Student Network Analysis: A Novel Way to Predict Delayed Graduation in Higher Education

Nasheen Nur[1], Noseong Park[2(✉)], Mohsen Dorodchi[1], Wenwen Dou[1], Mohammad Javad Mahzoon[1], Xi Niu[1], and Mary Lou Maher[1]

[1] University of North Carolina at Charlotte, Charlotte, NC, USA
{nnur,mdorodch,wdou1,mmahzoon,xniu2,m.maher}@uncc.edu
[2] George Mason University, Fairfax, VA, USA
npark9@gmu.edu

Abstract. We present a prediction model to detect delayed graduation cases based on student network analysis. In the U.S. only 60% of undergraduate students finish their bachelors' degrees in 6 years [1]. We present many features based on student networks and activity records. To our knowledge, our feature design, which includes conventional academic performance features, student network features, and fix-point features, is one of the most comprehensive ones. We achieved the F-1 score of 0.85 and AUCROC of 0.86.

Keywords: Network analysis · Student data · Risk prediction

1 Introduction

One of major strategic challenges that the U.S. higher education faces is timely completion of degree for college students [2]. Recent data from the National Center for Education Statistics shows that the majority (60%) of full-time undergraduate students take 6 years to earn a bachelor's degree [1]. As a result, higher education is under increasing pressure to demonstrate institutional effectiveness across a range of complicated factors [3]. According to [4], for instance, the U.S. government emphasizes the need of producing successful Science, Technology, Engineering, and Mathematics (STEM) graduates in a timely manner.

We propose a novel network analytic approach to predict at-risk students who fail to complete their degrees on time. Our approach is distinct from others due to the following two features: (1) We predict at-risk students early after 5-th semester; (2) In addition to classical academic features such as GPA and earned credits, we use various data from students' (extracurricular) activities to calculate student network features. We also define another type of features based on the same data, called fix-point features in our paper (see Fig. 1(a)). Throughout the analysis process, we have some interesting observations. At-risk students tend to have many weak connections, rather than a selected small

S. Isotani et al. (Eds.): AIED 2019, LNAI 11625, pp. 370–382, 2019.
https://doi.org/10.1007/978-3-030-23204-7_31

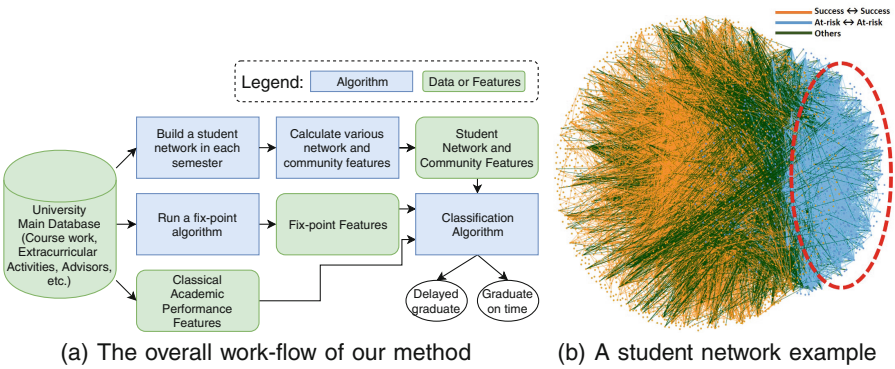

Fig. 1. (a) The overall work-flow of our presented framework. (b) Edge-colored student network after filtering out weak (i.e., edge weight $<90^{\text{th}}$ percentile) connections. Note that many at-risk students inside the dotted red circle do not have strong connections with successful students. (Color figure online)

number of strong connections so their network features (such as degree centrality, ego-network density and so forth) are distinctly different from that of successful students. In the future, we plan to intervene by strategically selecting at-risk students to consolidate their connections with successful students and to enhance their network features. We answer the following research questions in this paper:

1. Does including students' network features help predict at-risk students?
2. Does including students' network features help predict at-risk students in an earlier stage?
3. To what extent are the student network connections of successful students different from those of at-risks students over time?
4. Who are active participants of student communities & who are peripheral participants? How differently do successful and at-risk students behave in communities?

2 Related Work

During the past five years there has been an increase in research for improving learning and educational environments by leveraging analytics and the vast amount of data collected about the interaction of students with learning management systems (LMS). Course Signals [5–8] is an example of a learning analytic tool that not only classifies and identifies students at-risk but also provides interventions to improve student learning based on the analyzed data. This system processes student data from the Blackboard LMS to provide an early warning for students at-risk. Latest trends in learning analytics and knowledge (LAK) also shows a move towards sense-making from broad and general predictive models [9,10]. The LAK community is expanding and including broader

interdisciplinary research to scale up "from big data to meaningful data" [11]. Typical models to better understand student performance and risk are based on classical academic features such as GPA, course withdraw rate, high school GPA, standardized test scores, and so forth [5–8,12,13]. However, models of retention and risk/success analysis often neglect to lead to actionable knowledge [14] whereas some approaches have more focus on analytics that generate actionable knowledge rather than predictions of GPAs, assignment grades, etc.

There are also reports on how to incorporate network analysis to better understand student behavior and interactions [13,15–21]. Conventional network analytic research has generally focused on the deliberate behavior of each individual or groups but neglected the interlinked information between or among individuals or groups [22]. These network models are based on students' LMS or social media logs, e.g., who responds to whom in LMS tools or who likes whom in social media. The purpose of the network analyses includes identifying student groups and social networking behaviors that lead to risk or success [15,18]. For example, Romero [13] investigated interaction patterns among students in their LMS tools and created an unsupervised clustering method to detect course failures. In [18], they analyzed participation patterns in online discussions in order to reveal student clusters with leaders and peripherals. Authors of [21] presented that students' social involvement accumulated through academic activities is positively related to their academic performance. In [23], it was shown that students' co-enrollment networks follow power-law degree distributions and they predict course grades with simple network features whereas we predict a longer-term success with more comprehensive network features. The related works have clearly laid out the functionality of social network analysis and provided guidance for our study.

Our approach is distinct in two aspects: (i) We construct student networks using student records other than interactions recorded in the LMS or social media logs, and (ii) Our success measure is on-time graduation whereas some existing models focus only on course success and GPA.

3 Our Dataset

We collected data from a 13-year period and limited our analysis to undergraduate students who spent 8 or more semesters in our school and have selected computer science as their major at some point in their academic career. We chose on-time graduation as the measure of success, and built predictive models to identify students being at-risk of not graduating in six years. After excluding on-going students who enrolled in the past six years, the total number of students in our analysis is 2,552 where approximately 30% are at risk. We did not use self-reported data such as social media data and LMS logs. The benefit of not using LMS data is that not all Professors use the LMS in the same way so the analysis for all students in a major will not have consistent data. We collected student background information such as demographics and tests taken before admission; academic information such as major, courses taken, transferred courses, and advisers; extracurricular activities and participation in student organizations.

4 Proposed Prediction Method

We design and extract three types of features: academic features, student network features, and fix-point features. The first type is already widely used in the field of learning analytics, but to our knowledge, there are few works using the second and third types. Our main contributions lie in the feature engineering based on student networks and academic activities.

4.1 Academic Features

Academic features such as grade-point average (GPA) are very effective to identify at-risk students and had been used very widely in many works [5–8,12,13]. We adopted features related to personal information (age, citizenship, gender, primary ethnicity, etc.), high-school record (school rank, percentile, etc.), and academic progress (GPA, success rate in earning credits, the number of course withdraws, etc.).

4.2 Student Network Features

How to Build Student Network. We build a weighted student network in each semester. Edge weight value between two students represents the *cumulative intensity* of the connection by the time point we draw the network. Because it is cumulative, their intensity will increase as time goes by. We calculate the edge weight, denoted as $w(x, y, t)$ hereinafter, between two students x and y at a certain semester t as follows:

$$
\begin{aligned}
w(x,y,t) &= \exp\Big(\sum_i rescale(normalize(w_i(x,y,t)))\Big),\\
w_1(x,y,t) &= \sum_{t'\le t}\frac{C(x,t')\cap C(y,t')}{C(x,t')\cup C(y,t')}, \quad w_2(x,y,t) = \sum_{t'\le t} same\_activity(x,y,t'),\\
w_3(x,y,t) &= \sum_{t'\le t} same\_advisor(x,y,t'), \quad w_4(x,y,t) = same\_dept(x,y),\\
w_5(x,y,t) &= same\_major(x,y), \quad w_6(x,y,t) = same\_age\_high\_school(x,y),
\end{aligned}
\tag{1}
$$

where $C(x,t)$ is a set of courses taken by student x at semester t, $w_i(x,y,t)$ in the left-hand side means a cumulative value by t between two students x and y, and $same(x,y)$ or $same(x,y,t)$ in the right-hand side is an indicator function that returns 1 if two students x and y have the same (i) activity, (ii) advisor, (iii) department, (iv) major, or (v) high school with the same age and otherwise 0. Note that we do not consider time for high school record. Others do depend on time. For instance, $w_1(x,y,t)$, inspired from the Jaccard index[1], is to calculate the sum of the common course ratio between x and y until t. After normalization, w_i ranges in $[0,1]$ and after re-scaling, the mean of w_i at t becomes 0.5.

[1] The Jaccard index is a popular node similarity metric in networks based on the number of common neighbors divided by the sum of all neighbors.

Some weights have larger scales than others and may dominate the final weight value without the normalization and re-scaling. We prevent it by standardizing edge weights. In our definitions, thus, *each* w_i *becomes equally important*[2]. Activity describes most student-focused extracurricular clubs, sports, and programs at the college. The exponential function makes strong (i.e., large final weight) connections stronger. Therefore, a student network at t may consist of many edges that have relatively small weights and a small number of edges that have large weights. We draw a student network for each semester t and there are more than 40 networks created from 13 years of student records. In Fig. 1(b), we draw a student network using three edge colors: orange (among successful) and blue (among at-risks) between the same classes and dark green between different classes. Note that many at-risk students do not maintain strong connections with other successful students.

Basic Network Features. In network analysis, *centrality* comprises methods to measure the relative importance of nodes (i.e., students in our case). There exist many different centrality concepts such as degree centrality, closeness centrality [24], clustering coefficient [25,26], betweenness centrality [27,28], PageRank [29], and so forth. Among many, we select centrality measures that have enough discriminatory information to identify at-risk students after some statistical analyses, such as t-test and histograms. For almost all centrality concepts, there exist both unweighted and weighted versions. All centrality metrics used in this paper are weighted, unless otherwise stated.

Community-Based Features. Community detection is a long-standing research topic in network analysis. Sometimes it is used as a subroutine to solve other problems similar to our case [30–33]. We use overlapping community detection methods because one student can join multiple communities. We choose SLPA [34] as our base community detection method, considering its accuracy and popularity. After finding many overlapping communities in each N_t, we calculate the following features:

1. Let $Com(x) = \{Com_1, Com_2, \cdots\}$ be a set of communities that student x belongs to. Finally, we do MIN/MAX/AVG aggregations over the communities in $Com(x)$ for each type of network features(x).
2. In each N_t, a giant component means the biggest community. In many cases, the giant component is one of the most influential student groups and we check if a student is its member. After that, we calculate the ratio of such cases over time. The ratio of 1 for a student means that the student is a stable member of the giant components for his or her entire academic period.

[2] This is a very important fact about our network dentition. We do not focus on only courses but also many other aspects of academic life.

Ego-Network-Based Features. Ego-network(also called node-centric network) means an induced sub-network by one node and its neighbors [35]. From each node's (student's) ego-network, we extract its ego-network density and clustering coefficient [25,26] as network features. The density is formally defined as follows:

$$Density(x,t) = \frac{2\sum_{(a,b)\in Ego(x,t)} w(a,b,t)}{|Nei(x,t)| \cdot (|Nei(x,t)|-1)}, \tag{2}$$

where $Ego(.,t)$ returns an edge set of one's ego-network in N_t and $Nei(.,t)$ means a set of one's neighbors in N_t.

4.3 Fix-Point Features

Given a function $f(\cdot)$, a fix-point x means $x = f(x)$. fix-point calculation is used in various domains. One representative example is the stationary distribution of Markov chain, i.e., $\pi = \pi\mathbf{P}$, where $\mathbf{P}$ is transition matrix. In [29,36–41], authors defined a mutually recursive complex system of variables and their fix-point values are used to understand vertices. Defining a complex variable system requires domain dependent knowledge. We first introduce domain knowledge we gain from our educational experience and available data:

1. We think that courses/activities/undergraduate advisors *simultaneously* taken by many students share common characteristics. Suppose that course A and course B have many overlapping students in a semester. Those two courses may have some common characteristics.
2. A student's characteristic can be described by courses/activities/advisors that she/he had taken [42–44]. In the network features, we analyzed the interactions among students. Our fix-point features describe students from their course/activity/advisor records without considering other students.

Based on those intuitions, we define several variables that are mutually recursive as follows:

$$val(c_i,t) = \sum_{c_j} \frac{\#stu(c_i,c_j,t)}{\sum_{c_k} \#stu(c_k,c_j,t)} val(c_j,t), \quad val(s_i,t) = \sum_{c_j} take(c_j,s_i,t)\frac{1}{\#stu(c_j,t)} val(c_j,t)$$

$$val(a_i,t) = \sum_{a_j} \frac{\#stu(a_i,a_j,t)}{\sum_{a_k} \#stu(a_k,a_j,t)} val(a_j,t), \qquad + \sum_{a_j} take(a_j,s_i,t)\frac{1}{\#stu(a_j,t)} val(a_j,t)$$

$$val(v_i,t) = \sum_{v_j} \frac{\#stu(v_i,v_j,t)}{\sum_{v_k} \#stu(v_k,v_j,t)} val(v_j,t), \qquad + \sum_{adj} take(v_j,s_i,t)\frac{1}{\#stu(v_j,t)} val(v_j,t),$$

where c_i, a_i, v_i, and s_i represent course, activity, adviser, and student, respectively. $\#stu(x,y,t)$ returns the number of students who took two courses, activities, or advisers x and y together at semester t and $take(x,s,t) \in \{0,1\}$ is an indicator variable to denote if course, activity, or adviser x is taken by student s at semester t. Thus, $val(x,t)$ means an influence value each entity x has at semester t. As we construct a network N_t in each semester, these variables are defined for each semester too. We ignore department, major, degree information because they are too broad to be used in the variable definition.

If two courses, activities, or advisers have largely overlapping students, their values will be very similar because of the coefficient based on normalization. A student value is an aggregation of all the course/activity/advisor values so it will be solely decided by the courses/activities/advisers taken by the student (See Algorithm 1). It is an iterative method to update values. We check the convergence only for the student variables because they are what we are interested in. The converged student values are used as additional features. The proof of its convergence is removed in this paper due to space reasons.

4.4 Experiments

The time point of the data is Spring 2004 and the time point is Fall 2016. All students who graduated on or before Spring 2013 are in the train set and others are in the test set. The ratio of train:test is 77:23. We perform the grid search with 10-fold cross validation to find the best model. Many classifiers (including SVM, Random Forest, Decision Tree, AdaBoost, RBM, Bagging, Multi-Layer Perceptron, etc.) are tested. In the training set, two classes are slightly imbalanced, i.e., 63% successful and 37% at-risk, so we apply under/oversampling techniques [45] to make them balanced. In general, Random Forest works very well and all of the reported values were produced by it.

Input: Student network $N_t = (V, E)$, Course and Activity Records
Output: $val(s_i, t)$ for each student
Initialize $val(x, t) = \frac{1}{n}$ where n is the total number of courses, activities, advisers, or students depending on the type of x.
while *until the convergence of* $val(s_i, t)$ **do**
  Update $val(c_i, t)$; Update $val(a_i, t)$; Update $val(v_i, t)$; Update $val(s_i, t)$
return $val(s_i, t)$

Algorithm 1: Fix-point calculation algorithm for our complex variable system

Table 1. Prediction results

	F-1 Overall	AUCROC	Recall of At-risk	Recall of Successful	F-1 of Successful
	All students				
Academic Features	0.78	0.76	0.56	0.67	0.78
Network Features (w_1 only)	0.76	0.72	0.62	0.77	0.83
Network Features (w_2 only)	0.73	0.66	0.51	0.70	0.81
Network Features (w_3 only)	0.74	0.70	0.61	0.75	0.81
Network Features (w)	0.81	0.8	0.64	0.85	0.87
Academic + Network Features	0.84	0.86	0.69	0.87	0.89
Academic + Network + fix-point Features	**0.85**	**0.86**	**0.70**	**0.90**	**0.89**
	Early phase students				
Academic Features	0.8	0.75	0.5	0.89	0.8
Network Features (w_1 only)	0.8	0.71	0.54	0.85	0.86
Network Features (w_2 only)	0.75	0.66	0.51	0.84	0.82
Network Features (w_3 only)	0.76	0.68	0.55	0.87	0.83
Network Features (w)	0.8	0.77	0.55	0.79	0.79
Academic + Network Features	0.84	0.85	0.56	0.86	0.9
Academic + Network + fix-point Features	**0.85**	**0.85**	**0.58**	**0.88**	**0.9**

Prediction Results. We calculated our network features in various types of student networks whose edge weights are calculated with w_i only or with the combined weight w marked with "(w_i only)" or "(w)" as shown in Table 1, respectively. Note that some w_i is omitted in the table due to their performance inferior to others. Network features based on the combined weight w shows the best performance among them. Using only the academic features, we could recall 56% and 50% of at-risk students among all and early phase students. After adding network features, we could achieve the recall of 69% and 56% for the at-risk student class and after using all available features, they are improved to 70% and 58%. For other measures such as F-1 and AUCROC, our predictive model including all academic, network, and fix-point features outperform others in non-trivial margins. These results strongly teach us answers on our research questions **1** and **2**. That is, network features improve the overall prediction and in particular, during earlier periods.

Network Analysis. The degree centrality of a student in N_t is the sum of the edge weights to neighbors. At the beginning, we expected that successful students have more friends, thereby higher degree values. However, our observations disprove the hypothesis. In Table 2, we show degrees in various perspectives. We calculate average values for the top 50% and bottom 50% students in terms of $avg\_degree(\cdot)$ in each prediction class. Their average values are quite different, i.e., 2522.4 for successful v.s. **5258.1** for at-risk students (with p-value < 0.01).

Table 2. Average centrality,average community-based features and of two student classes. For space reasons, we list selected values. P-values are smaller than 0.01 only except the cases marked in boldface.

		Successful (entire period)	At-risk (entire period)	Successful (at 5th sem.)	At-risk (at 5th sem.)
Degree	All	2522.4	**5258.1**	**3106.5**	2856.6
	Top 50%	4487.3	**10143.0**	**5575.5**	5392.7
	Bottom 50%	**557.4**	373.1	**634.8**	311.7
Page Rank	All	0.00075	**0.00202**	0.00072	**0.00193**
	Top 50%	0.00118	**0.00365**	0.0011	**0.00346**
	Bottom 50%	0.00031	**0.00038**	0.00032	**0.00038**
Eigen.	All	0.0068	**0.02212**	0.0058	**0.02323**
Betw.	All	0.00079	**0.00456**	0.0006	**0.00391**
Close.	All	**0.4357**	0.4281	**0.4372**	0.4214
Min Degree	All	**553.3**	498.8	**568.8**	423.1
Min Eigen.	All	**0.5348**	0.5145	**0.5419**	0.4719
Giant.	All	0.0674	**0.087**	0.0665	**0.0929**
	Top 50%	0.1268	**0.1606**	0.122	**0.1713**
	Bottom 50%	0.0074	**0.0131**	0.0077	**0.0137**
Fix Point	All	0.000601	**0.000722**	0.001303	**0.001793**
	Top 50%	0.00115	**0.00178**	0.002344	**0.00541**
	Bottom 50%	**2.144450e-81**	1.561194e-81	**2.511162e-131**	1.818025e-131

Some at-risk students are exposed to more interactions than the majority of successful students. This can be interpreted in multiple ways, e.g., some at-risk students have too many student activities. At the same time, the bottom 50% at-risk students have much lower degrees than the bottom 50% successful students, i.e. **557.4** v.s. 373.1 (with p-value <0.01). This means that there also exist many at-risk students who do not interact with other students as much.

In the top 50% and bottom 50% cases of early phase students, successful students have higher average values than at-risk students but their significance level is low (p-value >0.01). However, successful students' average degree at 5th semester is larger than that of the entire period, i.e., 3106.5 v.s. 2522.4. This is possible when successful students (i) quickly stabilize their connections during their early academic periods and (ii) do not make many new connections in their late academic periods. However, at risk students need to take some courses many times in order to pass the course, and this shows up in our network as having many more connections because they take the same course more times than the successful student. At 5th semester, the average degree of the top 50% at-risk students is 5392.7 but it is improved to 10143 when considering their entire academic periods. This means that they interacted with many new students even in their late academic periods for student activities, courses, and advisers. Interestingly the bottom 50% of the at-risk students' average degree does not change significantly over time, i.e., 311.7 at 5th semester v.s. 373.1 during entire academic periods. They are consistently isolated from others. All these findings support research question **3**, that successful and at-risk student show different behavior over time. For other network features, we also hypothesized before calculating them that successful students have better values. However, our results show counter-evidences in some features. Because some at-risk students (e.g., the top 50% at-risk students in the previous degree centrality analysis) keep making new connections (rather than staying in a community), they are bridges over communities and as a result, their PageRank, betweenness, and eigenvector centrality values will be higher than other successful students. In the bottom 50% case, we could not observe significant differences.

We also tested many other centrality metrics such as closeness centrality [24], leverage centrality [46], clustering coefficient [25], and so on. For some of them, we did not observe significant differences between the two student classes. In Table 2, the higher average values of community features for successful students implies they play more important roles in communities than at-risk students. This is also well matched with the at-risk students' high betweenness centrality results which means that they are bridges over communities rather than core members. Moreover, they are more likely to be members of the giant components than successful students since they interact with many communities and thus end up with peripheral positions in the communities. Interestingly, for the bottom 50% at-risk cases, their average degree centrality is lower than that of the bottom 50% successful students but their average percentage of the membership to giant components is higher. This implies that those at-risk students may visit many communities but do not make many connections in the communities.

The ego-network density and clustering coefficient are complementary to each other. Ego-network density can be high if some edge weights are large. For clustering coefficient, however, some large edge weights cannot solely lead to the clustering coefficient of 1. Only when ego-network is a complete network, its clustering coefficient becomes 1. Because of this property, we expected larger clustering coefficients from successful than at-risk students. Successful students' ego-networks tend to be more complete than that of at-risk students. Ego-network density will be small if one does not maintain long-term connections in various activities. Thus, we hypothesized that successful students may have dense ego-networks. In all cases, successful students have better ego-network density than at-risk students, i.e., **3.634** vs. 3.073 and **3.707** vs. 2.706. This observation is well aligned with all the previous analyses because the ego-network density results also imply that successful students and their neighbors maintain strong connections. These evidences provide support for our research question **3** and **4**.

Our result also shows that at-risk students have higher fix-point values in general. We think that it is because of the same reason that at-risk students make many short-term connections with others. In early phase, the pattern is unique (supporting research question **3**). In the degree centrality, for instance, successful students' degree values are higher than that of at-risk students in the early period. However, for fix-point values at-risk students already show higher values. This means that we can catch students exposed to too many early connections. In all cases, p-values are smaller than 0.01.

5 Key Observations, Future Work and Conclusions

We start by building a network that represents the connections between students and a mutually recursive variable system using data collected by the university to solve a problem in higher education. The network is constructed using the data stored in the University student management system and does not rely on access to social media data or consistent use of LMS data. Our prediction of success or risk achieves F-1 = 0.85 and AUCROC = 0.86. Our student network analysis teaches us two very important insights, that is (i) At-risk students establish disorderly connections while successful students keep strengthening their existing connections, (ii) Successful students have high GPA neighbors and their ties are strong. The density of successful students' ego-networks are stable regardless of time period. Our degree centrality results say that some at-risk students keep making new connections and their ego-networks do not become dense or complete. Our community-based features also support that successful students are core community members where at-risks students reside in periphery.

We think that this network model of students can identify effective intervention points at an early stage for at-risk students. It might be their natural characteristics to make connections in such a way. But by helping them maintain long-term and stable connections, we believe that at-risk students can improve their success probabilities. Moreover, considering aspects of gender, race, ethnicity, generational social class, student body demographics, geographic location of

institution, and socio-economic status of students, are also large factors when determining how long it takes a student to graduate. We are in process of collecting these data and also the data from our LMS logs. We think that it is essential to consider those factors when setting up intervention plans and creating compact student networks.

Acknowledgements. This work is supported by the National Science Foundation under Grant No. 1820862. Noseong Park and Mohsen Dorodchi are the co-corresponding authors.

References

1. National Center for Education Statistics: Table 326, 10 (2016)
2. Tinto, V.: Research and practice of student retention: what next? J. Coll. Stud. Retent. Res. Theory Pract. **8**, 1–19 (2006)
3. Suskie, L.: How can we address ongoing accreditation challenges? Assess. Updat. **28**, 3–14 (2016)
4. Acton, R.K.: Characteristics of STEM success: a survival analysis model of factors influencing time to graduation among undergraduate stem majors (2015)
5. Arnold, K.E.: Signals: applying academic analytics. Educ. Q. **33**, n1 (2010)
6. Arnold, K.E., Pistilli, M.D.: Course signals at purdue: using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 267–270. ACM (2012)
7. Campbell, J.P., Oblinger, D.G., et al.: Academic analytics. EDUCAUSE Rev. **42**, 40–57 (2007)
8. Campbell, J.P.: Utilizing student data within the course management system to determine undergraduate student academic success: an exploratory study. ProQuest (2007)
9. Dawson, S., Gašević, D., Mirriahi, N.: Challenging assumptions in learning analytics. J. Learn. Anal. **2**, 1–3 (2015)
10. Dorodchi, M., Bendict, A., Desai, D., Mahzoon, M.J.: Reflections are good!: analysis of combination of grades and students' reflections using learning analytics. In: Proceedings of the 49th ACM Technical Symposium on Computer Science Education, pp. 1077–1077. ACM (2018)
11. Merceron, A., Blikstein, P., Siemens, G.: Learning analytics: from big data to meaningful data. J. Learn. Anal. **2**, 4–8 (2015)
12. Jayaprakash, S.M., Moody, E.W., Lauría, E.J.M., Regan, J.R., Baron, J.D.: Early alert of academically at-risk students: an open source analytics initiative. J. Learn. Anal. **1**, 6–47 (2014)
13. Romero, C., López, M.I., Luna, J.-M., Ventura, S.: Predicting students' final performance from participation in on-line discussion forums. Compu. Educ. **68**, 458–472 (2013)
14. Gašević, D., Dawson, S., Siemens, G.: Let's not forget: learning analytics are about learning. TechTrends **59**, 64–71 (2015)
15. De Laat, M., Lally, V., Lipponen, L., Simons, R.-J.: Investigating patterns of interaction in networked learning and computer-supported collaborative learning: a role for social network analysis. Int. J. Comput.-Support. Collab. Learn. **2**, 87–103 (2007)

16. Blackmore, C.: Social Learning Systems and Communities of Practice. Springer, London (2010). https://doi.org/10.1007/978-1-84996-133-2
17. Shum, S.B., Ferguson, R.: Social learning analytics. J. Educ. Technol. Soc. **15**, 3–26 (2012)
18. Takaffoli, M., Zaïane, O.R., et al.: Social network analysis and mining to support the assessment of on-line student participation. ACM SIGKDD Explor. Newsl. **13**, 20–29 (2012)
19. Mohamad, S.K., Tasir, Z.: Educational data mining: a review. Procedia - Soc. Behav. Sci. **97**, 320–324 (2013). The 9th International Conference on Cognitive Science
20. Adraoui, M., Retbi, A., Idrissi, M.K., Bennani, S.: Social learning analytics to describe the learners' interaction in online discussion forum in moodle. In: The 16th International Conference on Information Technology Based Higher Education and Training (2017)
21. Gašević, D., Zouaq, A., Janzen, R.: Choose your classmates, your GPA is at stake! The association of cross-class social ties and academic performance. Am. Behav. Sci. **57**, 1460–1479 (2013)
22. Carolan, B.V.: Social Network Analysis and Education: Theory, Methods & Applications. Sage Publications, Thousand Oaks (2013)
23. Gardner, J., Brooks, C.: Coenrollment networks and their relationship to grades in undergraduate education. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 295–304 (2018)
24. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**, 215–239 (1978)
25. Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. Proc. Natl. Acad. Sci. U.S.A. **101**, 3747–3752 (2004)
26. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
27. Anthonisse, J.M.: The rush in a directed graph. Stichting Mathematisch Centrum. Mathematische Besliskunde, pp. 1–10 (1971)
28. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
29. Lawrence, P., Sergey, B., Rajeev, M., Terry, W.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab (1999)
30. Li, Y., Martinez, O., Chen, X., Li, Y., Hopcroft, J.E.: In a world that counts: clustering and detecting fake social engagement at scale. In Proceedings of the 25th International Conference on World Wide Web, pp. 111–120. International World Wide Web Conferences Steering Committee (2016)
31. Mavroforakis, C., Valera, I., Gomez-Rodriguez, M.: Modeling the dynamics of learning activity on the web. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1421–1430. International World Wide Web Conferences Steering Committee (2017)
32. Hoang, M.X., Dang, X.-H., Wu, X., Yan, Z., Singh, A.K.: GPOP: scalable group-level popularity prediction for online content in social networks. In: Proceedings of the 26th International Conference on World Wide Web, pp. 725–733. International World Wide Web Conferences Steering Committee (2017)
33. Chaturvedi, S., Castelli, V., Florian, R., Nallapati, R.M., Raghavan, H.: Joint question clustering and relevance prediction for open domain non-factoid question answering. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 503–514. ACM (2014)

34. Xie, J., Szymanski, B.K., Liu, X.: SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: Data Mining Workshops, pp. 344–349. IEEE (2011)
35. Burt, R.S.: Models of network structure. Ann. Rev. Sociol. **6**, 79–141 (1980)
36. Gibson, D., Kleinberg, J., Raghavan, P.: Inferring web communities from link topology. In: Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space–Structure in Hypermedia Systems: Links, Objects, Time and Space–Structure in Hypermedia Systems, HYPERTEXT 1998, pp. 225–234. ACM (1998)
37. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 1998, pp. 668–677 (1998)
38. Miller, J.C., Rae, G., Schaefer, F., Ward, L.A., LoFaro, T., Farahat, A.: Modifications of kleinberg's hits algorithm using matrix exponentiation and web log records. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 444–445. ACM (2001)
39. Li, L., Shang, Y., Zhang, W.: Improvement of hits-based algorithms on web documents. In: Proceedings of the 11th International Conference on World Wide Web, WWW 2002, pp. 527–535. ACM (2002)
40. Kang, C., Park, N., Prakash, B.A., Serra, E., Subrahmanian, V.S.: Ensemble models for data-driven prediction of malware infections. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM 2016, pp. 583–592. ACM (2016)
41. Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., Subrahmanian, V.S.: Rev2: fraudulent user prediction in rating platforms. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, pp. 333–341. ACM (2018)
42. Diaz, D.P., Cartnal, R.B.: Students' learning styles in two classes: online distance learning and equivalent on-campus. Coll. Teach. **47**, 130–135 (1999)
43. Picciano, A.G.: Beyond student perceptions: issues of interaction, presence, and performance in an online course. J. Asynchronous Learn. Netw. **6**, 21–40 (2002)
44. Astin, A.W.: Student involvement: a developmental theory for higher education. J. Coll. Stud. Pers. **25**, 297–308 (1984)
45. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18**, 1–5 (2017)
46. Joyce, K.E., Laurienti, P.J., Burdette, J.H., Hayasaka, S.: A new measure of centrality for brain networks. PLoS One **5**, e12200 (2010)