Lu et al. *EPJ Data Science* (2022) 11:10
https://doi.org/10.1140/epjds/s13688-022-00322-0

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Academic failures and co-location social networks in campus

Shan Lu[1], Jichang Zhao[2*] and Huiwen Wang[2,3]

*Correspondence:
jichang@buaa.edu.cn
[2]School of Economics and Management, Beihang University, Beijing, China
Full list of author information is available at the end of the article

Abstract

Social network structures have an additional effect on success other than individual factors. Whether this is true for failure, the opposite of success, still remains unclear. This paper focuses on the associations between academic failures, that defined by not passing course exams, and campus social networks based on students' co-location occurrences. A novel method with statistical validation is first established to infer co-location social networks reliably from campus behaviors recorded by smart cards. It is found that network structures indicated by degree and clustering are indeed significantly associated with academic failures and failed students usually locate in small, dense but heterogeneous clusters. While their connections are more stable over time, failed students seem to be not positively influenced by those non-failed ones and even demonstrate a failure momentum. In fact, further exploitations imply that failed students have less incentives to adjust their social structures when compared to the non-failed ones. Additionally, we show that messages reflected in co-location social networks and behavioral activities indeed help predict failures and the network snapshot at mid-term offers competent prediction power on individual academic failure such that interventions in pursuit of avoiding failures can be applied within a substantial time window. Our findings underline the importance of understanding co-location social networks beyond failure in educational settings.

Keywords: Social network; Co-location network; Academic failure; Clustering; Campus behaviors; Social contagion

1 Introduction

Social networks and individual structural positions are shown to affect the selection on human behaviors (Borgatti et al. [1]) and even result in outcomes such as career success (Seibert et al. [2]), investment success (Nanda et al. [3]), online market success (Ansari et al. [4], Aral and Walker [5]) and academic success (Stadtfeld et al. [6]). For instance, social network provides access of information and resources that enhance work performance in one's career (Seibert et al. [2]); a central position in the co-investment network offers access advantage to gain entry into promising deals (Nanda et al. [3]); social contagion that relies on social networks benefits product adoption in marketing (Ansari et al. [4], Aral and Walker [5]). Apparently, the pathways of social network to success have attracted enormous and pervasive interests in previous efforts.

Springer

Success, however, is never the only part of human life. Failure, the exact contrary of success, takes profound occupations in various circumstances. However, a fine-grained analysis of the relationship between social networks and failures, especially the detailed picture of whether specific network structures are associated with failures, is still missing. Meanwhile, although the definition of success can be easily delimited as promotion in career (Seibert et al. [2]), substantial profit in investment (Nanda et al. [3]) or considerable popularity in marketing (Ansari et al. [4]), a general meaning of failure can not simply be elaborated as the opposite of the above. In particular, a failure should be defined as failing to pass or complete the goal in a specific scenario. While profit didn't reach a threshold, or the advertisement didn't reach a targeted popularity score can be defined as a failure, the corresponding threshold or targeted popularity need to be determined beforehand, which leads to the ambiguity of a failure. For some situations, the definition of failure is explicit. Yin et al. [7] classified a failure of grant application as an unsuccessful attempt, a failure of venture as failed to obtain an exit within five years after the first investment by venture capital firms, a failure of terrorist attack as failed to claim casualties. Similarly, for academic performance in campus, a failure can be explicitly and incontrovertibly defined as failing to pass course exams. Because of this convenience, we restrict ourselves to discuss the association between social networks of students and academic failures in this paper. Understanding this has the potential to considerably enhance scholars' knowledge of the role of social networks in failures, even leads to a better knowledge on the influence pathway of social structures on academic failures.

The main practical challenge with implementing insights of social structure's associations with academic failures is that they require access to students' social networks, which is not always feasible, especially on a large scale. Prior works conducted multiple surveys to measure social connections as well as self-reported measures of socioeconomic background, abilities and motivations, etc. (Dokuka et al. [8], Stadtfeld et al. [6]). Despite the elaborate research design, the high cost of surveys, especially for long-term tracking, has reluctantly restricted the number of subjects to be less than 300 in most cases (Stadtfeld et al. [6]). Yet, a small number of subjects are more likely to dismiss the true difference among students that lies in social networks. What is more challenging is that obtaining social relationships through questionnaires only relies on repetitive answers from subjects, whose possible memory errors or subjective deviations cannot be fundamentally avoided (Junco [9]). Thus, although survey-based studies capture important aspects of how social networks might contribute to academic performance, we know relatively little about how structural traits are associated with the academic failures in a more realistic campus social network.

An alternative and promising way is to use objective measures that are available in terms of digital footprints to construct social networks, such as online social media and offline behavior records (Kassarnig et al. [10], Yang et al. [11]). Notably, the selection on particular campus behaviors emerges from real social networks, offering an opportunity to capture the true connections amongst a fixed set of students. Students with connections tend to have similar offline activities, in particular, the ones with stronger social relationships act together at the same time more frequently. Therefore, high-frequency co-occurrences can be seen as proxies of real offline social relationships. Unlike studies whose data are mainly collected from self-reports and questionnaires and suffer from a small sample size and social desirability, the collection of smart cards data is unobtrusive and can serve as such

co-occurrence activities (Credé et al. [12], Poropat [13], Wald et al. [14]). Yang et al. [11] verified the strong correlation between the social network inferred from smart-card data and self-report friendship data, implying the reliability of the proxy. Difficulty of this approach lies in the determination of the critical value of co-occurrence frequency, that is, how many times in a certain period of time can be the reference of a true social connection. Yang et al. [11] thought the frequency to be 5, however, this critical value is too dependent on the specific campus environment to be extended to broader cases. This leads to the primary purpose of the present study, that is, constructing a reliable social network from offline campus activities to probe students' social connections. And here we do not seek to infer individual links with high accuracy but to seek a behavior-specific network model which can be informative of social network structures that relevant to academic performance. To fulfil this, we collect digital records from students' smart cards in six semesters, from which a bipartite network of "locations and students" is established with edges indicating students showing up at the same location at the same time. In the meantime, we apply statistical validation method on the bipartite network to infer the social connections among students, enabling us to obtain the offline social network without choosing the critical value of co-occurrence frequency in a data-driven manner.

More importantly, we hope to provide a potential demonstration of the role of social capital, defined as social structures that facilitate the actions of individuals (Coleman [15]), in academic failures. Stadtfeld et al. [6] concluded that social integration within the student cohort is strongly associated with academic success and that this association is not merely explained by the fact that students who are perceived as clever are more likely to be nominated as friends or study partners. They further pointed out that students who remained isolated in the network performed worse and were at higher risk of academic failure, even when taking into account various individual and socioeconomic factors. We seek to relate structural properties of a node in network to the number of failures. The key to this relation is to recognize that academic failures might be affected by students' social capital. We first find students who failed usually have large clustering and tend to connect to non-failed students. That's to say, failed students locate in small, dense but heterogeneous clusters. While their connections are more stable over time, a result of stable co-occurrence with a set of other students possibly induced by limited social willingness, there is no evidence for failed students to be positively influenced by non-failed students. It is further revealed that failed students have less incentives to adjust their social structures as compared to those of non-failed, implying their unwillingness to realign for improvement. Weak ties theory has suggested that information possessed by members of a clique prone to be either shared quickly or already redundant with the information hold by the other members (Granovetter [16]). However, our results unveil the possibility that avoiding individuals' failures, especially academic failures, is not as easy as developing information flow since the positive effect is difficult to be shared even within a stable clique of social network.

The third purpose of this paper is to explore whether failures can be predicted in advance, with the help of information embedded in co-location social networks. Despite the fact that academic performance prediction in recent years has gained much attention with the rapid growth of online learning platforms (Fei and Yeung [17], Zafra et al. [18], Thai-Nghe et al. [19]), answer to whether offline social structures can alert future behavioral outcomes such as academic failures is yet ambiguous. Forecasting failures at latter stages

through early offline campus activities may prepare educators for proactive intervention to avoid failures or correct behaviors according to the predicted outcome. Specifically, we follow the emergence of co-location social networks within a cohort of students who developed social networks during the academic year. While this network is validated to be indicative of academic performance, the network snapshot in the middle of a semester helps predict individual academic failure as competently as the network snapshot at the end of a semester. This in turn indicates that social structure carries informative messages on failures and interventions are able to be applied within a substantial time window before a failure happens.

Our analysis is performed in four steps. Firstly, patterns of academic failures in both cross-sectional and longitudinal are presented to provide necessary understandings on failure activities. Secondly, students' co-location social network in campus is built based on statistical validation on "location and student" bipartite network, accompanied by rich details on analysis of behavior characteristics and network positions of failure students. Thirdly, we relate positions in co-location social networks to academic failures to explore the potential pathways of social structures on failures. Lastly, we verify that academic failures can be predicted based on co-location social network at an earlier stage, which provides opportunities for proactive interventions.

2 Materials and methods

2.1 Data

We collect data from one university with 2588 undergraduates who entered this university in Sep 2014. For the propose of privacy protection, all the students' information was anonymous. Both the student name and student number in our raw data are already pseudonymous. The institutional review board of the university approved the study. The data range from 9/1/2014 and 7/15/2017, covering six terms for their first three years of study. We select data of six semesters according the school calendar. Each semester consists of eighteen weeks, in which the last two are exam weeks. The dataset consists of three types of data.

Digital records of daily activities from campus smart cards The smart cards basically record all the potential activities on campus, such as payment for showers in lavatory, meals in the cafeteria, shopping at the campus kiosk and borrowing books from library. Specifically, each record contains the student pseudonymous ID, timestamp, location and payment value (if there's any), which offers detailed information of students' potential social networks. Four indicators are formalized, namely, total expenditure, the frequency of having breakfast, the frequency of showering and the number of books borrowed from library on a semester granularity, so as to associate these basic patterns of students' campus life with academic performance at the end of semesters. Lu et al. [20] found that these attributes indeed contribute to academic performance either in their compounded manner or dynamic trend. We therefore control these factors in regression analysis.

Internet usage We also collect the total time spent on the Internet from the Internet centre of the university on a semester granularity. While watching videos, playing games and surfing the Internet might result in a large amount of time of Internet using, searching study materials or doing assignments may also lead to a long time of Internet using. Nevertheless, we also consider Internet usage as a control variable.

The raw data are extracted from the university data center with student identities coded for protecting privacy. Except for the above indicators, students' gender and ages are also kept as control variables. Students' academic performances, in terms of course grades that can be criterion of passing exams or not, are also collected.

2.2 Statistical validation on social connections in co-occurrence network

In construction of students' co-location social networks, it is necessary to introduce a statistical method for inferring ties among students that are truly proxies for offline relationships in campus from the bipartite system that include students (**S**) and locations (**L**). In particular, we focus on the projected network on student set **S**. The adjacency projected network is obtained by linking together those vertices of **S** which share at least a common neighbor element of locations **L** in the bipartite system. Details of the statistical method can be found in (Tumminello et al. [21]) and are also briefly presented below.

The method first decomposes the bipartite system into subsystems. Each subsystem $\mathbf{B}_k$ consists of all the $N_{\mathbf{L}}^k$ nodes of set **L** with a given degree k and of all the nodes from set **S** linked to them, where $N_{\mathbf{L}}^k$ nodes in $\mathbf{B}_k$ are homogeneous because they all have the same degree k. This indicates that these locations have the same probability of being visited by students. In other words, the statistical validation method treats the places with few people and places with lots of people separately to infer significant co-occurrence activities. Denote the set of nodes in **L** with a certain degree k as $\mathbf{L}_k$. Then student i and j of set **S** have $N_{i,j}^k$ common neighbors in set $\mathbf{L}_k$, that is, they 'meet' $N_{i,j}^k$ times in the subsystem $\mathbf{B}_k$. Denote the degree of nodes i and j in $\mathbf{B}_k$ as N_i^k and N_j^k, respectively. Under the hypothesis that elements i and j randomly connect to the nodes in $\mathbf{L}_k$, the probability that students i and j share X neighbors in $\mathbf{L}_k$ is given by the hypergeometric distribution, i.e.,

$$H\left(X \mid N_{\mathbf{L}}^k, N_i^k, N_j^k\right) = \frac{\binom{N_i^k}{X}\binom{N_{\mathbf{L}}^k - N_i^k}{N_j^k - X}}{\binom{N_{\mathbf{L}}^k}{N_j^k}}. \tag{1}$$

Equation (1) allows us to associate a p-value $p(N_{i,j}^k)$ with the actual number $N_{i,j}^k$ of neighbors that elements i and j share, i.e.,

$$p\left(N_{i,j}^k\right) = 1 - \sum_{X=0}^{N_{i,j}^k - 1} H\left(X \mid N_{\mathbf{L}}^k, N_i^k, N_j^k\right). \tag{2}$$

Since both each pair of students and links between each pair of students i and j in different subsystems that with different k need to be tested, we are performing the multiple hypothesis testing. Thus we use the very conservative Bonferroni correction, i.e. setting the level of statistical significance as $s = 0.01/N_t$ for multiple hypothesis testing, where N_t is the total number of tests. It is obvious that the setting of s is dependent on N_t which arises from the number of nodes in the network. Thus the level of statistical significance is fully determined by real data. In this way, we avoid subjective determination of co-occurrence threshold and the method can be easily extended to a more general case.

If $p(N_{i,j}^k) < s$ then there's a link between student i and j for the specific subsystem $\mathbf{B}_k$. We then summarize all validations obtained in the projected adjacency network and associate

with the link between i and j a weight equal to the total number of subsystems $\mathbf{B}_k$s in which the relationship between i and j has been statistically validated. If the weight of a link turns out to be zero then the link is removed. The resulting network is the aimed statistically validated network. Note that during the validation, if two students pass the statistical test in any of the subsystem, they are considered as a 'meet', where the number of 'meets' (edge weights) are ignored. Nevertheless, we have shown in the following analysis that the unweighted networks alone offer rich information to examine the associations with academic failures.

Students who either had no consumption records or had never 'met' others frequently enough to pass the statistical validation are obviously absent or isolated in the network. The number of this kind of students is basically very small, they are neglected since most existing studies in social networks focus on individuals who are observable with available data or belong to the largest component of network.

Of course the obtained statistically validated network depends on the way we set the statistical threshold s. We had examine a less restrictive p-value correction methods called False Discovery Rate (FDR) as suggested by (Tumminello et al. [21]), however, the network topological indicators from FDR validation are barely significant in regression analysis, implying the validated networks contain too much noise to approximate the true social interactions. We therefore limit ourselves to results from validation based on Bonferroni correction.

3 Results

3.1 The patterns of academic failures

Courses grades in Chinese university usually range from 0 to 100, and those below 60 are considered *failed*. Since the requirement of a bachelor degree is passing exams to earn necessary credits, a student failed a course is thus considered as a *failure*. Figure 1(a) describes the long tail effect of the failure of students: most students who failed only failed

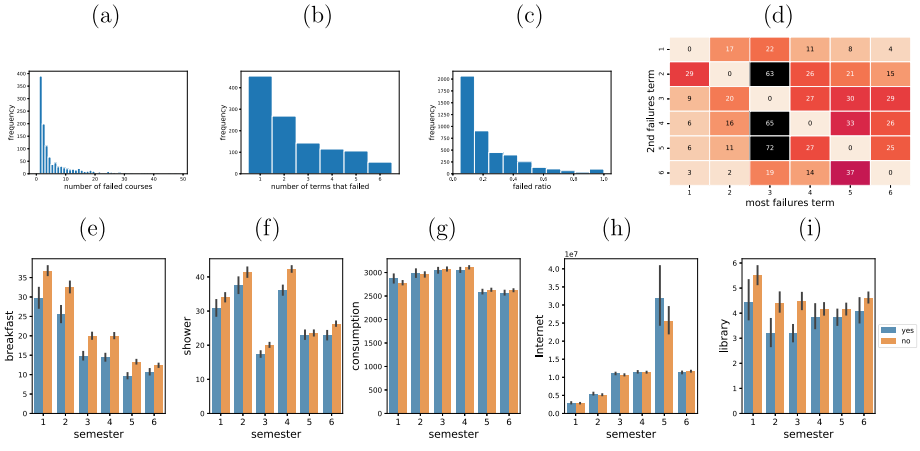

Figure 1 (**a**) The distribution of the number of failed courses of students. (**b**) The distribution of the number of terms that students have failed courses. (**c**) The distribution of the ratio of the number of failed courses to the corresponding number of courses for each student. (**d**) The joint frequency of terms for student having the largest number and the second largest number of failed courses among six terms. (**e**)–(**i**) The heights of bars represent the average levels for these indicators. Error bars (black) show the 95 percent confidence intervals of mean estimation. The blue bars labeled by 'yes' represent failed students and the orange bars labeled by 'no' represent non-failed students

once, whereas a minority failed students have a large number of courses that failed. Additionally, while some students failed in only one term, others would fail in more than one term and even failed in every term in the observation time window, see Fig. 1(b). Figure 1(c) further demonstrates that the distribution of ratios of failed courses to courses student took in each term, again showing that there are some students who had severe problems in passing exams. Although these students are all considered as failed in their academic records, the difference of the number of failed courses explicates the variety of "failure" in academic performance, offering strong endorsement for the number of failed courses to be one of the focuses in the following analysis. What's more, for those who failed in more than one term, it is most likely that the term they failed the most and the second are consecutive, see Fig. 1(d). This implies momentum in academic failure, which also cannot be neglected in profiling determinants of academic failures and similar observations in success of artists, film directors and scientists have been documented in Liu et al. [22].

Exploiting the longitudinal nature of our data, we investigate how the two kinds of students–students who failed and who did not–behave on campus in every semester. In particular, we compare the average of the number of breakfasts, the number of showers, total consumption, total time spent on Internet and the number of books borrowing from library, see Fig. 1(e)–(i). Results show that students who did not fail have higher numbers of breakfasts, showers and books borrowing than those who fail courses over six semesters, indicating that they are more likely to keep daily routines and diligent in studies, in sync with prior works (Cao et al. [23], Yang et al. [11], Lu et al. [20]). It is also worth mentioning that failed students and non-failed students have similar frequency of consumption in campus (see Fig. S1 (Additional file 1)), implying that building co-location network is likely to capture their campus activities of both kinds of students, contrary to the intuition that a failed student is usually recognized as a more indoorsy person. The lack of difference between the two groups in total consumptions and time spent on Internet suggests these two indicators might not be determinants in academic performance. Nevertheless, we use these behavior factors as control variables in the following regression models.

3.2 Emergence of co-location social network

The idea of inferring social ties is based on the fact that friends often have meals and shop together, and the chances of friends should be significantly larger than that of strangers appearing at the same smart card sensor machine simultaneously. We define two students "meet" each other based on students' co-occurrence at the same location within a small time interval. We first build bipartite network to model such co-occurrence from the digital records of daily activities. Let us consider the bipartite network $\mathbf{B}$ in which links connect the $N_{\mathbf{L}}$ elements of location set $\mathbf{L}$ to the $N_{\mathbf{S}}$ elements of student set $\mathbf{S}$. If one location l is visited by at least two students (node s_i and s_j) between time t and $t + \Delta t$, then location l at time t is seen as a node (node l_t), and indirect edges between s_j and l_t as well as s_i and l_t are added, shown in Fig. 2(a). While different locations are regarded as different nodes, the same location at different time are also viewed as different locations from a rigorous spatial temporal perspective. We find that the co-occurrence within a large Δt is more likely to bear random "meets" into the bipartite network, and an overly small Δt may cause a network that is too sparse to capture the stable social connections. We therefore set $\Delta t = 60$ seconds in this paper. Results for $\Delta t = 90$ seconds are very similar (see Fig. S2).

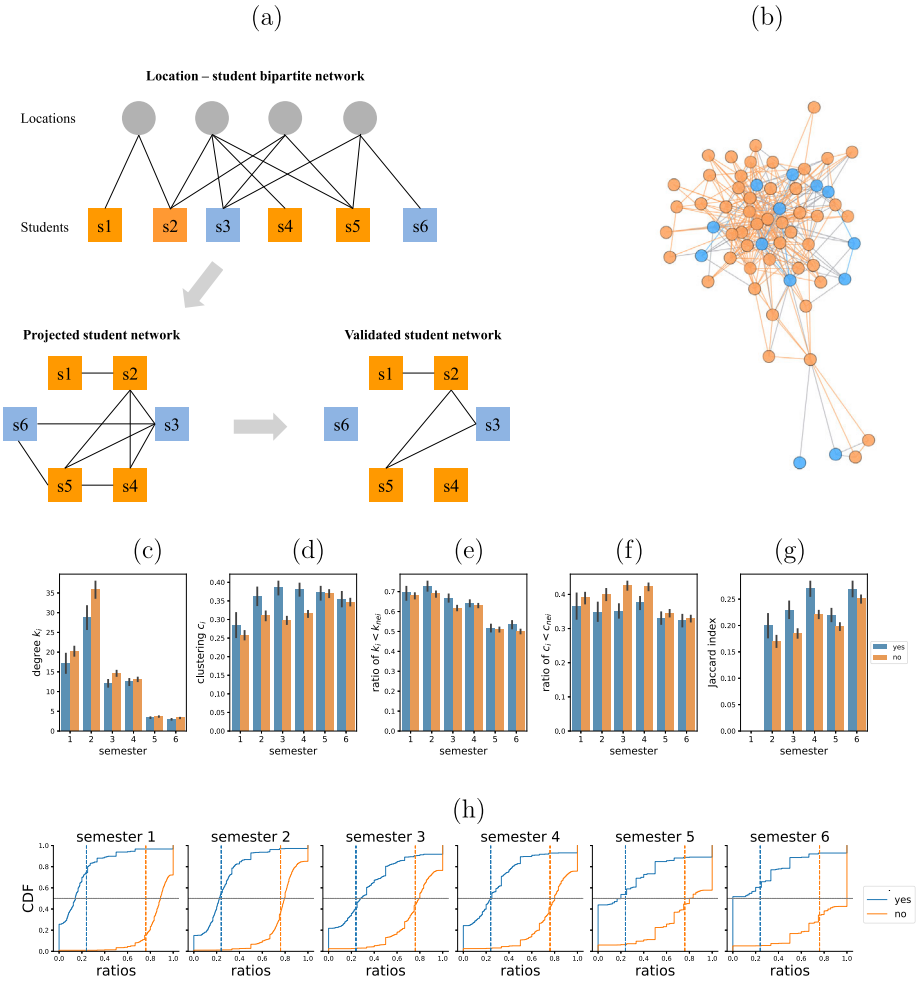

Figure 2 Co-location social network structures between students who failed and those who did not. (**a**) Network modeling procedure. (**b**) An example of inferred students network. For visualization, only one student with her/his neighbors are shown in the graph. Failed students are in blue and non-failed students are in orange. Edges in orange and blue indicate homogeneous connections and those in grey indicate heterogeneous connections. (**c**)–(**g**) The heights of bars represent the average levels for these indicators. Error bars (black) show the 95 percent confidence intervals of mean estimation. Jaccard index of a student in X is defined as the number of overlapping edges of his ego-network between two consecutive semesters divided by the average of number of edges in these two semesters, which measures the stability of a student's connections in different snapshots of campus network. (**h**) The blue line shows the CDF of $ratio_{\mathbf{yes}}$. The orange line shows the CDF of $ratio_{\mathbf{no}}$. The dashed lines are benchmark ratios. The blue dashed line is $ratio_{\mathbf{yes-baseline}}$. and the orange one is $ratio_{\mathbf{no-baseline}}$. The horizontal grey dashed line is where the CDF is equal to 0.5

The location-student bipartite network is then projected into a student co-location social network. The most straight way to fulfil this is to link students that have at least one common neighbor. However, sharing common neighbors can only indicate that they did "meet" rather than "know" each other. While we have tried to naively define students "know" each other according to a number of common neighbors greater than a threshold, implying they met repeatedly, we found that the projected networks are too sparse. We then turn to statistical validation method that aims to statistically infer each link of the projected network against a null hypothesis of random co-occurrence of common neighbors. Specifically, we adopt the validation method from (Tumminello et al. [21]) to com-

plete the projection, as mentioned in Sect. 2.2, since it takes into account the degree heterogeneity of elements of both locations and students, which makes it a general method for the projection of different bipartite networks. The method calculates the statistically significance of two students sharing the same neighbors in real bipartite network and then compares it with the corresponding statistically significance under the assumption of random bipartite network. If the former is larger, it reveals that there're high possibilities that the two students are acquaintance rather than strangers that randomly visit the same location at the same time. More details have been provided in Materials and methods. Figure 2(a) illustrates the network modeling procedure and Fig. 2(b) presents an example of subgraph of the validated students network. The basic information of the networks for six semesters are shown in Table S1. Although we conduct analysis for every semester, the absolute difference of focal indicators over time is not the main focus of this paper. Instead, we examine whether the variables of interests are consistently associated with academic failures. The validated networks are shown to possess the exclusive properties of social networks, i.e., positive assortativity coefficient and a far higher degree of clustering than the corresponding random model (Newman and Park [24]), as shown in Table S2, which in turn demonstrates the reliability of this co-location social network modeling procedure with respect to capturing offline social connections among students.

Except individual factors, like students' intelligence (Deary et al. [25]), their willingness to work hard (O'Hare [26]), and their socioeconomic background (Heckman [27]), that may determine academic performance, social science research has shown that access to social support is also an important factor in academic success (Goddard [28]), and that individuals' positions in social networks can explain success in many domains of life (Granovetter [16], Seibert et al. [2], Burt [29]). As such, we consider *degree* and *clustering coefficient* (or *clustering*, hereafter we use the two terms interchangeably) as individual-level indicators to summarize students' local structures in a social network. Degree is the number of ties of a node in a network, which gives an overall centrality of students from a systematic perspective, taking both weak ties and strong ties into account. The impact of the number of ties is that more ties per node lead to high probability of information exchange, where students communicate and learn from each other. Clustering is a measure of the extent to which nodes in a graph tend to cluster together, assessing the density of a student's ego network from a local perspective. A high clustering can render a node to receive communication from network members, but may also imply information redundancy since a small groups of nodes are connected and have access to the same information (Muller and Peres [30]).

We observe that students who failed had lower degrees but higher clustering from Fig. 2(c) and (d), respectively. Figure 2(e) and (f) then investigate the homogeneous connections of students in terms of connecting to those who have larger degrees and clustering coefficients than themselves. It is shown that the failed students have a bit higher potential to link with nodes with larger degrees and have a lower potential to link with nodes with greater clustering coefficients than non-failed students. That is to say, compared with non-failed students, failed students have smaller degrees and greater clustering coefficients than their neighbors, suggesting they prone heterogeneous linkages in co-location social networks. Besides, failed students' connections with others are more likely to be kept in the following term than non-failed students', see Fig. 2(g). Figure 2(h) further demonstrates the different characteristics of failed students and non-

failed students in homophily. Homophily, the preference to be near others who are like you, is widely discussed in social network studies (Borgatti et al. [1], Bednar [31]). Here in the co-location networks, we examine the extent to which the students connect to others with the same academic performance, i.e., passing or failing the exams. Specifically, suppose student i is a failed student, the ratio of his/her failed neighbours in the network is $ratio_{\text{yes}}$ = # failed students i connects/# all i's neighbours. Suppose student i is a non-failed student, then the ratio of his/her non-failed neighbours in the network is $ratio_{\text{no}}$ = # non-failed students i connects/# all i's neighbours. Each student has one ratio calculated and Fig. 2(h) compares the cumulative distribution function (CDF) of $ratio_{\text{yes}}$ (blue) with the CDF of $ratio_{\text{no}}$ (orange). The two dashed lines are the ratios of failed-students and non-failed students in the network, i.e., $ratio_{\text{yes–baseline}}$ = # failed students in the network/# all students in the network (blue), and $ratio_{\text{no–baseline}}$ = # non-failed students in the network/# all students in the network (orange). These two ratios are not equal to 0.5 as there are overall less failed students than non-failed students. However, they reflect the expected probability for a student to connect to failed students or non-failed students if this student has no preference in connecting to the two groups. These two benchmark ratios are regarded as a result of the randomization of students' connections, when there's no difference between failed students and non-failed students in connecting to others. We therefore compare them with the individual ratios calculated from the co-location networks. Since the majority of students are non-failed students, a student, whether he/she failed or not, tends to interact with non-failed students because very few students, relative to the non-failed ones, failed. It is indeed the case that the CDF of $ratio_{\text{yes}}$ quickly grows while the CDF of $ratio_{\text{no}}$ slowly grows as the ratio increases in all semesters. However, when looking at the dashed lines, it is shown that the majority of failed students' ratios of homogeneous connections (blue lines) are greater than the expected level for a student connecting to failed students (dashed blue lines), indicated by the fact that most intersections between CDFs of $ratio_{\text{yes}}$ (blue lines) and the expected ratios (blue dashed lines) are larger than 0.5 on the y-axis (grey dashed line). Although semester 3 is an exception, the intersection does not deviate too much from 0.5. In addition, the majority of non-failed students' ratios of homogeneous connections are higher than the expected level of a student connecting to non-failed students, indicated by the intersections between the CDFs of $ratio_{\text{no}}$ (orange lines) and the expected values (orange dashed lines) are less than 0.5 on the y-axis. Therefore, Fig. 2(h) indicates that most failed students have lower homogeneous connection ratios than the random configuration. The homophily among failed students are thereby relatively less than that among non-failed students. Stadtfeld et al. [6] pointed out that social relations like positive interaction and friendship can be beneficial. However, even though the way we construct students' social network cannot guarantee the connections are positive, our results cast the doubt on the peer effects because failed students are more likely to interact with non-failed students, and yet there's no evidence that they gain positive influence from these non-failed students, especially considering the existing of momentum in academic failures (see Fig. 1(d)). We will further examine the association between students' co-location social interactions and their course failures by using linear regression models in next section.

3.3 Local structures in co-location social networks are associated with academic failures

Despite the blunt comparison through data visualization, whether the network structures matter for academic failures need further investigation. As such, we explore the failure factors of students in negative binomial regression and logistic regression models, where the response is the number of failed courses and whether or not a student had any failed course in a semester, respectively. While the logistic regression only divides students into failed group and non-failed group, the negative binomial regression takes a step further to differentiate how many courses a student failed. Therefore, the logistic regression tells the relationships between interested variables and status of failure. Meanwhile, the negative binomial regression evaluates failure factors with respect to the frequency of failure events. As controls, we include basic demographic variables (age, gender) and behavioral variables (the numbers of having breakfasts, showers, total consumptions, books borrowing and time spent on Internet). To access whether co-location social network structures are useful to describe academic performance, we put degree and clustering, two variables that have already been shown to distinguish out failed students (Fig. 2), into regression models. The correlations between these input variables are shown in Fig. S3.

Figure 3 presents the results of the two regression models over six semesters. The number of breakfasts is an important explanatory variable in both models. The estimated coefficients are below zero, which implies that the higher the number of having breakfast the lower the probability of failures is. Similarly, the number of showers is mostly negative and significant in logistic regression but not in negative binomial regression, suggesting it might be able to discriminate academic failures but not how many failures. Besides, the

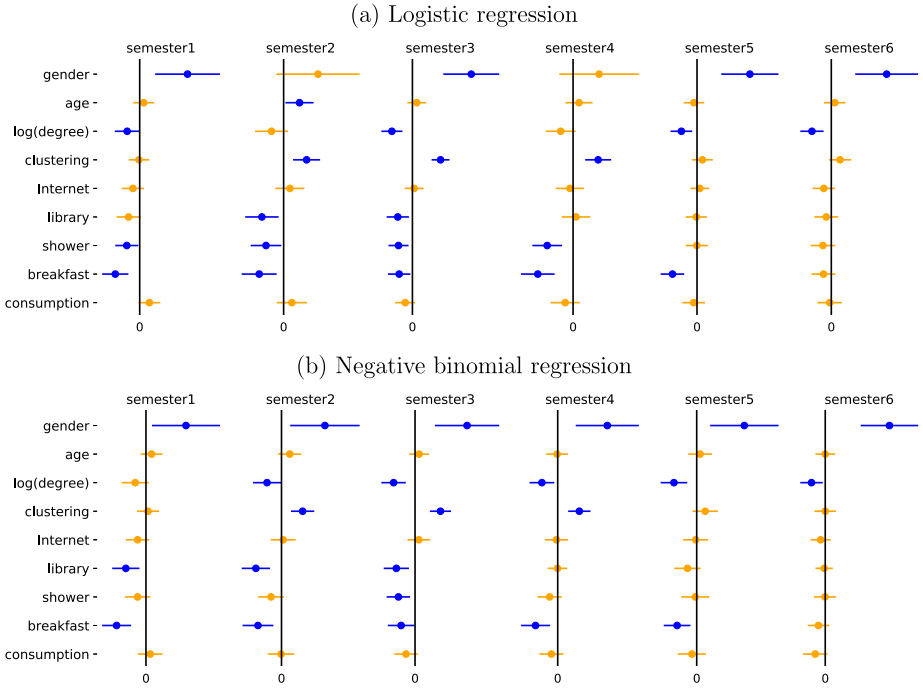

Figure 3 Individual and network factors explaining failure. The dots are coefficients estimates, with bars indicate 90 percent confidence interval. The bars that cover zero, colored in yellow, refer to insignificance estimates using t-test under 0.1 significance level, while those in blue refer to significant estimates

number of books borrowed from library are significantly negative in the first three terms. Although the patterns throughout the six semester vary a bit, the overall significance of the three behavioral variables highlights their capacity in explaining academic failures.

The potential importance of social network is highlighted by the significance of log(degree) and clustering. Soffer and Vazquez [32] pointed out that degree and clustering coefficient are strongly correlated with each other, thereby possibly yielding the multicollinearity for regression model. Here since we use log(degree) instead of degree considering the distribution of degree is highly right skewness, the correlations between the two network attributes are smaller than 0.31. Further robustness check experiments also verify that the significance and signs of regression coefficients when putting them separately in both logistic and negative binomial models stay the same with Fig. 3.[1] Specifically, the negative signs of regression coefficients for log(degree) demonstrate students who have many connections in social network prone to academic success. On the contrary, the positive signs of regression coefficients for clustering reveal those who lie in a small but connected part of social network tend to not only fail but also fail more courses. Despite the panel attributes of the six semesters, we argue that the co-location social networks arose from off-line campus behaviors in each term are independent. In either regression model, clustering is significant from the second to the fourth semester. This is probably because students in these semesters, compared with the first semester when students just enrolled in university and the last two semesters when students stepped into senior year, are more likely to form stable and local within-campus linkages in co-location social networks. Figure S1 also implies that the students' number of smart-card records declines in the last two semesters, indicating the potential of campus activities' reductions. Therefore, it should be noted that the associations between clustering in co-location social networks and academic failures alter with the heterogeneity of semesters. To be more specific, the associations between co-location social networks in campus and academic performance are not of significance in all semesters, especially for those of senior years. It is possible that the social network of students might evolve with terms going, in particular, the out-campus parts eventually grow and dominate. Because of the data unavailability, the present study only focused on the parts within campus, however, it would be interesting to exploit these associations from the perspective of out-campus in future work. Figure S4 gives the robustness check of regression analysis when taking the number of failed courses in the previous semester, reflecting historical individual academic performance, into account. Not surprisingly, the regression coefficients of this variable are significantly positive, implying academic failures indeed have momentum, in line with existing findings in Tamhane et al. [33]. Despite of this, the significance of clustering emphasizes the fact that the associations between local social structures and academic failures are potent from the perspective of co-location social networks in campus.

Overall, we found no evidence of social contagion, i.e., failed students are influenced by their non-failed neighbors. The reasons might be that failed students resist to be included in broader social relationships, resulting in a high clustering attributes. On one hand, their connections might not be effective influential linkages in terms of academic performance. On the other hand, a high clustering indicates that they might possess too much redundant social connections to inject substantial academic influence.

[1] The results are upon request from authors.

3.4 The failures of students can be predicted at early stages

The findings of significant variables in regression analysis provide important support to establish predicting models for identifying failed students at an early stage so that interventions can be introduced to avoid such failure. To understand the usefulness of our input features for prediction, we construct two benchmark models in which the predictor set is deliberately simplified. The first benchmark model (benchmark 1) drops the two network-related variables, i.e., degree and clustering, from our predictor set in order to identify the role of co-location network by comparing it with our model. The second benchmark model (benchmark 2) only uses gender and age, two very basic variables, to detect the role of all behavioral variables extracted from smart-card data. Since the non-failed students are much more than failed students, we first divide both groups of observations into five non-overlapping parts with equal size, where four parts are regarded as train set and the other one is test set, respectively. We then adopt upsampling for the failed students with replacement in train set to guarantee that the number of non-failed students equals to the number of failed students, based on which the logistic regression model is built. We calculate accuracy (acc) and Area under the ROC curve (AUC) of this model on test set, where the ratio of the number of failed students to the number of non-failed students is consistent with the original dataset, ensuring the evaluation on predictive power of the model meaningful in realistic scenarios. Results are shown in Fig. 4.

Not surprisingly, the predicting ability improves when the time span increases, however, it is worth noticing that information in the first 91 days has already offer as competent predicting accuracy as that in the model based on data of the whole semester. This in turn suggests that inferring social networks from campus activities does not necessarily require data of a whole term. In particular, the significance of clustering is spotted throughout the term (see Fig. S5–S6), which also supports the opinion that failures of students can be predicted in advance.

In terms of both accuracy and AUC, it is shown that our model performs better than two benchmark models. The significantly worse performance of benchmark 2 demonstrates the behavioral information, arose from smart-card data, is indeed of value for academic performance prediction. Our model is also found to be a bit better than benchmark 1, implying that the messages delivered by the co-location social networks also help prediction. Meanwhile, better predictive powers are found in semesters that have more signifi-

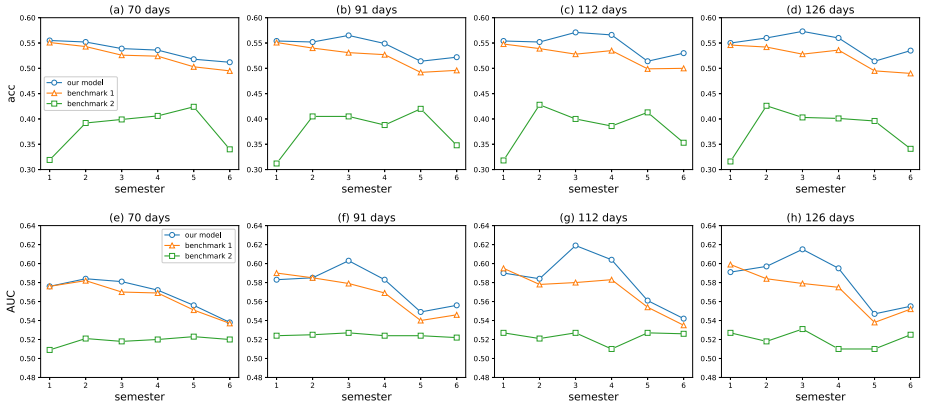

Figure 4 Comparisons on model prediction

cant variables in regressions shown in Fig. 3. With all the above discussions, however, it is worth noticing that the early signal is not very strong. This sheds lights on the difficulty of predicting academic failures by behavioral features that only originate from campus activities. Besides, the smart-card data, though provide abundant details of students' traces in campus, can only capture occurrences at registered places. We have further discussion on this limitation in the next section.

4 Discussion

In this paper, we unveil the co-location social structures beyond failures. Our results highlight that the scale of co-location social relations indicated by node degree may not be always significantly beneficial as expected, while the local density represented by node clustering are shown to have a crucial impact on individual academic failure. The main contribution of the current investigation is to provide empirical evidence regarding the co-location social network related determinants of failures. Studies on success have shown that social network indeed benefits success. For instance, social network provides access of information and resources that enhance work performance, power, reputation, and empowerment in one's career (Seibert et al. [2]). But whether failure is driven by the opposite logic to success is doubtful. Ansari et al. [4] pointed out a cohesive ego network where members are richly connected contribute to online success of music artist in long-term, which to some extent is contrary to our finding that clustering is positively related to failures. This again sheds lights on the fact that since the opposite of a "success" is not necessarily always a "failure", implications derived from success can not be directly generalized to the understanding of failures.

What's more, we find that the failed students lie in small but dense corners of co-location social networks and their connections with others are more likely to be kept in the following terms than non-failed students' (see Fig. 2). As mentioned above, the failed students are more deeply embedded in local structures revealed from high clustering coefficients. The higher constancy of their connections thus highlights the stableness of local structures. Though these local ties seemingly connect failed students to those who didn't fail and are stable over time, we found that they rarely help failed students and the failure even demonstrates a momentum. Indeed, as Stadtfeld et al. [6] have pointed out, not all connections are positive social connections that could evolve into collaboration ties to promote academic performance, especially here the connections of network originate from campus behavior activities. The failed student may act together and build stable relationships with the non-failed students, but these relationships do not seem to bring positive changes to academic failures. The lack of social contagion, in the presence of low homophily and high clustering, casts doubt on the explanation of individual's outcomes as a function of her social environments through influence processes or leveraging processes (Borgatti et al. [1]) in the context of academic failures.

The comparison of changes in node degree and clustering from mid-term period to end-term period suggests the behavioral adaptation of failed students is much weaker than non-failed students. Specifically, Table 1 shows the t-test for the means of changes of degree (or clustering coefficient) over the semester for students who failed and who did not, respectively. It is found that the changes of degree for students who failed and who did not are indeed different, at least in the first four terms. Those who didn't fail tend to change a bit more in co-location social interactions at the end of a term. The clustering coefficients, on the contrary, exhibit no significant difference between the two populations,

Table 1 T-test for the means of changes of degree (or cluster coefficient) for students who failed and who did not

var	Term	End days-start days	Mean of changes		Test statistic (p-value)
			Students who fail	Students who did not fail	
Degree	1	126-70	9.56	11.78	−2.25(0.02)
	1	126-91	4.81	6	−2.13(0.03)
	1	126-112	1.12	1.42	−1.31(0.19)
	2	126-70	15.77	19.94	−3.68(0.0)
	2	126-91	8.02	10.5	−4.0(0.0)
	2	126-112	1.21	1.46	−1.26(0.21)
	3	126-70	5.95	8.4	−5.21(0.0)
	3	126-91	3.16	4.88	−5.82(0.0)
	3	126-112	0.91	1.47	−4.31(0.0)
	4	126-70	4.72	5.53	−2.1(0.04)
	4	126-91	2.08	2.59	−2.25(0.02)
	4	126-112	0.27	0.46	−1.98(0.05)
	5	126-70	0.93	1.07	−1.65(0.1)
	5	126-91	0.51	0.59	−1.35(0.18)
	5	126-112	0.12	0.13	−0.39(0.7)
	6	126-70	0.79	0.86	−0.92(0.36)
	6	126-91	0.37	0.39	−0.4(0.69)
	6	126-112	0.1	0.11	−0.27(0.79)
Clustering	1	126-70	0.03	0.03	−0.14(0.89)
	1	126-91	0.02	0.02	−0.49(0.62)
	1	126-112	0.01	0.01	0.21(0.84)
	2	126-70	0.05	0.04	0.54(0.59)
	2	126-91	0.03	0.02	0.69(0.49)
	2	126-112	0.01	0	0.58(0.56)
	3	126-70	0.06	0.02	2.63(0.01)
	3	126-91	0.02	0.02	0.25(0.8)
	3	126-112	0.01	0	0.39(0.69)
	4	126-70	0.05	0.03	1.75(0.08)
	4	126-91	0.02	0.01	1.2(0.23)
	4	126-112	−0.01	0	−1.4(0.16)
	5	126-70	0.02	0.03	−0.76(0.45)
	5	126-91	0.03	0.02	0.42(0.67)
	5	126-112	0.01	0.01	−0.21(0.83)
	6	126-70	0.08	0.05	1.6(0.11)
	6	126-91	0.06	0.04	1.39(0.16)
	6	126-112	0.01	0.01	−0.43(0.67)

We have also built co-location social networks at different time points, including the first 70, 91, 112 days of each term. The changes of degree and cluster coefficients for nodes from these time points to the final snapshot (the end of a semester, 126 days) are then calculated to roughly indicate the evolution of social interactions. This is a two-sided test for the null hypothesis that two independent samples have identical average (expected) values. In the test, we first test using Levene test to see if the populations have identical variances by default. Different t-tests are then applied according to the result of Leven test when populations have either identical variances or different variances.

implying the local structures of nodes in co-location networks are most likely to be stable throughout a term. The results imply that the adaptation mechanism, which nodes become homogeneous as a result of experiencing similar social environments, only occurs to non-failed students. The failed students, again, with most connections to non-failed students, haven't developed similar adapting behaviors as their neighbors have. This echoes with our finding that when failed students who locate in structures where redundancy is high (high clustering), which usually suggests information tends to remain within clusters, the expected effect is not transmitted from non-failed students to failed students.

Our findings have important instruction meaning for detecting academic failures and providing personalized intervention and guidance at the early stage. Maintaining consistency with previous research on the spread of academic achievements, the nature of social

relationships is crucial for the transmission of specific types of information and behavior in social networks (Dokuka et al. [8]). To reduce the risk of failure, a low clustering with high degree for a student's ego co-location social network is preferred. Centola [34] had experimentally shown that high clustering causes a faster behavior spread. Unfortunately, we find that success is not easy to spread even with high clustering in terms of academic achievement. Therefore, there should be precautions to stimulate the positive effects from non-failed students to failed students, which requires a social network that guarantees positive social contagion instead of the one obtained from daily routine with limited social influence. Note that our discussions focus on academic failures, the methodology and results could be extended to other scenarios and integrating and examining these findings into more fields is also of great interest in the future.

The limitation of this study lies in the fact the co-location social networks are inferred based on high-frequency co-occurrences in smart-card system. While not all places in campus are registered into this system, the obtained co-location social networks inevitably face the risk of being a biased proximity of real social networks. Nevertheless, our data have largely covered most of daily routines in campus, and we have endeavoured to minimize the perplexity from off-campus activities through focusing on the first three years of undergraduate life, when students normally stay in campus to take courses. Additionally, similar to that in Godoy-Lorite and Jones [35], when it is extremely costly, or even impossible, to have direct access to the students' social networks, we do not seek to infer individual links with high accuracy but to seek a behavior-specific network model which can be informative of social network structures and more importantly, is relevant to academic performance. Besides, although we focus on the association between network characteristics and academic performance, we are not suggesting that the network is the reason behind their performance. Instead, our network construction method based on co-occurrence in smart-card system potentially captures the behavioral differences between failed and non-failed students, which are fundamentally part of reasons on why they failed. However, it is hard for us to locate the posited mechanism without intervention. With more data available, understanding this mechanism would be a promising direction in future explorations. Another limitation is the issue of missing control variables in regression analysis. When more variables such as study time, motivation, high-school GPA and etc. are available in data, a systematic analysis is expected to offer additional insights on the present topic.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-022-00322-0.

Additional file 1. Supplementary information (PDF 317 kB)

Acknowledgements
The authors also thank Mr. Yang Yang for his valuable suggestions on prediction models.

Funding
This research was financially supported by the National Natural Science Foundation of China (Grant Nos. 72001222 and 71871006). SL also thanks the support from the Emerging Interdisciplinary Project of CUFE, the Program for Innovation Research, and the disciplinary funding in Central University of Finance and Economics. JCZ also thanks the support from educational reform project of School of Economics and Management, Beihang University (Grant No. 430271).

Abbreviations
AUC, Area under the ROC curve; acc, accuracy; FDR, False Discovery Rate; GPA, Grade point average; ROC curve, Receiver operating characteristic curve.

Availability of data and materials
Data are upon request from the authors.

Declarations

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SL: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. JZ: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. HW: Conceptualization, Writing – review & editing. All authors read and approved the final manuscript.

Authors' information
Shan Lu received her Ph.D. degree in Statistics in 2019 and B.S. degree in Industrial Engineering in 2014 from Beihang University. She is currently an associate professor at School of Statistics and Mathematics, Central University of Finance and Economics. Her research interests include complex data analysis, machine learning and network science. She is a member of Financial Sustainable Development Research Team in Central University of Finance and Economics.
Jichang Zhao received his B.E. and Ph.D. degrees from Beihang University in 2008 and 2014, respectively. He is currently an associate professor at the School of Economics and Management, Beihang University. He is also a senior research scientist at the Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC). His research interests include computational social science and complex systems.
Huiwen Wang received her B.Sc. degree from Beihang University (BHU), China, in 1982, DEA MASE, from Paris XI, France, in 1989, and Ph.D. degree in Engineering System from BHU in 1992. She is currently a professor in School of Economics and Management (SEM), Beihang University. Also, she is director of SEM Academic Degrees Committee, and director of Research Center of Complex Data Analysis in BHU. Prof. Wang received National Science Fund for Distinguished Young Scholars. Her general area of research is statistics and data analysis, with a recent focus on multivariate analysis for high-dimension complex data. She is an IASC member, a member of National Statistics Teaching Materials Review Committee, executive director of China Marketing Association, editorial member of Journal of Symbolic Data Analysis.

Author details
[1]School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China. [2]School of Economics and Management, Beihang University, Beijing, China. [3]Key Laboratory of Complex System Analysis, Management and Decision (Beihang University), Ministry of Education, Beijing, China.

Publisher's Note

Received: 23 June 2021 Accepted: 9 February 2022 Published online: 23 February 2022

References
1. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895
2. Seibert SE, Kraimer ML, Liden RC (2001) A social capital theory of career success. Acad Manag J 44(2):219–237
3. Nanda R, Samila S, Sorenson O (2020) The persistent effect of initial success: evidence from venture capital. J Financ Econ 137(1):231–248
4. Ansari A, Stahl F, Heitmann M, Bremer L (2018) Building a social network for success. J Mark Res 55(3):321–338
5. Aral S, Walker D (2011) Creating social contagion through viral product design: a randomized trial of peer influence in networks. Manag Sci 57(9):1623–1639
6. Stadtfeld C, Vörös A, Elmer T, Boda Z, Raabe IJ (2019) Integration in emerging social networks explains academic failure and success. Proc Natl Acad Sci 116(3):792–797
7. Yin Y, Wang Y, Evans JA, Wang D (2019) Quantifying the dynamics of failure across science, startups and security. Nature 575(7781):190–194
8. Dokuka S, Valeeva D, Yudkevich M (2020) How academic achievement spreads: the role of distinct social networks in academic performance diffusion. PLoS ONE 15(7):e0236737
9. Junco R (2013) Comparing actual and self-reported measures of Facebook use. Comput Hum Behav 29(3):626–631
10. Kassarnig V, Mones E, Bjerre-Nielsen A, Sapiezynski P, Lassen DD, Lehmann S (2018) Academic performance and behavioral patterns. EPJ Data Sci 7(1):10
11. Yang Z, Su Z, Liu S, Liu Z, Ke W, Zhao L (2020) Evolution features and behavior characters of friendship networks on campus life. Expert Syst Appl 158:113519
12. Credé M, Roch SG, Kieszczynka UM (2010) Class attendance in college: a meta-analytic review of the relationship of class attendance with grades and student characteristics. Rev Educ Res 80(2):272–295
13. Poropat AE (2009) A meta-analysis of the five-factor model of personality and academic performance. Psychol Bull 135(2):322
14. Wald A, Muennig PA, O'Connell KA, Garber CE (2014) Associations between healthy lifestyle behaviors and academic performance in US undergraduates: a secondary analysis of the American college health association's national college health assessment II. Am J Health Promot 28(5):298–305

15. Coleman JS (1990) Foundations of social theory. Harvard University Press, Cambridge
16. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
17. Fei M, Yeung D-Y (2015) Temporal models for predicting student dropout in massive open online courses. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), IEEE, Los Alamitos, pp 256–263
18. Zafra A, Romero C, Ventura S (2011) Multiple instance learning for classifying students in learning management systems. Expert Syst Appl 38(12):15020–15031
19. Thai-Nghe N, Drumond L, Horváth T, Schmidt-Thieme L et al (2011) Multi-relational factorization models for predicting student performance. In: KDD Workshop on Knowledge Discovery in Educational Data (KDDinED). Citeseer, pp 27–40
20. Lu S, Zhao J, Wang H (2021) MD-MBPLS: a novel explanatory model in computational social science. Knowl-Based Syst 223:107023
21. Tumminello M, Micciche S, Lillo F, Piilo J, Mantegna RN (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6(3):e17994
22. Liu L, Wang Y, Sinatra R, Giles CL, Song C, Wang D (2018) Hot streaks in artistic, cultural, and scientific careers. Nature 559(7714):396–399
23. Cao Y, Gao J, Lian D, Rong Z, Shi J, Wang Q, Wu Y, Yao H, Zhou T (2018) Orderliness predicts academic performance: behavioural analysis on campus lifestyle. J R Soc Interface 15(146):20180210
24. Newman ME, Park J (2003) Why social networks are different from other types of networks. Phys Rev E 68(3):036122
25. Deary IJ, Strand S, Smith P, Fernandes C (2007) Intelligence and educational achievement. Intelligence 35(1):13–21
26. O'Hare L (2009) Improving academic achievement: impact of psychological factors on education. Ir J Psychol 30(3/4):233
27. Heckman JJ (2006) Skill formation and the economics of investing in disadvantaged children. Science 312(5782):1900–1902
28. Goddard RD (2003) Relational networks, social trust, and norms: a social capital perspective on students' chances of academic success. Educ Eval Policy Anal 25(1):59–74
29. Burt RS (2004) Structural holes and good ideas. Am J Sociol 110(2):349–399
30. Muller E, Peres R (2019) The effect of social networks structure on innovation performance: a review and directions for research. Int J Res Mark 36(1):3–19
31. Bednar J (2021) Polarization, diversity, and democratic robustness. Proc Natl Acad Sci 118(50):e2113843118
32. Soffer SN, Vazquez A (2005) Network clustering coefficient without degree-correlation biases. Phys Rev E 71(5):057101
33. Tamhane A, Ikbal S, Sengupta B, Duggirala M, Appleton J (2014) Predicting student risks through longitudinal analysis. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1544–1552
34. Centola D (2010) The spread of behavior in an online social network experiment. Science 329(5996):1194–1197
35. Godoy-Lorite A, Jones NS (2021) Inference and influence of network structure using snapshot social behavior without network data. Sci Adv 7(23):eabb8762