In-class social networks and academic performance: how good connections can improve grades

Luiz Gomes-Jr[1]

[1]DAINF – UTFPR – Curitiba – PR – Brazil

gomesjr@dainf.ct.utfpr.edu.br

Abstract. *Understanding how different variables affect student performance is an important requirement for improving educational practices. Since humans are highly social beings, social factors should play a significant role in the academic context. This paper analyzes the impact on academic performance of social indicators such as students friendship circle and in-class clustering. The analysis is based on data from six different classes of the topic Databases taken by students of computing-related majors. We assessed students' friendship circle in terms of density (sociability) and also quality (grades) of their friends. The paper shows results with strong, statistically relevant relationships between the social factors and student performance. Among other results, the analysis indicates that (i) students with higher social capital tend to perform better, and (ii) students with friends with higher grades have better chances of recovering from a low exam grade.*

1. Introduction

Understanding the factors that influence academic performance is a challenging endeavor. There are multiple variables in different levels, from genetics to individual history, from instructor's traits to group dynamics. Determining how these variables affect each other and contribute to academic performance is a requirement for better educational methods and policies.

This paper analyzes the relationship between social connections and academic performance. The main challenge in social network analysis is to gather the data representing social connections between people. This is especially challenging in the academic environment, since there is no established repository for this type of information. Previous studies have derived this information from smart-card use around a campus or from course enrollment records. In this paper we use a more direct approach, employing self-reported relationships to build the social network graph for several classes. We argue that this approach allows for the creation of more reliable social graphs, improving the accuracy of the inferences.

This paper analyzes the impact on academic performance of social indicators such as student's friendship circle and group dynamics factors such as in-class clustering. We employ several complex network measurements to quantify these indicators (Section 2). The analysis is based on data ($n = 148$) from six different classes on the topic Databases taken by students of computing-related majors (Section 3). The paper shows results with strong, statistically relevant correlations between the social factors and student performance (Section 4). We employ a linear regression model and several statistical inferences to support the following findings:

- Classes with more connections (denser graphs) tend to have higher average grade
- Students with stronger social capital (network centrality) tend to perform better
- Students that have friends with higher average grade have better chances of improving their own grades as the term progresses

We expect that these results can help educators evaluate the role of social interactions in academic performance. The results could provide evidence to support the investment in practices that foster a healthy social environment in the academic context.

2. Related work

The lack of reliable data to build social network graphs in the academic contexts limits the research on the topic. One approach to circumvent the problem is to infer social relationships from other data sources. Yao et al. [Yao et al. 2017] resorted to time and location-stamped data of students using their smart-cards on campus facilities. From the collected data, the social relationships were inferred based on co-occurrence of events between students. The researchers analyzed data from multiple locations (e.g. cafeteria, library etc.). The analysis showed associations between the grades of students and those of their inferred social circles. The data was used to build a label propagation model to predict student grades based on the grades of their peers, achieving around 40% accuracy. The researchers do not present analysis on social network measurements and their impacts on academic performance.

Gasevic et al. [Gašević et al. 2013] assess the relationship between performance and social circles in an online education scenario. The researchers used co-enrollment in courses as a proxy for social connection. The authors emphasize that this type of data is easy to obtain. However, it is highly questionable that co-enrollment data is correlated with social bonds, especially in an online university. The research fails to show significant influence on grades of even basic centrality measures such as node degree. In this paper we use self-reported information on social bonds between students which, we believe, produces a more realistic social graph.

Castilho et al. [Castilho et al. 2014] focus on students' preferences when forming intra-class groups for course assignments. The authors use group formation data from an undergraduate course and combine it with social interaction data from Facebook. The analysis shows the importance of social status (popularity) and social interactions when students select their assignment peers. In this paper we do not focus on analysing group formation, but this could be a future research direction.

A large body of research investigates the impact of the *use* of social network sites on academic performance (see [Doleck and Lajoie 2018] for a review). The goal is to determine whether the time spent in such sites could affect student performance. Even though the majority of papers show that site usage has a negative impact on performance, the field is still far from reaching a consensus. Here we do not consider the use of social network sites and instead focus on real, self-reported relationships between students of each class.

The field of complex networks have provided tools and models for the analysis of social networks in general [Borgatti et al. 2009, da F. Costa et al. 2011]. The analysis of these networks is based on algorithms that derive measurements that capture properties

of the underlying graph (see [da F. Costa et al. 2007] for a review on network measurements). Usually, these measurements quantify properties for the nodes or for the entire graph. In this paper we apply several measurements to quantify social characteristics of students (node level) and classes (graph level). Table 1 provides a short description of the main measurements used in this paper. For formal definitions we refer the reader to [da F. Costa et al. 2007].

Table 1. Descriptions of the main measurements used

Measurement	Informal description	Level
degree	number of neighbors of a node	node
eigenvector centrality	influence/connectedness based on a node's degree and, recursively, the influence of its neighbors	node
betweenness centrality	ratio of shortest paths that pass through a node; tends to be higher for nodes that are bridges between clusters	node
closeness centrality	the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph; the more central a node is, the closer it is to all other nodes	node
average neighbor degree	average degree of the neighbors of a node	node
clustering	tendency of neighbors of a node to be connected among themselves	node
assortativity	tendency of a network to have its nodes connected to similar nodes (in general in terms of degree)	graph
average shortest path length	average of the number of nodes in the shortest paths between all pairs of nodes	graph
global efficiency	the efficiency of a pair of nodes in a graph is the multiplicative inverse of the shortest path distance between the nodes; the average global efficiency of a graph is the average efficiency of all pairs of nodes	graph

3. Data collection and cleaning

The dataset used for the analysis is based on data from six classes on the topic Databases taken by undergraduate students between 2016 and 2018. The students are from computing-related majors, namely Computer Engineering and Information Systems.

Collecting social data is a challenging task, but in this case it was simplified by the nature of the practical assignment given by the instructor to all of these Databases classes: to build and analyze the class social network. To provide students with the data needed to complete the assignment, the instructor implemented a simple social network application on which the students have to enter their data in the beginning of the term. The application collects data on friendship connections and also personal preferences (music and movies). This analysis only used data on friendship connections. The students were aware that the

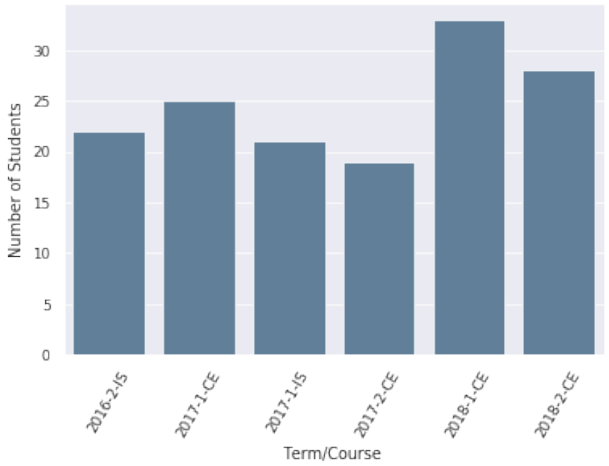

Figure 1. Distribution of students per class

data would be used for analysis. Only the instructor had access to individual data and only aggregated/anonymized data is reported here.

For each of the classes, the social graph was build and only the largest connected component was retained to represent the class social network. The graph was used to compute several complex network measurements (Table 1). Figure 2 shows an example of network for a class. Each node is a student (name anonymized) with the size of the node representing its *eigenvector centrality* score.

Figure 2. Class network example

The graph measurements were then integrated with the student performance data provided by the instructor. Students missing performance or social data (most likely class drop-outs) were excluded from the dataset. The final dataset contains 148 students distributed in classes as shown in Figure 1.

4. Data analysis

The data analysis reported here comprises four main steps: (i) exploratory analysis of class-level variables to assess the influence of social graph measurements on the average

performance of the classes; (ii) exploratory analysis of student-level variables to assess the influence of social capital measurements in the performance of students; (iii) building of a linear regression model with variables from the previous steps to quantify the influence of the variables; and (iv) analysis of the quality of connections (in term of grades) and how they help students recover from low grades. These steps are described in details in the following sections.

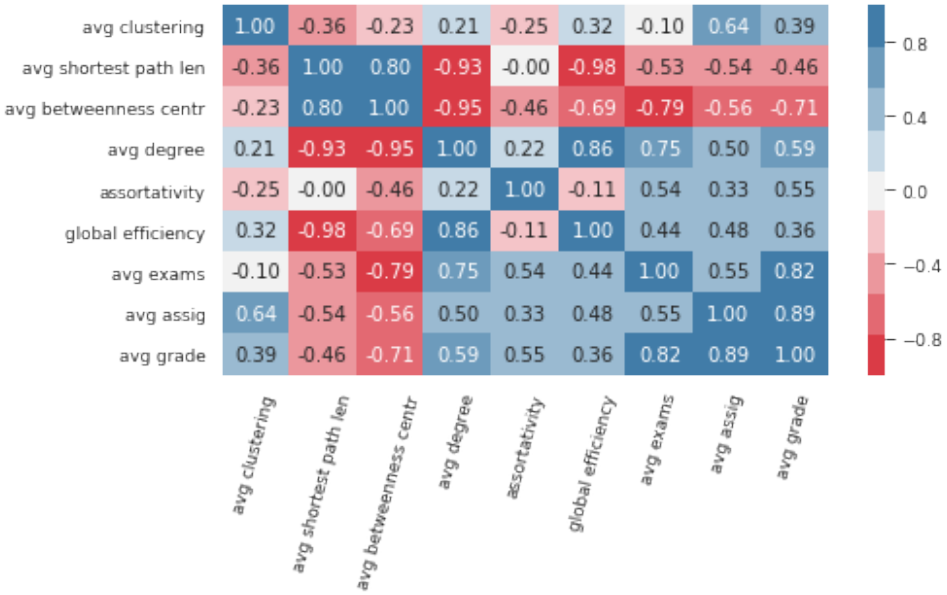

Figure 3. Correlations between class-level variables

4.1. Class-level variables

The class-level variables are intended to capture social characteristics of a given class, such as average number of friends, tendency to form clusters, etc. The class-level variables included are: average clustering, average shortest path length, average betweenness centrality, average degree, assortativity, global efficiency. The class-level variables for the grades are: average exams (average grades considering the two exams), average assignments (grade for the course assignment), and average grade (average final grade, considering exams, assignment, quizzes and exercises).

Results from the class level analysis have lower statistical relevance since there is less data (6 classes) for the inferences. The dataset shows very interesting patterns that should be interpreted carefully due to the lack of data. The heatmap in Figure 3 shows the correlation between the variables. The correlations show that, in general, the more connected a class is, the better its grades. The correlation between average degree and average grade on exams is 0.75 ($p = 0.087$), plotted in Figure 4. There are positive correlations between other variables that capture graph density, such as global efficiency and average clustering. The average shortest path length is negatively correlated with the grades for a similar reason: longer shortest paths mean less dense graphs. Average betweenness centrality is also negatively correlated with grades, which may be due to higher numbers of bridge nodes indicating isolated communities in the graph.

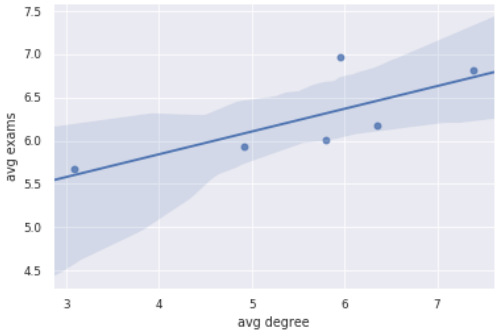

Figure 4. Class average degree vs. average grade for exams

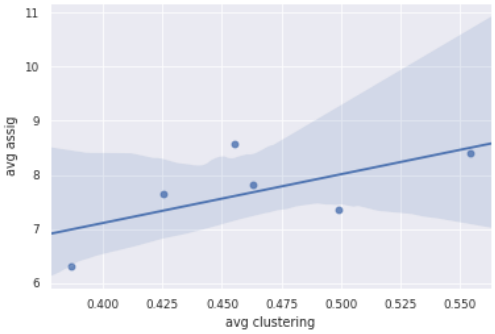

Figure 5. Class average clustering vs. average grade for assignments

Interestingly, average clustering is positively correlated with assignment grades (correlation: 0.64, $p = 0.167$, shown in Figure 5) but negatively correlated with exam grades (correlation: -0.10, $p = 0.849$). Despite the low statistical significance, it might be true that strongly connected social circles are more important for assignment grade (as explained previously, the assignments in the course are group assignments).

4.2. Student-level variables

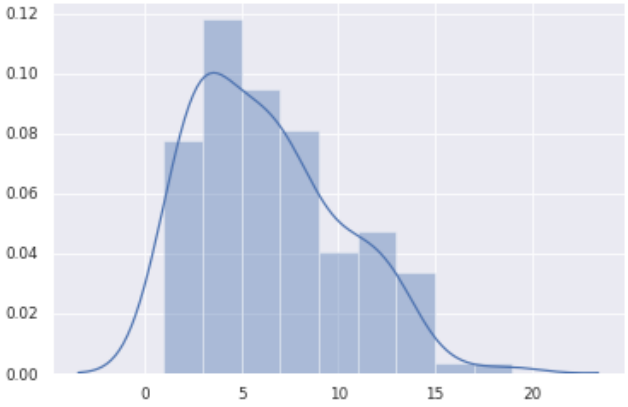

Figure 6. Degree distribution for students in the dataset

The student-level variables are intended to capture characteristics of individual's social network in the context of the class. The main student-level variables included are: average neighbor degree, betweenness centrality, closeness centrality, clustering, degree, eigenvector centrality. Results from the student level analysis have better statistical relevance since there is more data for the inferences.

We also included the class-level variables in the student data to assess whether being in a class with certain characteristics would influence individual performance. Overall, the correlations with class-level variables were weak. The strongest correlations were

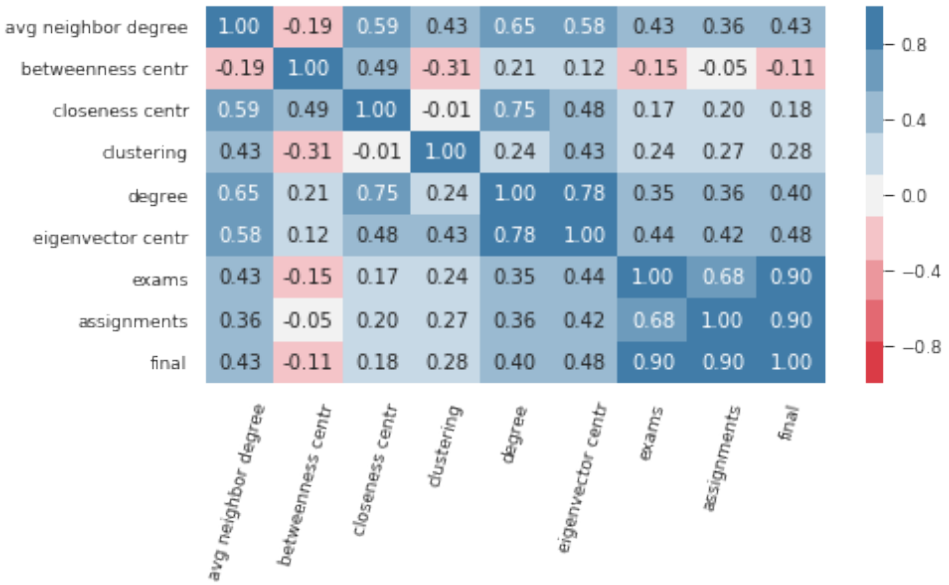

Figure 7. Correlations between student-level variables

with assortativity (0.21) and average betweenness centrality (-0.23), both considering final grades.

Figure 6 shows the degree distribution for the dataset (with the x axis representing the number of connections and the y axis their probabilities). The classes are small for any speculation about the shape of the distribution, but the figure shows the general trend in social networks of many nodes having few connections and few nodes having many connections.

The heatmap in Figure 7 shows the correlation between the student-level variables. There is a significant correlation ($p < 0,01$) between students' final grade and both *eigenvector centrality* (correlation: 0.48) and *average neighbor degree* (correlation: 0.43), suggesting that well connected students tend to have higher grades. The (weak) negative correlation between grades and *betweenness centrality* could be due to students that are in-between clusters feeling left out or having difficulties relating or being affiliated to friend groups.

Figure 8 shows a plot of individual students according to *eigenvector centrality* (normalized) and final grade. A regression line has been added for reference. The overall correlation between the variables seem pertinent. Some interesting patterns can be seen in the plot, such as a line of individuals around grade 2, probably indicating students that took too many credits or did not intend to take the course seriously from the beginning (common problems in this university). It is reasonable to assume that social factors should not play an important role in these cases. Another pattern emerges for students with very low *eigenvector centrality*, which are distributed regularly among the grades. These are probably students from other years or courses that did not know many of their peers. It is also reasonable to expect that these students would have different factors influencing their performance. Removing these two patterns from the dataset would make the correlation stronger, but since it is hard to gather data to determine the underlying reasons, we kept

all data points in the analysis.

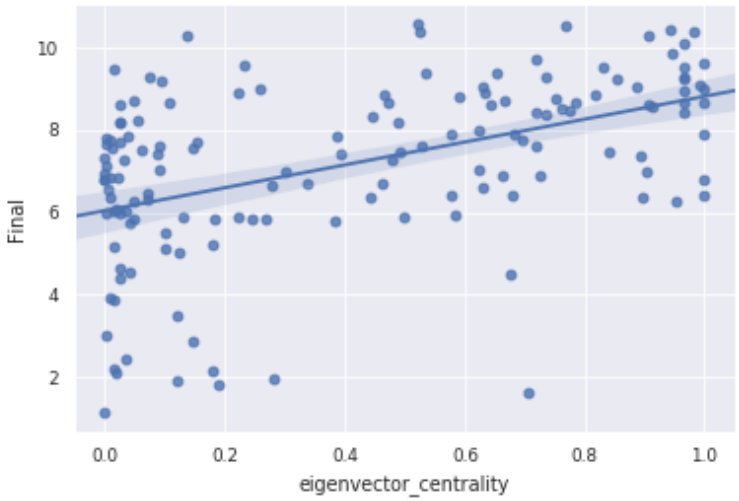

Figure 8. Student eigenvector centrality vs. final grade

4.3. Linear Regression Model

To assess the relationship between the social network variables and student performance, we built three linear regression models using *ordinary least squares* for parameter estimation. We built multiple models to evaluate whether the variables would have a different impact depending on the type of performance evaluation (grades for exams, assignments, or final grades).

To allow for comparison between models, we had the use a single criteria for feature selection. We opted to select variables with strongest combined correlation with our target variables. Therefore, for each variable, we aggregated its correlation with all targets (exams, assignments, final). We then ordered the list of variables according to strength of combined correlation. In the next step we eliminated redundant variables (e.g. for eigenvector centrality and degree, which are highly correlated, we kept only the most correlated with the target). The final list of independent variables used to build all models is: *eigenvector centrality*, *average neighbor degree*, *clustering*, *assortativity*, *average degree*, *average shortest path length*, *betweenness centrality*, *avgerage clustering*, *global efficiency*.

We used the Python module StatsModels[1] for model construction. All models achieved an R-squared value of over 0.3. Table 2 shows the results with parameter coefficients and p-values for each variable and for each target.

Several variables achieved statistical relevance for $p < 0.05$, but we focus the discussion at the more strict $p < 0.01$ level. The models confirm the influence of *eigenvector centrality* in all target variables. The 2.16 coefficient influencing final grade represents a difference of 1.45 points between students in the 25 and 75 percentile for eigenvector centrality.

[1] www.statsmodels.org

Table 2. Variables and coefficients for the three linear regression models

	Exams		Assignments		Final	
Variable	**Coef.**	**p**	**Coef.**	**p**	**Coef.**	**p**
Eigenvector Centrality	**1.76**	**0.00**	**1.98**	**0.00**	**2.16**	**0.00**
Average Neighbor Degree	**0.24**	**0.01**	0.24	0.05	0.24	0.02
Clustering Coefficient	-0.63	0.31	-0.29	0.70	-0.48	0.44
Assortativity	8.81	0.04	**18.11**	**0.00**	**15.83**	**0.00**
Average Degree	1.39	0.08	1.51	0.11	**1.97**	**0.01**
Avg. Shortest Path Length	28.56	0.04	39.87	0.02	**40.86**	**0.00**
Betweenness Centrality	-3.83	0.03	-1.76	0.40	-3.28	0.06
Average Clustering	12.81	0.05	**27.30**	**0.00**	**24.45**	**0.00**
Global Efficiency	131.35	0.05	**200.53**	**0.01**	**190.39**	**0.01**

Average neighbor degree is also significantly associated with all targets. The 0.24 coefficient represents a 0.24 increase in grade for each extra averaged degree score in a student's friendship circle.

The regression models did not confirm the impact of *clustering*. The reason might be the strong correlation between *clustering* and other variables such as *eigenvector centrality* and *average neighbor degree*.

Assortativity, *average degree* and *average clustering* are class-level variables that seem associated with academic performance. More connected classes with more homogeneous connections seem to lead to better grades.

4.4. Assessing friend's grades influence

The linear regression model showed a clear relationship between students' performance and the strength of their social connections. We now focus on analyzing the influence of the quality (in terms of grades) of the relationships.

Figure 9 shows a heatmap with the correlations between students grades (exam 1, exam 2 and final grades) and the grades of their friends (mean, maximum and minimum grades). The most important trend that can be observed is how the grades of the students tend to become more correlated as the term progresses: grades for Exam 1 have a correlation of 0.37 with friend's average for Exam 1, for Exam 2 this correlation increases to 0.52, and for the Final grade (including assignments) the correlation is 0.59. All correlations have significance level $p < 0.01$. This clearly shows that the grades of friends tend to influence a student and that influence becomes more prevalent as the term progresses (either because the student knows more about the performance of their peers or because they interact and share more knowledge).

It can also be seen from the heatmap that the strongest correlations are with the mean of the friends' grades, followed by the maximum grade among friends. This suggests that friends with lower grades tend to have a lower influence in performance.

We also calculated the correlation between the magnitude of improvement for a student $(E2 - E1)$ and his/her deficit compared to friends average $(E1\_friends\_mean - E1)$. The correlation is 0.45 $(p < 0.01)$, indicating that the larger the difference in

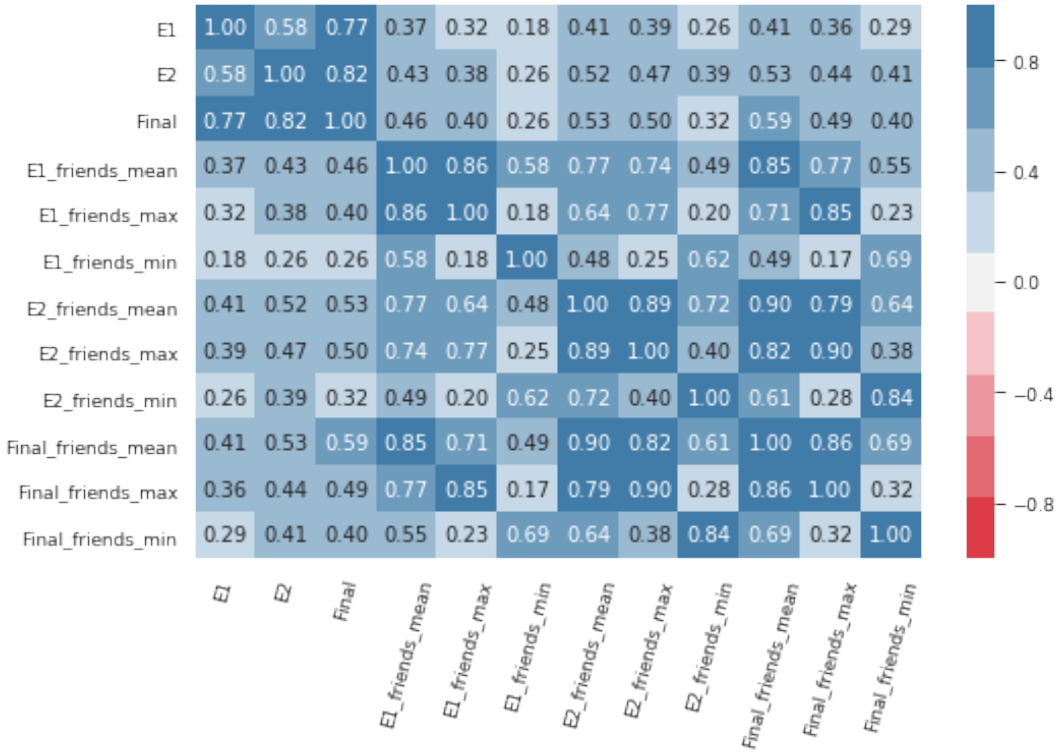

Figure 9. Correlations between student performance and their friends' performance

performance in E1, the larger the improvement for a given student.

Finally, we wanted to measure how much having a circle of friends with good academic performance can help a student in practice. We selected the students with lower grades in Exam 1 (lower than the passing grade of 6). This new dataset has $n = 62$. We then compared their performance based on the performance of their friends. Figure 10 shows the performance of the students with a low Exam 1 grade. Among these students, those with friends that performed well in the Exam 1 are represented in blue lines. It can be seen in the graph that these students tend to improve their grades. The students with friends that performed poorly in the Exam 1 (yellow lines) show no pattern for the other evaluations (some improve, some do not).

We run ANOVA tests on the averages of the students' improvement on Exam 2 for the two groups: (i) students with friends that did well on Exam 1 (grade larger than 6) and (ii) students with friends that also performed poorly in Exam 1 (< 6). Students with friends that performed poorly on Exam 1 only improved by 0.5 on average on Exam 2 ($p < 0.01$). Students with friends that performed well, in contrast, improved by 1.9 points on average – almost a 4 fold gain when compared with the other group. This suggests that having friends with good academic performance has a direct impact on students grade.

5. Discussion and Conclusion

Previous works had already established associations between students' friendship circles and academic performance. In this paper we explored similar questions using more precise data on students' social graphs. This allowed us to capture subtle but important

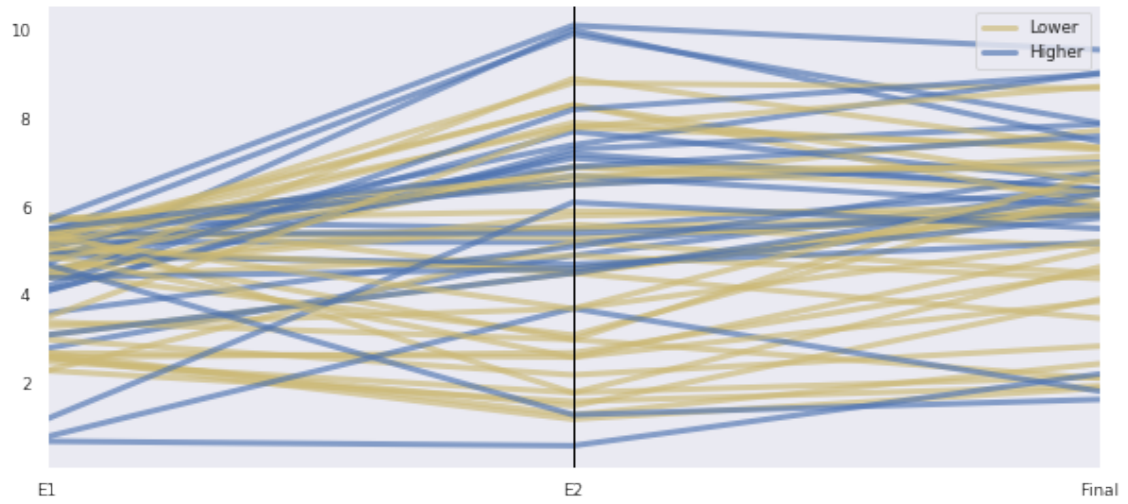

Figure 10. Performance of students that had a low grade in Exam 1

aspects of the relationship between social capital (in terms of graph measurements) and performance. In our opinion, the two most important new patterns detected are (i) the importance of complex network dynamics, and (ii) that friendship influence in performance seems to be a direct and time-dependent factor.

The importance of network dynamics in social metrics is suggested by the analysis of the correlation of social measurements with grades. The correlation between node degree (number of friends) and final grade was 0.43, which is smaller than the correlation between eigenvector centrality and final grade (0.48). Eigenvector centrality is a measurement that captures connectedness in a recursive fashion, in a model that encompasses random walk dynamics. Intuitively, the measurement produces higher values for nodes that are connected to highly connected nodes (recursively). This indicates that the relationship between social capital and grades is indeed a property of the complex social dynamics represented in the graph.

The other important aspect captured was that the influence of friends in a student's performance in not a constant. The analysis of the correlations and of the improvements from low grades show that the influence tends to grow as the term progresses. The data shows that students tend to normalize their grades with those of their friends with better performance. This does not seem to occur in the other direction – meaning that friends with poor performance have less influence on their peers.

This paper also confirms the relationship between social capital and performance. To quantify the influence we built three linear regression models, accessing each of the three grades (exams, assignment, final). All models achieved a R-squared value of over 0.3, which we consider a significant explanatory strength given that social variables are only marginal factors in student performance. In general, most of the performance of a student is more likely to be explained by variables like his/her previous performance, which is in turn associated with social, developmental and genetic factors. Of course, sociability is also highly influenced by the same factors. These variables are, however, virtually impossible to control for.

We believe that the findings presented here can help education professionals in assessing the importance of social factors in academic performance. This could then be used to guide policies for improving social interactions in the academic context.

In future work we expect to gather more data from other classes and universities to enable better inferences for class-level variables (the data collection application is available for other instructors). We also intend to analyze the factors influencing group formation and performance in the practical assignment. In terms of improving the model, we expect to integrate better criteria to identify drop-outs and correlated variables.

References

Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916):892–895.

Castilho, D., de Melo, P. O. S. V., Quercia, D., and Benevenuto, F. (2014). Working with friends: Unveiling working affinity features from facebook data. In Adar, E., Resnick, P., Choudhury, M. D., Hogan, B., and Oh, A. H., editors, *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

da F. Costa, L., Oliveira Jr, O., Travieso, G., Rodrigues, F., Boas, P., Antiqueira, L., Viana, M., and Rocha, L. (2011). Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics*, 60:329–412.

da F. Costa, L., Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

Doleck, T. and Lajoie, S. P. (2018). Social networking and academic performance: A review. *EAIT*, 23(1):435–465.

Gašević, D., Zouaq, A., and Janzen, R. (2013). "choose your classmates, your gpa is at stake!": The association of cross-class social ties and academic performance. *American Behavioral Scientist*, 57(10):1460–1479.

Yao, H., Nie, M., Su, H., Xia, H., and Lian, D. (2017). Predicting academic performance via semi-supervised learning with constructed campus social network. In Candan, K. S., 0002, L. C., Pedersen, T. B., Chang, L., and Hua, W., editors, *Database Systems for Advanced Applications - 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part II*, volume 10178 of *Lecture Notes in Computer Science*, pages 597–609. Springer.