Information Sciences 375 (2017) 155–170

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

The strength of the work ties

Douglas Castilho[a,b], Pedro O.S. Vaz de Melo[b], Fabrício Benevenuto[b,*]

[a] Federal Institute of Southern Minas Gerais, Poços de Caldas, Brazil
[b] Federal University of Minas Gerais, Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 8 October 2015
Revised 23 August 2016
Accepted 5 September 2016
Available online 15 September 2016

Keywords:
Strength of ties
Facebook data
Team formation
Working affinities

ABSTRACT

College students often have to team up for class projects, and they select each other based not only on past performance (e.g., grades) but also on whether they have friendship ties (e.g., whether they trust each other). There has not been any study on the relationship between team formation for class projects and social media. To fix that, we ask two group of university students to tell us with whom they wish to work. Afterward, we gathered their online Facebook data and tested the predictors of team formation. We found that self-organized selection of team members does not strongly depend on past grades, but rather on Facebook-derived proxies for tie strength, popularity, extroversion and homophily. These results have important theoretical implications for the team formation literature and practical implications for online educational platforms.

1. Introduction

During our lives, we perform collaborative tasks in a wide and diverse range of activities. In fact, it is part of our routine to select or be selected by someone to do a collaborative task. Selecting students to participate in a school project, hiring employees to a company, picking up players for a football friendly match and selecting colleagues to approach a research problem are just a small sample of decisions involved in collaborative activities that most of the people eventually do in their lives. Given this context, we ask: *what factors influence such decisions, i.e., what factors are determinant for selecting/repelling someone for a given collaborative task?* Without much thought, one could answer this fundamental question by saying the proficiency (or the skill) of a person to do a given task determines if she/he will be selected for a collaboration in this task. Although we agree that proficiency may plays an important role in the decision, we again ask: is proficiency the only determinant factor? If not, is the proficiency even the main factor?

In this paper we take the first steps towards answering these particular questions. Together with the proficiency, we propose that social behavior have a strong impact on selecting collaborators. Such analysis is now possible because, unlike decades ago, time expensive surveys are not the only available methodology to capture social behavior anymore [39,40]. With the growing popularity of the Internet and their applications, almost everyone have their social interactions registered in an online social network [23]. Online social networks such as Facebook and Google+ are able to mimic the real social environment into a virtual one [21,24,30–32,47,48]. Analyzing how people behave in these virtual social environments may tell how they behave in the real one and, therefore, verify if their social behavior impact on their collaboration decisions.

To verify if social behavior impact on such decisions, we conduct an experiment in a very particular scenario. First, we conduct a sociometric test [35] on two classes of undergraduate students, in which individuals were asked if they would like

* Corresponding author:
E-mail addresses: douglas.braz@ifsuldeminas.edu.br (D. Castilho), olmo@dcc.ufmg.br (P.O.S. Vaz de Melo), fabricio@dcc.ufmg.br (F. Benevenuto).

http://dx.doi.org/10.1016/j.ins.2016.09.009
0020-0255/

to work with every other student in class. Then, using a Facebook application we developed, we gathered data containing a number of social features about their profiles and their interactions that can potentially describe their social behavior towards the other students. Our scenario is appropriate because college students often have to team up for class projects. While it may look simple and natural the way students choose their groups in a classroom, we believe that the process that determine their choices involves a complex mix of social attributes and knowledge skills in order to create a team that is both successful and pleasant to work. They select each other based not only on past performance (e.g., grades) but also on whether they get along (e.g.,whether they trust each other).

Our analysis on this data unveils a number of interesting findings. First, using the student grades to infer individuals skills, we discovered that the most skilled students were not always preferred, indicating that social capital does play a very important role to determine their choices. Then, we further investigate a number of features extracted from the Facebook data related to the strength of the friendship, the popularity of the individual on Facebook, if she is extroverted, and her similarity with other students. Our analysis unveils at least six features extracted from Facebook that are more informative than grades to determine the willingness of students to work together.

Moreover, in order to verify if the answers the students gave in the sociometric test can be predicted, we compared several state-of-the-art machine learning algorithms when their task is to classify the type of relationship two students have: positive, neutral or negative working affinity. We tested each algorithm using three possible datasets: (i) one containing only proficiency features, (ii) one containing only social features, and (iii) one containing both. This comparison revealed that combining social features with proficiency features can significantly improve the classification accuracy and performs reasonably well, with an accuracy of approximately 70%. Moreover, it also revealed that social features are consistently the more relevant features in this task, although the importance of each feature varies slightly between classes of students. These findings suggest that, although there are social features which almost universally indicate whether two students are willing to work together, their importance may vary from scenario to scenario.

Finally, although our findings were drawn from a very particular classroom scenario, they have broader implications. For instance, they show the importance of building up a wide and diverse personal profile when the aim is to be selected for a given collaborative task, i.e., there are characteristics different from proficiency one should improve to be selected. Also, for the team formation problem, our findings show that online social network data can indicate if two individuals would like or not to work together and, as it is well known, social affinity is desirable for achieving maximum performance of a team. Our findings can also leverage several online applications, such as team and collaboration recommendation systems that highlights potential fruitful collaborations and hides collaborations between potential conflictual relationships.

2. Related work

We review previous studies on team formation, online *vs.* offline behavior, and efforts that attempt to identify behavioral features from online data. These studies help us to identify the Facebook-derived features that are expected to be associated with team formation.

2.1. Team formation

There is a broad literature related to team formation. Most of the literature has focused on the problem of how to identify the members of a group who are collectively best suited for solving a specific task. Wi *et al.* [46], for example, modeled this problem as an integer programming problem to find an optimal match between individuals and requirements. Agustín-Blas *et al.* [1], instead, proposed to partition the staff-resource matrix in a way that all members of a team share the most accurate knowledge of the team's resources. Those approaches, however, do not consider whether team members are likely to enjoy fruitful personal relationships. To fix that, recent efforts have proposed to augment existing approaches with members' temperament [16] and with interpersonal attributes [11].

Recently, with the emergence of Massive Open Online Courses (MOOCs), students are experiencing global exposure and opportunities for connecting and interacting with millions of people all around the world. Sinha *et al.* [41] approaches the problem of team formation in this new kind of environment to enable course instructors to automatically group students in teams that have fairly balanced social connections with their peers. Wen et al. [45] studied what factors could identify successful virtual teams in NovoEd MOOCs and found that team leaders play a central role in determining team performance.

Despite the importance of all these studies on team formation, they do not propose to augment those traditional approaches with online features derived from social media sites, as our work aims to do.

2.2. Online vs. offline behavior

A lot of research work has gone into understanding to which extent online social network data can be used to infer offline behavior. Jones *et al.* [24] used a similar methodology to our work to gather social features from Facebook to infer offline behavior. They run a survey to Facebook users asking those individuals the name of their best friends. They then correlated this survey data with the number of public and inbox messages those individuals exchanged with their Facebook friends. They showed that public communication is as informative as inbox messages are to infer tie strength. Xiang *et al.* [47] proposed a model for predicting tie strength from Facebook interactions and number of common friends. Xie *et al.* [48]

studied the behavioral features associated with Twitter users who happen to be classmates or friends in real life. Manson *et al.* [32] analyzed how college students express affection to their close friends on Facebook, and identified 30 main ways to express affection. Gilbert *et al.* [20] developed a model to predict the strength of the tie, based in social data. They collected Facebook information about the relationship in different strength dimensions (intimacy, intensity, duration, social distance, services, emotional support and structural). This model was also validated by Luarn and Chiu [31], who found that the variables in the dimension of emotional intensity had stronger effects than other interaction variables.

In a different direction, Frezzo *et al.* [17] created a micro-world to characterize complex social interactions that mimic important aspects of the real offline world. Also, Vaz de Melo *et al.* [44] showed that offline behavioral patterns, namely encounters in university campuses, may not be good predictors of online behavior, namely Facebook friendships. Complementary to the above efforts, our work considers a novel scenario in which features extracted from online social data can be useful for a relevant task.

2.3. Identifying behaviorial features

Human behavior is constantly registered in online social network interactions and also in the underlying structure of friendship connections [4]. The idea of inferring user behavior based on online social network data has been used on different tasks, usually associated with the use of machine learning [7,15]. Interesting tasks include the identification of malicious and opportunistic behavior [3,12] and the identification of users geolocation [25] or other profile aspects [34].

Efficient teams are usually composed by people who complement themselves, i.e., who add different set of abilities to it. In this context, the term social capital is probably the closest concept that captures these different abilities that individuals bring to a team. It usually stands for the ability of people to secure benefits just by being members of specific social groups or by occupying specific advantageous positions in a social network [13,38]. For instance, individuals who belong to multiple groups tend to transmit valuable information from one group to another. In sociology and marketing studies, social capital has been often used to explain why specific individuals are more likely to come across new job opportunities [22]. More recently, it has been also associated with group effectiveness [37], in other words, the higher the group social capital resources, the higher the group effectiveness. The above efforts inspired some of the features we used in this paper, although intrinsically derived from the underlying social structure among classmates and their behavior expressed in Facebook.

Related to this paper, Gee *et al.* [18] investigated six million Facebook users' data found that in the process of finding a job, most people are helped through one of their numerous weak ties, but a single stronger tie is significantly more valuable at the margin. In a previous effort [8], we approached the problem of identifying the relevant features extracted from Facebook data to infer working affinities among students. In this paper we extend our previous effort by leveraging the identified features towards a predictive classification task. Thus, our effort consists of effectively quantifying the discriminative power of the unveiled features and we also attempt to validate our methodology by reproducing our first experiment using data from two classes of students instead of only one.

3. Methodology and datasets

In this section, we describe our methodology and datasets. We prepared a very particular experimental scenario. First, we selected two classrooms of undergraduate students of an anonymous university of an anonymous country. Then, through a sociometric test [35], we asked to each student of these two classrooms if they would like to work together with each other student of his/her same classroom. To analyze and understand their answers, we collected the information about their performance in class, i.e. their grades, and also several pieces of information about how they socially interact with the other students of the classroom. The latter is a set of online interactions collected through a Facebook application developed for this particular purpose. These datasets are very appropriate to address the questions we posed because (i) each student has answered the question about every other student in class and (ii) all of them know each other in person and fairly well, since they are supposed to see each other at least twice a week. In the next sections we describe the details of this data collection process. More details about the methodology and the datasets can be seen in the Supplementary Information material.

3.1. The sociometric test

Sociometry is a quantitative method for measuring social relationships [35]. The sociometric test can be applied in any circumstance in which you want to understand the relationships within a group. From this knowledge it is possible, for instance, to reorganize the connections, the distribution of tasks, to define new leaders, among other applications [6]. In general, the sociometric test consists of a questionnaire to each member of a group of people. From the questionnaire is built the sociogram, that is basically the mapping of the social network of the group.

In our experiment, the sociometric test was applied to understand the existing dynamics in a group of people when they are supposed to collaborate to perform group tasks. For this, we selected two classrooms: one class with 31 and another with 19 undergraduate students of an anonymized university of an anonymized country. Then, we applied a questionnaire to each student containing the following question: "Would you like to work with this person?". After this question, the survey shows to the participant a list containing the names of all classmates. In front of each name was a blank space where the

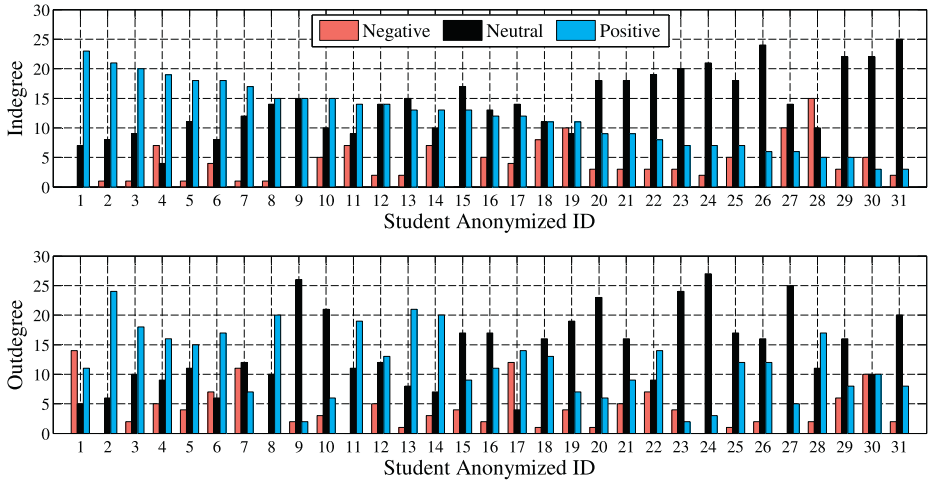


Fig. 1. Individual scores attributed and received by each participant of the sociometric test (Class 1).

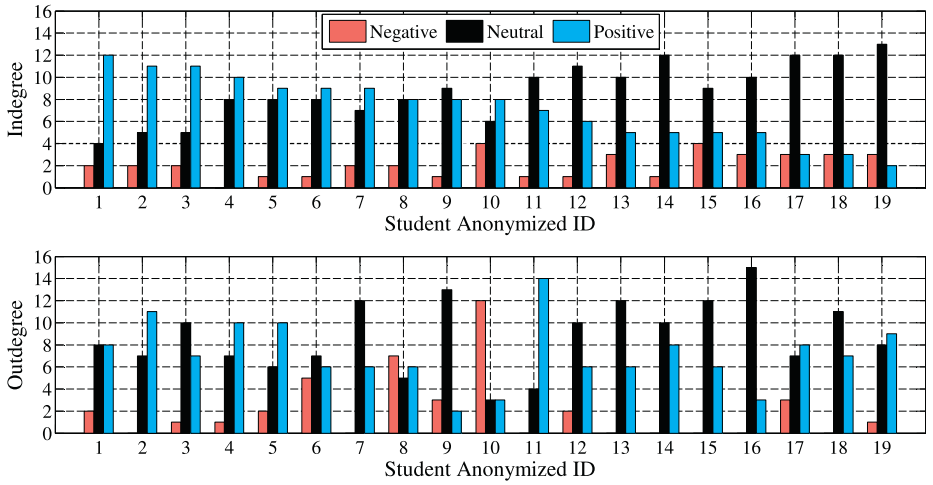


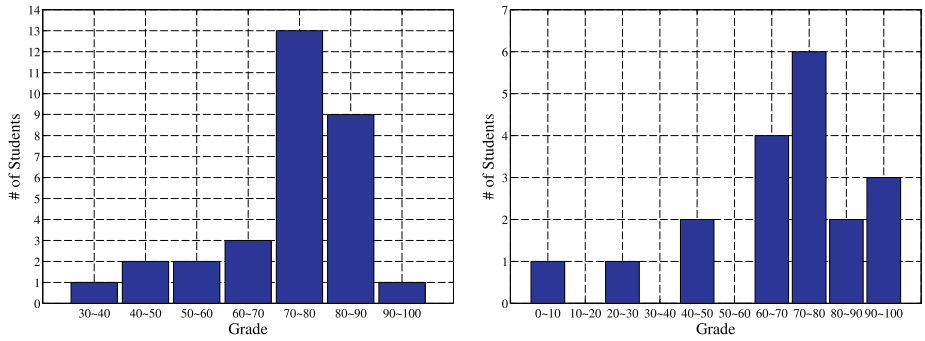
Fig. 2. Individual scores attributed and received by each participant of the sociometric test (Class 2).

participant had the opportunity to check one of the following responses: “YES”, “NO” or “INDIFFERENT”. When the answer is “YES”, it indicates that the student would be interested in running some group activity with the individual in question. When the answer is “NO”, the student rejects the idea of performing some group activity with the individual. Finally, when the answer is “INDIFFERENT”, the student is indifferent to that particular individual.

Thus, we have three different types of relationships $(i \rightarrow j)$ between students i and j. First, the relationship can be positive, i.e. $(i \rightarrow j) = 1$, indicating the interest of student i to work with the student j. Second, the relationship can be negative, i.e. $(i \rightarrow j) = -1$, indicating that individual i has no interest in working with j. Finally, the relationship can be neutral, i.e. $(i \rightarrow j) = 0$, when the individual i is indifferent with respect to individual j. Since the survey was administered to all students in class and each student answered the survey with respect to all the other students, we have two complete sociogram, that consists of 930 answers among the 31 students and 342 answers among the 19 other students.

This complete sociogram can also be seen as a complete directed signed graph $G_S(V, E_S)$ where the set of nodes V is composed by the students and the set of directed edges E_S are the answers. In Figs. 1 and 2 we show the outdegree and indegree of each student in G_S grouped by the sign of the edge. We can note different sorts of profiles. For instance, in class 1 there are students who received and gave a lot of positive edges (e.g. student 2) and also students who are indifferent for and towards most of the class (e.g. student 24). Moreover, there are students that are not negative towards anyone (e.g. student 11) and also a student who received a negative answer by almost half of the class (student 28). Next, we investigate if the students grades' can explain these results.

It is important to point out that all experiments were conducted in classes composed by students in the initial stage of their graduation courses. All tasks during this period are basic by nature, i.e., they do not require any specialized skill that some students may have and others may not. The grade is the only metric they have to differentiate their skills among



(a) Class 1 (b) Class 2

Fig. 3. Grade's histogram.

Table 1
Correlation between grades and positive, negative and neutral in/out degree.

	Spearman Coeficient		p-value	
degree	Class 1	Class 2	Class 1	Class 2
$indeg^{+}$	**0.4727**	**0.4572**	0.0073	0.0491
$indeg^{-}$	-0.2543	-0.0191	0.1674	0.9381
$indeg^{0}$	-0.3433	**-0.5175**	0.0586	0.0232
$outdeg^{+}$	-0.0363	-0.3081	0.8461	0.1993
$outdeg^{-}$	-0.0471	**0.5178**	0.8014	0.0231
$outdeg^{0}$	-0.0373	-0.1303	0.8421	0.5949

them, and that is the reason we use it. Moreover, during this stage of graduation, students know that it is very likely that a student who got good (bad) grades in a class is also going to get good (bad) grades in another. The sociometric test was used, among other reasons, to evaluate if these notions are as important as social affinities during this particular time of graduation.

3.2. Performance in class

In collaborative tasks, maybe the most used (or expected) strategy to pick collaborators is to select those who are the most proficient to do the task. For instance, consider scenarios where a company is hiring employees or two soccer captains are picking players in a match among friends. It is not an absurd to say that most people would guess that the most skilled ones would be selected first. Thus, in order to verify if and how much the proficiency of the students is related to the answers they give and receive we collect the grades they got for this particular class in the semester. In Fig. 3a and b we show the histogram of the grades obtained by the students, in a range from 0 (worst) to 100 (best). Observe that although most of the students have grades between 71 and 90, there are those who have failed in the course (grades bellow 60) and those who achieved an excellent performance (grades higher than 90).

To verify the impact of the grades in the answers gave by the students, we calculated the Spearman's rank correlation coefficient between the rank given by the grades and the rank given by the in and out degree of the students grouped by the sign of the edge. We use the terms *indeg* and *outdeg* to indicate the indegree and the outdegree, respectively. Moreover, we use the symbols + , 0 and - to indicate the positive, neutral and negative signs, respectively. The correlation results are showed in Table 1 for the classes 1 and 2. We also compute the p-values for testing the hypothesis of no correlation against the alternative that there is a nonzero correlation. For the Spearman's rank correlation, this is done by generating random permutations of the ranks and verifying the probability of having the given correlation coefficient from this random process. Observe in Table 1 that there is a significant correlation and low p-value between $indeg^{+}$and students' grades. From this, we can conclude that proficient students attract positive answers in the survey, i.e., students who choose to work with her/him. However, observing the other correlations, which are not significant, and p-values, which are high, we can conclude that a student's grade do not have a causal relationship to the number of negative and neutral answers she/he receive and, also, to the answers she/he gives. Thus, although the grades (or the proficiency) of the students have an impact in their answers, there is still a lot that they cannot explain.

3.3. Gathering facebook data

Thus, our conjecture is that some of the answers can also be explained by the position of the student in the social network formed by the students of this particular class. Consider, for instance, positive answers given by close friends or

negative answers given between groups of students that do not go along well. To what extent can an answer be guided by factors similar to these? To answer that, we collect the Facebook interactions of the students questioned in the survey. For this, we have developed an application that collects several information from their Facebook accounts, such as their friends in class, the number of inbox messages they exchange, their posts and respective comments, among others. It is important to point out that all students agreed to participate, and only data related to them was collected, i.e., we do not have any information from people outside the class.

A summary of the data we collected from Facebook can be seen in the Supplementary Information material. From initial observations, we noted that the occurrence of friendships on neutral edges is significantly lower than on positive and negative edges. Moreover, the average number of comments on shared links among negative edges is greater than on positive and neutral. As expected, we see that the average number of inbox message exchanged on positive edges is significantly higher than on neutral and negative edges. We also observed that the number of common interests is very low for the three edge classes. From these initial observations we see a potential impact of social interactions in the answers made by the students. We formalize and quantify this impact in the following sections.

3.4. Data limitations

In terms of the limitations of our datasets, we note that representativeness is a very challenging issue in our study, as in many empirical analyses. We here designed an experimental methodology that is as thorough as possible, given our practical constraints. We applied a sociometric test in two classes of undergraduate students, where all students agreed to participate and all of them have a Facebook account, allowing us to gather their online social interactions through a third-party Facebook application. We left as future work the design of experiments that covers larger classes of students from different backgrounds and countries. Furthermore, although our experiments are limited to classes of 31 and 19 students, the objects of study here are the relationships among these students, which correspond to 930 and 342 links labeled as positive, negative, or neutral. To ensure that our sample sizes are not too small to draw conclusions, in all analysis we applied statistical tests to verify if results are statistically meaningful. We left as future work the validation of our findings within different universities and in different scenarios, like companies.

It is also important to note that our Facebook dataset consists only of statistics about the interactions among the students who agreed to participate in our experiments. Our Facebook application could not collect the content of the messages exchanged by students due to limitations imposed by the ethics council of the university where we applied the sociometric test. This prevented us to explore a number of features, for instance, the aspects related to the sentiment expressed in the messages exchanged among the students.

Finally, we note that representativeness is an important but challenging issue in any empirical study, as ours. We here design an experimental methodology that is as thorough as possible, given our practical constraints. However, we acknowledge that it is impossible to generalize our findings without future studies. We hope that this work will encourage future efforts to apply our methodology across various universities and over more countries.

4. Social features

We have seen that although the target's proficiency is correlated with the decision of selecting or not this target for a collaboration, it cannot explain everything. Moreover, we have seen that particular Facebook interactions are more (or less) present in certain groups of signed edges, indicating that social behavior may also impact in the answers we got in the survey.

Thus, in this section we describe several social features that are able to influence the decision of selecting a person to collaborate. These features are directly extracted from the Facebook data that we collected. We modeled this data into an undirected graph $G_F(V, E_F)$ where the set of nodes V are the students (the same set of $G_S(V, E_S)$) and an edge exists between two students if they are friends in Facebook. We divide the proposed social features into two groups:

- **G1: Actor attributes**, which apply to the students and
- **G2: Link attributes**, which apply to the relationship between two students. Note that even if two students are not friends in Facebook their relationship will have a value for the attribute.

For the attributes in **G1**, we verify and quantify the influence using the same methodology we used to identify that the grades had an impact in the answers, i.e., we compute the Spearman's rank correlation coefficient between the rank given by the in and out degree of the students in G_S, grouped by the sign of the edge and the rank given by the attribute. For each attribute in **G2**, we compute the Cumulative Distribution Function (CDF) of the attribute grouped by the sign of the edge. If two CDFs (e.g. the CDFs for negative and positive edges) are significantly distinct, then we have a strong indication that the attribute is able to influence the answers. To verify if two CDFs are significantly distinct, we perform a two-sample Kolmogorov-Smirnov (KS) test to check if the two samples are drawn from the same distribution. If this hypothesis is rejected, we can conclude that the two distributions are significantly different and, therefore, the attribute may separate edge signs.

Table 2
Impact of the popularity in the answers given in the questionnaire applied. The values correspond to the Spearman's rank correlation coefficient (and the respective p-value) between the rank produced by the popularity metrics and the rank given by the in and out degree of the students in G_S, grouped by the sign of the edge.

	$popularity_1$		p-value		$popularity_2$		p-value	
degree	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
$indeg^+$	**0.46**	0.15	0.009	0.52	**0.49**	0.1	0.004	0.67
$indeg^-$	**0.36**	0.21	0.04	0.37	0.18	-0.03	0.32	0.87
$indeg^0$	**-0.74**	-0.18	0.0001	0.44	**-0.66**	-0.01	0.0003	0.94
$outdeg^+$	**0.58**	0.2	0.0007	0.39	**0.37**	0.21	0.03	0.38
$outdeg^-$	0.12	-0.09	0.52	0.69	0.21	-0.006	0.24	0.98
$outdeg^0$	**-0.64**	0.02	0.0001	0.931	**-0.46**	0.11	0.008	0.62

Table 3
Impact of extroversion metrics in the choice of partners on collaborative activities.

	$extroversion_1$		p-value		$extroversion_2$		p-value	
degree	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
$indeg^+$	0.03	0.12	0.85	0.6	**0.51**	0.13	0.002	0.59
$indeg^-$	**0.41**	0.01	0.01	0.95	0.23	0.13	0.212	0.59
$indeg^0$	-0.26	-0.07	0.14	0.75	**-0.72**	-0.14	0.003	0.55
$outdeg^+$	0.09	0.16	0.61	0.5	0.38	-0.05	0.031	0.8
$outdeg^-$	0.22	-0.002	0.22	0.99	0.26	-0.23	0.153	0.34
$outdeg^0$	-0.22	0.17	0.22	0.46	**-0.52**	0.43	0.002	0.06

4.1. G1: actor attributes

4.1.1. Popularity

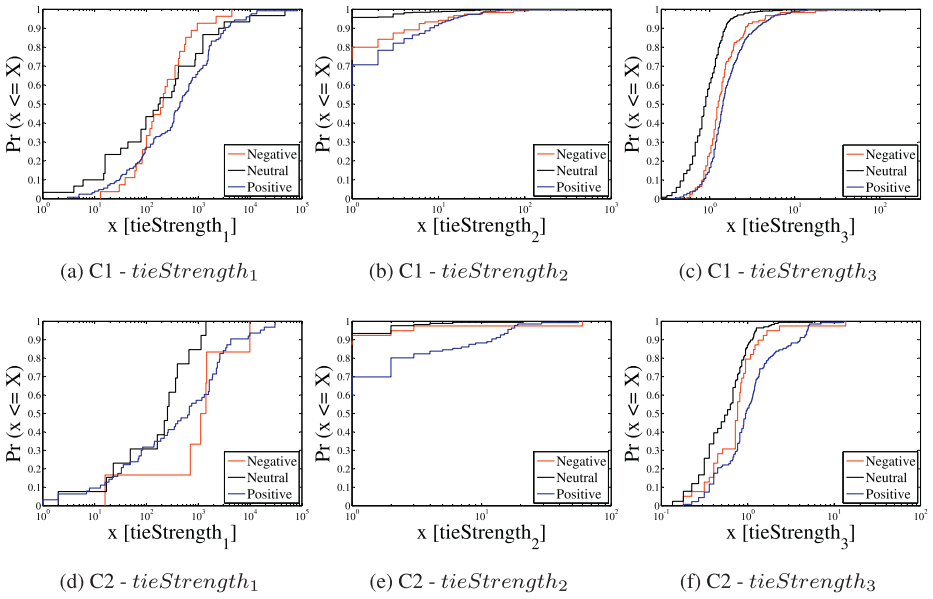
Here we investigate if popular students in class tends to attract a specific type of answer, e.g. positive edges. We calculate the popularity of a student in two ways. First, we define the metric $popularity_1(i)$ as the number of students in class that student i is friend on Facebook, i.e., $popularity_1(i) = degree_{G_F}(i)$. Moreover, we define the metric $popularity_2(i)$ as the number distinct students who posted activities in student i's Facebook page, e.g., comments on her/his links, likes on her/his photos, among others.

In Table 2 we show the Spearman's rank correlation coefficient between the rank produced by the popularity metrics and the rank given by the in and out degree of the students in G_S grouped by the sign of the edge. First, observe in Table 2 that while the $popularity_1$ metric is significantly correlated with the degree of the students for several signs and in both directions, the $popularity_2$ metric is significantly correlated only with the $indeg^0$. We say a correlation is *significant* when the p-values are lower than 0.05 (note the numbers in bold for significant correlations). In both cases, the strongest correlation is seen for $indeg^0$, i.e., the number of incoming neutral edges. Since it is negative, it indicates that students who are not popular tend to receive more neutral edges, i.e., people are usually indifferent towards them. Moreover, since the $outdeg^0$correlation is significant for the $popularity_1$ metric, we may also infer that students who are not popular also tend to vote "INDIFFERENT" more. On the other hand, by observing the correlation for the $indeg^+$ and $indeg^-$, it is curious that the more popular is a student, more she/he tends to receive negative and positive votes students. This could indicate that popular students are more well known by the class, so it is easier to make an strong point (negative or positive decision) about them. Assessing $outdeg^+$, it is possible to infer that curiously popular students tend to vote positive more.

4.1.2. Extroversion

Another feature that may impact in the students' decision is their level of extroversion. Extrovert people tend to enjoy human interactions and to be enthusiastic, communicative,assertive, and gregarious [14]. Here, we infer students extroversion based on the number of public interactions that individuals perform in other student's walls. We assume the extent to which an individual publicly interacts with others on Facebook measures how much social attention this individual is seeking, which represents the central feature of extrovert people [2]. We define two ways to measure if a student is extrovert: the metric $extroversion_1(i)$ as the number of public interactions that the student i published in others Facebook pages, e.g., comments on others' links, likes on others' photos, among others; and the metric $extroversion_2(i)$ as the number distinct students to which the student i posted public activities.

We show in Table 3 the Spearman's rank correlation coefficient between the rank produced by the extroversion metrics and the rank given by the in and outdegree of the students in G_S. For most of the results concerning the $extroversion_1(i)$ metric, we can note low correlations and high p-values. However, observe that there is a significant correlation between the $extroversion_1(i)$ metric and the $indeg^-$, which may indicate that the more an individual posts on other's walls, less other



(a) C1 - $tieStrength_1$ (b) C1 - $tieStrength_2$ (c) C1 - $tieStrength_3$

(d) C2 - $tieStrength_1$ (e) C2 - $tieStrength_2$ (f) C2 - $tieStrength_3$

Fig. 4. CDFs for the tie strength metrics grouped by their sign (Class 1 and Class 2).

students want to work with her. This could indicate that students who excessively post public comments on Facebook may also be intrusive, generating negative reactions from others. Another observation is that when a student posts an excessive number of messages to others, she may leave the impression that she spends an excessive time procrastinating on Facebook and, for this reason, would not be a good project mate. Concerning the $extroversion_2(i)$ metric, observe there are significant negative correlations for the $indeg^0$and $outdeg^0$. This could indicate that students who are not publicly active on Facebook are usually not well known by the others, generally attracting and generating neutral reactions. Moreover, we could understand that students who post public comments on a large number of Facebook pages tend to attract either positive or negative reactions, mostly positive, since the correlation is significantly positive with the $indeg^+$.

4.2. G2: link attributes

4.2.1. Strength of the tie

The *strength of the tie* measures how close two individuals are. As we mentioned before, there are several ways to compute that when online social network data is available. In this paper, we consider four metrics. First, we define the metric $tieStrength_1(i, j)$ as the total number of private inbox messages students i and j exchanged. Second, we define the metric $tieStrength_2(i, j)$ as the total number of public interactions students i and j exchanged, i.e., we count all the public activity student i posted on student j's profile page and vice-versa. Finally, we define the metric $tieStrength_3(i, j)$ as the tie strength metric proposed by [20], which is a linear model considering seven groups of variables extracted from Facebook. In this case, we use the same coefficients as shown in [20] and we considered only the data we have available, which allowed us to consider only three groups of variables: *Structural*, Intensity and Intimacy variables. In the group of *Structural Variables*, we used common musics, common groups, common interests, common movies and common friends. In the group of *Intensity Variables*, we used comments on photos, comments on links, comments on status updates, comments on albums, likes on photos, likes in links, likes in status updates and inbox messages. In the group of *Intimacy Variables*, we used tags on photos and status updates. Finally, we define the binary metric $tieStrength_4(i, j)$ as 1 if students i and j are friends on Facebook and 0 otherwise.

In Fig. 4 we show the CDFs for the first three tie strength metrics grouped by their sign, separated for Class 1 and Class 2. Also, in Table 4 we show the KS statistic for the two-sample Kolmogorov-Smirnov (KS) test to check if two tie strength distributions are drawn from the same distribution. The KS statistic is $D_{1,2} = \sup_x |CDF_1(x) - CDF_2(x)|$, where CDF_1 and CDF_2 are the empirical distribution functions of the first and the second sample, respectively. If this hypothesis is rejected (values in bold), we can conclude that the two distributions are significantly different and, therefore, the given tie strength attribute may separate edge signs. Conversely, if it is accepted, then the given tie strength attribute is probably not a good predictor of the edge sign.

First, observe that the shape of the curves are similar between the classes. In fact, from Table 4 we can see that in only two cases the result of the KS test differs between the two classes. Moreover, the $tieStrength_3$ metric is the one that best separates the three edge signs, since all hypothesis tests were reject with very low p-values. Concerning the $tieStrength_1$ metric, although it cannot distinguish the three curves very well, there are striking differences among edge signs. For

Table 4
The KS statistic for the KS test that the two samples come from the same distribution. Bold values indicate that the hypothesis was rejected and, therefore, the given tie strength attribute may separate edge signs. Significance of the test: $p-value < 0.05^{*}$, $p-value < 0.01^{**}$, $p-value < 0.001^{***}$, $p-value < 0.0001^{****}$.

	$tieStrength_1$		$tieStrength_2$		$tieStrength_3$	
edge pair	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Neg *vs.* **Neu**	0.20	0.12	**0.29****	0.03	**0.37******	**0.33****
Neg *vs.* **Pos**	**0.29***	**0.31****	**0.16***	**0.28***	**0.14***	**0.29****
Neu *vs.* **Pos**	0.22	**0.39******	**0.35******	**0.30******	**0.45******	**0.39******

instance, for Class 1, the neutral curve has approximately 50% of edges with less than 100 conversations, while for the negative and positive curves, that number is 30%. Moreover, for the $tieStrength_2$ metric, while approximately 96% of neutral edges has zero public interactions, that number decreases to 80% for negative edges and to 71% for positive edges. Concerning the $tieStrength_3$, observe that all three distributions have similar behavior, with the neutral curve being reasonably far apart from the others. Approximately 70% of the neutral edges have $tieStrength_3$ values smaller than 1, whereas for the positive and negative edges, these values represent approximately 18% of the edges. In summary, given that most of the KS tests failed and were consistent between the two classes, we conclude that the tie strength metrics may be used to explain the answers given by the students in the sociometric test.

For the $tieStrength_4$ metric, since it is binary, we simply compute the proportion of edges that have values $tieStrength_4 = 1$, i.e. are friends on Facebook, for each given sign. For the negative edges, the proportion is 40% in Class 1 and 31% in Class 2, while for the positive edges, the proportion is 46% and 49%, respectively. These values are significantly higher than the one for the neutral edges, that is 20% for both classes. This indicates that the $tieStrength_4$ metric has also a potential to differentiate neutral edges from positive and negative ones.

4.2.2. Homophily

Homophily is the tendency of individuals to associate and bond with similar others [33]. Individuals in homophilic relationships share common characteristics. To investigate the homophily in our context, we define three different metrics to measure the similarity on Facebook. The $similarity_1$ measures the similarity between two individuals in terms of the network topology. To capture the proximity between individuals, we apply the Jaccard Coefficient, which is able to measure the degree of overlap between node vectors, i.e., the neighbors of each node. Given two node vectors r_i and r_j representing the neighbors of students i and j in G_F, we define $similarity_1$ as:

$$similarity_1(i, j) = \frac{|r_i \cap r_j|}{|r_i \cup r_j|}$$

where r_i and r_j is the set of friends that the students i and j have on Facebook, respectively.

Second, we define the metric $similarity_2$ as a measure of the features that two students have in common on Facebook. To measure this we use the information about the movies and Facebook groups that two students have in common. We also collected information about music and interests, but since only a few pairs of students have items of such nature in common, we discarded these features in our analysis. Given two feature vectors mov_i and gr_i representing the movies and groups that a student i own, respectively, we define the $similarity_2(i, j)$ as:

$$similarity_2(i, j) = \frac{|mov_i \cap mov_j|}{|mov_i \cup mov_j|} + \frac{|gr_i \cap gr_j|}{|gr_i \cup gr_j|}$$

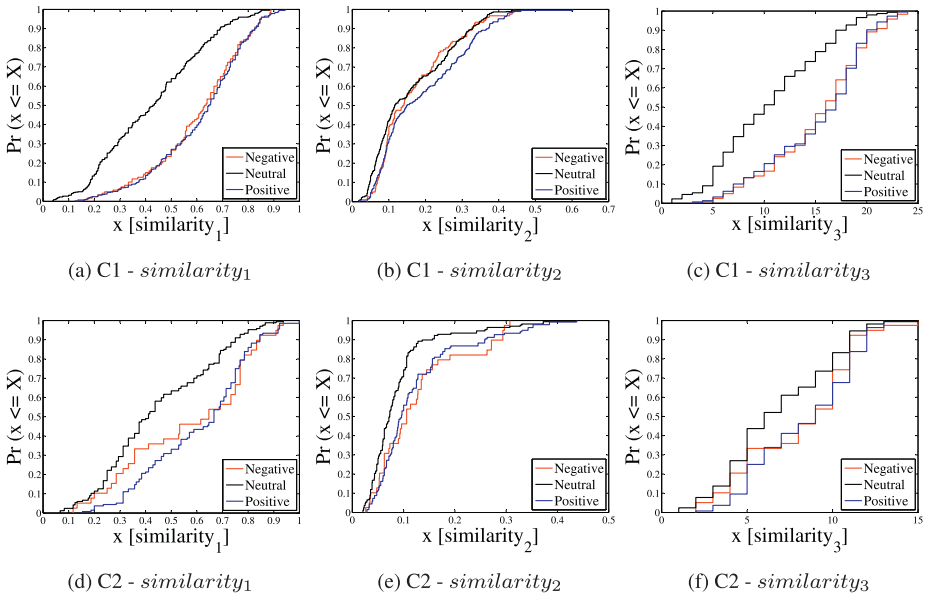
where we applied the Jaccard Coefficient in this two vectors of each students i and j and calculate the arithmetic average of these two values.

Moreover, we define the metric $similarity_3$ as the number of common friends that two different students share on Facebook. We do not add this data to $similarity_2$ because it is related to the network structure, i.e., it is not information that students share on Facebook. This feature only represents the number of common friends. Moreover, although $similarity_1$ and $similarity_3$ are similar metrics, the former measures how close two groups of friends are, while the latter measures the magnitude of their intersection.

In Fig. 5 we show the CDFs for the two homophily metrics grouped by their sign. Again, we show in Table 5 the KS statistic of the two-sample Kolmogorov-Smirnov (KS) test to check if two similarity distributions are drawn from the same distribution. Observe that the neutral distribution is clearly distinct from the other two, which are fairly similar. This shows that at the level of the network structure, neutral relationships have very different behavior from the positive and negative relationships.

5. Working affinities classification

In order to explain the behavior of working affinities through proficiency and social interactions, we present an analysis using machine learning algorithms to classify the relationships. This step is related to the last activity in the *Data Analysis*,



(a) C1 - $similarity_1$ (b) C1 - $similarity_2$ (c) C1 - $similarity_3$

(d) C2 - $similarity_1$ (e) C2 - $similarity_2$ (f) C2 - $similarity_3$

Fig. 5. CDFs for the similarity features grouped by their sign.

Table 5
The KS statistic for the KS test that the two samples come from the same distribution. Bold values indicate that the hypothesis was rejected and, therefore, the given similarity attribute may separate edge signs. Significance of the test: $p-value < 0.05^{*}$, $p-value < 0.01^{**}$, $p-value < 0.001^{***}$, $p-value < 0.0001^{****}$.

	$similarity_1$		$similarity_2$		$similarity_3$	
edge pair	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Neg *vs.* **Neu**	**0.39******	**0.30****	**0.15***	**0.30****	**0.39******	**0.25***
Neg *vs.* **Pos**	0.0645	0.1863	**0.1411***	0.10	0.07	0.11
Neu *vs.* **Pos**	**0.3928******	**0.3088******	**0.13****	**0.25*****	**0.38******	**0.20****

proposed in the methodology presented in Section 3. Using the characteristics presented in previous sections, we classify the relationships of students according to the three classes of work affinity: negative, neutral and positive. These classes correspond to the answers reported by students in sociometric test.

5.1. Classifiers

The methods used to classify are implemented and available through the software *Weka*[1]. *Weka* is a free software containing a collection of machine learning algorithms and methods for statistical analysis. We evaluated five machine learning models: *Multilayer Perceptron* (MLP), *Random Forest* (RF), *Naive Bayes* (NB), *k-Nearest Neighbors* (KNN) and *Support Vector Machines* (SVM). These algorithms are based on supervised learning and have as input a file, organized as follows: each row of the file corresponds to an observation; each column corresponds to a Facebook characteristic; the last column represents the working affinity label (positive, negative or neutral). We chose these algorithms because they are widely used and are significantly different among themselves.

5.2. Data treatment

Sociometric test dataset have direct edges labeled as positive, negative and neutral. However, there is a significant difference between the number of labels: 12.9% negative, 47.41% neutral and 39.69% positive in *Class 1*; 11.40% negative, 48.83% neutral and 39.77% positive in *Class 2*. A dataset is imbalanced if the classification categories are not equally represented, i.e., there is much more instances of some classes than others [9]. There are a number of solutions to this problem. We use two solutions that include different approach of resampling: oversampling and undersampling both as data level [10]. Undersampling technique consists of selectively choose a subset patterns from the majority class, keeping the minority class

[1] http://www.cs.waikato.ac.nz/ml/weka/.

with the original data [27]. Oversampling technique proposes that the minority class has an oversampling, creating "synthetic patterns", as an effective way to increase the classifier sensitivity to minority class [9]. In our experiments, we use the imbalanced data set and two approaches to treat the imbalance problem:

- *Random Undersampling*: It is a simple approach to sampling. Patterns of the majority class are randomly deleted until the ratio of the minority and majority class has a desired level. Despite its simplicity, the random undersampling has been empirically shown to be one of the most effective methods of sampling;
- *SMOTE Oversampling*: the SMOTE (***S**ynthetic **M**inority **O**ver-sampling **TE**chnique* - is simple, but efficient, outperforming random oversampling in various problems of low dimensionality [28]. In this algorithm, the minority class is increased by taking each sample and introducing synthetic examples throughout the dataset. Depending on how much one wants to increase the number of minority class patterns, k neighbors of each sample are selected. The basic implementation commonly uses 5 closest neighbors [9].

Undersampling and oversampling techniques used with appropriate classification algorithms produce satisfactory results [9,10,28,29]. Since it is not our objective exhausting the continuous improvement efforts of the classification results, we used these two sampling techniques and the imbalanced data to show the potential in each application. All classification results presented in this paper were obtained using the oversampling technique. The results with imbalanced and undersampling data are consistent with the oversampling results and they are presented in Supplementary Information material. Undersampling and oversampling techniques are applied only during the training period.

We have 930 tuples (or relationships) in class 1 and 342 tuples in class 2 to build the input matrix. The actor attributes were processed as follows: for a relationship ($i \rightarrow j$), the student *source* i answered about his work affinity for the student *target* j. Thus, the values of these features have been placed for both students *source* and *target*. The relationship features between students i and j (*source* and *target*) were placed as individual columns in the matrix. The output for the classification models are relationships labels ($i \rightarrow j$) resulting from the sociometric test. Moreover, the input values were normalized using the following formula:

$$norm_i = \frac{x_i - min}{max - min}$$

where min is the smaller value in C_k, C_k is a set with the feature k values, max is the bigger value in C_k and x_i is the value to be normalized. A summary of the data used in the experiments can be seen in Supplementary Information material.

5.3. Evaluation metrics

We evaluate our classification results using the following standard metrics:

- **Confusion matrix**: provides an effective measure of a classification model by showing the number of correct classifications versus the predicted ratings for each class. Each column of the matrix represents the instances of a class that was expected, as each line represents the actual instances of the class [43].
- **Precision**: the proportion of correct predictions in relation to the total number of predictions made for a class. Formally, it is defined as $precision = \frac{TP}{TP+FP}$, where FP (false positive) is the set of cases of class α that were incorrectly classified as being of this class, and TP (true positive) is the set of instances of class α that were correctly classified.
- **Recall**: the fraction of relevant instances that are correctly predicted. Formally, it is defined as $recall = \frac{TP}{TP+FN}$, where FN (false negative) corresponds to the set of relevant instances of class α that were incorrectly classified.
- **Micro-F1** and **Macro-F1**: First, we define the metric F1 (*F-Score* or *F-Measure*) as the harmonic mean between precision and recall, defined by $F1 = 2pr/(p+r)$, where p is the precision and r is the recall [49]. Thus, the micro-F1 is calculated by computing first the overall precision and recall values, and then calculating F1. Macro-F1 metric values are first computed obtaining the F1 results for each class in isolation and then averaging these results. The macro-F1 considers each class equally important, while the micro-F1 is related only to the number of instances correctly classified by the model, independent of the number of instances of each class.

5.4. Classification setup

We tested the classifiers using three different sets of features:

1. **Social features**: only the features derived from the social interactions collected from Facebook, such as the strength of the tie, homophily, extroversion and popularity.
2. **Proficiency features:** only the features that describe the proficiency of the subject, which is, in this case, the grade of the student.
3. **All features:** both the social and proficiency features.

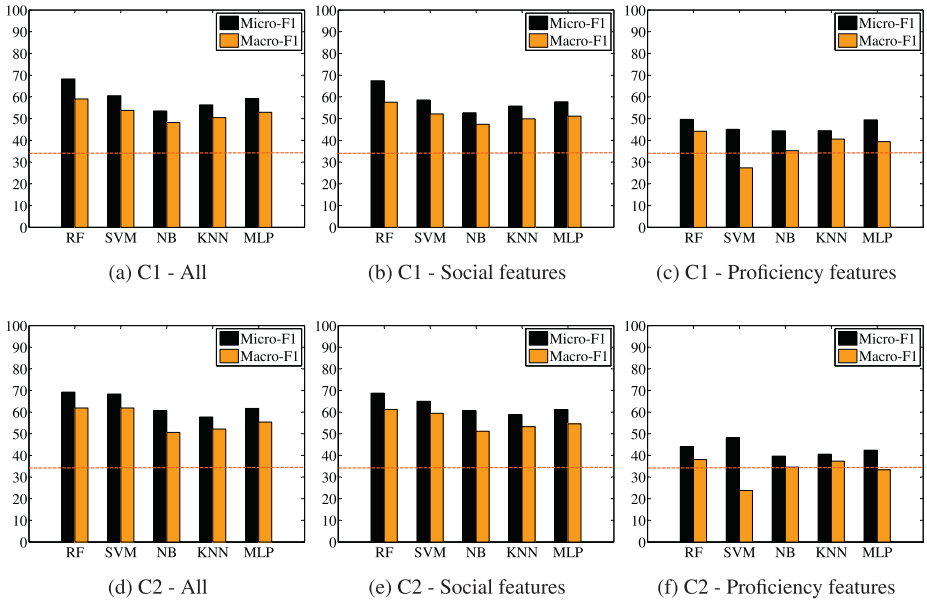


Fig. 6. Classification results.

Table 6
Confusion matrix for Class 1 (C1) and Class 2 (C2) with Random Forest classifier using all features.

	Predict					
	Neg (C1)	**Neg (C2)**	**Neu (C1)**	**Neu (C2)**	**Pos (C1)**	**Pos (C2)**
Negative	**28.89%**	**38.46%**	35.86%	40.85%	35.25%	20.68%
Neutral	6.81%	8.78%	**77.28%**	**74.93%**	15.91%	16.29%
Positive	10.42%	4.12%	19.30%	24.85%	**70.28%**	**71.03%**

The classification process is conducted using the *k-fold cross-validation* technique, which is also called rotation estimation. In this technique, the dataset D is randomly divided into k disjoint subsets $D_1, D_2, \ldots, D_k$ of approximately equal size, known as *folds* [19]. The classifier is trained and tested k times: every time $t \in \{1, 2, \ldots, k\}$ is trained in $D - D_t$ and tested in D_t. Oversampling technique was applied to all sets $D - D_t$ during the training period. The accuracy of cross-validation can be estimated from the total number of correct classifications, divided by the number of instances in the dataset [26]. In the experiments, the original database is partitioned into five subsets. In all experiments, *5-fold cross validation* was repeated 50 times with different seeds used to shuffle the original database. Thus, the results presented here are the arithmetic means of these 50 runs with their respective 95% confidence interval.

5.5. Classification results

In Fig. 6, we present the classification results for the two groups of students when each one of the three sets of features were considered. Values for Micro-F1 are equivalent to the accuracy of each classifier, while Macro-F1 is more sensitive to the hit rate of each class. First, observe that the classification results using social features are significantly better than when solely proficiency features were used. Also, note that adding proficiency features to the social features (first column) does not contribute much to improve the classification results using solely social features (second column). The highest scores for both C1 and C2 were achieved by the Random Forest (RF) algorithm using all features. In this case, for C1 the Micro-F1 and Macro-F2 are 0.682 and 0.59, respectively. For C2, the Micro-F1 and Macro-F2 are 0.692 and 0.619, respectively.

Fig. 7 shows the precision and recall results obtained in our experiments. Again, for both C1 and C2 the highest scores were achieved by the RF algorithm when all features were used. For C1, the precision was 0.675 and the recall was 0.683. For C2, the precision was 0.690 and the recall was 0.692.

Since the RF classifier was the one which achieved the best results in all cases, in Table 6 we show the confusion matrix for the RF classifier. Bold values âÇïâÇïrepresent instances that were correctly classified. Observe that, for all cases, the success rates of neutral edges are the highest, followed by the positive and negative edges.

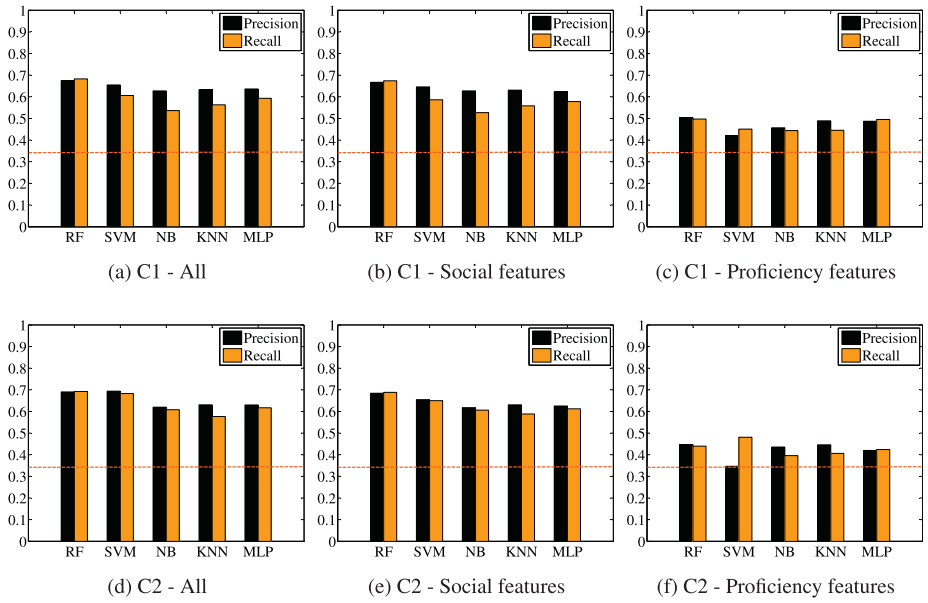


Fig. 7. Precision and Recall.

Table 7
Ranking of most important attributes, presented by the IG (*Information Gain*) Ranking and the χ^2 (*Chi-Squared*) Ranking.

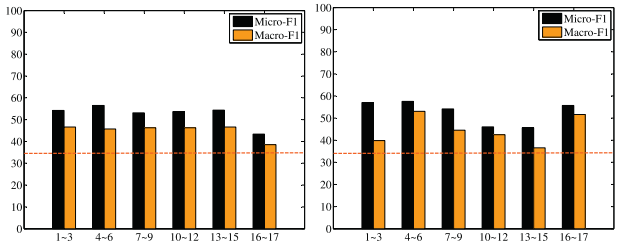
Description	**Rank IG** (Class 1)	**Rank IG** (Class 2)	**Rank χ^2** (Class 1)	**Rank χ^2** (Class 2)
$tieStrength_3$	1	5	1	5
$tieStrength_4$	2	2	2	4
$similarity_1$	3	7	3	7
$similarity_3$	4	14	4	11
$tieStrength_2$	5	3	5	3
$popularity_1(source)$	6	4	8	6
$extroversion_1(target)$	7	9	6	9
$tieStrength_1$	8	1	9	2
$popularity_2(target)$	9	17	10	17
$grade(target)$	10	13	7	13
$extroversion_2(target)$	11	11	11	14
$extroversion_2(source)$	12	10	13	10
$grade(source)$	13	12	12	12
$popularity_1(target)$	14	15	14	15
$extroversion_1(source)$	15	6	15	1
$popularity_2(source)$	16	16	16	16
$similarity_2$	17	8	17	8

5.6. The most important features

After having analyzed our Facebook-derived features separately, we now study how they collectively explain tie formation within a team, and, more importantly, which are the most predictive ones. To this end, we resort to two measures: the *Information Gain* and χ^2 (Chi Squared) coefficients [50]. Both of them are feature selection methods widely-used to identify the subset of the features that are most predictive in a classification.

We process the 930 and 342 tuples as shown in Section 5.2, in the form *source* student *i* decided what to do with *target* student *j* (each tuple comes with corresponding features and class grades), and we obtain the results in Table 7. We find that the two measures consider the very same features to be relevant. The most important result is that class grades are not that important: We need to go down C1's list at the 10th and 7th positions to find them and at the 13th positions in C2's list. By contrast, features related to tie strength, such as the Gilbert's proxy for tie strength [20], were consistently high ranked in terms of importance.

Finally, note that feature $extroversion_1(source)$ is the most important feature according with χ^2 for Class 2, what is a surprising result, since its rank is low for Class 1. This suggests that, although social features are proved here to be predictiors of working affinities, they should be used carefully. Distinct scenarios can be correlated, but their differences may lead to different prediction models. Thus, this show the importance of machine learning when tackling the task of predicting working affinities using social features.



(a) Class 1 (b) Class 2

Fig. 8. Impact of Reducing the Attribute Set.

A question that naturally arises from ranking features is: how good would the classification results be considering only a subset of the features? To answer that, we performed the classification task for six disjoint subsets of the features listed in Table 7. Each subset *ij* is composed by the features ranked in the *i*th to the *j*th position in the Information Gain rank described in Table 7. For instance, the subset 13 contains the most, the second most, and the third most important features in the Information Gain rank. Thus, in Fig. 8 we show the classification results for all these subsets of features. Observe that the Micro-F1 and Macro-F1 values using such subsets are significantly worse than the results using all features (Fig. 6). Nevertheless, note that even small subsets of features are able to generate significantly better results than chance (the dashed red line at 33%).

6. Conclusion and result applicability

This work proposes a methodology for predicting working affinities from online social networks interactions, which can potentially improve group formation algorithms. The proposed methodology consists of four well-defined steps. First, we extracted from a group of people the working affinities among them, which will be input to the experiment. For that, we applied a sociometric test in two classes of undergraduate students, in which individuals were asked whether they would like to work with each one of their classmates. Second, through a Facebook application we developed, we collected information about the online interactions within this group of people. Third, we establish a mapping between the online social interactions and the working affinities. This mapping is done through the proposed social features, which were presented in Section 4. Finally, we used machine learning algorithms to predict working affinities among students from these features. For this, an experiment was conducted in a very particular scenario. First, we applied a sociometric test in a class of undergraduates students, in which individuals were asked whether they would like to work with all other students in the class. Together with social features, we also investigated the impact of proficiency features, represented here by the grades of these students.

Although the dataset seems limited in terms of students, comprising one class of 31 students and one with 19 students, it is not in terms of relationships, which is the scope of this work. In total, we analyzed 1272 working affinities, where each one of these is described by a rich set of information. From this analysis, which involved both online and offline data, we verified that Facebook-derived features are more predictive of whom students wish to work with than proficiency features, i.e., the students' grades. The features related to tie strength, such as the Gilbert's proxy for tie strength [20], were consistently high ranked in terms of importance, suggesting the importance of bonding (as opposed to bridging) social capital in team formation [5]: as one expects, trust and social embeddedness (rather than presence of weak ties) are associated with willingness to team up. To see why this is of theoretical importance, consider that Facebook is a distal communication modality, in that, users are separated in space and time. Yet, our results suggest that the social-networking site resembles proximal communication between students embedded in the classroom's offline social network, and that is in line with recent studies on the relationship between offline and online interactions.

Concerning the applicability of our work, nowadays companies use dynamic cooperatives, where the harmonious interpersonal relationship is highly desirable by these institutions. Practices for harmonic grouping are used by human resources departments to maximize the performance of work teams, consequently increasing the productivity of these companies [36,42]. The social dynamics in a group of people can be analyzed and modeled using the methodology proposed by [35], the Sociometric Test. Although the realization of a Sociometric test to team up is efficient and effective, this method can be, in some circumstances, invasive and even uncomfortable. In this context, our proposed prediction models could be used to avoid the sociometric test and provide information of working affinities more confidently, thereby improving the performance in the execution of collaborative tasks.

These results are also of practical importance for educational sites such as Coursera[2] or universities that offer online courses. Given the large number of individuals in the world such sites serve, one effective way for them to team up students

[2] https://www.coursera.org/about.

at scale is to use the very same features we have studied here. In the future, we plan to repeat similar studies across classes in different countries to explore cross-cultural effects. After that, a real application that recommends teams out of Facebook accounts is in order.

Acknowledgments

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-14, and by individual grants from CNPq, CAPES, IFSULDEMINAS and Fapemig.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ins.2016.09.009.

References

[1] L.E. Agustín-Blas, S. Salcedo-Sanz, E.G. Ortiz-García, A. Portilla-Figueras, A.M. Pérez-Bellido, S. Jiménez-Fernández, Team formation based on group technology: a hybrid grouping genetic algorithm approach, Comput. Oper. Res. 38 (2) (2011) 484–495.
[2] M.C. Ashton, K. Lee, S.V. Paunonen, What is the central feature of extraversion? Social attention versus reward sensitivity, J. Person. Social Psychol. 83 (1) (2002).
[3] F. Benevenuto, T. Rodrigues, V.A. Almeida, J. Almeida, M. Gonçalves, K. Ross, Video Pollution on the Web, First Monday, 15, 2010.
[4] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, Characterizing user navigation and interactions in online social networks, Inf. Sci. 195 (15) (2012) 1–24.
[5] M. Burke, R. Kraut, C. Marlow, Social capital on facebook: differentiating uses and users, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, pp. 571–580.
[6] D.M. Bustos, The Sociometric Testing: Fundamentals, Techniques and Applications, Brasilense Publisher, 1979.
[7] F. Campuzano, T. Garcia-Valverde, J.A. Botia, E. Serrano, Generation of human computational models with machine learning, Inf. Sci. 293 (2015) 97–114.
[8] D. Castilho, P.V.d. Melo, D. Quercia, F. Benevenuto, Working with friends: unveiling working affinity features from facebook data, in: Proceedings of the International AAAI Conference on Web-Blogs and Social Media, Ann Arbor, MI, USA, 2014.
[9] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. (2002) 321–357.
[10] N.V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, ACM Sigkdd Explor. Newslett. 6 (1) (2004) 1–6.
[11] S.-J. Chen, L. Lin, Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering, IEEE Trans. Eng. Manag. 51 (2) (2004) 111–124.
[12] H. Costa, L.H. Merschmann, F. Barth, F. Benevenuto, Pollution, bad-mouthing, and local marketing: the underground of location-based social networks, Inf. Sci. 279 (2014) 123–137.
[13] D. Easley, J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World, 8, Cambridge Univ Press, 2010.
[14] H. Eysenck, Readings in Extraversion-introversion: Theoretical and Methodological Issues, Readings in Extraversion-introversion, Staples Press, 1970.
[15] F. Figueiredo, J. Almeida, M. Gonçalves, F. Benevenuto, Trendlearner: early prediction of popularity trends of user generated content, Inf. Sci. 349–350 (2016) 172–187.
[16] E.L. Fitzpatrick, R.G. Askin, Forming effective worker teams with multi-functional skill requirements, Comput. Indus. Eng. 48 (3) (2005) 593–608.
[17] D.C. Frezzo, K.E. DiCerbo, J.T. Behrens, M. Chen, An extensible micro-world for learning in the data networking professions, Inf. Sci. 264 (2014) 91–103.
[18] L.K. Gee, J.J. Jones, M. Burke, Social networks and labor markets: how strong ties relate to job finding on facebook's social network, J. Labor Econ. (2016) 686225.
[19] S. Geisser, Predictive Inference, 55, CRC Press, 1993.
[20] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 211–220.
[21] P.A. Grabowicz, J.J. Ramasco, E. Moro, J.M. Pujol, V.M. Eguiluz, Social features of online networks: the strength of intermediary ties in online social media, PLoS ONE 7 (1) (2012) e29358.
[22] M.S. Granovetter, The strength of weak ties, AmJ. Sociol. (1973) 1360–1380.
[23] J. Heidemann, M. Klier, F. Probst, Online social networks: a survey of a global phenomenon, Comput. Netw. 56 (18) (2012) 3866–3878.
[24] J.J. Jones, J.E. Settle, R.M. Bond, C.J. Fariss, C. Marlow, J.H. Fowler, Inferring tie strength from online directed behavior, PloS one 8 (1) (2013) e52168.
[25] D. Jurgens, T. Finethy, J. McCorriston, Y.T. Xu, D. Ruths, Geolocation prediction in twitter using social networks: a critical analysis and review of current practice, in: Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM), 2015.
[26] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI, 14, 1995, pp. 1137–1145.
[27] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: ICML, 97, 1997, pp. 179–186.
[28] A.Y.-c. Liu, The effect of oversampling and undersampling on classifying imbalanced text datasets, 2004. Ph.D. thesis, Citeseer.
[29] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. Part b: Cybern. 39 (2) (2009) 539–550.
[30] Y. Liu, J. Venkatanathan, J. Goncalves, E. Karapanos, V. Kostakos, Modeling what friendship patterns on facebook reveal about personality and social capital, ACM Trans. Comput. Human Interact. 21 (3) (2014) 1–20.
[31] P. Luarn, Y.-P. Chiu, Key variables to predict tie strength on social network sites, Internet Res. 25 (2) (2015) 218–238.
[32] D.H. Mansson, S.A. Myers, An initial examination of college students' expressions of affection through facebook, Southern Commun. J. 76 (2) (2011) 155–168.
[33] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Ann. Rev. Sociol. (2001) 415–444.
[34] A. Mislove, B. Viswanath, K.P. Gummadi, P. Druschel, You are who you know: inferring user profiles in online social networks, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, ACM, 2010, pp. 251–260.
[35] J.L. Moreno, Who Shall Survive?: A New Approach to the Problem of Human Interrelations, Beacon House Inc., 1953.
[36] F. Moscovici, Desenvolvimento Interpessoal: Treinamento Em Grupo, JosÉ Olympio, 1996.
[37] H. Oh, G. Labianca, M.-H. Chung, A multilevel model of group social capital, Acad. Manag. Rev. 31 (3) (2006) 569–582.
[38] A. Portes, Social capital: its origins and applications in modern sociology, in: Eric L. Leser (Ed.), Knowledge and Social Capital, Butterworth-Heinemann, Boston, 2000, pp. 43–67.
[39] M.J. Salganik, K.E.C. Levy, Wiki surveys: open and quantifiable social data collection, PLOS ONE 10 (5) (2015) e0123483.
[40] T. Silva, P.V. De Melo, J. Almeida, A. Loureiro, Large-scale study of city dynamics and urban social behavior using participatory sensing, Wireless Commun. IEEE 21 (1) (2014) 42–51.
[41] T. Sinha, Together we stand, together we fall, together we win: dynamic team formation in massive open online courses, in: IEEE Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), 2014, pp. 107–112.

[42] R.T. Sparrowe, R.C. Liden, S.J. Wayne, M.L. Kraimer, Social networks and the performance of individuals and groups, Acad, Manag. J. 44 (2) (2001) 316–325.
[43] J. Townsend, Theoretical analysis of an alphabetic confusion matrix, Percept. Psychophys. 9 (1) (1971) 40–50.
[44] P.O.V. de 0. Melo, A.C. Viana, M. Fiore, K. Jaffrès-Runser, F.L. Mouël, A.A. Loureiro, L. Addepalli, C. Guangshuo, RECAST: telling apart social and random relationships in dynamic networks, Perfor. Eval. 87 (2015).
[45] M. Wen, D. Yang, C.P. Rosé, in: Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22–26, 2015. Proceedings, chap. Virtual Teams in Massive Open Online Courses, Springer International Publishing, Cham, 2015, pp. 820–824, 2015.
[46] H. Wi, S. Oh, J. Mun, M. Jung, A team formation model based on knowledge and collaboration, Expert Syst. Appl. 36 (5) (2009) 9121–9134.
[47] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 981–990.
[48] W. Xie, C. Li, F. Zhu, E.-P. Lim, X. Gong, When a friend in twitter is a friend in life, in: Proceedings of the 3rd Annual ACM Web Science Conference, ACM, 2012, pp. 344–347.
[49] Y. Yang, An evaluation of statistical approaches to text categorization, Inf. Retr. 1 (1–2) (1999) 69–90.
[50] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: ICML, 97, 1997, pp. 412–420.