A Project Teams Creation Based on Communities Detection

Mikhail Semenov[1], Lev Bulygin[1], Elena Koroleva[1], and Dilmurat Tursunov[2]

[1] Tomsk Polytechnic University, Tomsk, Russia,
{sme, leb1, eak6}@tpu.ru,
[2] Ural State Pedagogical University, Yekaterinburg, Russia,
d_osh@rambler.ru

Abstract. The purpose of this study is to detect project teams in a group. A key point in considering group's relationships is the reciprocal influence, whereby group's members influence each other. There was conducted a survey based on reciprocal nomination method, and then a social network was constructed. Participants were first-year bachelor students of Tomsk Polytechnic University. Various social network analysis algorithms were used to cluster network in communities. The results of analysis were discussed with the teachers and students, and then detected community teams were adjusted within the key actors of group. The results of the study may be used to create project teams, which can make successful collective actions in educational projects.

Keywords: community detection, key actors, project team.

1 Introduction

A problem arising very frequently is how to identify new teams in an existing group. Community detection helps to understand the distribution of key social actors and their interrelations in the network [2]. At present, many community detection algorithms have been designed [1,3,8,13,15,19]. Much research has been conducted on social network analysis (SNA) using graph theory [10, 11, 14, 16]. One of the important results is the identification of sociometric features that characterize a network. However, it is still not clear, which algorithms are reliable and should be used in applications, because network is a single object, which cannot be simply splitted into a training and a test dataset [20].

Firstly, we need to define what is meant by a project team. A *project team* is a group of people who are able to act in concert and to achieve collectively the common goal. Team members have specific and unique roles, where the performance of each role contributes to achievement of the team's goal. In project teams members care about the success of other team's members because their own goal attainment is often inextricably bound to collective achievement [21].

This study is a part of one-year follow-up research of social network changes among the members of students group. The purpose of this study is to detect project teams in a group. We present an approach for project teams creation based on SNA methodology. This approach includes descriptive analysis, community structure identification and key actor analysis using graph theory.

The structure of the paper is following. In Section 2, we give short overview of related works. Then, in Section 3, we introduce the dataset, on which our approach has been tested. We calculate the measures of our networks, detect key social actors. In Section 4, we examined an influence of teams' interactions over time on the academic performance. A brief summary follows in Section 5.

2 Related Works

Pijl et al. [14] compared two methods for the assessment of students' friendship networks: the reciprocal nomination method and social cognitive mapping. In total, 190 participants took part in the experiment. The authors introduced types of isolated students in their study: a) a student with no reciprocated links at all (type 1), and b) a student with one reciprocated link (type 2). A cohesive subgroup defined as a group of at least three students a) who have more internal links than external links, b) are connected by some path to each of the group members and remain connected when up to 10% of the group is removed.

Rienties et al. [18] conducted a set of experiments in order to understand how students develop and maintain learning and friendship relations over time in a large classroom setting (200+ students). Students were put in 41 teams of 5 students on average. The results indicate that the instructional design might have a strong influence on how students work together in teams, how social learning and friendship interactions develop, and finally increase academic performance.

Pronin et al. [16] suggested grouping method to reorganize student groups using the SNA methodology. The problem was in reorganizing four existing groups of students into three new groups. The Girvan-Newman algorithm was used in order to create three new groups, and then authors adjusted new groups based on existed relations between students and the modularity score. This method may be used to create project teams for research classes or scientific labs.

Liu et al. [11] proposed the algorithm to measure the importance of the actors in network. This algorithm based on in-weighted degree and out-weighted degree of vertex and on considering the information of the directed edges. In contrast, D. Conway [4] introduced the method using the comparison of centrality's relative values such as eigenvector centrality and betweenness.

Lomi et al. [12] specified a model that permit estimation of the interdependent contribution of social selection and social influence to individual performance. The proposed stochastic model is based on the direct observation of connectedness between students. In their study authors focused on the effects of 75 participants on individual performance at the classroom level.

Ertem et al. [7] used SNA metrics in order to predict learning performance in terms of student's position in the network. Authors found a positive correlation between students' performance and six employed metrics: degree, eigenvector centrality, betweenness centrality, hub, authority and PageRank.

In contrast with modularity that was proposed by M. Newman [13], there was an accepted standard for the results of community detection. Yang et al. [20] found that a conductance and a triangle participation ratio could provide the best performance in characterizing the communities detection quality.

3 Experimental Evaluation

Research questions. Our one-year follow-up research is aimed to explore the community organization of the network and the influence of the structural factors of the network on academic performance of students. We formulated the following questions concerning the community organization of the network:

Q1. What is the network structure at the classroom level?

Q2. How to use the network structure to increase academic performance of students?

Dataset. In our experiments, a local social network was built on reciprocal nomination between 20 first-year bachelor students of Tomsk Polytechnic University during the fall semester of 2015. We used the direct-preference questions. The 20 students answered the four social network questions:

1. Name classmates with whom you spend free time.
2. Name classmates to whom you are applying for the information, related to academic activities.
3. Name classmates who could influence your academic performance.
4. Name classmates with whom you don't want to cooperate in framework of creative project.

The students were allowed to nominate up to four classmates. Participants were 20 first-year bachelor students who were 17 to 19 years old ($M = 18.5, SD = 0.35; 75\%\ male$). Data were collected on-line with the Google forms.

Data Preprocessing. Four square matrices A_1, A_2, A_3, and A_4 of size 20 were generated respectively on the basis of the questionnaire. In each adjacency matrix A_1, A_2, A_3 the element (i, j) is equal to 1 if row student i nominated the column student j, otherwise the element $(i, j) = 0$. In matrix A_4 the element (i, j) is equal to -1 if row student i nominated the column student j, otherwise the element $(i, j) = 0$. Then each matrix A_1, A_2, A_3 was summarized with the matrix A_4, and the binarization procedure was applied to the result: if the element (i, j) is less or equal than 0 then it gets set to 0, otherwise it is set to 1.

The social networks are represented by directed graphs $G_k = (V_k, E_k), k = 1, 2, 3$, where a set of vertices V_k includes $n = 20$ members of group, and a set of edges $E_k \subseteq V_k \times V_k$ presents the relation «reciprocal nomination» that corresponds to $k = 1, 2, 3$ questions. A key point in considering these relationships is the reciprocal influence, whereby team's members influence each other. The graph G_k is directed, i.e. every edge $(i, j) \in E_k$ links the source vertex i and the target vertex j. The direction of the edges is makes each adjacency matrix A_1, A_2, A_3 of the each directed graph G_1, G_2, G_3 non-symmetric because the source vertex defines nomination to the target vertex but not vice versa. The number of edges $m_k = |E_k| \leq n \cdot (n - 1), k = 1, 2, 3$.

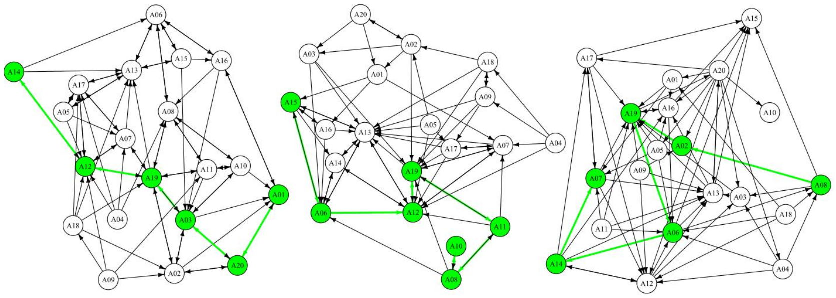

Fig. 1. Three social networks and the longest path (green)

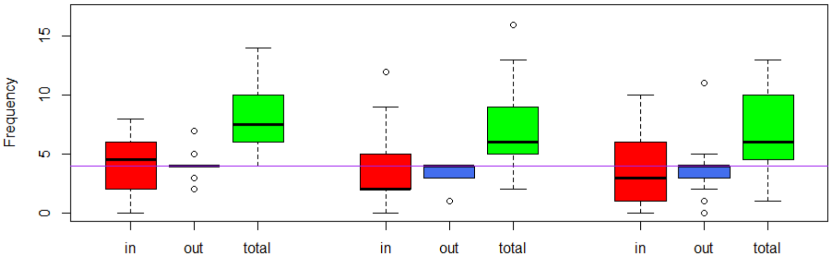

Fig. 2. Degree distribution: in-degree (red), out-degree (blue), total-degree (green) of vertices of the graphs G_1, G_2, G_3

Descriptive Network Statistics. To address research question **Q1** we calculated descriptive network statistics in order to define network structure at the classroom level. Figure 1 illustrates the original networks G_1, G_2, G_3, reflecting the structure of reciprocal nominations in the group. These networks G_1, G_2, G_3 have the identical vertex set V, $|V| = n = 20$, but different sets of edges E_1, E_2, E_3: $|E_1| = m_1 = 81$, $|E_2| = m_2 = 72$, $|E_3| = m_3 = 71$ respectively, in which one vertex represents an actor, and one edge denotes the nomination between any two actors. To each actor in the student's group labels were assigned: $A01, A02, \ldots, A20$. In our experiment, loops and multi-edges are not allowed.

It's seen that the *diameter* of graph G_2 equal to 6 because the longest path between actors $A15$, $A06$, $A12$, $A19$, $A11$, $A08$ and $A10$ takes 6 edges, the *diameter* of graphs G_1 and G_3 equal to 5 (Figure 1, green arrows), while the lengthes of *average path* between any vertex pairs are $\{2.166, 2.236, 2.190\}$ respectively. In our experiment, *density* range from 18%(0.186) to 21%(0.213) only, it can be explained by the limitation of the questionnaire, which recommended to nominate up to four actors. The *transitivity* range from 0.399 to 0.409 indicates the low level of intra-group interaction, the *reciprocity* range from 0.11 to 0.66.

A basic property of the vertices in a graph is their degree. Degree provides information on the position of actors and how they communicate. The in-degree $d^{in}(v)$ (out-degree $d^{out}(v)$) of vertex v is equal to the number of incoming (outgoing) edges. The in-, out-, total degree distribution of vertices of the graphs G_1, G_2, G_3 are shown in Figure 2. As we can see from Figure 2, the *modal interval* equal to 4, which has frequency more than any other interval.

Table 1. Descriptive Network Statistics

Network statistics	Definition	G_1	G_2	G_3	$\langle G_2 \rangle$
Average path	Shortest paths between all pairs of vertices	2.166	2.236	2.19	2.262
Transitivity	Fraction of edge pairs (i, j) and (j, k) in the graph such that (j, k) is also linked in the graph	0.399	0.409	0.400	0.449
Clustering coefficient [10]	Indicates how a vertex are embedded with neighbors	0.3763	0.3767	0.248	0.258
Reciprocity	Fraction of edges in the graph that go in both directions	0.666	0.472	0.11	0.166
Average degree	Average number of edges incident with vertices	4.05	3.6	3.55	3.6
SD in-degree	Standard deviation of in-degree	2.54	3.23	2.95	3.23
SD out-degree	Standard deviation of out-degree	0.94	0.75	2.24	0.75
Density	Ratio of the number of edges and the number of possible edges	0.213	0.189	0.186	0.189
Edges	Number of edges	81	72	71	72
Diameter	Longest path between pairs of vertices	5	6	5	5

In order to estimate clustering in our networks, we use the clustering coefficient. This coefficient is defined as the average value over all vertices v, of the vertex-specific clustering coefficients [10]

$$\mathrm{cl}(v) = \frac{(A + A^T)^3_{vv}}{2((d^{in}(v) + d^{out}(v))(d^{in}(v) + d^{out}(v) - 1) - 2(A^2)_{vv})},$$

where A is the adjacency matrix. In our case, clustering coefficient $\mathrm{cl}(v)$ is from 0.248 to 0.377.

To evaluate the significance of the network statistics (Table 1) the simulation of the random graphs was used. Based on topological properties of the each graph G_1, G_3, G_3 1000 random networks were generated to compute their average network statistics. Table 1 gives these statistics in the last column $\langle G_2 \rangle$ corresponding to the network G_2. The network G_2 has approximately the same average path length than 1000 random graphs of the same size (2.236 and 2.262 respectively), and the network G_2 has a clustering coefficient that is higher than the corresponding value of 1000 random graphs (0.3767 and 0.258 respectively).

Community Detection Algorithms. At present, many community detection algorithms have been designed [1, 3, 13, 15, 19]. In our experiments, we chose three community detection algorithms: edge betweenness algorithm, walktrap algorithm and optimal community algorithm. We used R realization of the algorithms and *igraph* software package [5].

Denote by $\mathcal{C} = \{C_1, C_2, \ldots, C_p\}$ a partition of a set of vertices V. We call $\mathcal{C}$ a *clustering* of a graph G and the C_i, which is required to be nonempty, *community*, $i = 1, 2, \ldots, p$. Following to the paper [1] $E(\mathcal{C}) = \bigcup_{i=1}^{p} E(C_i)$ is the set of *intracluster* edges, $m(\mathcal{C}) = |E(\mathcal{C})|$, and $E \setminus E(\mathcal{C})$ is the set of *intercluster* edges, $\bar{m}(\mathcal{C}) = |E \setminus E(\mathcal{C})|$.

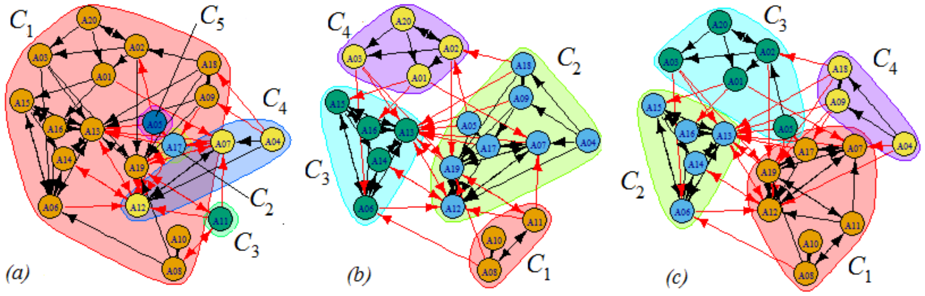

Fig. 3. Community structure identification: a) edge betweenness, b) walktrap, and c) optimal. An intracluster edge denotes a red arrow, an intercluster edge – a black arrow

Let us list major features of these algorithms. The edge betweenness algorithm [13] is based on calculating the betweenness (number of shortest paths between any two vertices which pass through this edge) of all edges in the graph and removing the edge with the largest betweenness score. This process is repeating on the resulting graph until no edges remain. A partition $\mathcal{C}$ of a set of vertices V can be computed in $O(nm^2)$ time. The walktrap algorithm is based on random walk process on a graph [15]. A walker is on a vertex and moves to a random vertex each time step. After a few steps (3–5) the walker is more likely to stay within the same community because there are only a few external edges. The walktrap algorithm uses the results of this random walk process to merge separate vertices in communities that minimizes distance from other vertices in the community. Time complexity of the walktrap algorithm is $O(mn^2)$ in the worst case. The optimal community algorithm [1] detects the community structure for a graph, by maximizing the modularity score over all possible partitions. The algorithm starts with the singleton community clustering and iteratively merges those two communities that yield a clustering with the best modularity. Time complexity of optimal community algorithm is exponential in the number of vertices $O(2^n)$.

Using algorithms mentioned above each graph G_1, G_2, and G_3 was divided into communities. Figure 3 shows results of network partitioning G_2 into clusters, which are denoted by different colors. Five communities were detected with edge betweenness algorithm, while using walktrap and optimal algorithm partitioned the graph G_2 into 4 communities. The numbers of actors in these communities are different: $V(\mathcal{C}^b) = \{14, 1, 1, 3, 1\}$, $V(\mathcal{C}^w) = \{3, 8, 5, 4\}$ and $V(\mathcal{C}^{op}) = \{7, 5, 5, 3\}$. We use subscripts $\{b, w, op\}$ to denote the clustering algorithm that was used. Figure 4 shows distribution of number communities detected with the edge betweenness algorithm, walktrap algorithm and optimal algorithm for 1000 random graphs of the same size as the network G_2. According to Figure 4, it is clear that the actual number of communities detected in the original network G_2 (4 communities for walktrap and optimal algorithms) would be considered typical from the perspective of random graphs, while using the edge betweenness algorithm partitioned 1000 random graphs into number communities from 1 to 14.

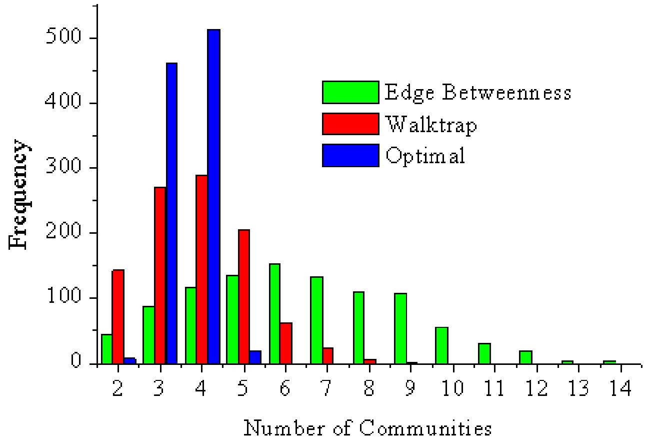

Fig. 4. Distribution of number of communities detected with the edge betweenness algorithm (green), walktrap algorithm (red), and optimal algorithm (blue) for 1000 random graphs

Since *modularity* was proposed by M. Newman [13], there was an accepted standard for the results of community detection [10, 20]

$$\text{mod }(\mathcal{C}) = (m(C) - E(m(C)))/4,$$

where $m(C) = |\{(u,v) \in E : u \in C, v \in C\}|$, $E(m(C))$ is the expected value of $m(C)$ under some model of random edge assignment.

The edge betweenness and walktrap algorithms are the hierarchical clustering algorithms, and the modularity score of the current clustering is stored after each time step. In the optimal community algorithm the highest modularity is defined after $(n-1)$ merges. The algorithms above use the modularity score to decide where to stop the splitting or merging. Another way to qualify the communities detection is to compute scoring function based on the internal connectivity, external connectivity and combination of internal and external connectivity of vertices set [8, 20]. To each network partitioning $\mathcal{C}^b$, $\mathcal{C}^w$, $\mathcal{C}^{op}$ from Figure 3 the number of intracluster edges $m(\mathcal{C})$ are given by diagonal elements and the number of intercluster edges $\bar{m}(\mathcal{C})$ between communities C_i and C_j are given by (i,j) elements, $i \neq j$, $i,j \in \{1,2,\ldots,p\}$ for various algorithms are represented by matrices:

$$E(\mathcal{C}^b) : \begin{pmatrix} \mathbf{39} & 3 & 4 & 13 & 2 \\ 3 & \mathbf{0} & 0 & 3 & 1 \\ 4 & 0 & \mathbf{0} & 2 & 0 \\ 13 & 3 & 2 & \mathbf{4} & 1 \\ 2 & 1 & 0 & 1 & \mathbf{0} \end{pmatrix}, \; E(\mathcal{C}^w) : \begin{pmatrix} \mathbf{4} & 5 & 1 & 0 \\ 5 & \mathbf{21} & 11 & 5 \\ 1 & 11 & \mathbf{15} & 4 \\ 0 & 5 & 4 & \mathbf{6} \end{pmatrix}, \; E(\mathcal{C}^{op}) : \begin{pmatrix} \mathbf{19} & 9 & 5 & 5 \\ 9 & \mathbf{15} & 5 & 2 \\ 5 & 5 & \mathbf{7} & 1 \\ 5 & 2 & 1 & \mathbf{4} \end{pmatrix}.$$

In the each matrix row and column sums belong to the number of edges incident on a given community. In the matrixes we bold the number of intercluster edges. Using data from matrixes we calculated the *conductance* [20]:

$$\text{con}(\mathcal{C}) = \bar{m}(C)/(2 \cdot m(C) + \bar{m}(C)),$$

where $\bar{m}(C)$ the number of edges on the boundary of community C, $\bar{m}(C) = |\{(u,v) \in E : u \in C, v \bar{\in} C\}|$. The conductance has a value between 0 (best score)

Table 2. Scoring Function

Partitions	G_1		G_2		G_3	
	mod ($\mathcal{C}$)	con($\mathcal{C}$)	mod ($\mathcal{C}$)	con($\mathcal{C}$)	mod ($\mathcal{C}$)	con($\mathcal{C}$)
$\mathcal{C}^b$	0.023	0.603	0.075	0.252	0.015	0.327
$\mathcal{C}^w$	0.311	0.296	0.314	0.220	0.085	0.154
$\mathcal{C}^{op}$	0.338	0.227	0.323	0.230	0.163	0.352

and 1 (worst score). Table 2 gives the value of scoring function: modularity, mod ($\mathcal{C}$) and conductance, con($\mathcal{C}$) for various partitions $\mathcal{C}^b$, $\mathcal{C}^w$, $\mathcal{C}^{op}$ and networks G_1, G_2, and G_3.

Community Structures Comparison. After getting the communities, the partitions were compared using various metrics, and results are presented in Table 3. *Normalized mutual information* (NMI) measure is based on the fact that if two partitions are similar to each other, then only a small amount of additional information is needed to infer one clustering assignment from the other [6]. The NMI measure, the *Rand index* (RI) have a value between 0 and 1, when the two partitions agree perfectly, these measure are 1 [17]. The *adjusted Rand index* (ARI) can yield negative values, and adjusted Rand index is more sensitive that the Rand index to measure agreement between two partitions [9]. As one can see from Table 3, the partitions $\mathcal{C}^w$ and $\mathcal{C}^{op}$ are similar to each other, while the partition $\mathcal{C}^b$ differs from the partitions $\mathcal{C}^w$ and $\mathcal{C}^{op}$ considerably.

Table 3. Community Structures Comparison Using Various Metrics

Pairs Partitions	NMI	RI	ARI
($\mathcal{C}^b$,$\mathcal{C}^w$)	0.335	0.500	−0.005
($\mathcal{C}^b$,$\mathcal{C}^{op}$)	0.271	0.495	−0.016
($\mathcal{C}^w$,$\mathcal{C}^{op}$)	0.717	0.816	0.494

Key actor analysis. Next to the analysis described above, we identified key social actors. We used the comparison of relative values of eigenvector centrality and betweenness centrality. The betweenness centrality gives a higher score to a vertex that sits on many shortest path of other vertex pairs, and it centrality usually refers to the access to novel information and control benefits. Eigenvector centrality gives a higher score to a vertex if it connects to many high score vertices. We calculated the linear regression model, Figure 5 shows a scatter plot of Eigenvector centrality as a function of Betweenness centrality. The equation for the line in Figure 5 is $y = 0.0102x + 0.2833$ (red line), this linear model was significant ($F = 14.118$, p-level $= 0.0014 < 0.05$).

Figure 5 shows each vertex's relative value of eigenvector centrality and betweeness, scaled by the value of the regression residuals, labels of actors scaled by the absolute value of residuals. D. Conway [4] has found that people with low eigenvector centrality but high betweenness centrality are important *gate keepers* between teams (actors $A08$, $A11$), while people (actors $A14$, $A17$) with high eigenvector centrality but low betweenness centrality has direct contact to *important* people (actors $A12$, $A13$, and $A19$).

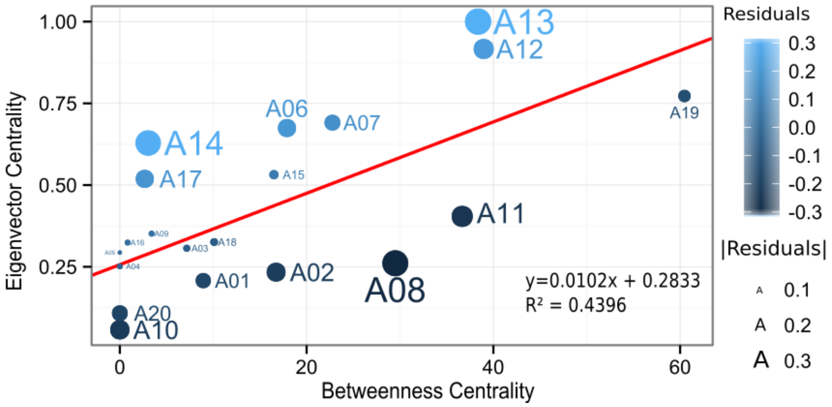

Fig. 5. Key actor analysis

The results of the community structure identification (Figure 3) and key actor analysis (Figure 5) were discussed with the teachers and students. A frequently mentioned disadvantage of the community structure based on the edge betweenness algorithm was providing more unbalanced partition than walktrap and optimal algorithms. The statistical characteristics of community sizes and their variations are: $M = 4.0, SD = 5.65$ (edge betweenness algorithm), $M = 5.0, SD = 2.16$ (walktrap), $M = 5.0, SD = 1.36$ (optimal algorithms).

Teachers recommended to separate the key actors $A12$ and $A19$ into the different teams. According to the community structure identification the following options are available: to move a key actor from C_2^w to C_1^w or from C_1^{op} to C_4^{op} (Figure 3). The conducted SNA modelling leads to the next result: the compromise between the modularity and the conductance is to move the actor $A19$ from C_2^w to C_1^w. In this case, as we expected, the modularity score decreased from 0.314 to 0.304, while the conductance increased from 0.22 to 0.263.

4 Effect of Network Structure on Academic Performance

As a result, four project teams were formed after the midterm (9th week of the fall semester 2015): $T_1 = \{A08, A10, A11, A19\}$, $T_2 = \{A04, A05, A07, A09, A12, A17, A18\}$, $T_3 = \{A06, A13, A14, A15, A16\}$, and $T_4 = \{A01, A02, A03, A20\}$. To address research question **Q2** we examined an influence of teams' interactions over time on the academic performance.

In our experiment, during the second period of the term (from 10th to 18th weeks) students from the experimental group (g_{exp}, $n_1 = 20$) additionally were meeting with the peers of their team once a week during a 2 hours tutorial in the classroom, and they worked on a project. It is notable that students cannot change teams during the experiment. A control group (g_{cont}, $n_2 = 21$) is not splitted into project teams and does not have any additional meetings.

Measurement of academic performance on the experimental group and the control group was collected at two time points: a) at the midterm ($p_1 = 9$ weeks),

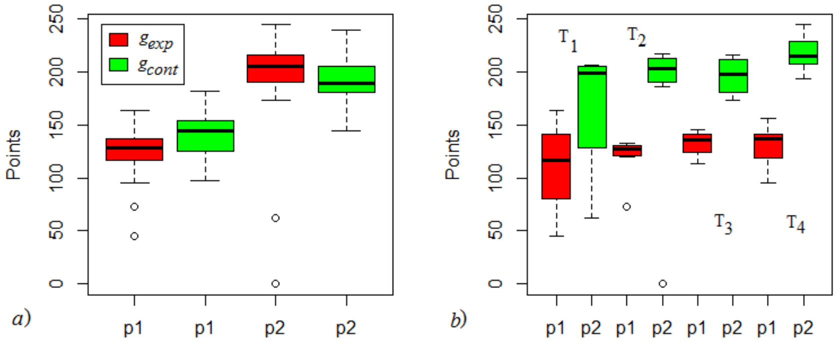

Fig. 6. Points distribution: a) the experimental (g_{exp}) and control (g_{cont}) groups, b) project teams (T_1, T_2, T_3, T_4) at time points (p_1, p_2)

b) at the end of term ($p_2 = 9$ weeks). We received this information directly from the Education Office. In both groups the each student could earn 480 points during the fall semester of 2015 (50% at the midterm, 50% at the end of the term). In our experimental group, average overall performance at the midterm was $M = 122.7$ points, $SD = 26.96$, range $R = [45, 163]$.In the control group, average overall performance at the midterm was $M = 140.09$ points, $SD = 22.14$, range $R = [97, 182]$.

Figure 6 gives descriptive statistics of academic performance (in points) to the experimental and control groups as well as the each team. There are outlier points in the dataset, these observations correspond to the points of the actor $A04$, who dropped out from the experiment after 11th week.

Firstly, the Shapiro-Wilk test was applied to check whether the distribution of dependent variable came from a normally distributed population. The dependent variable is the number of points at time points (at the midterm and at the end of term). At .05 significance level the null hypothesis was rejected and there is evidence that the distribution of points in the experimental group ($W = 0.87$, p-value $= 0.017 < 0.05$) is not from a normally distributed population, while the distribution of points in the control group ($W = 0.97$, p-value $= 0.73 > 0.05$) is from a normally distributed population. Hence, we decided to use the non-parametric statistics. It is clear from Figure 6 that the median of the experimental group (g_{exp}) at the first time point (p_1) is less than the median of the control group (g_{cont}) and vice versa at the second time point (p_2). At the first data point (p_1) in the experimental group the median was 128.5 points, while in the control group the median was 144 points, at the second data point (p_2) the experimental median was 205, the control median was 189.

We applied the Mann-Whitney-Wilcoxon criteria to test of the null hypothesis that students from the control group tend to have the larger value of academic performance (in points) than students from the experimental group. At the .05 significance level, the null hypothesis ($U = 137.5$, p-value $= 0.058 > 0.05$) was accepted at the first time point. We conducted a randomization test of no difference in population medians (null hypothesis) against a two tailed alternative, where the difference in sample medians is the test statistic. We created 5000 randomizations of the $n_1 + n_2 = 20 + 21 = 41$ observations. The two tailed probability under the null hypothesis is p-value $= 0.0376 < 0.05$, and 95% confidence limits are -14.01 and 14.0. The obtained median difference was -15.5 points, which clearly falls outside the interval. Thus we can reject the null hypothesis, and conclude that the median points of the control group is significantly greater than the median points of the the experimental group. At the second data point (p_2) we repeated the randomization test for comparing two medians, 95% confidence limits are -16.5 and 16.5. The obtained median difference was 16.0 points, which clearly falls inside the interval. We can not reject the null hypothesis, and we can expect that the influence of teams' interactions on the individual academic performance is positive.

The next set of statistical tests was applied to the experimental group. Firstly, we need to test that none of $k = 4$ teams stochastically dominates one another. The Kruskal-Wallis test was applied to decide whether the population medians on a dependent variable are the same across all levels of a factor. The factor has four levels: $1 = T_1$, $2 = T_2$, $3 = T_3$, and $4 = T_4$. The null hypothesis is that the medians are equal across the teams. At .05 significance level, we conclude that the medians are equal across the teams ($\chi^2 = 2.0$, p-value $= 0.57 > 0.05$) at the midterm. Secondly, for the comparison across repeated measures at the midterm and the end of the term the Friedman's test was used. It is used to test for differences between the two snapshot data when the dependent variable being measured is ordinal (ranks in our case). The null hypothesis that the distributions are the same across repeated measures was rejected ($\chi^2 = 16.2$, p-value $= 5.7 \cdot 10^{-5} < 0.05$). Hence, the distributions across repeated measures are different. There is evidence of the influence of teams' interactions on the individual academic performance.

5 Summary

Project teams are detected using various social network analysis algorithms. The key actor analysis allows us to identify individuals who have the strongest influence on other members of the group. The results of communities detection can be used in the educational process but require discussions with teachers and students. According to compromise between the SNA results and semantic recommendations of teachers and students, we have chosen the basic algorithm, and project teams were created. We found evidence of peer effects on academic performance. In the experimental group as a whole, as well as in the detected teams the academic performance increased in comparison with the control group.

The further research of our longitudinal study can be continued in the following directions. At first, it is community detection in terms of motifs, i.e. dyads, triads (two or three students are only connected to each other) as a subgraph with a fixed number of vertices and with a given topology. Such description allows us to identify complexity levels of a project to each team and different assessment methods of team performance. At second, it is an application of qualitative analysis of relations inside and outside project teams and assessment of potential predictive factors of relations.

References

1. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Transactions on Knowledge and Data Engineering 20(2), 172–187 (2008)
2. Carrington, P., Scott, J., Wasserman, S.: Models and Methods in Social Network Analysis. Cambridge University Press, New York (2004)
3. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks (2004), `http://www.arxiv.org/abs/cond-mat/0408187`
4. Conway, D.: Social network analysis in R (2009), `http://files.meetup.com/1406240/sna_in_R.pdf`
5. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal Complex Systems, 1695 (2006), `http://igraph.org`
6. Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification (2005), `http://arxiv.org/abs/cond-mat/0505245v2`
7. Ertem, Z., Veremyev, A., Butenko, S.: Detecting large cohesive subgroups with high clustering coefficients in social networks. Social Networks 46, 1–10 (2016)
8. Fortunato, S.: Community detection in graphs. Physics Reports 486, 75–174 (2010)
9. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2(1), 193–218 (1985)
10. Kolaczyk, E., Csardi, G.: Statistical Analysis of Network Data with R. Springer, New York (2014)
11. Liu, H., Qin, X., Yun, H.: A community detecting algorithm in directed weighted networks. Series Lecture Notes in Electrical Engineering (98), 11–17 (2011)
12. Lomi, A., Snijders, T., Steglich, C., Torly, V.: Why are some more peer than others? Evidence from a longitudinal study of social networks and individual academic performance. Social Science Research 40(6), 1506–1520 (2011)
13. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2), 26113 (2004)
14. Pijl, S.J., Koster, M., Hannink, A., Stratingh, A.: Friends in the classroom: a comparison between two methods for the assessment of students' friendship networks. Soc Psychol Educ (14), 475–488 (2011)
15. Pons, P., Latapy, M.: Computing communities in large networks using random walks (2005), `http://arxiv.org/abs/physics/0512106`
16. Pronin, A., Veretennik, E., Semyonov, A.: Formirovanie uchebnyh grup v universitete s pomoshju analiza socialnyh setej. Voprosy obrazovanija (3), 54–74 (2014)
17. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66, 846–850 (1971)
18. Rienties, B., Heliot, Y.F., Jindal-Snape, D.: Understanding social learning relations of international students in a large classroom using social network analysis. High Education (66), 489–504 (2013)
19. Rosvall, M., Bergstrom, C.T.: Maps of information flow reveal community structure in complex networks. PNAS 105(4), 1118–1123 (2008)
20. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS'12. Beijing, China (12-16 Aug 2012)
21. Zaccaro, S.J., Rittman, A.L., Marks, M.A.: Team leadership. The Leadership Quarterly (12), 451–483 (2001)