Working with Friends: Unveiling Working Affinity Features from Facebook Data

Douglas Castilho* **Pedro O.S. Vaz de Melo*** **Daniele Quercia†** **Fabrício Benevenuto***

*CS Dept., Federal University of Minas Gerais (UFMG), Brazil
†Yahoo Labs, Spain

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 get along (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 a group of university students to tell us with whom they wish to work, gather their online Facebook data, and test the predictors of team formation. We find that self-organized selection of team members does not strongly depend on past grades but rather on Facebook-derived proxies for tie strength, popularity, and homophily. These results have important theoretical implications for the team formation literature, and practical implications for online educational platforms.

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/avoiding someone for a given collaborative task?* Without much thought, one could answer this fundamental question by saying that the proficiency (or the skill) of a person in carrying out a specific task determines whether (s)he will be asked to team up. Although we agree that proficiency definitely plays an important role in the decision, we again ask: is proficiency the only factor at play? If not, is the proficiency even the most important factor?

In this paper, we take the first step towards answering those questions. In addition to proficiency, this work suggests that social behavior too might affect the choice of our collaborators. Testing such an hypothesis does not necessarily require to administer time-consuming surveys. With the growing popularity of the Internet and their applications, individuals' interactions are tracked on online social networks such as Facebook and Google+. Analyzing how people behave in those virtual social environments might well result into predictive behavioral features.

To verify to which extent online social behavior impacts team formation, we conduct an experiment with a class of undergraduate students. First, we administer a sociometric test to the class: in it, the students were asked whether they would like to work with every other student in the class. Then, using a Facebook application, we gathered data containing a number of social features about the students' profiles and their interactions. Our scenario is realistic, not least these students often have to team up for class projects. We posit that the process of choosing team mates involves a complex mix of social attributes and knowledge skills that produces teams that are likely to be successful and pleasant to work in. That is because, to put it simply, students might choose 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 students' grades as proxies for skills, we find that the most skilled students were not always preferred. Then, we further investigate a number of features extracted from the Facebook data: the strengths of a student's friendships, his/her popularity on Facebook, whether (s)he is extrovert, and his/her similarity with other students. We find that eight Facebook-derived features are more informative than grades to predict team formation.

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 other than proficiency that increase the likelihood of being selected. Also, for the team formation problem, our findings show that online social network data effectively suggests whether two individuals wish to work together or not. Also, our findings might support future online applications, such as team recommender systems on collaborative platforms.

Related Work

We review previous studies on team formation, social capital, and online *vs.* offline behavior. These studies will help us to identify the Facebook-derived features that are expected to be associated with team formation.

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.* (Wi et al. 2009), for example, modeled this problem as an integer programming problem to find an optimal match between individuals and requirements. Agustín-Blas *et al.* (Agustín-Blas et al. 2011), 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, researchers have proposed to augment existing approaches with members' temperament (Fitzpatrick and Askin 2005) and with interpersonal attributes (Chen and Lin 2004). There does not seem to be any work on team formation that proposes to augment those traditional approaches with online features derived from social media sites, as our work aims to do.

Social Capital

The term social capital has been used in a variety of contexts. 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 (Portes 2000; Easley and Kleinberg 2010). 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 (Granovetter 1973). More recently, it has been also associated with group effectiveness (Oh, Labianca, and Chung 2006).

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.* (Jones et al. 2013) administered a survey to Facebook users: they asked those individuals to name their best friends. They then related this survey data with the number of public and inbox messages among those individuals and corresponding best friends on Facebook. They showed that public communication is as informative as inbox messages are to infer tie strength. Xiang *et al.* (Xiang, Neville, and Rogati 2010) proposed a model for predicting tie strength from Facebook interactions and number of common friends. Xie *et al.* (Xie et al. 2012) studied the behavioral features associated with Twitter users who happen to be classmates or friends in real life. Manson *et al.* (Mansson and Myers 2011) analyzed how college students express affection to their close friends on Facebook, and identified 30 main ways to express affection.

Complementary to the above efforts, our work considers a novel scenario in which online social data can be useful.

Methodology and Data Set

In order to address the questions we posed, we prepared a very particular experimental scenario. First, we selected a classroom of undergraduate students of an anonymous university of an anonymous country. Then, through a sociometric test (Moreno 1953), we asked how each student of this classroom feels about working with every other student of this 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 data sets 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.

The Sociometric Test

Sociometry is a quantitative method for measuring social relationships (Moreno 1953). It was developed by psychotherapist Jacob L. Moreno in his studies of the relationship between social structures and psychological well-being. 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 (Bustos 1979). 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 a classroom of 31 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 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 relatioship 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 the complete sociogram, that consists of 930 answers among the 31 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 Figure 1 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, there are students who received and gave a lot of positive edges (e.g. student 2) and also students who are indifferent for and toward most of the class (e.g. student 24). Moreover, there are students that are not negative toward anyone (e.g. student 11) and also a student who received a negative answer by almost half of the class (student 28). Finally, note student 1. He/she is the most desirable work partner for all other students (most positive incoming edges), and also the most particular about who to work with (most negative outgoing edges). It is not hard to find people like this one in collaborative environments, but as we see here, there is not a clear rule to dictate the decisions made by the students. Next, we investigate at which extent the students grades' are able explain these decisions.

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 Figure 2 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 on the answers given 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. Observe in Table 1 that there is a significant correlation and low p-value between $indeg^+$and the students' grades. From this, we can conclude that proficient students attract positive answers in the survey,

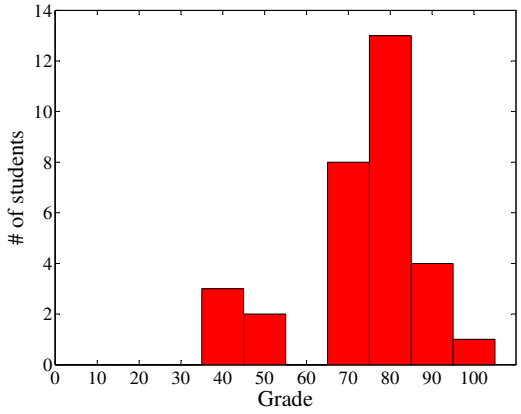

Figure 2: Grade's histogram.

i.e., students who choose to work with her/him. However, observing the other correlations, which are not signifcant, and p-values, which are high, we can conclude that a student's grade does 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.

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

Degree	Spearman Coeficient	p-value
$indeg^+$	**0.4727**	0.0073
$indeg^-$	-0.2543	0.1674
$indeg^0$	-0.3433	0.0586
$outdeg^+$	-0.0363	0.8461
$outdeg^-$	-0.0471	0.8014
$outdeg^0$	-0.0373	0.8421

It is worth mentioning that although the students may not be fully aware of the other students' grades, we strongly believe that they are good indicators for the perception of the proficiency a student has from the other students of the class. We believe that for three main reasons. First, the grades usually reflect the behavior of the student in class, i.e., good students tend to participate, to help others in excercises, to deliver tasks in time etc. These behaviors (and the opposite) are perceived by the students. Second, the students of this particular class know each other for years, i.e., they have a fair idea of which student is likely to have a good or bad performance in class. Third, the grades are usually shared among the students so they can compare their scores. If they not know the grades of everyone, it is very likely that they know which are the students with the highest (or lowest) grades.

Gathering Facebook Data

Since the grades cannot explain all the decisions, our conjecture is that some of the answers can also be explained by

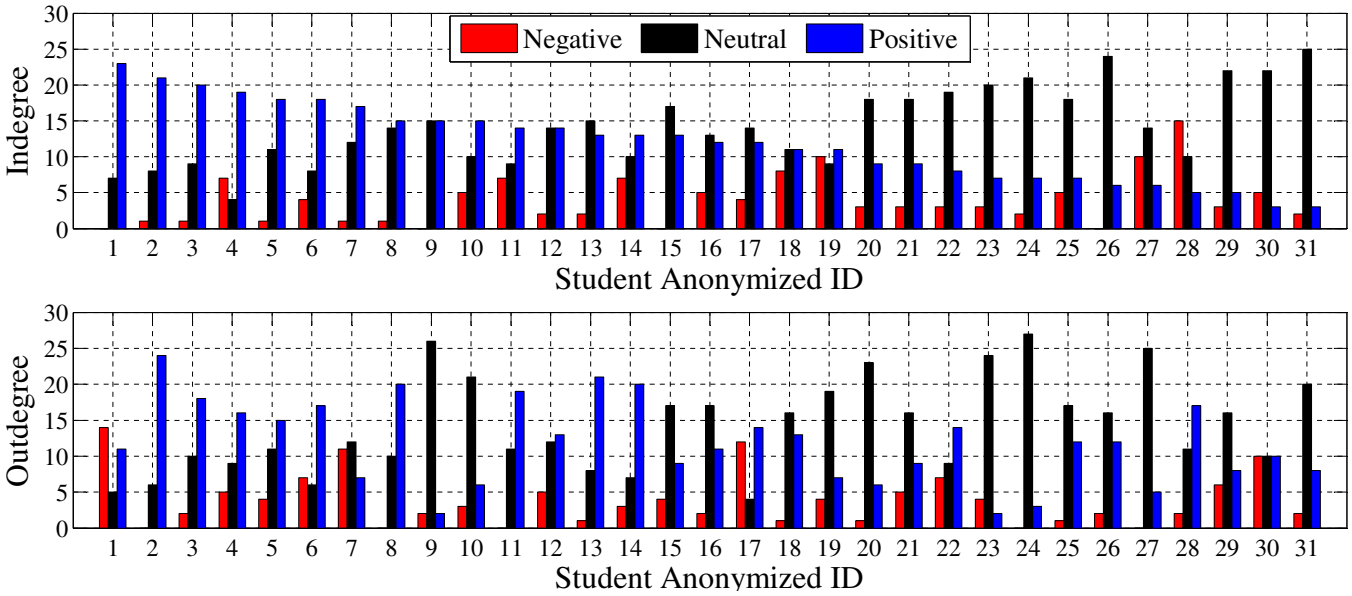

Figure 1: Individual scores attributed and received by each participant of the sociometric test.

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 an answer can 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 public 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 Table 2, all grouped by the sign of the edge. First, observe that the occurrence of friendships on neutral edges is significantly lower than on positive and negative edges. Moreover, it is curious to see that the average number of comments on shared links among negative edges is greater than on positive and neutral. Nevertheless, as expected, we see that the average number of inbox messages exchanged on positive edges is significantly higher than on neutral and negative edges. Finally, observe 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.

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 a class of undergratuate 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 a class of 31 students, the objects of study here are the relationships among these students, which correspond to 930 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 data set consists of only 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 concil 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.

Social Features

Preliminaries

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

Table 2: Features collected from Facebook, grouped by the sign of the edge.

Features	Positive Ocurrence	Ocurrence per Positive Edge	Negative Ocurrence	Ocurrence per Negative Edge	Neutral Ocurrence	Ocurrence per Neutral Edge
Number of friends	168	0.46	48	0.40	86	0.20
Number of inbox messages	248141	672.47	13127	109.39	70829	166.06
Number of tags	47	0.13	9	0.08	5	0.01
Comments on photos	496	1.34	41	0.34	166	0.38
Comments on links	255	0.69	171	1.43	339	0.77
Comments on status updates	439	1.19	75	0.63	554	1.26
Comments on albums	4	0.01	0	0	0	0
Films in common	844	2.29	210	1.75	812	1.84
Groups in commom	2846	7.71	833	6.94	2555	5.79
Interests in commom	16	0.04	8	0.07	8	0.02
Musics in commom	681	1.85	256	2.13	935	2.12
Likes in photos	84	0.23	32	0.27	54	0.12
Likes in links	63	0.17	16	0.13	24	0.05
Likes in status updates	46	0.12	11	0.09	4	0.01

in the survey. Thus, if one desires to construct a model to predict the answers given in the questionnaire, which features he/she should use?

Thus, in this section we describe several social features that are able to influence the decision of selecting a person to collaborate, i.e., features that could be incorporated to a model for predicting the sign of the edges in $G_S(V, E_S)$. These features are directly extracted from the Facebook data 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 social features we propose into two groups:

- **G1: Actor attributes**, which characterize the students and
- **G2: Link attributes**, which characterize 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 the attributes in **G2**, we verify and quantify the influence by computing 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.

G1: Actor Attributes

Popularity Here we investigate if popular students in class tend 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., $popularity_1(i) = degree(i) \in G_F$. 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.

Table 3: Impact of the popularity in the answers given in the questionnaire. 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
$indeg^+$	**0.46**	0.009	**0.49**	0.004
$indeg^-$	**0.36**	0.04	0.18	0.32
$indeg^0$	**-0.74**	0.00001	**-0.66**	0.00003
$outdeg^+$	**0.58**	0.0007	**0.37**	0.03
$outdeg^-$	0.12	0.52	0.21	0.24
$outdeg^0$	**-0.64**	0.0001	**-0.46**	0.008

In Table 3 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 that both $popularity_1$ and $popularity_2$ metrics are significantly correlated with the degree of the students for several signs and in both directions. 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 toward them. Moreover, since the $outdeg^0$correlation is significant for both metrics, 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. This shows that popular students are more well known by the class, so it is easier to make a strong point (negative or positive decision) about them. By analyzing $outdeg^+$, it is possible to infer that, curiously, popular students tend to vote positive more.

Extroversion Another feature that may impact in the students' decisions is their level of extroversion. Extrovert people tend to enjoy human interactions and to be enthusiastic, communicative, assertive, and gregarious (Eysenck 1970). There are well known ways of measuring extroversion (Rocklin and Revelle 1981), but since they rely on questionnaires and sophisticated tests, we use our Facebook data to infer how extrovert a student is. Here, we infer students' extroversion based on the number of public interactions they perform in other students' 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 (Ashton 2002). We define two ways to measure if a student is extrovert: the metric $extrovertion_1(i)$ 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 $extrovertion_2(i)$ as the number distinct students to which the student i posted public activities.

We show in Table 4 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 $extrovertion_1(i)$ metric, we can note low correlations and high p-values. However, observe that there is a significant correlation between the $extrovertion_1(i)$ metric and the $indeg^-$, which may indicate that the more an individual posts on others' walls, less other students want to work with her. This suggests that students who excessively post public comments on Facebook may also be intrusive, generating negative reactions from others. Another conjecture 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 $extrovertion_2(i)$ metric, observe there are significant negative correlations for the $indeg^0$and $outdeg^0$. This suggests that students who are not publicly active on Facebook are usually not well known by the others, generally attracting and generating neutral reactions. Moreover, it shows that students who post public comments on a large number of Facebook pages tend attract either positive or negative reactions, mostly positive, since the correlation is significantly positive with the $indeg^+$.

Table 4: Impact of outgoiness metrics in the choice of partners on collaborative activities.

	$extrovertion_1$	p-value	$extrovertion_2$	p-value
$indeg^+$	0.035	0.85	**0.516**	0.002
$indeg^-$	**0.417**	0.01	0.230	0.212
$indeg^0$	-0.266	0.14	**-0.726**	0.0003
$outdeg^+$	0.093	0.61	0.386	0.031
$outdeg^-$	0.225	0.22	0.262	0.153
$outdeg^0$	-0.225	0.22	**-0.521**	0.002

G2: Link Attributes

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 (Gilbert and Karahalios 2009). In this case, we use the same coefficients as shown in (Gilbert and Karahalios 2009) and we considered only the data we have available: *Structural Variables* are common musics, common groups, common interests, common movies and common friends; *Intensive Variables* are comments on photos, comments on link, comments on status update, comments on album, like in photos, like in links, like in status update and inbox messages; and *Intimacy Variables* are represented only by tag on photos and status update. 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 Figure 3 we show the CDFs for the first three tie strength metrics grouped by their sign. Observe that the $tieStrength_1$ could not distinguish very well the distribution of the three curves, but we can infer some results such as neutral distribution have about 50% of the edges with lower than 100 conversations, while for negative and positive that number is 30%. The $tieStrength_2$ is able to better distinguish the three distributions, and we can clearly see that almost 96% of the neutral distribution has zero public interaction. Moreover, for the positive and negative distributions, this number is also quite high, with values of approximately 71% and 80% respectively. Analyzing the $tieStrength_3$, we see that all three distributions have similar behavior, with the neutral being reasonably far apart from the others. Approximately 70% of the neutral edge have $tieStrength_3$ values smaller than a strength value 1, whereas for the positive and negative distributions these values represent approximately 18% of the edges. Thus, we can conclude that these metrics of tie strength may explain the behavior of the positive, negative and neutral attributes, because the behavior is different for the different classes of edges.

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%, while for the positive edges, the proportion is 46%. These

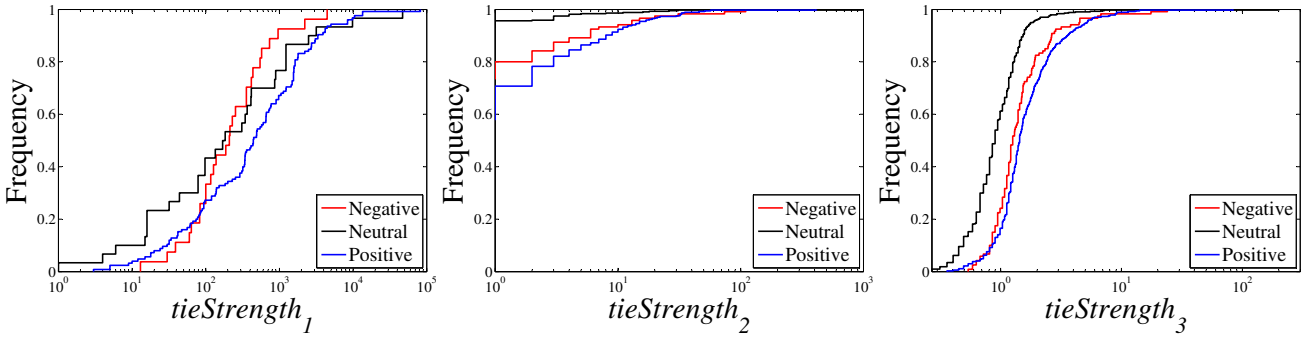

Figure 3: CDFs for the tie strength metrics grouped by their sign.

values are significantly higher than the one for the neutral edges, that is 20%. This indicates that $tieStrength_4$ metric has a potential to differentiate neutral edges from positive and negative ones.

Homophily The homophily is the tendency of individuals to associate and bond with similar others (McPherson, Smith-Lovin, and Cook 2001). Individuals in homophilic relationships share common characteristics. To investigate the homophily in our context, we define two 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 that we use the information about the movies and Facebook groups that two students have in common. 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.

In Figure 4 we show the CDFs for the two homophily metrics grouped by their sign. Observe that from $similarity_1$ we can clearly differentiate the neutral distribution of the other two. This shows that at the level of the network structure, neutral relationships have very different behavior from the positive and negative relationships. Through this metric is impossible to separate the negative from the positive relationships because they have similar behavior. However the metric $similarity_2$ distributions of the three relationships have very similar behavior, making it difficult their separation.

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 previous $similarity_2$ because it belongs to the network structure, and not an information that students share in Facebook. This feature only represents the number of common friends.

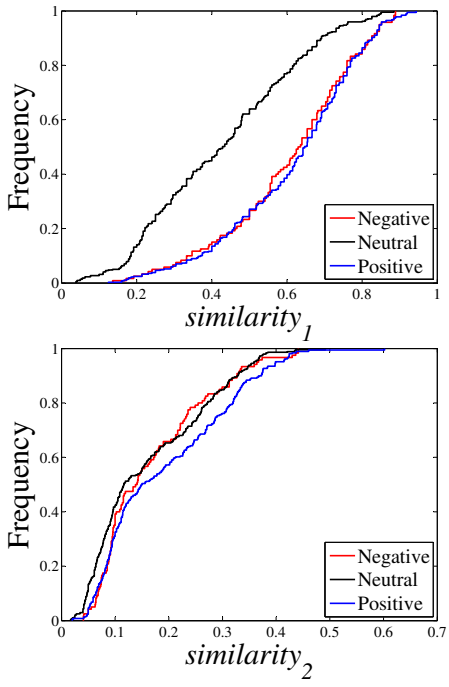

Figure 4: CDFs for the similarity features grouped by their sign.

The Most Important Factors

After having analyzed our Facebook-derived features separately, we now study how they collectively explain tie formation within a team, and, more importantly, which ones are more predictive than others. To this end, we resort to two

Table 5: A resume of the features analyzed in this study

Feature	***Category***	***Description***
$Grade$	Proficiency	Grades of the students
$popularity_1$	Popularity	Number of friends the student have on Facebook
$popularity_2$	Popularity	Number of distinct students who posted activities in the student's Facebook page
$extrovertion_1$	Extroversion	Number of public interactions that the student published in others' Facebook pages
$extrovertion_2$	Extroversion	Number of distinct students to which the student posted activities
$tieStrength_1$	Tie Strength	Total number of private inbox messages exchanged between the two students
$tieStrength_2$	Tie Strength	Total number of public interactions exchanged between the two students
$tieStrength_3$	Tie Strength	Tie strength metric proposed by (Gilbert and Karahalios 2009)
$tieStrength_4$	Tie Strength	Binary metric, 1 if the students are friends on Facebook and 0 otherwise
$similarity_1$	Similarity	The similarity between the two students' neighbors vectors
$similarity_2$	Similarity	The similarity between the vectors of movies and groups the two students have on Facebook
$similarity_3$	Similarity	Number of common friends that the two students share on Facebook

measures: the *Information Gain* and χ^2 (Chi Squared) coefficients (Yang and Pedersen 1997). 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 tuples 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 6. We find that the two measures consider the very same features to be relevant. That is because the Spearman rank correlation coefficient between the ranks generated by the two measures is as high as 0.9810 with p-value $4.2894 * 10^{-12}$. The most important result is that class grades are not that important: we need to go down the list at the 10th and 13th positions to find them. By contrast, the most predictive feature is the proxy for tie strength that has been tested most extensively in the literature (Gilbert and Karahalios 2009), and that speaks to the external validity of our results. Also, social features such as pairwise similarity between users are more predictive than grades.

Table 6: Ranking of most important attributes, presented by the IG (*Information Gain*) Ranking and the χ^2 (*Chi-Squared*) Ranking

Description	***IG Rank***	***IG Value***	χ^2 ***Rank***	χ^2 ***Value***
$tieStrength_3$	1	0.194	1	226.01
$tieStrength_4$	2	0.181	2	220.00
$similarity_1$	3	0.151	3	184.51
$similarity_3$	4	0.150	4	181.10
$tieStrength_2$	5	0.098	5	120.01
$popularity_1(source)$	6	0.084	8	100.93
$extrovertion_1(target)$	7	0.084	6	116.81
$tieStrength_1$	8	0.083	9	98.95
$popularity_2(target)$	9	0.079	10	96.17
$Grade(target)$	10	0.075	7	104.31
$extrovertion_2(target)$	11	0.073	11	91.05
$extrovertion_2(source)$	12	0.069	13	82.56
$Grade(source)$	13	0.065	12	89.44
$popularity_1(target)$	14	0.048	14	62.14
$extrovertion_1(source)$	15	0.040	15	46.70
$popularity_2(source)$	16	0.035	16	44.39
$similarity_2$	17	0.022	17	27.30

Conclusion

Compared to class grades, Facebook-derived features are more predictive of whom students wish to work with. The most important of those features is Gilbert's proxy for tie strength (Gilbert and Karahalios 2009), suggesting the importance of bonding (as opposed to bridging) social capital in team formation (Burke, Kraut, and Marlow 2011): as one expects, trust and social embeddedness (rather than presence of weak ties) are associated with willingness to team up. These results have established, for the first time, the relationship between offline team formation and online interactions. 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. These results are also of practical importance. For example, take education sites such as Coursera[1] partnering with top universities to offer free courses online. Given the large number of individuals in the world such sites serve, one effective way for them to team up students 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

[1] `https://www.coursera.org/about`

to explore cross-cultural effects. After that, a real application that recommends teams out of Facebook accounts is in order.

Acknowledgments

This work was funded by author's individual grants from CAPES, CNPq, and Fapemig.

References

Agustín-Blas, L. E.; Salcedo-Sanz, S.; Ortiz-García, E. G.; Portilla-Figueras, A.; Pérez-Bellido, Á. M.; and Jiménez-Fernández, S. 2011. Team formation based on group technology: A hybrid grouping genetic algorithm approach. *Computers & Operations Research* 38(2):484–495.

Ashton, Michael C.; Lee, K. P. S. V. 2002. What is the central feature of extraversion? social attention versus reward sensitivity. *Journal of Personality and Social Psychology* 83(1).

Burke, M.; Kraut, R.; and Marlow, C. 2011. Social capital on facebook: Differentiating uses and users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 571–580. ACM.

Bustos, D. M. 1979. *The Sociometric Testing: fundamentals, techniques and applications*. Brasilense Publisher.

Chen, S.-J., and Lin, L. 2004. Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *Engineering Management, IEEE Transactions on* 51(2):111–124.

Easley, D., and Kleinberg, J. 2010. *Networks, crowds, and markets*, volume 8. Cambridge Univ Press.

Eysenck, H. 1970. *Readings in Extraversion-introversion: Theoretical and methodological issues*. Readings in Extraversion-introversion. Staples Press.

Fitzpatrick, E. L., and Askin, R. G. 2005. Forming effective worker teams with multi-functional skill requirements. *Computers & Industrial Engineering* 48(3):593–608.

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 211–220. ACM.

Granovetter, M. S. 1973. The strength of weak ties. *American journal of sociology* 1360–1380.

Jones, J. J.; Settle, J. E.; Bond, R. M.; Fariss, C. J.; Marlow, C.; and Fowler, J. H. 2013. Inferring tie strength from online directed behavior. *PloS one* 8(1):e52168.

Mansson, D. H., and Myers, S. A. 2011. An initial examination of college students' expressions of affection through facebook. *Southern Communication Journal* 76(2):155–168.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 415–444.

Moreno, J. L. 1953. *Who shall survive?: A new approach to the problem of human interrelations*. Beacon House Inc.

Oh, H.; Labianca, G.; and Chung, M.-H. 2006. A multilevel model of group social capital. *Academy of Management Review* 31(3):569–582.

Portes, A. 2000. Social capital: Its origins and applications in modern sociology. *LESSER, Eric L. Knowledge and Social Capital. Boston: Butterworth-Heinemann* 43–67.

Rocklin, T., and Revelle, W. 1981. The measurement of extroversion: A comparison of the eysenck personality inventory and the eysenck personality questionnaire. *British Journal of Social Psychology* 20(4):279–284.

Wi, H.; Oh, S.; Mun, J.; and Jung, M. 2009. A team formation model based on knowledge and collaboration. *Expert Systems with Applications* 36(5):9121–9134.

Xiang, R.; Neville, J.; and Rogati, M. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web*, 981–990. ACM.

Xie, W.; Li, C.; Zhu, F.; Lim, E.-P.; and Gong, X. 2012. When a friend in twitter is a friend in life. In *Proceedings of the 3rd Annual ACM Web Science Conference*, 344–347. ACM.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, 412–420.