Social Network Analysis and Data Mining: An Application to the E-Learning Context

Camilo Palazuelos, Diego García-Saiz, and Marta Zorrilla

Dept. of Mathematics, Statistics, and Computer Science, University of Cantabria
Avenida de los Castros s/n, 39005, Santander, Spain
{camilo.palazuelos,diego.garcia,marta.zorrilla}@unican.es

Abstract. With the increasing popularity of social networking services like Facebook or Twitter, social network analysis has emerged again. Discovering the underlying relationships between people—as well as the reasons why they arise or the type of those interactions—and measuring their influence are examples of tasks that are becoming to be paramount in business. However, this is not the only field of application in which the use of social network analysis techniques might be appropriate. In this paper, we expose how social network analysis can be a tool of considerable utility in the educational context for addressing difficult problems, e.g., uncovering the students' level of cohesion, their degree of participation in forums, or the identification of the most influential ones. Furthermore, we show that the correct management of social behavior data, along with the use of the student activity, helps us build more accurate performance and dropout predictors. Our conclusions are drawn from the analysis of an e-learning course taught at the University of Cantabria for three consecutive academic years.

Keywords: social network analysis, data mining, e-learning.

1 Introduction

Social network analysis (SNA), which consists in generating patterns that allow identifying the underlying interactions between users of different platforms, has been an area of high impact in the last years. The appearance of social networking services, such as Facebook or Twitter, has caused a renewed interest in this area, providing techniques for the development of market research using the activity of the users within those services.

However, SNA techniques do not just concentrate on social networks, but also focus on other fields, such as marketing (customer and supplier networks) or public safety [7]. One of the fields in which they are also applied is education [11]. Thanks to SNA, it is possible to extract different parameters from the student activity in online courses, e.g., the students' level of cohesion, their degree of participation in forums, or the identification of the most influential ones. This kind of analyses might be helpful for teachers to understand their students' behavior, and as a consequence, help them to get better results.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 651–660, 2013.

SNA is also useful for generating new data as attributes, which can be subsequently applied data mining techniques to obtain student behavior patterns. In the educational field, there is a well defined area called *educational data mining* [12]. Building accurate performance and dropout predictors, which help teachers to prevent students from failing their subjects, is one of the main problems tackled in this area. For this purpose, classification techniques, by means of prediction models, are usually applied to uncover the students' behavior, e.g., amount of time dedicated to accomplish certain tasks or activity in forums, that results in a pass, a fail, or a dropout. For the issue of prediction, SNA provides a new useful framework that might improve the accuracy of those models. In this paper, in which we analyze an e-learning course taught at the University of Cantabria (UC) for three consecutive academic years, we show how SNA helps to uncover behavior patterns and build models that predict the performance and dropouts of students accurately.

The paper is organized as follows. Section 2 provides background and reviews the state of the art in the use of SNA in the educational field. Section 3 describes the characteristics of the academic course under study and presents the datasets generated for the experiments. Section 4 discusses the results obtained. Finally, Sect. 5 summarizes the most important conclusions of our work and draws our future lines of research.

2 Background and Related Work

SNA is the methodical study of the relationships present in connected actors from a social point of view. SNA represents both actors and relationships in terms of network theory, depicting them as a *graph* or *network*, where each *node* corresponds to an individual actor within the network, e.g., a person or an organization, and each *link* symbolizes some form of social interaction between two of those actors, e.g., friendship or kinship. Although social networks have been studied for decades [13,14], the recent emergence of social networking services like Facebook or Twitter has been the cause of the unprecedented popularity that this field of study has now.

Despite the fact that there are antecedents in the SNA literature that push back its origin to the end of the nineteenth century, Romanian psychiatrist Jacob Moreno—who, in the early 1930s, became interested in the dynamics of social interactions—is widely credited as the founder of SNA. In his seminal book [8], Moreno established the foundations of *sociometry*, a field of study that later became SNA. Since then, an extraordinary variety of SNA techniques has been developed, allowing researchers to model different types of interactions, e.g., movie actors [15] or sexual contact networks [6], and giving solution to very diverse problems, e.g., detection of criminal and terrorist patterns [7] or identification of important actors in social networks [9,10].

In order to estimate the prominence of a node in a social network, many centrality measures have been proposed. The research devoted to the concept of centrality addresses the question “Which are the most important nodes in a social network?” Although there are many possible definitions of importance, prominent nodes are supposed to be those that are extensively connected to other nodes. Generally, in social networks, people with extensive contacts are considered more influential than those with comparatively fewer contacts. Perhaps, the most simple centrality measure is the *degree* of a node, which is the number of links connected to it, without taking into consideration the direction of the links. If we take into account that direction of the links, a node has both *indegree* and *outdegree*, which are the number of incoming and outgoing links attached to it, respectively. There are more complex centrality measures, such as the *betweenness* [4] of a node, which is equal to the number of shortest paths from all nodes to all others that pass through such a node, as well as *authorities* and *hubs* [5]; a node is an authority if its incoming links connect it to nodes that have a large number of outgoing links, whereas a node is a hub if its outgoing links connect it to nodes that have a large number of incoming links.

From a node-level point of view, centrality measures constitute a very useful tool for the inference of the importance of nodes within a network. Due to their own nature, some of them, e.g., betweenness, cannot be trivially calculated, so that *network-level metrics*, which can be computed more easily and provide helpful information by considering the network as a whole, can be used for complementing the aforementioned centrality measures. One of these network-level metrics is the *density* of the network, which measures the number of links within the network compared to the maximum possible number of links. The *diameter* of the network is also a useful network-level metric; it is defined as the largest number of nodes that must be traversed in order to travel from one node to another. Other meaningful network-level metric is the *number of connected components* of the network, i.e., the number of subnetworks in which any two nodes are connected to each other by paths without taking into consideration the direction of their links. Finally, the last metric to be mentioned is *reciprocity*; it occurs when the existence of a link from one node to another triggers the creation of the reverse link.

There are some applications of SNA to the educational field. Brewe et al. [2] used a multiple regression analysis of the Bonacich centrality for evaluating the factors that influence participation in learning communities, e.g., students’ age or gender, and Crespo and Antunes [3] proposed a strategy to quantify the global contribution of each student in a teamwork through adaptations of the PageRank algorithm. SNA can also provide relevant information that can be used in educational data mining tasks, such as predicting performance or dropouts. For instance, Bayer et al. [1] used the following centrality measures for the prediction of dropouts: degree, indegree, outdegree, and betweenness. They came to the conclusion that these measures improve the accuracy of classification models in comparison with the sole use of demographic and academic attributes, e.g., students’ age, gender, or number of finished semesters.

3 Data Characterization

Our case study uses data from a virtual course hosted on Blackboard/WebCT, entitled "Introduction to Multimedia Methods." This course was taught for three consecutive academic years (2007–2008, 2008–2009, and 2009–2010) at the UC. It was designed by means of web pages and included some video tutorials, Flash animations, and interactive elements. Students had to complete four exercises and an ordinary final exam (not online). In particular, the forum—used, in this paper, for building the social network of interactions between instructors and students—was mainly used for making questions about the organization of the course, its contents, and deadlines by students, as well as answering students' doubts and make announcements by the instructor. The average number of students enrolled in the course was 70, of which more or less the half followed the course up to the end, whereas the rest dropped out. The students' profile was diverse, coming from different degrees, such as Computer Science, Mathematics, and even History.

We created three datasets with the student activity data from Blackboard and social behavior from the social networks analyses performed with *ORA. The attributes related to the student activity are: (i) total time spent in the course, (ii) total number of sessions performed, (iii) average time spent and number of sessions performed per week, and (iv) number of messages read and written in email and forums. The SNA attributes chosen are the following centrality measures: (i) degree, (ii) indegree, (iii) outdegree, (iv) betwenness, (v) authority, and (vi) hub, as well as (vii) *top3* (if a node is ranked in the top 3 nodes in some of the previous centrality measures, *top3* is true; otherwise, it is false), and (viii) percentage in *top3*, i.e., the number of *top3*s that are set to true for a certain node. The three datasets have 194 instances, i.e., one instance per student, being the difference between them the values of the class attribute. For each student, `performance.dat` indicates whether he or she passed the subject, `dropout.dat` whether he or she dropped out the course, and `mixed.dat` whether he or she passed, failed, or dropped out.

4 Experimentation

The methodology followed in this paper includes the tasks listed below:

1. Extraction and transformation of the student activity data from Blackboard;
2. Generation of social networks with *ORA using the questions and answers present in forums;
3. Analysis of social attributes for the educational field;
4. Selection of meaningful social attributes for prediction;
5. Development of classifiers for predicting performance and dropouts;
6. Discussion and conclusions.

Figure 1 shows graphically the steps performed in our case study.

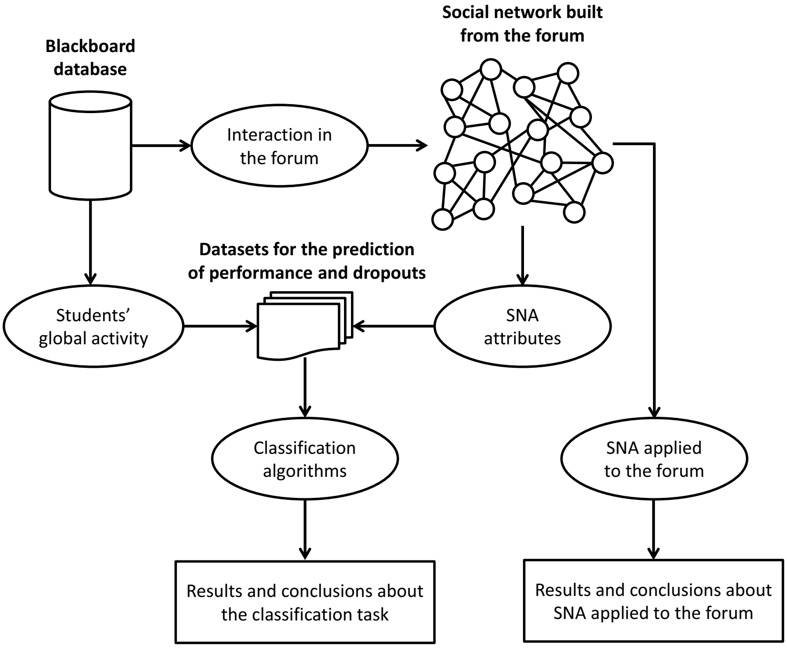
Blackboard database
Social network built from the forum
Interaction in the forum
Datasets for the prediction of performance and dropouts
Students' global activity
SNA attributes
Classification algorithms
SNA applied to the forum
Results and conclusions about the classification task
Results and conclusions about SNA applied to the forum

Fig. 1. Process of the experimentation

Figure 2 depicts the network of interactions between the instructors and the students of the course "Introduction to Multimedia Methods" taught in 2008–2009 at the UC. In this course, we found a single connected component and a diameter of 11, as well as low values of density (0.07) and reciprocity (7% of the links were reciprocal). In the previous and the subsequent courses (2007–2008 and 2009–2010), we found three and one connected components, and diameters of 19 and 16, respectively, as well as low values of density (0.05 and 0.06) and reciprocity (12% and 8%). A possible explanation for the low values detected of both density and reciprocity is that the instructor answered to the questions in the forum faster than students, preventing these from helping each other.

As can be observed, the node with more links is that corresponding to the main instructor. This means that, in this course, most interactions in the forum occur between the instructor and the students, whereas it is less frequent that those interactions occur between the students themselves. Thus, the forum is mainly used in two different ways, as the instructor pointed out: (i) students make questions—about the contents or the organization of the course—that should be answered by the instructor and (ii) the instructor makes important announcements. This kind of interactions are better showed in a graphical way by using SNA, making it easy for the instructor to interpret them, i.e., if the forum was used for a concrete activity, it could be helpful to ensure its good performance.

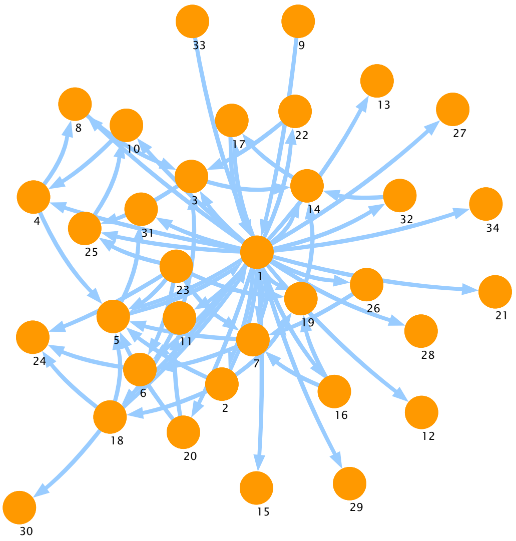

Fig. 2. Network of interactions between the instructor and the students of the course "Introduction to Multimedia Methods" taught in 2008–2009 at the UC

These conclusions can be better understood by analyzing the node centrality values exposed in Table 1. On the one hand, the instructor (node 1) has the highest values of degree and outdegree. Moreover, the difference in outdegree between the instructor and the second and the third ranked users is very high. This also happens to the betweenness and hub centrality measures. Thus, we can conclude that the instructor is the user that answered the great majority of messages posted by the students in the forum. On the other hand, the highest indegree and authority values correspond to nodes 2, 3 and 6. These students are the users that posted more messages in the forum. As a matter of fact, these three students scored the best in the course. Thus, with this analysis, we can conclude that students with a high number of interactions in the forum are likely to get good scores, a fact to be analyzed using data mining techniques. The instructor can have a better understanding of the students' behavior and improve their participation in the course. Similar results were obtained with the other two academic years, 2007–2008 and 2009–2010.

Firstly, we studied which SNA attributes were more relevant for the classification of the three datasets described in Sect. 3. We first used the attribute selection algorithm named *CfsSubSetEval* provided by Weka, which selects the best attributes for all classifiers. As a result, the most important attribute is *top3*, i.e., there is not other attribute that is considered most important than *top3*. The same was corroborated using association rules. We run the Apriori algorithm with the `mixed.dat` dataset as input—discretized with *PKIDiscretize*—and one of the rules showed that if *top3* is true, then students pass the subject with a confidence of 70%. On the one hand, by using *ClassifierSubSetEval* as attribute

Table 1. Rankings of top 3 nodes for different centrality measures

	Top 1		**Top 2**		**Top 3**	
	Node ID	Value	Node ID	Value	Node ID	Value
Degree	1	166	3	39	5	35
Indegree	3	36	5	33	6	17
Outdegree	1	157	2	6	23	6
Betweenness	1	505	17	151	14	142
Authority	3	0.93	5	0.76	6	0.41
Hub	1	1.41	23	0.03	2	0.03

selection algorithm with J48 as base classifier, we found that for this concrete classifier, the most important attribute is not only *top3*, but also the betweenesss and authority centrality measures. On the other hand, by using naïve Bayes as base classifier, the most important attributes were the degree, authority and hub centrality measures. Thus, we conclude that, for different models, different attributes are the most important ones for the prediction task.

Next, we proceeded to analyze whether SNA metrics might be suitable to predict performance and dropouts. For this task, we built six classifiers with the three datasets described in Sect. 3, but only using social data, i.e., without considering activity data. The classification algorithms chosen for this purpose were: J48, random forests, naïve Bayes, Bayesian networks, JRip, and Ridor. As can be observed, these algorithms follow different paradigms: J48 is based on trees, random forests are based on trees and meta-learning, naïve Bayes and Bayesian networks follow a Bayesian approach, and finally, JRip and Ridor are based on association rules. The implementations of these algorithms are those offered by Weka and they were run with the default parameters. Table 2 shows the average accuracy achieved by each algorithm using 10-fold cross-validation.

Table 2. Accuracy percentage in `performance.dat` and `dropout.dat`

	`performance.dat`	`dropout.dat`
J48	71.10	71.00
Random Forests	71.90	73.10
Naïve Bayes	70.32	71.00
Bayesian Networks	71.10	74.61
JRip	70.10	73.58
Ridor	68.00	74.01

As can be seen, the accuracy achieved for predicting dropouts (`dropout.dat` dataset) is higher than 70% in all cases, reaching a value of nearly 75% with Bayesian networks. Regarding performance, only the Ridor algorithm obtained an accuracy below 70%, being the rest of results better than this. So, in the light of the results, we can draw that SNA attributes can reasonably predict student performance and dropouts separately.

Finally, we compared the accuracy achieved by these classifiers—when run with the mixed.dat dataset as input—for assessing how SNA attributes contribute to improve the prediction task. Firstly, we built classifiers by only considering activity data, and before that, by using the whole data set, i.e., activity and social data. Likewise, we also tested whether using discretized SNA attributes would improve the accuracy. The discretization was performed by the *PKIDiscretize* algorithm offered by Weka. Table 3 shows the accuracy achieved in each case.

Table 3. Accuracy percentage in mixed.dat

	Activity Attributes	**Act. and SNA Attributes**	**Act. and SNA Discr. Attributes**
J48	77.20	79.79	80.31
Random Forests	78.75	80.31	80.31
Naïve Bayes	65.29	65.29	65.29
Bayesian Networks	80.83	81.87	76.68
JRip	83.42	77.20	80.83
Ridor	78.23	79.79	78.75

The results show that 4 of the 6 classifiers improved their accuracy when SNA attributes were used, achieving a significant improvement of 2.59% with J48. That improvement was higher, 3.11%, when discretized SNA attributes were used. Naïve Bayes got neither better nor worse accuracy, and only JRip got worse, but the models generated by this algorithm did not use SNA attributes, so that these results might be worse due to the randomness of the 10-fold cross-validation process.

In short, we can conclude that SNA attributes are useful for improving both students' performance and dropout prediction. Figure 3 depicts one of the classification models generated by J48 using the mixed.dat dataset with SNA attributes without discretization. Here, we can observe that the improvement in the classification task is due to the use of two SNA centrality measures: indegree and authority.

5 Conclusions

In this paper, we applied SNA techniques for the analysis of the interactions between the students of a course taught at the UC for three consecutive academic years. Furthermore, we used data mining methods for the prediction of the students' performance and dropouts. SNA in the educational field can be a powerful framework for the analysis of the students' social behavior and the relationships between them and their instructors. Forums are one of the tools in which this kind of analysis can be applied.

Moreover, we found that SNA can be useful for instructors to better understand how students use the forum, concluding that they often use this tool for

```
average_time_per_week <= 63: DROPOUT
average_time_per_week >  63
|   number_of_messages_written_in_the_forum <= 0
|   |   average_number_of_sessions_per_week <= 3
|   |   |   number_of_messages_read_in_the_forum <= 52: PASS
|   |   |   number_of_messages_read_in_the_forum >  52: DROPOUT
|   |   average_number_of_sessions_per_week >  3
|   |   |   total_time <= 1962: PASS
|   |   |   total_time >  1962
|   |   |   |   total_time <= 2000: DROPOUT
|   |   |   |   total_time >  2000: PASS
|   number_of_messages_written_in_the_forum >  0
|   |   number_of_messages_written_in_the_forum <= 8
|   |   |   normalized_indegree <= 0.013
|   |   |   |   number_of_messages_read_in_the_forum <= 87: PASS
|   |   |   |   number_of_messages_read_in_the_forum >  87
|   |   |   |   |   number_of_email_messages_read <= 24
|   |   |   |   |   |   number_of_messages_written_in_the_forum <= 4
|   |   |   |   |   |   |   indegree <= 1
|   |   |   |   |   |   |   |   normalized_authority <= 0.029
|   |   |   |   |   |   |   |   |   number_of_email_messages_read <= 7: FAIL
|   |   |   |   |   |   |   |   |   number_of_email_messages_read >  7: PASS
|   |   |   |   |   |   |   |   normalized_authority >  0.029: PASS
|   |   |   |   |   |   |   indegree >  1: FAIL
|   |   |   |   |   |   number_of_messages_written_in_the_forum >  4: PASS
|   |   |   |   |   number_of_email_messages_read >  24: FAIL
|   |   |   normalized_indegree >  0.013: DROPOUT
|   |   number_of_messages_written_in_the_forum >  8: PASS
```

Fig. 3. A J48 model for the `mixed.dat` dataset

making questions about the contents or the organization of the course, but it is seldom utilized for answering the questions posted by other students, which are responded by the instructor. As a matter of fact, we showed that the students who posted more questions and were answered the most are indeed the same students that, at the end of the course, achieved higher scores. This fact is better reflected with data mining analysis.

Regarding the classification of the students' performance and dropouts, we can conclude that, for both goals, SNA attributes are very useful since classification models of different kinds of classifiers achieve accuracy values over 70% with them. Moreover, using SNA attributes, along with the attributes containing the students' behavior, produces an improvement in models with respect to only using the former. In some cases, this improvement is higher than 2% of accuracy.

In a near future, we wish to extend our work in many directions. Firstly, we would like to test the effectiveness of our approach by using more e-learning courses with different characteristics. Our intention is to experiment on courses with a higher number of students, even though the number of instances in educational data mining datasets is usually limited. Also, we will attempt to apply the classification models presented in this paper to predict the hypothetical students' performance in early stages of the courses. Finally, we would like to consider other SNA metrics and algorithms, e.g., PageRank, as well as other data mining techniques aside from classification, e.g., association or clustering.

Acknowledgments. This work is partially supported by the Ministry of Education, Culture, and Sport of the Government of Spain, under grant Beca de colaboración (2012–2013), and by the University of Cantabria, under a Ph.D. studentship (2011–2015).

References

1. Bayer, J., Bydzovská, H., Géryk, J., Obšıvac, T., Popelınský, L.: Predicting Dropout from Social Behaviour of Students. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 103–109 (2012)
2. Brewe, E., Kramer, L., Sawtelle, V.: Investigating Student Communities with Network Analysis of Interactions in a Physics Learning Center. Physical Review Special Topics–Physics Education Research 8(1), 010101 (2012)
3. Crespo, P., Antunes, C.: Social Networks Analysis for Quantifying Students' Performance in Teamwork. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 234–235 (2012)
4. Freeman, L.: A Set of Measures of Centrality based on Betweenness. Sociometry 40(1), 35–41 (1977)
5. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)
6. Klovdahl, A., Potterat, J., Woodhouse, D., Muth, J., Muth, S., Darrow, W.: Social Networks and Infectious Disease: The Colorado Springs Study. Social Science & Medicine 38(1), 79–88 (1994)
7. Krebs, V.: Mapping Networks of Terrorist Cells. Connections 24(3), 43–52 (2002)
8. Moreno, J.: Who Shall Survive? Beacon House (1934)
9. Palazuelos, C., Zorrilla, M.: FRINGE: A New Approach to the Detection of Overlapping Communities in Graphs. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part III. LNCS, vol. 6784, pp. 638–653. Springer, Heidelberg (2011)
10. Palazuelos, C., Zorrilla, M.: Analysis of Social Metrics in Dynamic Networks: Measuring the Influence with FRINGE. In: Proceedings of the 2012 EDBT/ICDT Workshops, pp. 9–12 (2012)
11. Rabbany, R., Takaffoli, M., Zaïane, O.: Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques. In: Proceedings of the 4th International Conference on Educational Data Mining, pp. 21–30 (2011)
12. Romero, C., Ventura, S.: Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 40(6), 601–618 (2010)
13. Scott, J.: Social Network Analysis: A Handbook. SAGE Publications (2000)
14. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences. Cambridge University Press (1994)
15. Watts, D., Strogatz, S.: Collective Dynamics of Small-world Networks. Nature 393(6684), 440–442 (1998)