value $w(teacher, Dave) = T_D(teacher, Dave)$ is high since their marks were similar. In (c) a new assessment by Dave on ex_2 is added

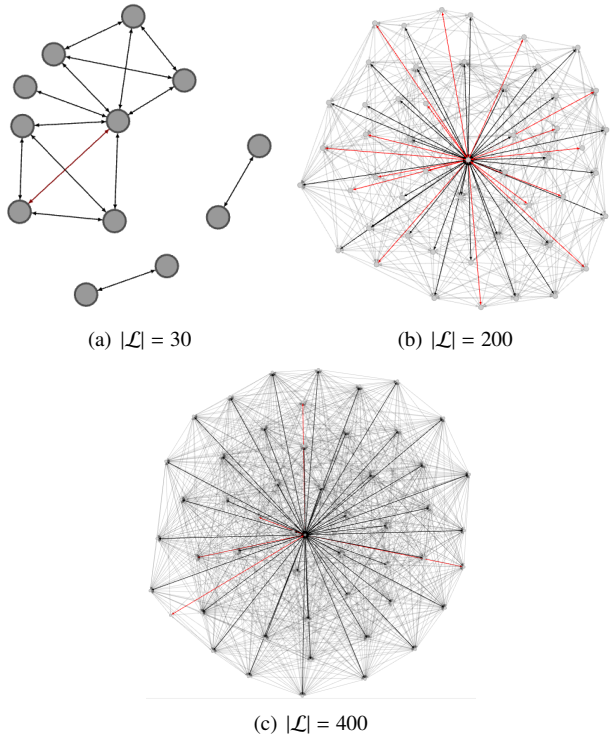
[2] T_I is based on a fuzzy-based similarity relation *sim* (Equation 1) and fulfilling the $\otimes$-Transitivity property: $sim(u,v) \otimes sim(v,w) \leq sim(u,w)$, $\forall u, v, w \in V$, where $\otimes$ is a t-norm (Godo and Rodríguez, 2008).



(a) $\mathcal{L}=\{\mu_{ex_1}^{teacher}=\langle 5,5\rangle\}$ (b) $\mathcal{L}=\{\mu_{ex_1}^{teacher}=\langle 5,5\rangle,\mu_{ex_1}^{dave}=\langle 6,6\rangle\}$

(c) $\mathcal{L}=\{\mu_{ex_1}^{teacher}=\langle 5,5\rangle,\mu_{ex_1}^{dave}=\langle 6,6\rangle,\mu_{ex_2}^{dave}=\langle 2,2\rangle\}$ (d) $\mathcal{L}=\{\mu_{ex_1}^{teacher}=\langle 5,5\rangle,\mu_{ex_1}^{dave}=\langle 6,6\rangle,\mu_{ex_2}^{dave}=\langle 2,2\rangle,\mu_{ex_2}^{patricia}=\langle 8,8\rangle\}$

Fig. 1. Trust graph example 1.



(a) $|\mathcal{L}|$ = 30 (b) $|\mathcal{L}|$ = 200

(c) $|\mathcal{L}|$ = 400

Fig. 2. Trust graph example 2

to $\mathcal{L}$ with no consequences in the graph construction. In (d) student Patricia adds an assessment on ex_2 that allows to build a direct trust between Dave and Patricia and an indirect trust between the teacher and Patricia, through Dave. The suggested assessments generated with COMAS in (d) are: $\langle 5, 5\rangle$ for exercise 1 (which preserves the teacher's assessment) and $\langle 3.7, 3.7\rangle$ for exercise 2 (which uses the weighted aggregation of the students assessments with $\omega = 1$).

Note that the trust graph built from $\mathcal{L}$ is not necessarily connected. Figure 2 shows an example of a trust graph of a particular learning community involving 50 students and a teacher. When S has a history of 30 assessments ($|\mathcal{L}| = 30$) we observe that not all nodes are connected (Figure 2 (a)). As the number of assessments increases, the trust graph becomes denser (Figure 2 (b)) and eventually it gets completely connected. In Figure 2 (c) we see a complete graph.

4. Evaluation

An assessment service is successful if teachers agree with the suggested assessments and the computation time scales up to large numbers of students. Therefore, evaluating an assessment service needs to measure: 1) the level of agreement between teacher's opinion and the suggested assessments, and 2) that this quality does not degrade with large number of students. For the first measurement, we describe in Section 4.2 the results of an experiment over real data. For the second measurement, we show in Section 4.3 that the algorithm scales for a large number of *simulated* students. Section 4.1 opens with the experimental platform.

4.1. Experimental Platform

In our experimental platform we consider a course or classroom where students submit the assignments. The teacher marks all assignments (we consider this set the ground truth) and students mark some of the assignments of their fellow students.

From the existing literature on peer assessments, most works (e.g. Piech et al. (2013); de Alfaro and Shavlovsky (2013); Walsh (2014); Wu et al. (2015)) do not bias the final computation towards the opinion of any particular member of the community. In other words, every body plays an equal role in the assessments. This is a radical difference between COMAS and the existing literature, except for Collaborative Filtering (CF). CF recommends contents to users based on their previous ratings (which constitute their profile), and selects only those opinions from the community that are relevant to the profile of a particular user. As such, it biases the overall recommendation for that particular user. We adapt CF to our problem by considering the teacher and students profile as the set of their previous assessments (rates). The items being rated are therefore the submitted assignments. In the case of the CF algorithm, the suggested assessments are the recommendations that Collaborative Filtering generates for the *teacher*. Community Assessment, on the other hand, calculates the suggested assessments relying on the direct and indirect trust values of the teacher on its students, which allows us to capture the reliability of student assessments from the point of view of the teacher.

Experimental evaluation runs as follows on every benchmark:

- First, a small subset of teacher assessments a_τ is chosen

randomly. We add these teacher assessments to CF and COMAS.

- Second, we perform several iterations where:
  - We add one student assessment to CF and COMAS, chosen randomly.
  - We generate suggested marks with the available information. We set $\omega = 3$ to give more importance to the assessments of highly trusted students and less importance to less trusted ones (as explained in Subsection 3.3).
  - We count the number of suggested marks generated by CF and COMAS. Note that initially the algorithms may be unable to suggest assessments for all assignments. As more teacher or student assessments are introduced, more suggested assessments can be calculated.
  - We calculate the error of the suggested marks with respect to the ground truth for CF and COMAS. The range of the error is $[0, 1]$ and it is defined over the set of assignments $\mathcal{A}$ accordingly:

$$Error = \frac{\sum_{\alpha \in \mathcal{A}} sim(\mu_\alpha^\tau, \mu_\alpha)}{|\mathcal{A}|} \quad (6)$$

    where μ_α^τ is the teacher assessment and μ_α is the suggested assessment provided by COMAS or CF for a given assignment $\alpha \in \mathcal{A}$. When a suggested assessment for a given assignment can not be calculated (ignorance) we assume a default assessment with mark $\lambda/2$ in all evaluation criteria.

4.2. Evaluation with Real Data

In this section, we present experiments performed over real data coming from two English language classrooms (30 14 years old students). The classrooms belong to the Secondary School 'Torras i Bages' in L'Hospitalet de Llobregat, near Barcelona. Two different tasks were given to the classroom: an English composition task and a song vocabulary task. A total of 71 assignments were submitted by the students and marked by the teacher.

Student assessments took place at their computers room. Students used Google forms to answer a defined set of evaluation criteria (instructional rubrics) and assessed their fellow students during a 1 hour period. A total of 168 student assessments were completed by the students (each student assessed on average 2.4 assignments). The assessments were done using rubrics specifically devised for the tasks. One rubric considered 3 evaluation criteria: focus, coherence and grammar conventions. The second rubric considered 5 criteria: in-time submission, requirements, precision, quantity and lyrics. All criteria had marks varying from 1 (very bad) to 4 (very good). Students had to mark each criteria on their own, without consulting the teacher. Some of the criteria were objective (fulfillment of date of submission, quantity of wh-question words that appeared on a given song), but others were more subjective (focus, coherence, originality).

Tables 1, 2 and 3 present detailed results averaged over 50 executions. We consider a subset of teacher assessments $a_\tau = 4$ (teacher assessments are 5.6% of the total number of required assessments, the rest of teacher assessments are used only to calculate the error). Table 1 shows the improvement of COMAS over CF considering the number of final marks generated. We notice that the improvement is remarkable, ranging between 78.80% and 104.65%. This highlights COMAS's first point of strength in outperforming CF: increasing the number of assessments that can be calculated. Table 2 shows the improvement of COMAS over CF in terms of the error generated. Here, the error is calculated over the entire set of assignments, including assignments that receive the default mark. This highlights COMAS's second point of strength in outperforming CF: decreasing the error of the assessments calculated. Finally, Table 3 shows the error of COMAS and CF when the assignments that receive default marks are not considered. In this case, the set of assignments considered for COMAS is different from the set of assignments considered for CF. As such, we cannot compute COMAS's improvement over CF. Nevertheless, it is obvious from the results presented by Table 3 that the improvement of COMAS over CF is not as impressive as those of Table 2. Nevertheless, as simulated data will illustrate shortly, this improvement increases for slightly larger numbers of student assessments. We also note that COMAS is able to maintain a similar error with respect to CF even when its set of final marks is much larger, as shown in Table 1.

Table 1. Number of final marks generated with $a_\tau = 4$.

Average number of assessments per student	CF	COMAS	Improvement
1	10.8	19.7	82.41%
1.5	17.2	35.2	104.65%
2	24.1	44.1	82.99%
2.4	28.3	50.6	78.80%

Table 2. Error with $a_\tau = 4$, measured with respect to the entire set of assignments (including assignments that receive default assessments)

Average number of assessments per student	CF	COMAS	Improvement
1	0.4185	0.3850	8.00%
1.5	0.3955	0.3265	17.45%
2	0.3713	0.2939	20.85%
2.4	0.3563	0.2674	24.95%

Figure 3 presents further results for $a_\tau = \{2, 4, 6, 8\}$ (teacher assessments vary from 2.8% to 11.2% of the total number of assessments needed). On the left side the precision improvement of COMAS over CF is shown. In the center, we can see the number of final assessments provided by each algorithm on every iteration. The right side graphics are commented later.

Table 3. Error with $a_\tau = 4$, measured with respect to different sets of assignments for CF and COMAS (as assignments that receive default assessments are not considered, and these assignments are different for CF and COMAS)

Average number of assessments per student	CF	COMAS
1	0.1884	0.1996
1.5	0.1944	0.2070
2	0.2014	0.2036
2.4	0.1996	0.1999

Results show that the number of final marks produced by COMAS is significantly higher than the ones provided by CF while the precision increases. We attribute this effect to the indirect trust relations that COMAS is able to calculate based on student's interaction. Indirect trust values allow COMAS to take more information into account when calculating the final mark, which permits the calculation of a larger number of marks with a higher accuracy. These metrics improve as more student assessments are considered. As more teacher assessments are provided (higher a_τ), the number of final marks produced also increase in both COMAS and CF.

Finally, on the right side of of Figure 3 we present a measure of COMAS's error for indirect trust values. We calculate such measure by removing a particular direct link from the trust graph and calculating the alternative indirect trust link. Then, we calculate the distance between the direct and the indirect trust values. We do this for every possible direct link. This error gives us an idea of the accuracy of the generated trust graph. We can observe how the error diminishes as the number of student assessments increases in all scenarios. This indicates that indirect trust values in fact become more accurate over time.

At the end of the experiments, we were able to calculate a ranking of students based on the direct trust measure between the students and the teacher. Overall, the teacher's feedback about the adjustment of such ranking compared to her experience in the classroom was very positive, providing the teacher also with an additional metric about the performance of the classroom with respect to student assessments.

We must take into consideration the fact that these peer-assessment exercises were the first to be performed by the majority of the students, who were still learning how to use the tool. So far, the results are getting better and better, as students are becoming more familiar with the rubrics and Google forms, and they are becoming less shy and more assertive in their assessments. This experiment will continue to be carried out during the whole school year, with different rubrics presented to the students.

4.3. Evaluation with Simulated Data

In this section we extend the results observed in data coming from real classrooms to larger classrooms. For this, we simulate experiments with one teacher and 100 students.

We define the marks given in the simulation by building assessment functions for the teacher and the students. We follow the same distribution of mark distances between the teacher and the students observed in real data. We define which student assesses the assignments of which other student in our simulation by considering different social network structures among students.

4.3.1. Assessment Functions

For every assignment $\alpha \in \mathcal{A}$, we define the assessment function $f_\tau : \mathcal{A} \rightarrow [0, \lambda]^n$. This function essentially describes what mark would the teacher give to α, if s/he decided to assess it. The values of this function are generated randomly.

For every assignment $\alpha \in \mathcal{A}$, we also define an assessment function for each student $\rho \in \mathcal{S}$, $f_\rho : A \rightarrow [0, \lambda]^n$, such that: $sim(f_\rho(\alpha), f_\tau(\alpha)) \leq d_\rho$, where $d_\rho \in [0, 1]$ specifies how close the student's assessments are to that of the teacher. Therefore every student in the simulation is characterized by this closeness, which defines the quality of their assessments. We sampled d_ρ from the distance probability distribution between the teacher and the students in the real dataset.

4.3.2. Social Network Generation

We assume students in online learning communities prefer to assess the assignments of students they know, or have a certain social relationship with, as opposed to picking random assignments. For instance a social relation will be born between two students if they interact with each other, say by collaboratively working on a project together or by chatting on the forum of the course.

We define a social network as a graph $\mathcal{N}$ where the set of nodes are the members of the learning community: the teacher and students, and edges represent their social ties. We rely on such social networks to simulate which student will assess the assignment of which other (neighboring) student.

We clarify that social networks are different from the trust graph of Section 3. While the nodes of both graphs are the same, edges of the social network represent social ties, whereas edges in the trust graph represent how much does one referee trust another in judging the others' work.

Several models have been proposed to simulate social networks. Topological and structural features of such networks have been explored in order to understand which generating model resembles best the structure of real communities (Ferrara and Fiumara, 2011; Phelps, 2013). Here we followed three different approaches:

- **Random Network**. The Erdős-Rényi model (Erdős and Rényi, 1959) for random networks consists of a graph containing n nodes connected randomly. Each possible edge between two vertices may be included in the graph with probability p and may not be included with probability $(1 - p)$. In addition, in our case there is always an edge between the node representing the teacher and the rest of nodes, as the teacher knows all of its students (and may eventually mark any of them). The degree distribution of random graphs follows a Poisson distribution.

  Figure 4(a) shows an example of a random graph with nodes and $p = 0.5$ and its degree distribution. Note that the point with degree 50 represents the teacher node while the rest of the nodes degree fit a Poisson distribution.

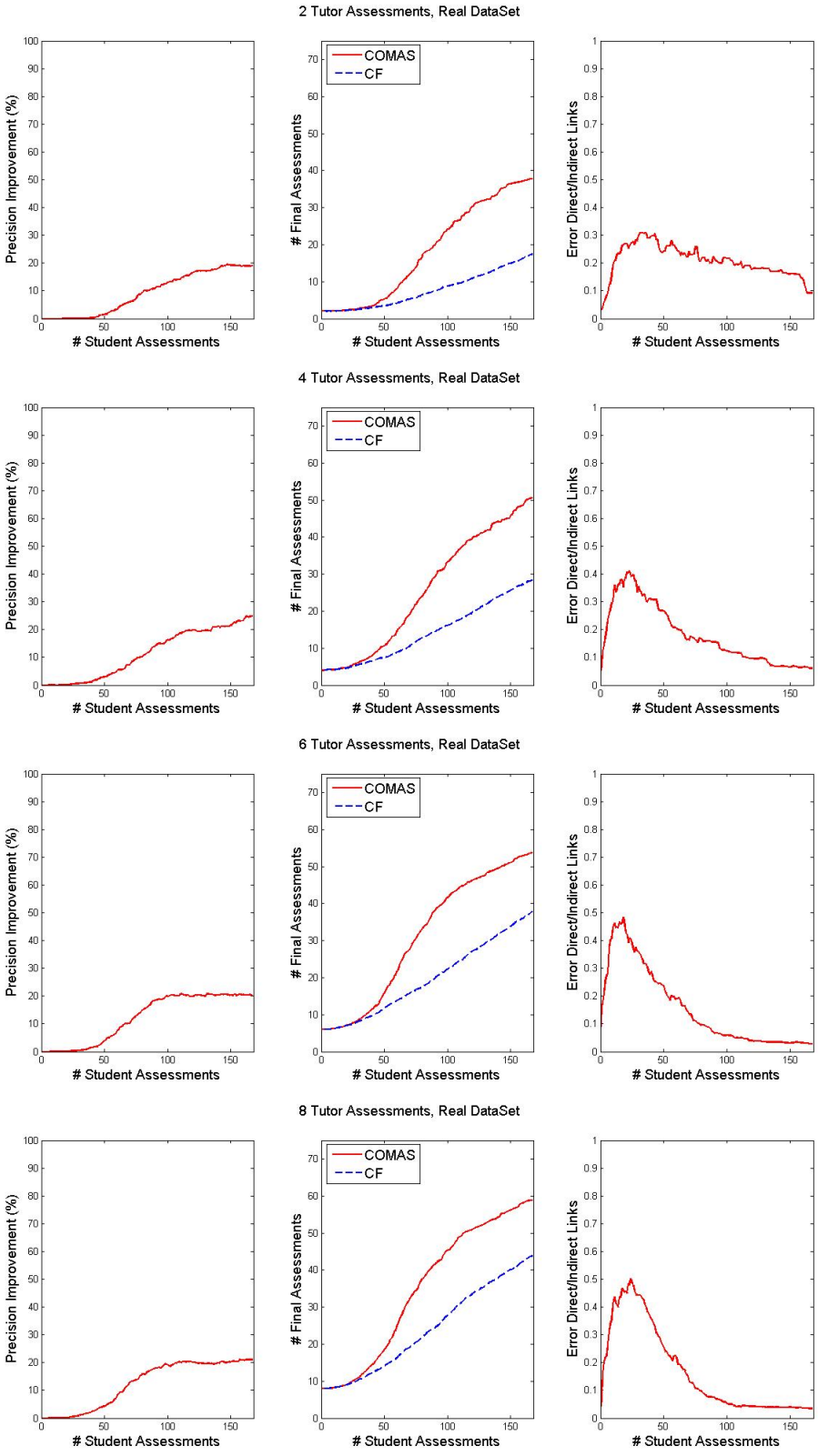


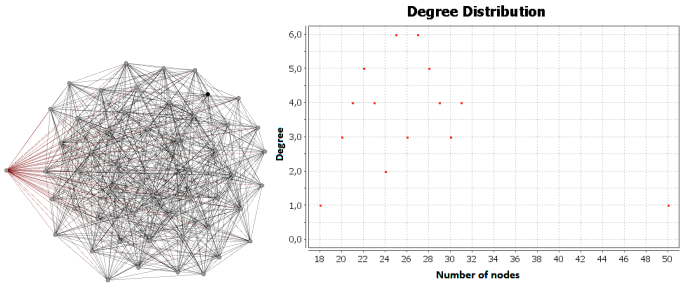
Fig. 3. Experiments with Real Data

- **Power Law Network**. The Barabási-Albert model (Barabási and Albert, 1999) for power law networks base their graph generation on the notions of *growth* and *preferential attachment*. Nodes are added one at a time. Starting with a small number of initial nodes, at each time step we add a new node with m edges linked to nodes already part of the network. In our experiments, we start with $m + 1$ initial nodes. The edges are not placed uniformly but preferentially in proportion to the degree of the network nodes. The probability p that a new node is connected to a node i already in the network depends on the degree k_i of node i, such that: $p = k_i / \sum_{j=1}^{n} k_j$. As above, there is also always an edge between the node representing the teacher and the rest of nodes. The degree distribution of this network follows a Power Law distribution.

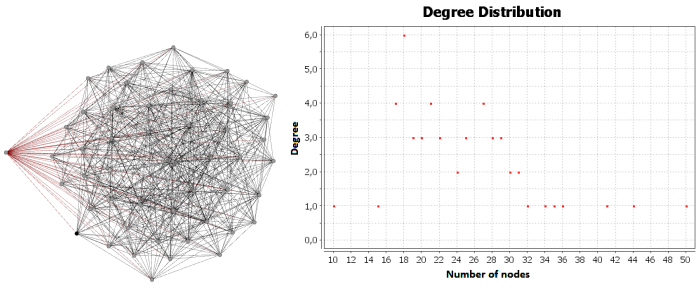
  Figure 4(b) shows an example of a power law graph with 51 nodes and $m = 16$ and its degree distribution. The point with degree 50 describes the teacher node while the rest of the nodes closely resemble a power law distribution. Recent empirical results on large real-world networks often show, among other features, their degree distribution following a power law (Ferrara and Fiumara, 2011).

- **Cluster Network**. As our focus is on learning communities, we also experiment with a third type of social network: a cluster network which is based on the notions of *groups* and *hierarchy*. Such networks consists of a graph composed of a number of fully connected clusters (where clusters may represent classrooms or similar pedagogical entities). Additionally, as above, all the nodes are connected with the teacher node.

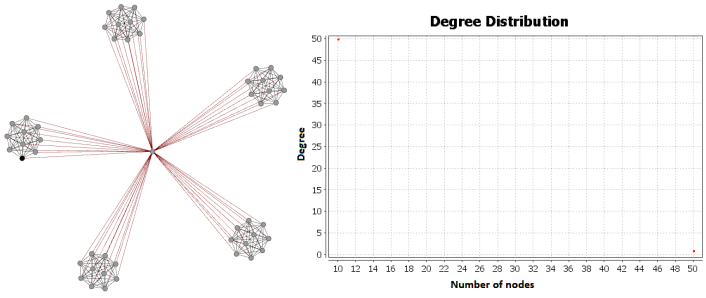
  Figure 4(c) shows an example of a cluster graph with 51 nodes, 5 clusters of 10 nodes each and its degree distribution. The point with degree 50 describes the teacher while the rest of the nodes have degree 10, since every student is fully connected with the rest of the classroom.



(a) Random Network (approx graph density 0.5)



(b) Power Law Network (approx graph density 0.5)



(c) Cluster Network (approx graph density 0.2)

Fig. 4. Social Network generation examples

4.3.3. Evaluation

We run the experiments in the three types of social networks introduced earlier: random social networks (with 100 nodes, $p = 0.5$, and approximate density of 0.5), power law networks (with 100 nodes, $m = 32$, and approximate density of 0.5), and cluster networks (with 100 nodes, 5 clusters of 20 nodes each, and approximate density of 0.2).

We perform 500 iterations of student assessments (each student marks 5 other fellow students in average). In every iteration, one student assessment is chosen randomly, where the evaluated student and the referee are neighbors in the social network. We have three evaluation criteria and maximum mark $\lambda = 10$.

Detailed results for $a_\tau = 5$ (5 % of the total number of required assessments) in the Random social network case appear in Tables 4,5 and 6, averaged over 50 executions. Further results with $a_\tau = \{5, 10, 15, 20\}$ (teacher assessments vary from 5% to 20% of the total number of required assessments) are presented in all social network topologies on Figure 5. On the left side we show the precision improvement and on the right side the number of final assessments.

Table 4. Number of final marks generated with $a_\tau = 5$.

Average number of assessments per student	CF	COMAS	Improvement
1	9.3	12.2	31.18%
2	20.4	56.6	177.45%
3	36.2	92.5	155.52%
4	53.2	98.6	85.34%
5	68.1	99.6	46.26%

Results observed in real data are mostly maintained in the synthetic data when scaling to larger social networks (where we note that similar results were observed in all topologies). In other words, Table 4 (similar to Table 1) highlights COMAS's first point of strength in outperforming CF: remarkably increasing the number of assessments that can be calculated. Table 5 (similar to Table 2) highlights COMAS's second point of strength in outperforming CF: remarkably decreasing the error of the assessments calculated. However, Table 6 illustrates

Table 5. Error with $a_\tau = 4$, measured with respect to the entire set of assignments (including assignments that receive default assessments)

Average number of assessments per student	CF	COMAS	Improvement
1	0.4981	0.4969	0.24%
2	0.4951	0.4656	5.96%
3	0.4906	0.3886	20.79%
4	0.4755	0.3440	27.66%
5	0.4569	0.3205	29.85%

Table 6. Error with $a_\tau = 4$, measured with respect to different sets of assignments for CF and COMAS (as assignments that receive default assessments are not considered, and these assignments are different for CF and COMAS)

Average number of assessments per student	CF	COMAS
1	0.2506	0.2494
2	0.2484	0.2356
3	0.2433	0.2041
4	0.2342	0.1781
5	0.2227	0.1626

that even when the error of COMAS and CF is calculated by not considering the assignments that receive default marks, COMAS still achieves remarkably lower error than CF for cases when the average number of student assessments is larger than or equal to 3. We were not able to observe this behaviour in the results of the real dataset (Table 3) because the average number of students' assessments in the real dataset was less than 3.

5. Conclusion

The paper has presented a novel algorithm that suggests assessments for assignments in the education domain based on students' assessments. We have experimentally showed that the algorithm works well in a real setting, and that it scales for large numbers of students. Furthermore, we illustrate that COMAS outperforms the infamous CF algorithm in two different ways: (1) by remarkably increasing the number of assessments that can be calculated, and (2) by remarkably decreasing the error of the assessments calculated.

The application presented in this paper is specially useful in the context of MOOCs, where there is a low number of teacher assessments and students are encouraged to interact and assess one another. Direct and indirect trust measures can then be calculated among students and COMAS can suggest assessments accordingly. Experimental results show our method is able to calculate a significant number of assessments with a low error in cases where teacher information is limited.

We foresee several lines of future work. Error indicators can be designed and displayed to the teacher managing the course, for instance to inform the teacher about the assignments that have not received any assessments yet, or the suggested marks that are considered unreliable. For example, a suggested mark

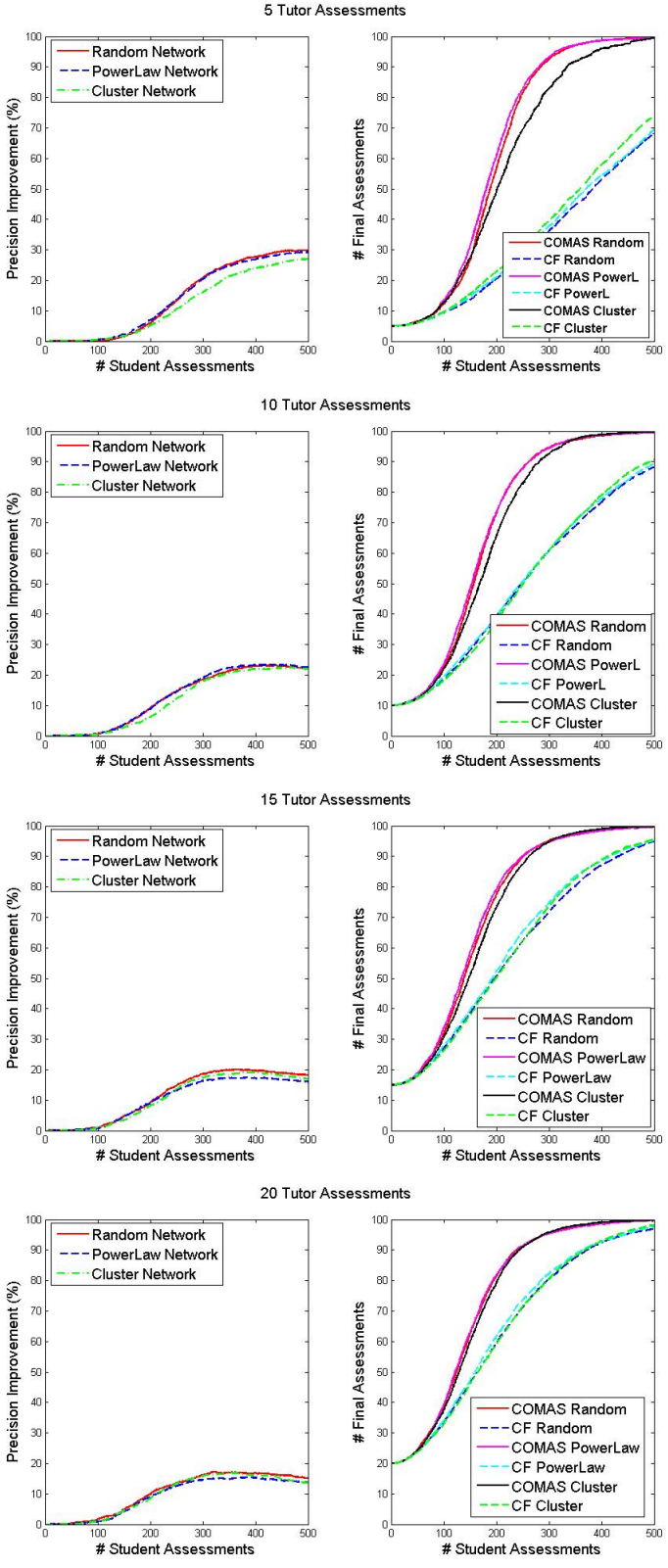


Fig. 5. Experiments with Simulated Data on different social network topologies of 100 students. It is hard to distinguish between different topologies in the COMAS and CF case because differences are minimal.

on a given assignment may be considered unreliable if all, or the majority, of students providing assessments for that assignment are considered not trustworthy as their trust falls below a preselected threshold. Alternatively, a reliability measure may also be assigned to the computed trust measure T_D. For instance, if there is only one assignment that has been assessed by τ and ρ, then the computed $T_D(\tau, \rho)$ will not be as reliable as having a high number of assignments assessed by τ and ρ. As such, some reliability threshold may be used that defines what is the minimum number of assignments that both τ and ρ need to assess for $T_D(\tau, \rho)$ to be considered reliable. Providing such error indicators can help the teacher decide whether to assess more assignments, which would then result in improving the error or it may increase the set of suggested assessments. Finally, if the error reaches a satisfactory level for the teacher, the teacher can decide to endorse and publish the marks. In addition, the trust measures themselves can be an indicator for the teacher of the student's learning progress over time.

Another interesting line for future work could focus on highlighting the missing connections in the trust graph, that if introduced they would improve the graph's connectivity, maximize the number of direct edges, or decrease the error. The question that follows then is: what assignments should be recommended to which students to assess next, such that the trust graph and the overall assessment outcome would improve? Additionally, future work may also study different approaches for calculating indirect trust values between referees. In this paper, we use the product operator. We suggest to study a number of operators, and run an experiment to test which is most suitable. To do such a test, we may calculate the indirect trust values for edges that do have a direct trust measure, and then see which approach for calculating indirect trust gets closest to the direct trust measures.

Finally, a graphical representation of the consensus status of the class, as the one proposed in Wu et al. (2015) for a social network, could be useful as a pedagogical tool. Students could see graphically how close/far their assessments are with respect to the teacher and their students, and teachers could visually identify assignments with discordant opinions. This tool could complement the reliability ranking of students discussed in Section 4.2, generated for the teacher and based on the direct trust measure between the students and the teacher.

6. Acknowledgments

This work is supported by the CollectiveMind project (Spanish Ministry of Economy and Competitiveness, under grant number TEC2013-49430-EXP), the MILESS project (Spanish Ministry of Economy and Competitiveness, under grant number TIN2013-45039-P) and the PRAISE project (funded by the European Commission, under grant number 388770).

References

de Alfaro, L., Shavlovsky, M., 2013. Crowdgrader: Crowdsourcing the evaluation of homework assignments. Thech. Report 1308.5273, arXiv.org .

Barabási, A., Albert, R., 1999. Emergence of scaling in random networks. Science .

Bickmore, T.W., Vardoulakis, L.M.P., Schulman, D., 2013. Tinker: a relational agent museum guide. Autonomous Agents and Multi-Agent Systems .

Erdős, P., Rényi, A., 1959. On random graphs. Publicationes Mathematicae .

Ferrara, E., Fiumara, G., 2011. Topological features of online social networks. Communications in Applied and Industrial Mathematics .

Godo, L., Rodríguez, R., 2008. Logical approaches to fuzzy similarity-based reasoning: an overview. Preferences and Similarities .

Hannon, V., 2009. 'Only connect!' : a new paradigm for learning innovation in the 21st century. Centre for Strategic Education occasional paper ; no. 112, September 2009, Centre for Strategic Education, East Melbourne, Vic.

Jenkins, H., 2009. Confronting the challenges of participatory culture: Media education for the 21st century .

de Jonge, D., Osman, N., i Gui, B.R., Sierra, C., 2014. Electronic institutions for community building (v2). Deliverable D3.2. PRAISE - Practice and peRformance Analysis Inspiring Social Education. Http://www.iiia.csic.es/praise/files/deliverables/PRAISE_D3.2.pdf.

Lu, J., Zhang, Z., 2012. Understanding the effectiveness of online peer assessment: A path model. Journal of Educational Computing Research 46, 313–333. doi:10.2190/EC.46.3.f.

Phelps, S., 2013. Emergence of social networks via direct and indirect reciprocity. Autonomous Agents and Multi-Agent Systems .

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D., 2013. Tuned models of peer assessment in moocs. Proc. of the 6th International Conference on Educational Data Mining (EDM 2013) .

Shardanand, U., Maes, P., 1995. Social information filtering: Algorithms for automating "word of mouth", ACM Press. pp. 210–217.

Stepanyan, K., Mather, R., Jones, H., Lusuardi, C., 2009. Student engagement with peer assessment: A review of pedagogical design and technologies, in: Spaniol, M., Li, Q., Klamma, R., Lau, R. (Eds.), Advances in Web Based Learning ICWL 2009. Springer Berlin Heidelberg. volume 5686 of *Lecture Notes in Computer Science*, pp. 367–375. doi:10.1007/978-3-642-03426-8_44.

Topping, K., 1998. Peer assessment between students in colleges and universities. Review of Educational Research 68, 249–276. doi:10.3102/00346543068003249.

Walsh, T., 2014. The peerrank method for peer assessment, in: Schaub, T., Friedrich, G., O'Sullivan, B. (Eds.), ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014), IOS Press. pp. 909–914. doi:10.3233/978-1-61499-419-0-909.

Wu, J., Chiclana, F., Herrera-Viedma, E., 2015. Trust based consensus model for social network in an incompletelinguistic information context. Applied Soft Computing .

Yee-King, M., Confalonieri, R., de Jonge, D., Hazelden, K., Sierra, C., d'Inverno, M., Amgoud, L., Osman, N., 2013. Multiuser museum interactives for shared cultural experiences: an agent based approach. Proc. of 12th International Conference on Autonomous Agents and Multiagents Systems .