Statistics and Its Interface Volume 11 (2018) 191–200

Social network analysis based on canteen transaction records*

Yuewen Liu†, Ke Xu, Xiangyu Chang‡,§,
Dehai Di, and Wei Huang

College students' social network could influence their academic performance, attrition, and even mental health. Unfortunately, it is not easy to collect college students' social network information. The existing methods (e.g., collecting data via survey, online social network web sites and phone call records) suffer from various shortcomings. In this paper we present a case study in which the college students' social network is extracted from their canteen transaction records. In detail, we empirically collect a canteen transaction data set which has 4.5 million transaction records of 16 thousand undergraduate students during one semester (112 days). We propose a systematic method to extract the canteen social network. Based on the extracted network, we calculate some network attributes for each student, and employ regression analysis to study the relationships between students' network attributes and academic performance. The findings of this case study encourage us to build more rigorous statistical methods to extract social network from transaction records, and to examine the effects of network attributes.

AMS 2000 subject classifications: Primary 91D30.
Keywords and phrases: Social network analysis, Canteen transaction records, Academic performance.

1. INTRODUCTION

This paper presents a case study of social network analysis, where the social network is extracted from college students' canteen transaction records, and then the students' network attributes (e.g., degree, closeness centrality, betweeness centrality and clustering coefficient) are found to be significantly correlated with their academic performance.

College students' social network is critical in several perspectives. First, according to the literature, students' social network activity significantly impacts their academic performance [20, 10, 17, 27], and then influences their future career development [13]. A student with poor academic performance may encounter more difficulties in pursuing promising job positions and research opportunities [13]. Second, students' social integration is found to be strongly correlated with student retention [18, 4]. A lot of colleges suffer from high student attrition rates [18, 23]. It is reported that the four-year degree completion rates in most higher education institutions in US are only around 50% [23]. A better understanding of college students' social network could be beneficial for reducing the student attrition rate. Third, students' social network is also correlated with their mental health. Feelings of social isolation may be present and could result in suicidal feelings, thoughts and even suicide attempt [1, 24]. Exploring students' social network and identifying the possibly at-risk students as early as possible may help these students or even save their lives.

To study the correlation between college students' behavior and their social network by statistical models (e.g., the stochastic block model [22], latent space model [9] and mixture graph model [8]), we need to collect their social network data comprehensively; however, this is not easily done, especially on a large scale. There are generally three types of methods for collecting social network data in the literature. The most traditional method is survey, i.e., asking students to self-report the list of their friends [2]. The advantage of survey is that, researchers can also collect some subjective factors (e.g., personality traits) when collecting social network information. However, the survey method suffers from several limitations such as expensive and time consuming data collection process, low participate rate and self-report bias [25, 21]. Moreover, the students can only report their major relations in a relatively short term, rather than all the relations and the network dynamics in a relatively long term. The second method is to collect college students' social network from some public social network web sites, such as Facebook [14]. However, online social networks are different from the real-life social network [5, 19], thus may not be able to fully reflect students' real college life. The third method is to collect students' phone call records. As we all know, phone call records are extremely private, thus are not easy to be collected on a large scale [15].

The limitations of the existing methods drive us to conduct this case study, to explore a method to collect students' social network information utilizing some conven-

*We thank the AE and a referee for their helpful comments and suggestions which greatly improved our paper.
†Liu was partially supported by the National Natural Science Foundation of China (Project No. 71301128, 71331005, 91546119) and the China Postdoctoral Science Foundation (Project No. 2014M560795, 2015T81039).
‡Chang was partially supported by the National Natural Science Foundation of China (Project No. 11401462, 61502342, 61603162) and the China Postdoctoral Science Foundation (Project No. 2015M582630).
§Corresponding author.

tional large scale data sets. Inspired by [18], we choose canteen transaction records as our data source. Compared with the aforementioned methods, canteen transaction records are relatively general (most of the contemporary colleges have adopted smart card systems in their canteens), easy to collect (commonly all the canteen transaction records are in one single database table in the database systems), comprehensive (canteen transaction records cover nearly all the students' entire time in college), less sensitive (canteen transactions happen in public places and have no "content" information) and relatively easy to analyze (the canteen transaction records are in consistent and simple format). In this paper, we propose a novel method to extract college students' social network information from canteen transaction records, assuming that two students who frequently purchase meals adjacently in the same queue are more likely to be socially related. We design a mechanism to decide the threshold frequency of canteen "queue-adjacent" events over which a canteen "friend" relation is identified. Compared to some other institutionally caused social relations (such as classmates), the canteen "friend" relations reflect relatively active and close relations, thus are more likely to have strong correlations with students' behavior. Using the canteen network information, we are able to calculate students' network attributes to reflect their "positions" in the college social network.

To evaluate our method, we collect a canteen transaction data set with around 4.5 million canteen transaction records from a famous university in mainland China. The data set covers around 16 thousand undergraduate students' canteen transaction records during one semester (112 days). We extract the social network and conduct the exploratory social network data analysis. We also conduct the regression analysis to verify the relationships between the students' network attributes and their academic performance. The data analysis results show that our method is reasonable in extracting social network from canteen transaction records.

The remainder of this paper is organized as follows: Section 2 introduces the method of extracting social network. Section 3 describes the canteen transaction data set, the data analysis procedure and the results. Section 4 discusses the findings and concludes this paper.

2. METHODOLOGY

In this section, we present the method to extract social network from canteen transaction records.

2.1 Extract social network

To extract the social network from canteen transaction records, we consider the queue sequence when students purchase from the same POS machine. It is common to observe that two or more students choose to purchase meals adjacently in the same queue so that they can chat with each other, especially when they have to wait for a long time during the canteen peak time. If two students frequently purchase meals adjacently in the same queue, we may expect that the two students are more likely to be friends or be linked socially.

The detailed method is intuitive. Firstly, we identify the "adjacent" student pairs in the canteen queues from the canteen transaction records. Secondly, we analyze how frequently each student pair appears. Then we need to decide the threshold of appearance frequency over which we recognize a canteen "friend" relation. By comparing the observed network with a generated random network, we develop a mechanism to decide the threshold value according to false edge rate.

Formally, we define the following two terms: *queue-adjacent event* and *canteen friend relation*.

Queue-adjacent Event: We define a canteen queue-adjacent event exists when two students purchase from the same POS machine "adjacently". Suppose there are N students in total. For the sake of simplicity, assume each student takes K meals each day, makes one transaction in each meal. If a student i makes the kth transaction in day d, let l_{idk} denote the transaction POS machine and t_{idk} denote the transaction time. Suppose this transaction is the s_{idk}th transaction of POS machine l_{idk} in the day. Then we define a binary variable δ_{ijdk}:

$$\delta_{ijdk} = \begin{cases} 1 & \text{if } l_{idk} = l_{jdk}, |s_{idk} - s_{jdk}| \le s^* \\ & \text{and } |t_{idk} - t_{jdk}| \le t^*, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

δ_{ijdk} indicates whether the students i and j purchase from the same POS machine l_{idk} "adjacently": there are no more than $(s^* - 1)$ students between them in the queue, and the time interval between their transactions is no more than t^*.

Canteen Friend Relation: Obviously, most of the canteen queue-adjacent events happen by chance. In most cases, a student does not know the student before or after him/her in the queue. Therefore, only if there are multiple canteen queue-adjacent events between two students, we consider there is a canteen friend relation. We define the number of canteen queue-adjacent events between two students as:

$$w_{ij} = \sum_{d,k} \delta_{ijdk}.$$

Suppose the threshold of the number of queue-adjacent events is w^*, equal to or above which we consider there is a canteen friend relation. Then the indicator of a canteen friend relation can be represented as:

$$a_{ij} = \begin{cases} 1 & \text{if } w_{ij} \ge w^* \\ 0 & \text{otherwise,} \end{cases}, \tag{2}$$

and the canteen friend relation adjacent matrix can be represented as $A = (a_{ij}) \in \mathbb{R}^{N \times N}$, where N is the total number of students.

2.2 Determine the critical value w^*

We still need to determine the critical value w^*. It is intuitive that the higher the critical value is, the more likely a canteen friend relation will be true (i.e., the connected two students have a real social relation), but the less number of canteen friend relations we will generate. Therefore, an appropriate w^* should tradeoff two targets: (1) to exclude as many false relations as possible; and (2) to cover as many true relations as possible.

The Observed Network: Given a critical value w^*, we can generate an *observed network* from a canteen transaction data set using Eq. (2) defined in the previous section. Denote the set of the edges in the *observed network* given w^* as $\mathcal{O}_{w^*}$, then the number of edges $|\mathcal{O}_{w^*}|$ can be counted from the canteen transaction data set. If we can classify $\mathcal{O}_{w^*}$ into two subsets: the true edge set (the edges which represent true relations, denoted by $\mathcal{T}_{w^*}$), and the false edge set (the edges which are caused by randomness, denoted by $\mathcal{F}_{w^*}$), then we are able to calculate a false edge rate $\theta_{w^*} = \frac{|\mathcal{F}_{w^*}|}{|\mathcal{O}_{w^*}|}$. However, since it is impossible to classify the edges in $\mathcal{O}_{w^*}$, we need to estimate the number of false edges $|\mathcal{F}_{w^*}|$ which are generated by randomness.

The Random Network: Now we build a theoretical model to generate a *random network*. Assume there is only one POS machine (imagine we connect all the queues of all the POS machines one by one to produce a long queue), and the queue sequence is random and independent among meals. Under these assumptions, the probability of two students i and j having a canteen queue-adjacent event in one meal is roughly $\frac{2s^*}{N}$ (if we further consider multiple canteens and POS machines as well as time intervals between transactions, the probability could be even smaller). Based on this probability, δ_{ijdk} becomes a Bernoulli random variable with $\mathbb{P}(\delta_{ijdk} = 1) = \frac{2s^*}{N}$. Suppose there are D days, then the number of canteen queue-adjacent events between student i and j is $w_{ij} = \sum_{d=1}^{D}\sum_{k=1}^{K}\delta_{ijdk}$. Obviously, w_{ij} follows a Binomial distribution $\mathcal{B}(KD, \frac{2s^*}{N})$, with the following probability mass function:

$$p(w) = \binom{KD}{w}\left(\frac{2s^*}{N}\right)^w\left(1-\frac{2s^*}{N}\right)^{KD-w}.$$

According to Eq. (2), the probability of having a canteen friend relation between students i and j can be calculated by $P(w^*) = \mathbb{P}(a_{ij} = 1) = \mathbb{P}(w_{ij} \geq w^*) = 1 - \mathbb{P}(w_{ij} < w^*)$, thus we have

$$P(w^*) = 1 - \sum_{w=1}^{w^*-1} p(w).$$

With the probability $P(w^*)$, we are able to build a *random network* following the ER random graph model [7]: whether each pair of nodes have a connection is a Bernoulli random variable with probability $P(w^*)$. Denote the set of the edges in the *random network* given w^* as $\mathcal{R}_{w^*}$, then we have the expectation of $|\mathcal{R}_{w^*}|$ as:

$$E(|\mathcal{R}_{w^*}|) = \frac{N(N-1)P(w^*)}{2}. \tag{3}$$

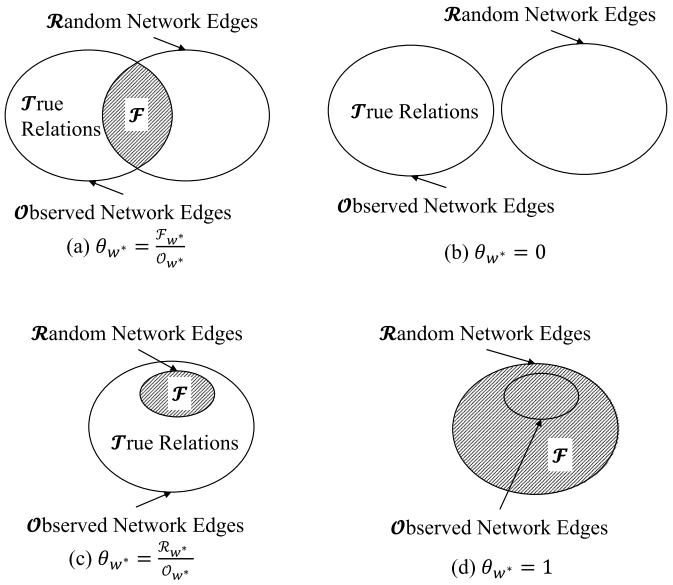


Figure 1. False edge rate.

The False Edge Rate: To estimate the false edge rate θ_{w^*}, we need to compare the *observed network* and the *random network* of the same critical value w^*. First we assume that the edges in the *random network* are all false edges. As shown in Figure 1(a), $\mathcal{F}_{w^*} = \mathcal{O}_{w^*} \cap \mathcal{R}_{w^*}$, the false edge rate should be $\theta_{w^*} = \frac{|\mathcal{O}_{w^*} \cap \mathcal{R}_{w^*}|}{|\mathcal{O}_{w^*}|}$. Unfortunately, we still do not know the intersection of $\mathcal{O}_{w^*}$ and $\mathcal{R}_{w^*}$.

Consider three extreme cases: (1) the edges of two networks have no intersection, i.e., $\mathcal{O}_{w^*} \cap \mathcal{R}_{w^*} = \emptyset$, then $\theta_{w^*} = 0$, as shown in Figure 1(b); (2) the *random network* edge set is a subset of the *observed network* edge set, i.e., $\mathcal{R}_{w^*} \subseteq \mathcal{O}_{w^*}$, then $\theta_{w^*} = \frac{|\mathcal{R}_{w^*}|}{|\mathcal{O}_{w^*}|}$, as shown in Figure 1(c); and (3) the *observed network* edge set is a subset of the *random network* edge set, i.e., $\mathcal{O}_{w^*} \subseteq \mathcal{R}_{w^*}$, then $\theta_{w^*} = 1$, as shown in Figure 1(d). Summarize the three extreme cases, we have $\theta_{w^*} \in [0, min(1, \frac{|\mathcal{R}_{w^*}|}{|\mathcal{O}_{w^*}|})]$.

Now we relax the assumption that "all the edges in the *random network* are false edges", i.e., there are some edges which represent true relations in the *random network*. Under this relaxed assumption, the maximum false edge rate should be less than $\frac{|\mathcal{R}_{w^*}|}{|\mathcal{O}_{w^*}|}$ in Figure 1(c), and less than 1 in Figure 1(d). Therefore, the upper bound of the false edge rate could still be $min(1, \frac{|\mathcal{R}_{w^*}|}{|\mathcal{O}_{w^*}|})$.

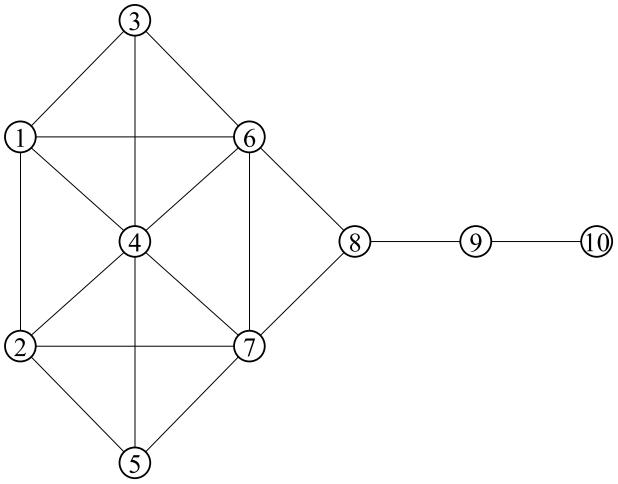


Figure 2. Kite network.

We use this upper bound as the estimation of the false edge rate. Operationally, we have

$$\hat{\theta}_{w^*} = \begin{cases} \frac{E(|\mathcal{R}_{w^*}|)}{|\mathcal{O}_{w^*}|} & \text{if } |\mathcal{O}_{w^*}| > E(|\mathcal{R}_{w^*}|), \\ 1 & \text{if } |\mathcal{O}_{w^*}| \leq E(|\mathcal{R}_{w^*}|) \end{cases} \tag{4}$$

where $|\mathcal{O}_{w^*}|$ can be counted from the canteen transaction data set, and $E(|\mathcal{R}_{w^*}|)$ can be calculated using Eq. (3). An appropriate w^* should be selected to satisfy $\hat{\theta}_{w^*} < \alpha$, where α is the confidence level (e.g., $\alpha = 0.01$), and at the same time to keep as many edges as possible.

2.3 Social network attributes

Based on the topology of the extracted social network, we are able to calculate a variety of network attributes for each student [11, 6]. In detail, four attributes which reflect a student's "position" in the canteen social network are computed. To facilitate the interpretations, we use a famous example "Kite network" [12] to illustrate the computation of the attributes, as shown in Figure 2.

The most commonly used attribute in a network is degree, defined as $DE_i = \sum_{j=1}^{N} a_{ij}$, $i \in \{1, ..., N\}$. A node's degree reflects the number of nodes which directly connect to the node. For example, in Figure 2, Node 4 has the highest degree ($DE_4 = 6$), while Node 10 has the lowest degree ($DE_{10} = 1$). In the canteen social network, the degree DE_i indicates student i's number of canteen friends, and reflects the student's social integration level. Since a higher social integration level commonly correlates to a better academic performance [20], we also expect a positive correlation between a student's degree and academic performance.

Another group of commonly studied network attributes is centrality [11, 16]. Centrality measures the extent to which a node approaches the "center" of a network. There are two commonly adopted measurements of centrality: the closeness centrality and the betweenness centrality.

Suppose there are some paths between Node i and Node j, where a path is a group of linked edges which connect the two nodes. The path length is defined as the number of edges on a path, and the shortest path is the path with the minimum path length. For example, in Kite network, the paths between Node 1 and Node 5 include {12,25}, {14,45}, {16,67,75}, {13,36,68,87,75}, etc. The shortest paths (also named as geodesic paths) are {12,25} and {14,45}. The distance d_{ij} between Node i and Node j is defined as the length of the shortest path(s) between Node i and Node j when $i \neq j$, or 0 when $i = j$. Then the closeness centrality of Node i is $CC_i = \frac{N}{\sum_{j=1}^{N} d_{ij}}$, $i \in \{1, ..., N\}$, which is the inverse of the average distance between Node i and all the nodes. In the canteen social network, the larger the closeness centrality is, the shorter the average distance between a student and all the other students will be.

The betweenness centrality is defined as $BC_i = \sum_{i \neq j \neq k} \frac{g_{jk}^i}{g_{jk}}$, $i \in \{1, ..., N\}$, where g_{jk} is the number of the shortest path from Node j to Node k, while g_{jk}^i is the number of the shortest path from Node j via Node i to Node k. Intuitively, betweenness centrality is the percentage of the shortest paths via Node i comparing to all the shortest paths. Betweenness centrality reflects the ability of a student to connect to different communities.

The last attribute is the clustering coefficient. Let Node i directly connect to some other DE_i nodes, i.e., Node i's degree is DE_i. The network among the DE_i nodes is also called Node i's ego network. It is easy to verify that there are at most $\frac{DE_i(DE_i-1)}{2}$ edges among the DE_i nodes. Assume that there are in fact E_i edges among the DE_i nodes, then Node i's clustering coefficient can be defined as $CL_i = \frac{2E_i}{DE_i(DE_i-1)}$, $i \in \{1, ..., N\}$. Based on this definition, a student's clustering coefficient reflects the density of the student's ego network. For example, in the Kite network, Node 1's ego network is constructed by { 2,3,4,6}, $DE_1 = 4$, $E_1 = 4$, thus $CL_1 \approx 0.667$.

Next section will empirically show that the degree, closeness centrality, betweenness centrality and clustering coefficient calculated based on the canteen social network are significantly correlated with students' academic performance.

3. EMPIRICAL STUDY

3.1 Data sets

The data sets in this empirical study are collected from a famous university in mainland China. The university granted us the privilege of using the data for research purposes only. All the identification-related fields in the data sets are encrypted to protect the students' privacy. In detail, two data sets are used in this study. The first one is a canteen transaction data set, including 4,461,327 canteen transaction records of 16,050 undergraduate students, during the 2014 fall semester (from September-8-2014 to December-28-2014, 16 weeks, 112 days). The fields in this data set include:

student ID (encrypted), gender (GE), class ID (encrypted), date, time, POS (point of sale) machine ID (encrypted), and transaction value.

The second data set is the undergraduate students' academic performance data set in the same semester. The academic performance is measured by the average of the students' course scores (in 100 point scale), weighted by course credit. It is obvious that the average score weighted by course credit is similar to the Grade Point Average score (GPA), but reflects more details (in 100 point scale, rather than in 4 grade point scale). Hereafter we name the average score weighted by course credit as Score Point Average (SPA). The academic performance data set covers 16,236 undergraduate students. The fields of the academic performance data set include: student ID (encrypted), enrollment year (EY), and SPA.

The encryption methods for student IDs in both two data sets are the same, thus we are able to connect the two data sets together via encrypted student IDs. However, the students in the two data sets are not perfectly matched to each other, due to international exchange programs and other unknown reasons (thusly some students may have no canteen transaction records or SPA scores). Moreover, there are quite a number of holdover students in the university. These students take courses with other students who are a year or two older than they are. To reflect this fact, we define a variable "class year" (CY) to indicate the enrollment year of a class. Practically, we use the enrollment year of the majority of students in a class as the "class year". Then we are able to calculate (1) a student's grade (GR), which is "$2015 - CY$" (since the data sets were collected in 2015); and (2) the difference between a student's class year and enrollment year, named as "year difference" (YD). The year difference variable is a non-negative integer: 0 indicates a normal student, while a positive number (from 1 to 3) indicates a holdover student.

3.2 Network extraction

The students' canteen social network information is extracted following the method proposed in Section 2. First we need to determine the s^* and t^*. According to both observations in canteens and interviews with students, we find that it is a common practice that two or three students purchase meals together so that they can chat with each other in queues. Moreover, compared with the extracted network when $s^* = 1$ or $s^* = 3$, the extracted network when $s^* = 2$ has the largest number of edges, and thusly has the most complete information. Based on these considerations, we set $s^* = 2$. As to t^*, according to our statistics, more than 80% of the time intervals between two adjacent transactions on the same POS machine are within 60 seconds. The network given $t^* = 60$ covers more than 93% of the students. Therefore, $t^* = 60$ is large enough for extracting the canteen social network.

We then need to decide the critical value w^*. Since there are around 16 thousand students' transaction records for 112 days in the data set, following subsection 2.1, we have the Table 1: as w^* increases, the estimated false-positive rate $\hat{\theta}_{w^*}$ decreases sharply. When $w^* = 4$, the estimated false-positive rate is $0.005 < 0.01$, which is acceptably small. Therefore, we choose $w^* = 4$ to be the threshold value for identifying a canteen friend relation.

Table 1. Determine the critical value w^*.

w^*	$P(w^*)$	$E(\lvert\mathcal{R}_{w^*}\rvert)$	$\lvert\mathcal{O}_{w^*}\rvert$	$\hat{\theta}_{w^*}$
1	0.0806	10314035	5023421	1
2	0.0033	425942	774171	0.550
3	0.0001	11775	153898	0.077
4	**0.0000**	**244**	**52914**	**0.005**
5	0.0000	4	30160	0.000

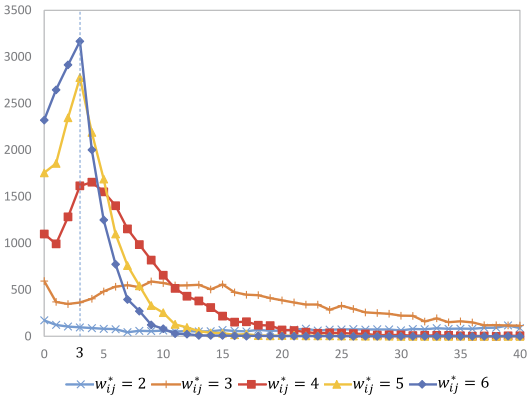


Figure 3. The degree distributions. The distributions when $w^* = 4, 5, 6$ peak around 3, which may reflect the other 3 undergraduate students in the same dorm.

To further verify the appropriateness of the threshold value w^*, we calculate some network indices using different w^*. The network indices are listed in Table 2. According to Table 2, as the threshold value w^* increases, (1) the numbers of connected nodes ($dNodes$) and edges ($Edges$), percentage of connected nodes ($perc$), network density ($Density$) and the average node degree ($avgDE$) all decrease; (2) each pair of connected nodes are more likely to have the same gender (sG increases), the same class (sC increases), the same class year (dCY decreases), and similar SPA scores ($dSPA$ decreases). In other words, the higher the threshold value w^* we choose, the less students will be connected in the network, and the more homogenous each pair of connected students will be. By choosing the threshold value $w^* = 4$, we generate a network with 14,951 (93.15%) connected nodes and 52,914 edges. The network is sparse, with the density 0.0004 and the average degree 6.59.

As we know, the degree distribution of a network is commonly not a normal distribution, thus we also depict the degree distributions given different threshold values w^* in Figure 3. The degree distributions show that the modes of degree when $w^* = 4, 5, 6$ are around 3, which means that a typical student is more likely to have 3 canteen friend

Table 2. Network indices. There are 16050 nodes in the network. The abbreviations are: cNodes*—the number of connected nodes,* Edges*—the number of canteen friend relations,* perc*—the percentage of the connected nodes,* avgDE*—the average node degree,* Density*—the network density,* sG*–the rate of whether two connected nodes have the same gender,* sC*—the rate of whether two connected nodes belong to the same class,* dCY*—the average of the difference between two connected nodes' class year, and* dSPA*—the average of the difference between two connected nodes' SPA.*

w^*	*cNodes*	*Edges*	*perc*	*avgDE*	*Density*	*sG*	*sC*	*dCY*	*dSPA*
1	16038	5023421	0.9993	625.97	0.0390	0.67	0.01	1.08	10.68
2	15879	774171	0.9893	96.47	0.0060	0.70	0.02	0.97	10.31
3	15457	153898	0.9631	19.18	0.0012	0.74	0.06	0.83	9.94
4	**14951**	**52914**	**0.9315**	**6.59**	**0.0004**	**0.78**	**0.13**	**0.66**	**9.55**
5	14295	30160	0.8907	3.76	0.0002	0.82	0.25	0.46	9.18
10	11887	13499	0.7406	1.68	0.0001	0.95	0.59	0.07	8.20
50	1984	1141	0.1236	0.14	0.0000	0.96	0.73	0.03	7.02
100	212	107	0.0132	0.01	0.0000	0.95	0.79	0.02	6.43

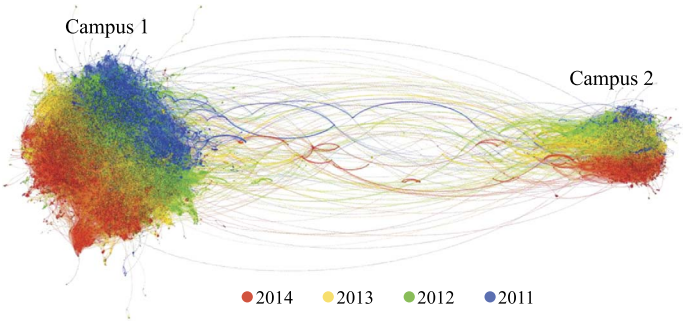
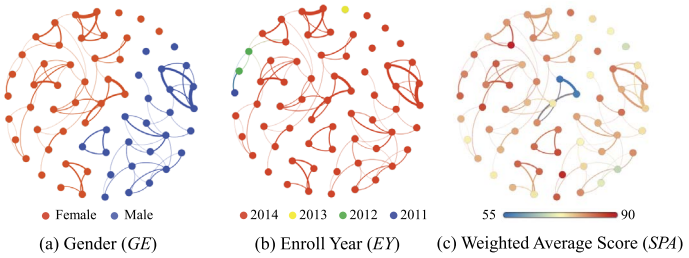


Figure 4. The canteen social network. Each dot represents one student, and each line between two dots represents a canteen friend relation.



(a) Gender (*GE*) (b) Enroll Year (*EY*) (c) Weighted Average Score (*SPA*)

Figure 5. The social network of a randomly selected class.

relations. This is consistent with the fact that every 4 undergraduate students share one dorm in the university. From this perspective, the threshold $w^* = 4$ is also appropriate.

3.3 Network visualization

To get an intuitive understanding of the extracted network, Figure 4 visualizes the network using *Gephi* (with "Force Atlas 2" layout), and colors the nodes by class year. Interestingly, the network is naturally clustered into two big components, which correspond to the two campuses of the university. Both of the two components show clear blocks of class years, which reflect that the students with the same class year are more likely to have meals together. More interestingly, the block of 2014 (year 1, red) is "neighbored" with the block of 2013 (year 2, yellow), then 2013 is "neighbored" with year 2012 (year 3, green), and then 2012 is "neighbored" with 2011 (year 4, blue). This pattern shows that students are more likely to have connections with students a grade older or younger than themselves, than they are with the other older or younger students.

To further investigate the network in detail, we randomly choose one class and draw the social network among the students in Figure 5, and then color the social network by gender (GE), enrollment year (EY) and weighted average score (SPA), separately. It is easy to observe that there are several strongly connected cliques of 3-4 students in the network, which probably reflect the dorm mate cliques. Figure 5(a) shows that the students of the same gender have meals together. Figure 5(b) shows that the enrollment year of the majority students (red) is 2014, i.e., the class year (CY) is 2014. The four students in the other enrollment years (yellow, green and blue) are relatively isolated from the class. Figure 5(c) reflects that each pair of connected students generally have similar weighted average scores (SPA). In summary, Figure 5 reflects the homophyly and/or the social influence of the network. These direct observations in Figure 5 are consistent with the indices listed in Table 2, as well as our understanding of the life habits of undergraduate students. Figure 5 also shows the reasonability of our network extraction method.

3.4 Regression analyses

The relationships between students' network indicators and academic performance are studied in this subsection using regression analysis. Before conducting the regression analysis, we need to drop some groups of "special cases" from the data set, to make the results more clear. The first group of "special cases" are the year four students. A large number of fourth year students have failed courses in the previous three years, and have had to take make-up exami-

Table 3. Descriptive statistics. $N = 11272$.

Variable	Description	Min	Max	Mean	SD
SPA	Weighted average of scores.	0.75	98.75	77.24	11.69
GE	Gender: male–1, female–0.	0	1	0.71	-
GR	A student's grade.	1	4	-	-
YD	The difference of class year and enroll year.	0	3	0.04	0.25
DE	Degree centrality.	0	49	7.09	5.39
CC	Closeness centrality.	0	1	0.18	0.08
BC	Betweeness centrality, logarithm transformed.	0	14.89	8.32	3.51
CL	Clustering coefficient	0	1	0.20	0.24
nT	The number of transactions.	1	1060	297.44	133.26
vT	The average value of transactions.	0.86	12.36	3.88	0.81

Table 4. Correlation matrix. The mark $**$ indicates the correlation is significant at 0.01 level (2-tailed), and the mark $*$ indicates the correlation is significant at 0.05 level (2-tailed). $N = 11272$.

IV	*SPA*	*GE*	*GR*	*YD*	*DE*	*CC*	*BC*	*CL*	*nT*
GE	-.18**								
GR	.01	.02*							
YD	-.35**	.02*	-.10**						
DE	.15**	.25**	-.03**	-.09**					
CC	.07**	.10**	.01	-.06**	.16**				
BC	.20**	.19**	-.06**	-.10**	.64**	.14**			
CL	-.01	-.08**	-.13**	-.02*	-.16**	-.03**	-.34**		
nT	.25**	.30**	-.11**	-.12**	.60**	.20**	.58**	-.17**	
vT	-.17**	.21**	.06**	.07**	-.25**	-.05**	-.23**	.13**	-.34**

nations in their fourth year. The make-up examinations are different from the normal examinations, and should be removed from the calculation of *SPA*. Unfortunately, we are not able to identify those students with make-up examinations. To avoid the including of the students who have taken make-up examinations, we remove the year-four students from the data sets. The second group of "special cases" is the special program students. There are some special programs, such as the "juvenile program" (which directly recruits students from junior high school), the "experimental program" (which recruits top students from all specialties), etc. These special programs are not the same as the normal programs. The university marked these special programs so that we are able to remove these students from the data set. Third, there are a handful of "students" who have too many canteen friends far more than we expected. For example, a "student" has 113 canteen friends. We suspect that the corresponding canteen ID card is used by multiple students, rather than the card owner only. We drop 7 IDs whose canteen friends are more than 50, because commonly there are no more than 50 students in a class. After the "special cases" are removed, we have 11,272 students remaining.

Table 3 shows the descriptive statistics of the variables which will be used in the regression. The variables *DE*, *CC*, *BC* and *CL* are calculated based on the extracted network. Note that we logarithm transformed the *BC* value since it follows a semi-power law distribution and is heavily left skewed.

The correlation matrix is reported in Table 4. The correlation matrix shows that, the social network indices, except *CL*, are all significantly correlated with the weighted average score (*SPA*).

Then we conduct regression analyses to study the relationships between students' social network attributes and *SPA*. The linear regression is chosen since the dependent variable *SPA* is a continuous variable, and approximately follows a normal distribution. The regression results are illustrated in Table 5. In detail, we conduct four regressions as follows:

Model 1: Benchmark First, we show a benchmark model which only covers the demographic variables. The regression model is as follows:

$$SPA = c + \alpha_0 GE + \alpha_1 GR + \alpha_2 YD + \epsilon.$$

The results of the benchmark model show that the female students' *SPA* are better than the male students, which is consistent with the findings in the literature [26]; not surprisingly, the holdover students' *SPA* are much less than the normal students.

Model 2: Effects of social network Based on the benchmark model, we add the four network attributes. Then we have the regression model 2:

$$\begin{aligned} SPA &= c + \alpha_0 GE + \alpha_1 GR + \alpha_2 YD + \alpha_3 DE \\ &+ \alpha_4 CC + \alpha_5 BC + \alpha_6 CL + \epsilon \end{aligned}$$

Table 5. Regression results. The dependent variable is SPA. *The marks ** and * indicate $p<0.01$ and $p<0.05$ separately.*

IV	Model 1	Model 2	Model 3	Model 4
c	81.44**	74.53**	74.49**	72.01**
GE	-4.29**	-5.55**	-5.69**	-6.67**
GR	-0.20*	0.02	0.04	0.28**
YD	-16.25**	-14.85**	-14.76**	-13.99**
DE		**0.12****	**0.43****	**0.14***
CC		**6.48****	**6.08****	**2.90***
BC		**0.59****	**0.45****	**0.23****
CL		**1.78****	**1.30****	**1.52****
DE^2			**-0.01****	**-0.01****
nT				0.02**
vT				-0.01
N	11272	11272	11272	11272
df	3	7	8	10
F	656.22**	384.252**	340.68**	337.85**
$Adj.\ R^2$	0.149	0.192	0.194	0.230

The results show that, there is a positive relationship between the number of canteen friends (DE) and SPA. This finding is consistent with the findings in the literature that social integration is positively connected with academic performance [20]. The closeness centrality (CC), which reflects the average distance of a student to the other students in the university, is positively correlated with SPA score. This indicates that the more closer a student is to the "center" of the university network, the higher SPA score the student will have. The regression coefficient of betweeness centrality (BC) is also significantly positive, which indicates that if a student can bridge the gaps between several communities (which will put them on a shorter path between the communities), the student will have a better SPA. This result may reflect the benefit of the "structural hole" [3] in the college social network. Finally, we find that the clustering coefficient (CL) is also positively correlated with the SPA score. The clustering coefficient reflects the connectivity of a student's canteen friends. This finding shows that, the more connections between a student's friends, the higher the student's SPA will be. We also conduct ANOVA between Model 1 and Model 2. The results show that Model 2 is significantly different from Model 1 ($F=141.73$, $p<0.001$).

Model 3: Nonlinear effects of degree We expect a reversed U-shape relationship between the degree (DE) and SPA: a moderate number of friends should be beneficial to a student's academic performance. If a student has no friend, it may show that the student is having troubles getting involved in college life; however, if a student has too many friends, it may detract from their studying and negatively affect their course scores. Therefore, we add a quadratic term of degree in the regression model 3:

$$\begin{aligned} SPA &= c+\alpha_0 GE+\alpha_1 GR+\alpha_2 YD+\alpha_3 DE \\ &+ \alpha_4 CC+\alpha_5 BC+\alpha_6 CL+\alpha_7 DE^2+\epsilon. \end{aligned}$$

The regression coefficient of DE^2 is significantly negative, which confirms our expectation that there would be a reversed U-shape relationship between degree and SPA. However, the ANOVA results between Model 2 and Model 3 show that there is no significant difference ($F=1.12$, $p=0.29$).

Model 4: Control of confounding effects There could be some confounding effects in the regression models. The social network is extracted from the canteen transaction records, thus a student needs to have some canteen transaction records, otherwise there will be no canteen relationships extracted. Hence, there should be a high correlation between the degree and the number of transactions (0.60, see Table 5). We need to control the number of transactions in the regression model, so we add the variable nT into the regression. We also control the average value of transactions (which may indicates a student's economic status) into the regression model. The improved regression model is:

$$\begin{aligned} SPA &= c+\alpha_0 GE+\alpha_1 GR+\alpha_2 YD+\alpha_3 DE \\ &+ \alpha_4 CC+\alpha_5 BC+\alpha_6 CL+\alpha_7 DE^2 \\ &+ \alpha_8 nT+\alpha_9 vT+\epsilon. \end{aligned}$$

With the control variables in the model, the coefficients of D, C, and BC decrease, as compared to Model 3. However, these coefficients remain the same sign and are still significant, which shows that the effects of social network attributes persist with the control of transaction frequency. The ANOVA between Model 3 and Model 4 also shows that the two models are significantly different ($F=282.03$, $p<0.001$).

3.5 Robustness check

We also conduct some more regressions using model 4 to check the robustness of the regression results, as shown in Table 6. First we add the year 4 students and the "special program" students into the regression, indicated by "RC 1" in Table 6. With a larger sample (N=14,347), the results of "RC 1" are consistent with the results of model 4, but with a smaller adjusted R^2. We then extract the network using $w^*=3$, which generates more connections between the students, and then use the network attributes to run the regression. As illustrated by the "RC 2" column in Table 6, the regression coefficients of degree (DE) and closeness centrality (CC) are not significant. This indicates that $w^*=3$ may not be a proper threshold value, since it brings too many "false" relations into the network. Lastly, we extract the network using $w^*=5$. The regression results are listed in column "RC 3". The regression coefficients are consistent with Model 4, except that the clustering coefficient (CL) is not significant. This could be caused by the removal of too many "true" relations among the student's friends at $w^*=5$. Note that we use the same sample of Model 4 in "RC

Table 6. Robustness check. The dependent variable is SPA. *The robustness check is based on Model 4 in Table 5. The marks ** and * indicate* $p < 0.01$ *and* $p < 0.05$ *separately.*

IV	RC 1	RC 2	RC 3
c	71.41**	71.43**	72.75**
GE	-6.70**	-6.67**	-6.59**
GR	0.58**	0.27**	0.26**
YD	-11.91**	-14.00**	-13.99**
DE	**0.12***	**0.06**	**0.52****
CC	**4.03****	**2.99**	**2.45****
BC	**0.23****	**0.36****	**0.11****
CL	**1.66****	**1.87****	**0.47**
DE^2	**-0.01****	**-0.00****	**-0.04****
nT	0.02**	0.02**	0.02**
vT	0.04	0.01	-0.03
N	14347	11272	11272
df	10	10	10
F	375.81**	338.51**	340.09**
Adj. R^2	0.207	0.230	0.231

2" and "RC 3" to make the regression results comparable to the results of Model 4. The robustness check reflects the consistency of regression results among different situations, and the properness of our method to extract network.

4. DISCUSSION

This study shows some initial efforts to extract social network from canteen transaction records, and connect the social network with students' academic performance. The results of this paper are descriptive but promising. Along this line, there are some directions for future research.

The first research direction is to develop a more rigorous method to extract social network. Although we have validated a relatively rigorous method in this paper, there is still some information we ignored. For example, if two students have similar life habits and food preferences (especially the food few students prefer), they are likely to have canteen queue-adjacent events even when they do not know each other. By analyzing students' life habits and food preferences, the network can be extracted more precisely.

Secondly, this study presents the correlations between students' social network attributes and academic performance. However, the underlying mechanisms of these correlations are still not clear. There are several possible mechanisms in which social network correlates to a student's academic performance: (1) homophyly, i.e., the students connect to each other because they are similar; (2) social influence, i.e., the connected students influence each other; (3) network structure, i.e., the network position of a students may reflects the social capital, structural hole, etc. In future, statistical models should be built to distinguish these three mechanisms.

Lastly, in this study we used some network attributes in our regression models. The network attributes are calculated based on the whole network, thusly they violate the assumption of linear regression that the observations of independent variables should be "independent" from each other. However, we were not able to avoid introducing network attributes in regressions. Our findings encourage future research to build rigorous statistical models and develop corresponding theories to fix the problem.

Received 23 July 2016

REFERENCES

[1] Bearman, P. S. and Moody, J. (2004). Suicide and friendships among American adolescents. *American Journal of Public Health* **94** 89–95.

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

[3] Burt, R. S., Jannotta, J. E., and Mahoney, J. T. (1998). Personality correlates of structural holes. *Social Networks* **20** 63–87.

[4] Cabrera, A. F., Nora, A., and Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education* **64** 123–139.

[5] Cummings, J. N., Butler, B., and Kraut, R. (2002). The quality of online social relationships. *Communications of the ACM* **45** 103–108.

[6] Daudin, J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18** 173–183. MR2390817

[7] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs *Magyar Tud. Akad. Mat. Kutató Int. Közl* **5** 17–61. MR0125031

[8] Gazal, S., Daudin, J., and Robin, S. (2012). Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation* **82** 849–862. MR2929296

[9] Hoff, P., Raftery, A., and Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098. MR1951262

[10] Jain, T. and Kapoor, M. (2015). The impact of study groups and roommates on academic performance. *Review of Economics and Statistics* **97** 44–54.

[11] Kolaczyk, E. D. and Csárdi, G. (2014). *Statistical Analysis of Network Data with R*, Springer, New York. MR3288852

[12] Krackhardt, D. (1990). Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. *Administrative Science Quarterly* **35** 342–369.

[13] Kuncel, N. R., Hezlett, S. A., and Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology* **86** 148–161.

[14] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks* **30** 330–342.

[15] Mayer, J., Mutchler, P., and Mitchell, J. C. (2016). Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academy of Sciences* **113** 5536–5541.

[16] Newman, M. (2010). *Networks: An Introduction*, Oxford university press, Oxford. MR2676073

[17] Parker, J. (2012). Does Living Near Classmates Help Introductory Economics Students Get Better Grades? *The Journal of Economic Education* **43** 149–164.

[18] Ram, S., Wang, Y., Currim, F., and Currim, S. (2015). Using Big Data for Predicting Freshmen Retention. *in International Conference on Information Systems*.

[19] Reich, S. M., Subrahmanyam, K., and Espinoza, G. (2012). Friending, IMing, and hanging out face-to-face: overlap in adolescents' online and offline social networks. *Developmental Psychology* **48** 356–368.

[20] Richardson, M., Abraham, C., and Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin* **138** 353–387.
[21] Sarker, F., Tiropanis, T., and Davis, H. C. (2014). Linked data, data mining and external open data for better prediction of at-risk students. *in Control, Decision and Information Technologies (CoDIT), International Conference on, IEEE* 652–657.
[22] Snijders, A. and Nowicki, K. (2001). Estimation and Prediction for Stochastic Block Structures. *Journal of the American Statistical Association* **1** 75–100. MR1947255
[23] Sweeney, M., Rangwala, H., Lester, J., and Johri, A. (2016). Next-Term Student Performance Prediction: A Recommender Systems Approach. *arXiv preprint arXiv:1604.01840*.
[24] Trout, D. L. (1980). The role of social isolation in suicide. *Suicide and Life Threatening Behavior* **10** 10–23.
[25] Wang, R., Harari, G., Hao, P., Zhou, X., and Campbell, A. T. (2015). SmartGPA: how smartphones can assess and predict academic performance of college students. *In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM* 295–306.
[26] Zhou, Y.-X., Zhao, Z.-T., Li, L., Wan, C.-S., Peng, C.-H., Yang, J., and Ou, C.-Q. (2014). Predictors of first-year GPA of medical students: a longitudinal study of 1285 matriculates in China. *BMC Medical Education* **14** 87–95.
[27] Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics* **85** 9–23.

Yuewen Liu
Department of Information Management and E-business
School of Management
Xi'an Jiaotong University
China
E-mail address: liuyuewen@xjtu.edu.cn

Ke Xu
Department of Business Statistics and Econometrics
Guanghua School of Management
Peking University
China
E-mail address: xk0566@163.com

Xiangyu Chang
Center of Data Science and Information Quality
Department of Information Management and E-business
School of Management
Xi'an Jiaotong University
China
E-mail address: xiangyuchang@gmail.com

Dehai Di
Department of Information Management and E-business
School of Management
Xi'an Jiaotong University
China
E-mail address: ddh@xjtu.edu.cn

Wei Huang
Department of Information Management and E-business
School of Management
Xi'an Jiaotong University
China
E-mail address: whuang@xjtu.edu.cn