Modeling Teamwork through Social Networks
Pedro Terras Crespo
Instituto Superior Técnico,
Technical University of Lisbon

Abstract. Teamwork plays an important role in education, aiming for developing a set of skills only achievable through social interaction. The advances of educational data mining already allow for predicting students performance, but the accuracy of those methods is deeply impaired when teamwork results have to be estimated. The social nature of teamwork opens a window over the problem, since existing algorithms for the analysis of social networks are indeed effective. In this paper we propose to model teamwork through social networks, and make use of PageRank and its variants to improve the prediction of team results. In particular, we describe two different graph-based models to represent the interaction among students and explore three algorithms to analyze those models. Experimental results show that the more informed are the methods the better is the prediction, despite the model used.

Keywords. Educational data mining, Social network analysis, Teamwork, Prediction

INTRODUCTION
Predicting students’ results, along with student modelling, is one of the fundamental goals of educational data mining [1]. Actually these two issues are deeply connected, and improving one of them almost directly implies improving the second one.

The literature on both topics is vast, and the results are admirable. Nevertheless, there are some evidences on the difficulty of predicting or modelling students when teamwork is present [2]. Actually, this issue was rarely addressed in the data mining literature ( [3], [4]), relying almost completely on studies in the area of psychology [5], [6]. There are several concurrent issues to understand team working: skills and knowledge of individual members, the potential social cohesion among team members, and the task management ability, are the keys to improve results [7].

Undeniably, the first issue has been addressed by the fields of student modelling and performance prediction, but the second and third issues remain mostly unsolved by the rare approaches developed. The development of methods for social networks analysis, however, brings a new insight into the exploration of the social relations, revealing to be a powerful tool to model those relations.

In this paper, we describe teamwork using two different graph-based models, and show that these models satisfy the most important properties of social networks. Additionally, we explore PageRank [8] and two variants to assess how do these methods contribute to identify a score per each student, determining his value from the team point of view. The usefulness of this new score is assessed through the comparison of the accuracy of classification methods trained with and without these new scores.

The rest of the paper is organized as follows: next, we briefly overview the work done in the area of teamwork prediction and the different approaches based on social network analysis performed over educational data. Following this, we described two graph-based models to represent teamwork and then, we demonstrate the applicability of our proposal. In order to do that, we perform a deep analysis of both models, first through a statistical point of view, and then by applying ranking algorithms to identify a score for each student, that is used for prediction purposes. The paper concludes with a summary of the results achieved, along with a critical analysis and some guidelines for future work.
BACKGROUND

The concept of Social Network, typically seen as interactions between individuals, has become extremely popular in the last decade due to its huge application in the online domain in websites such as Facebook, Twitter or LinkedIn. Although the structure of such Social Networks can be represented in terms of a graph, they often enclose tremendous amounts of information in terms of the linkage between individuals and content shared among them [9].

Social networks were applied to educational data just a couple of years ago, mainly dedicated to produce recommender systems through collaborative filtering techniques. In this context, the major effort is on recommending the most adequate set of learning resources for a specific student [10]. The contributions in this area are vast; see for example [11] and [12] to name a few. Other social-network based systems were proposed for education, either for analysing the educational communities, for modelling students, through the understanding of their communication [13] or for detecting collaboration patterns [14].

Despite this variety, social based approaches proposed for team recommendation are scarce with [15] and [4] some of the exceptions. However, none of them discuss neither evaluate different models for representing social networks, which does not allow to fully understand if different models will lead to better results. A detailed survey of the educational data mining approaches based on social network analysis can be found in [16].

GRAPH-BASED MODELS FOR REPRESENTING TEAMWORK

A social network is defined as a network of interactions or relationships, where nodes usually represent actors and the edges consist of the relationships or interactions between these actors [9]. A social network can be represented in terms of a graph $G = (V, E)$ defined as a set of vertices or nodes $V$ and edges $E$.

Teamwork is a particular case of a social interaction, since in this context students establish a mutual relationship, sharing common goals and tasks. Teamwork plays an important role in the personal development, and is almost present in every educational program. In particular, students enrol in a variety of teams along their studies, either because the team does not perform well or just due to different academic pairs followed by team members, resulting in a set of entangled social relationships among several students. From this reality, it is natural to accept that teamwork in general can be modelled as a social network, where students are the main actors, represented as nodes in the graph, and relations among them are represented through its edges.

Despite the structure inherent to the graph that represents the network, it is usual to associate additional data to it, in particular to each node. This data, usually known as content-based component, contains other kinds of non-linkage data such as profile attributes, text, multimedia information or user-tags.

From the prediction point of view, all data that is available and may help to improve accuracy should be used. In this manner, data such as students’ average grade and corresponding standard deviation, or the number of failures can be used for that purpose. More informed data, such as the grades obtained by each team or individually, should also be of great importance but are not usually available, neither for training nor for prediction. Indeed, it is not expected that some student remember all his team results, individual grades per subject, and other detailed data.

Moreover, as usual in data mining: the more informed is the data, the better is the prediction accuracy. In this manner, it is expected that a richer content-based component in the social network should improve the prediction of team results.
Having established the connection between teamwork and the concept of social networks, we propose two different models for representing teamwork. These models combine both structural and content-based components, and result from the exploration of some properties of graphs, allowing for representing different aspects of students behaviour.

**Unipartite graph of students**

Our first model mostly follows the interpretation of the general social network definition as described above where students are represented by nodes and the social interactions by edges in a graph $G = (V, E)$. Since the model will be based on a unique type of nodes, the graph will be unipartite. However, some additional remarks are needed.

Graphs can be undirected if the edges do not imply a sense of direction (e.g. a friendship is a mutual interaction) or can be directed if a sense of direction is implied (e.g. one person makes a phone call to another). If the graph is undirected, the number of nodes adjacent to a node $i$ is called its degree and is represented by $d(i)$. On the other hand, if the graph is directed, the in-degree of the node $i$ represents the number of inbound connections and is denoted by $d_{in}(i)$ while the number of outbound connections is denoted by $d_{out}(i)$, and is called the out-degree.

Seeing that social interactions within teams are mutual between all members, our model will be based on an undirected graph.

But, graphs can also be categorized according to its kind of edges. A graph can be unweighted if all the edges have the same importance or can be weighted. In the second case, it is usual to have multiple edges between a single pair of nodes: each one with a different weight (e.g. duration of phone calls or monetary amounts for some transaction).

In our case, we propose two variations for unipartite graph based models: the first one using an unweighted graph, and the second one based on weighted graphs allowing for representing several works performed by the same team.

Although the structure proposed is simple, it introduces a problem of loss of information. Firstly by representing teamwork as an unweighted graph, we lose the information on how much work has been done between the students. Secondly by representing the student grades through the mean on their average, we lose the context of where they were obtained such as the team and the subject.

Second, weighted graphs continue to represent the different interactions through a single connection, loosing the context of the different teams involved. It is interesting to note that this model is not able to explicit represent teams with more than two members; despite it is able to denote all interactions performed between any pair of students. Any available data, such as the grade obtained by the team is discarded.

**Bipartite graph with students and groups**

Another interesting feature is the overall structure of the graph. As been said, if all the nodes belong to the same class the graph is said to be unipartite; if there are multiple classes of nodes the graph is said to be multipartite. In multipartite graphs edges can only be drawn between nodes of different types. Bipartite graphs are a particular case with two classes of nodes, satisfying a set of particular properties [17].

Our second model explores bipartite graphs, where nodes can both represent students and teams. The team nodes will serve as interaction mediators between student nodes participating in the same social group. Introducing team nodes allows for solving the problem of having multiple interactions between the same students while keeping the resulting graph as an unweighted graph. In this new
structure we have two kinds of nodes, being in the presence of a bipartite graph (a multipartite graph
with two classes of nodes). In this type of graph edges can only be drawn between different classes,
meaning that students will now only have edges connecting to team nodes. The resulting graph will
remain unweighted and undirected. Introducing the mediator class, team nodes, allows not only for a
change in the structure of the social network, but also allows the network to contain finer-grained data
regarding the students’ classification from each interaction.

Next, we will evaluate the proposed models, accounting for their completeness on the amount of
information retained, the ease of exploration and the impact on the prediction results.

MODELS EVALUATION
In this chapter we will apply and evaluate both models to represent teamwork. The evaluation is
performed by combining social network analysis and data mining techniques, in order to assess how
these models contribute to improve the prediction of team results. For this purpose we collected and
analyse real student data serving as a case study to validate and refine our research strategy.

The dataset is drawn from a set of students enrolled in the academic program of Information
Systems and Computer Engineering at Instituto Superior Técnico, Technical University of Lisbon in
Portugal. In this program teamwork is part of the evaluation methodology in almost every single
course, making it an excellent source for collecting team data and predicting their performance.

From this academic environment we collected a sample containing approximately 1700
evaluations of over 550 unique students. This corresponds to the data of 8 subjects over 2 years. Each
student record contains individual grades, team grades and final grades for each enrolment at a given
course, as described is the previous chapter, to be the content-based component of the models.

Statistical Analysis
One of the simplest strategies to analyse a social network is through statistical analysis, which focuses
on examining some statistical properties, in particular static properties for describing the structure of
the graph at a fixed point in time [18].

The first main property concerns the diameter of the graph defined as the maximum shortest-path
distance between any reachable pair of nodes (i.e. the minimum number of hops that takes to travel the
graph from one end to the other). In real-world graphs where temporal or spatial proximity can be
overlooked, the distance of a graph tends to follow a well-known pattern referred as small world
phenomenon or six degrees of separation. According to this theory, first introduced by Milgram [19],
each pair of actors in the planet is separated at most by six degrees of separation.

To prevent the high probability that particular cases of long chains will dispel the graph diameter
from the actual pattern in the rest of the graph, the effective diameter containing a fraction of all
connected node pairs (usually 90%) is the metric often used.

After modelling the sample according to the unipartite model (graph of students), we obtained a
social network with a maximum diameter of 13 hops, an effective diameter of 7 hops and an average
diameter of 5.5 hops. The fact that the effective diameter exceeds the six degrees of separation may be
due to restrictions in the number of elements per teamwork imposed by each academic course that
doesn’t exist in real-world social networks. However taking into consideration that the average
diameter is lower than six hops, we consider the resulting graph to follow an approximation of the
’small-world phenomenon’.

When modelling the sample according to the bipartite representation (graph of students and
teams) the metrics mentioned above roughly doubled their values, obtaining a maximum diameter of
26 hops, an effective diameter of 14 hops and an average diameter of 10 hops. This is however closely related with the edge decomposition that was made when using the bipartite structure. Each one of the edges from the first model has been decomposed into two separate edges and an intermediary team node, in the bipartite model, causing the distance between the student nodes to also double when weighting every edge with the unit value of one. However, if the edges were to be weighted with the value of 0.5, the distance distribution should remain equivalent to the one presented for the first model.

The second main static property described, called the Heavy-tailed Degree Distribution, concerns the degree distribution of each node (i.e. the number of edges present on each node) and stipulates that the distribution obeys a power law of the form \( f(d) \propto d^{-\alpha} \), with \( \alpha > 0 \) and \( f(d) \) representing the fraction of nodes with a degree of \( d \).

Analysing the Heavy-tailed Degree Distribution, we can see that the degree distribution of both representations obeys the power law \( f(d) \propto d^{-\alpha} \). The high number of nodes with a low degree could be explained by the arrival of new students to the academic program, dropouts or even failure to find other students to complete a full team.

**Ranking Analysis**

Ranking algorithms are one of the foundation stones for the social networks’ structural analysis. Link prediction, spam detection and collaborative filtering are examples of techniques that rely on ranking algorithms. Ranking a graph is actually establishing a measure of similarity between the nodes.

The PageRank algorithm [8], developed by Larry Page and Sergey Brin in 1998, is known worldwide as a tool for ranking web pages in order of their importance. PageRank is based on the principle of random walks on unipartite directed and unweighted graphs, meaning that information is diffused from one node to another [20]. The amount of importance transferred from a node \( j \) to a node \( i \) depends on both the importance of node \( j \) and the number of nodes inheriting the transfer. In the traditional PageRank algorithm the initial value of importance is the same for all nodes in a graph and the importance inherited from a node to its neighbors is made in equal portions. The rank of a node is then determined by combining the initial value of importance with the amount of importance inherited through the linkage structure through the means of a scalar called the damping value. In order to improve the accuracy of rankings, some PageRank adaptations also exist.

The Personalized PageRank [21] adaptation allows differentiating the initial value of importance of each node of the graph through the use of a personalization vector. If the nodes are differentiable in other than their linkage structure, then Personalized PageRank also differentiates them in terms of their initial importance.

The Focused PageRank adaptation [22] enables the diffusion of information between nodes to be made selectively according to personalized transition probabilities between nodes. If a node has neighbor preferences in terms of their linkage, then Focused PageRank allows for the information to be inherited in different portions.

The PageRank results shown in this section were calculated separately with damping factors of 0.85 (default) and alternatively of 0.15 despite not finding any significant differences between them. During the calculation of the PageRank results we did not guarantee that the sum of personalization vectors for a single node totalizes one, because we do not wish to interpret the result as a probabilistic distribution, but rather as a relative value of importance among students.

When exploring the unipartite model through the traditional PageRank algorithm the results have shown that there is a clear correlation between the node’s average degree and the score obtained,
meaning that the higher the degree of a node, the higher the expected score. The analysis of PageRank’s results leads us to conclude that a purely structural analysis based on the typical representation of a social network might not be suitable for capturing the value of each student in team working.

The next step is to try to influence the node scores through the use of personalized vectors containing additional data regarding the students’ grades, according to the Personalized PageRank adaptation. In this method we used the final average grade of each student, as the values for the PageRank vector. The best results with the Personalized PageRanks were achieved with a damping factor of 0.15. The damping factor regulates the amount of information inherited from other nodes but a formula to determine the best value for this parameter does not exist.

When applying the Personalized PageRank algorithm to the unipartite model using the final average grades to populate the personalization vector containing additional information about each student, the correlation between the degree of a node has faded and almost seems to no longer apply. The relationship between the Personalized PageRank and the average of team grades is now clearly correlated, proving that the use of Personalized PageRank vectors with content-based data has greatly influenced the process of structural analysis, proving to be better suited for capturing the value of each student in team working.

With a weighted analysis our goal is to overcome the unweighted graph restriction imposed on the representation structure by both traditional and personalized PageRank. Our major problem regarding this issue was that we were not able to account for the amount of work that has been done between each one of the individuals in the previous PageRank algorithms. Students that have worked multiple times together were being treated the same way as those who had only worked together once and therefore were being penalized on their ranks.

When exploring the bipartite model, the amount of work done between students is already captured in the network structure when using either the traditional PageRank or the Personalized PageRank algorithms. Once again, the correlations between the ranks and both node degree and average team grades produced similar results, proving to be characteristic of the algorithms themselves and not so much of the network structure.

**Prediction**

In order to validate our claim, we tested how much the accuracy of a classifier improves, when it is trained with social scores obtained through the described algorithms. In particular, we compare the accuracy of classifiers trained exclusively over social scores, without scores and combining scores and other data, as the available one in the network content-based component. In order to perform such comparison we trained decision trees, with the C4.5 algorithm [23]. The choice of this algorithm derives from its ability to handle both nominal and numerical attributes for the attributes in the observations. Its main drawback is not being able to handle numeric values for the class attribute, and as such, we were forced to round up the numerical team results and use it to denote the set of representative nominal classes ranging from 0 to 20 or alternatively from F to A, according to the most popular grading systems.

In order to do the comparison, we selected the best results achieved through cross-validation, and the average of several possible parameterizations on the attributes used, such as ordering the students among teams according to its scores, either ascending or descending.

From the best parametrizations identified, it is then possible to compare the accuracy of the classifiers trained exclusively from isolated attributes. **Fig. 1** presents this comparison putting ranking
scores obtained from the unipartite (left) and bipartite models (right) against the content-based popular attributes. From their analysis, it is clear that when used in isolation the results obtained on predicting team results are quite satisfactory. The PageRanks score, which reflects a purely structural analysis, has shown to be able to achieve similar results as any of the content-based attributes when applied to the unipartite model, and has even shown an improvement when applied to the bipartite model.

Influencing the structural analysis with content-based attributes through the Personalized PageRank algorithm has also achieved improvements in all scenarios, overcoming PageRank results. The same thing happens when using Focused PageRank but just slightly.

The comparison of the accuracy obtained when using the different network models can be seen in Fig. 2. We note that the accuracy over bipartite model has revealed to be slightly, but constantly, better than over the unipartite network model, either in capturing the amount of work done between students or in achieving the best accuracy levels.

Despite the results achieved, using only isolated attribute as the only source of information would hardly achieve the best possible results in predicting team grades, especially when there are other sources available that could contain complementary information. Although using more information when building the predictive model typically results in a higher percentage of correctly classified observations, it also increases the probability of adding redundant information. The decision trees’ ability for dealing with the presence of redundant information was one of the most influential factors for choosing them as the supervised classification prediction model.

For determining the best possible results, we have combined the data types described as the content-based component for usage as sources of information for generating the decision tree while varying several attribute parameterizations such as data ordering, grading system and attribute types. In those cases where ranking scores were used, we also determined the ideal value for the damping factor.

Although the accuracy obtained over bipartite network model presented better results when using ranking scores in isolation, both models presented similar results when combined with other sources of data. Fig. 3 highlights the results comparison between the most relevant source data combinations using a 5 fold cross validation evaluation. Having domain information present on every data combination allowed us to focus on the impact of the ranking scores while achieving the best possible results. Analyzing the obtained comparative results we can reach two conclusions regarding the overall accuracy on the predictive model.

- The higher the number of attributes used for training, the higher the obtained accuracy.
- Combining ranking scores with any other attribute allows for a significant increase in the classifiers’ accuracy.

It is important to distinguish, as before, among public and private data: the first is public and generally available, as the final grade of a student after finishing some course, while private data refers to data such as team or individual grades. Indeed, it is difficult for a student remember this kind of data, and it is difficult to accept the requirement of such values in order to perform any prediction. Our results show that by using a combination of all available attributes (public, private and ranking scores) as the training dataset reaches the best results, that can be use as a target. Moreover, our results demonstrates that replacing private data by ranking scores when combined with public data achieved over 85% of accuracy, which compares with the target of 87%.

**CONCLUSIONS**

Teamwork plays a fundamental role in education, aiming for the development of particular skills on students, like collaboration, leadership and confidence. Though its importance, it did not deserve the
corresponding attention in the area of educational data mining yet.

The ultimate goal of modelling teamwork is undoubtedly to be able to understand and predict if a specific team will succeed. Modeling teamwork as a social network may not be as straightforward as simply connecting students who worked together, due to having to deal with the problem of the loss of

---

**Fig. 1** Accuracy of classifiers trained exclusively with isolated attributes, over unipartite (left) and bipartite models (right)

**Fig. 2** Comparison of classifiers accuracy when trained over unipartite and bipartite models

**Fig. 3** Comparison of classifiers accuracy when trained over different datasets
information. Designing richer team models can lead to improvements in predicting team value, however the prediction accuracy is closely related with the ability of extracting the extra information present in the team models through the available ranking algorithms. In order to choose among team models, we have combined available data and ranking scores, derived from the different models and several ranking algorithms for training a classifier for predicting team results.

Experiments in real data have demonstrated our claims. First, the number and nature of data attributes used to train the prediction model directly influences the classifier accuracy. Second, ranking scores can be used alone for training such classifiers, achieving better results than any other data attribute alone (from the attributes presented). Third, ranking scores may replace private data (data usually unavailable) without losing accuracy. Fourth, among the ranking algorithms used, Focused Personalized PageRank achieved the best results when exploring unipartite graphs. Finally, bipartite graph based models seem to slightly overcome unipartite models, either in terms of the information explicitly represented or in the prediction accuracy achieved.

Although the prediction of the unknown value for teamwork results presents by itself an interesting research category on educational data mining, the results obtained can also prove to be useful in other research categories such as recommender systems, especially when the techniques involved are based on social network exploration. As a result, the team recommender system we proposed is an example of a potential application of teamwork prediction.

ACKNOWLEDGMENT
This work is partially supported by Fundação para a Ciência e Tecnologia under research project educare (PTDC/EIA-EIA/110058/2009).

REFERENCES


