DECISION SCIENCES INSTITUTE

Predicting Retention by Analyzing Student Networking

Josephine M. Namayanja
University of Massachusetts Boston
Email: josephine.namayanja@umb.edu

Roger Blake
University of Massachusetts Boston
Email: roger.blake@umb.edu

ABSTRACT

This study uses graph theory to analyze a co-curricular program of activities in order to conceptualize and quantify student engagement and in turn enable the degree to which student engagement impacts retention to be evaluated.

KEYWORDS: Community Detection, Graph Modelling, Social Networking, Peer Engagement, Student Retention

INTRODUCTION

Student retention has an enormous impact on colleges and universities that goes well beyond the financial implication. Retention and student attrition affects the reputation, rankings, culture, enrollment decisions, and ultimately the viability of an educational institution. Even with the recognition of how important retention is and the considerable efforts devoted to it, the percentage of students who enroll and complete their undergraduate degree has changed little over time (NCES, 2005): the nationwide percentage of students who complete a four year degree is only 35.3% for public four year institutions, and 40.7% for those that are private (ACT, 2010). It is not surprising that researchers have increasingly turned their attention to this topic.

Institutional engagement and student peer engagement have long been theorized as two of the primary influences on student retention (Tinto, 2006). Neither form of engagement is readily quantified. It is not surprising that the majority of studies aimed at predicting student retention consider factors more readily quantifiable and amenable to analysis, such as SAT scores, high school GPAs, and socio-economic factors, such as in research by (Delen, 2010). The ability to quantify student engagement and use it to predict retention has not been well-studied, and can have significant benefits in practice.

To the extent the impact of institutional engagement has been studied, it has tended to be quantified by measures such as the number of visits to student support services or tutoring centers. Peer engagement is even more difficult to quantify, as a result, few studies have addressed the impact it has on student retention. Our study contributes by developing a measure of peer (student) engagement and demonstrates the potential explanatory power it can have for predicting student retention.

The study of social networks can be used to identify and measure and evaluate interactions of actors in the form of peer engagement. A social network is described as a set of interconnected individuals (Wasserman & Faust, 1994). The objective of our research is first, to identify patterns in student networks that are engaged in co-curricular activities, and second, to show that those patterns are capable of student retention.

BACKGROUND

For this study, we are examining social networking in the context of the Management Achievement Program (MAP) at the College of Management in University of Massachusetts Boston. MAP started in 2006 and is designed to foster engagement and build competencies for academic success. It is designed to foster both peer and institutional engagement by providing the opportunity for students to personally synthesize their academic and professional goals and experiences, and at the same time increase involvement in the College and local business communities. MAP consists of a range of activities designed to support experiential learning including seminars, career workshops, seminars, guest speakers, on-site visits to companies, job fairs, and networking events.

MAP is a graduation requirement for all students in the college, and to meet that requirement students can select from a range of events offered each semester. Each MAP event is designated with a certain number of MAP "miles" as an award for participation, and to meet their requirement, each student must reach a certain number of MAP "miles". However, students' participation and types of events they attend vary over time. This creates an opportunity to detect patterns of networking among students and in the interests of specific groups. The focus of this paper is to detect communities over time. Community detection aims at grouping nodes based on relationships among them to form strongly linked subgraphs (Wang et al, 2015). Here a community can be defined by a group of students represented as nodes, who attend similar activities. The data for our analysis consists of attendance records for MAP activities. Hence, we propose a graph modelling approach where we create an undirected graph based on the attendance records. Here the students are represented as nodes and the similarity in their MAP activities denotes the edges. Our goal is to detect clusters of nodes in the graph.

The communities that emerge can be associated with specific MAP activities and other underlying characteristics. For example, students in Accounting may participate in similar activities, while those in Marketing may also participate in an exclusive set of activities. However, there could be a possibility that such disparate cliques (Mokken, 1979) merge to form clubs (Mokken, 1979) connected by an edge or a set of edges. Such clubs can be formed by cross-disciplinary peer engagement where students from different majors are linked together to form even further natural divisions in the network called communities. Communities are used to depict nodes/students that are more connected together than to the rest of the network. For this we propose to utilize graph metrics, particularly modularity to evaluate intra-community density versus inter-community density. Here, we detect the number of communities over time, which may vary in size.

Additionally, we study the impact of student participation rate on communities detected. Our assumption is that, as the number of MAP events increases and the overall student participation increases, there will be more student interactions that will boost peer engagement. Additionally, such peer engagement can also be defined in terms of centrality measures such as average betweenness centrality and average degree centrality ((Freeman, 1979; Kosorukoff, 2011) in relation to modularity to assess the role of key nodes in building dense or sparse communities.

In general, identifying such communities can be useful as they may help to uncover unknown functional units such as information networks in social networks (Newman, 2006b). Additionally, such communities can also be used in predictive modeling to determine what type of activities a given student is likely to attend. More so one can associate the patterns detected with characteristics such as a student's choice of a major and potential rate of attrition. The practical implications of detecting community structures can be used to target appropriate activities to a specific group of students. This is turn can be used to support student retention within academic programs by developing more specialized programs for students in similar majors. Furthermore, the detection of overlaps amongst communities can be used to design cross-disciplinary programs within MAP as well as academic courses to attract different groups of students and thus continue to enhance experiential learning as described in (Blake & Gutierrez 2011).

METHODS

Utilization of MAP

MAP was not developed for the purpose of gathering data for analyzing retention; it is a program designed to develop and foster the professional demeanor of its participants. The rationale for introducing MAP was that professionalism can best be developed and demonstrated when participants engage in co-curricular activities, in contrast to the assimilation of content delivered only through lectures or seminars. The philosophy of MAP is based on a philosophy embedded in how the program is defined:

> "An engaging and comprehensive program designed to develop and enhance each student's professional demeanor, build competencies for academic success, increase involvement in the College and local business communities, and allow the opportunity for students to personally synthesize their academic and professional goals and experiences."

MAP is a requirement for all undergraduate students in business administration at a state university. Students in MAP are required to select and participate in events and activities designed to build professional and career skills. Among the types of events and activities regularly offered are career workshops, seminars, forums, company visits, presentations by senior executives, student clubs, and service learning activities such as volunteer work.

Upon admission into the college, since transfer students enter with a varying number of credits, each participant's record is assessed to determine the number of MAP miles that will be required to graduate. Each MAP event offers a number of miles depending on the involvement and initiative it requires. For the majority of events, participants can earn 50 miles; an example of such an event would be a workshop on interviewing or resume writing. Events such as participating in a college-wide case competition can earn as many as 200 miles.

Some events also require participants to record their reflections to summarize what they learned and how they envision applying it to their academic or professional careers. These written responses are also a source of valuable information that might be used in conjunction with the data in this study, but that is beyond the scope of this paper and part of future work.

The MAP program operates as follows. After matriculation and prior to their first semester, each participant is given a bar-coded identification card and access to a web-based portal for reviewing upcoming events, registering for specific events, entering reflections on events attended, and checking the status of MAP miles and accounts. The attendance and participation of MAP events is recorded through a bar-code scanning system which is regularly uploaded to a back-end system. This is because these events are independent of the course schedules and are often held outside of classrooms or off-site, A sample of attendance records is provided in Table 1. It should be noted that the Student IDs provided in the sample are anonymized.

Table 1. Management Achievement Program (MAP) Sample Attendance Records

Student ID	Student start semester	Event	Event Type	Event Date	Event Miles
21425	Fall 2010	Business Ettiquette – How to drool in an interview	Seminar	8/24/2010	50
21425	Fall 2010	InterviewTrak Seminar – How not to drool in an interview	Workshop	8/5/2010	50
21425	Fall 2010	Investment Club Membership Fall 2007	Student Club	10/1/2010	50
21425	Fall 2010	Investment Club Presents: Emerging Economies	Student Club	9/24/2010	50
21425	Fall 2010	Investment Club Spring 2007	Student Club	10/22/2010	50
52129	Spring 2009	Boston Globe Tour – Come see a losing business up close	Workshop	3/26/2011	75
52129	Spring 2009	COMETT - Fall 2008	Company tour	2/12/2011	75
47874	Spring 2010	Interview Track – How to hire beautiful women	Seminar	5/24/2010	50
47874	Spring 2010	Myers-Briggs/Joyce Morgan	Workshop	9/18/2010	75
47874	Spring 2010	Prep for Career Fair and primp for your hair	Career Cafe	10/4/2010	50
54644	Fall 2008	MIcrosoft	Workshop	12/8/2010	50
54644	Fall 2008	Spring 2009 Career Expo	Workshop	5/19/2011	100
54644	Fall 2008	Student Activities Fair	Workshop	10/14/2010	50
54645	Fall 2008	Dynamic Resume, too bad you don’t have one	Workshop	10/20/2010	50

MAP was implemented in the fall of 2006 and the number and type of events held each semester has grown significantly. For this study, we selected a time window from Fall 2006 to the end of Spring 2010. They were 614 events offered during this time period and the total participation was 9,925 attendees.

Graph Modelling

We first select a set of temporal bins B = {b_1…b_x}, where each temporal bin b_i represents an Academic year. The beginning of the Fall Semester is considered the beginning of the Academic Year and the end of the Summer semester in the following calendar year marks the end. For this study, we select four temporal bins, specifically AY2006-2007, AY2007-2008, AY2008-2009, and AY2009-2010. In our future work, we plan to include more time periods as data becomes available.

For each temporal bin we only select the Student ID and the Event ID. Hence, we omit all other attributes captured in the attendance records and create the graph of relationships between students based on the events they have attended. Each Student ID becomes a node or vertex and Event ID that two given Student IDs attend establishes the relationship between them and forms an edge. Essentially, this models a graph.

A graph is made up of nodes or vertices and lines called edges that connect them. It is defined as

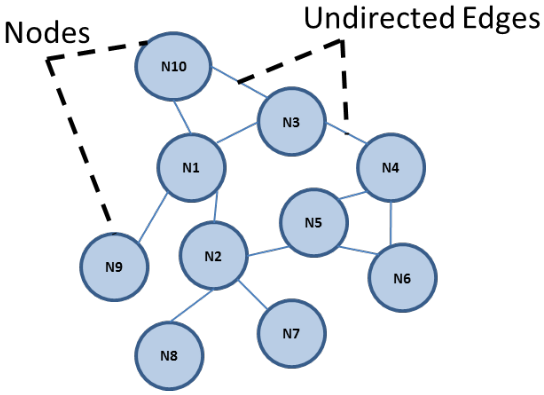
Given a graph G = (N, E), where *N* is the number of nodes in the network, and *E* is the number of edges such that a_{ij} = 1 if nodes *i* and *j* are connected by an edge and a_{ij} = 0 otherwise. The connectivity between nodes is represented through an adjacency matrix A. The matrix A is a square matrix as show by an example in Figure 1.

Figure 1: Adjacency Matrix

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
N1	0	1	1	0	0	0	0	0	1	1
N2	1	0	0	0	1	0	1	1	0	0
N3	1	0	0	1	0	0	0	0	0	1
N4	0	0	1	0	1	1	0	0	0	0
N5	0	1	0	1	0	1	0	0	0	0
N6	0	0	0	1	1	0	0	0	0	0
N7	0	1	0	0	0	0	0	0	0	0
N8	0	1	0	0	0	0	0	0	0	0
N9	1	0	0	0	0	0	0	0	0	0
N10	1	0	1	0	0	0	0	0	0	0

A graph can be directed or undirected; for this study, we focus on an undirected graph G where G does not identify the direction of the edge between nodes as shown in Figure 2.

Figure 2: Undirected Graph



Given x temporal bins, the result is a set of graphs representing each temporal bin. Additionally, we consider an unweighted graph where the nodes and edges are not assigned weights. The graph is created regardless of the map miles requirement and number of events students have attended. So if a pair of students has one or more events in common, only a single unweighted edge is established. For example, if a pair of students, represented as nodes *a* and *b* attend one event and students *a* and *c* attend five events, we only establish an edge between each respective pair of students which is considered an unweighted edge. However, for our future work, we plan to apply a weighted approach, where each edge is assigned a weight as a function of the number of events and miles per event between two students. This can be used to determine the strength of the relationship between nodes or community of nodes. We provide a summary of the nodes and edges per temporal bin in Table 2.

Table 2: Summary of Nodes and Edges per Temporal Bin

Temporal Bin	Nodes	Edges
AY2006-2007	193	3834
AY2007-2008	682	44758
AY2008-2009	870	79117
AY2009-2010	935	68800

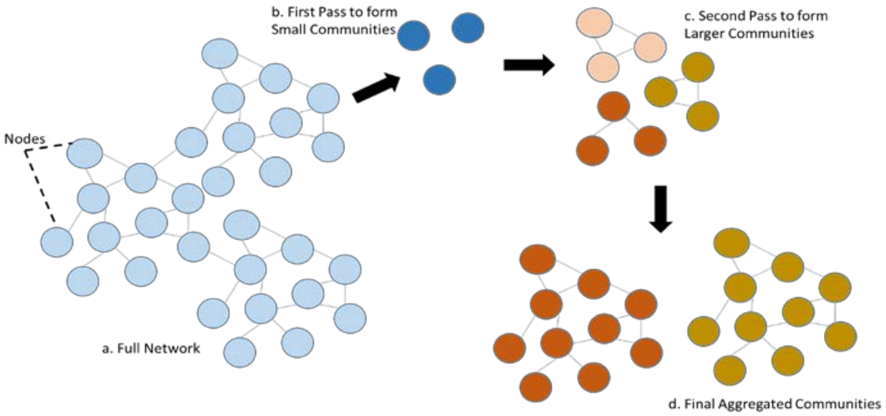
It should be noted that the size of the temporal bins varies and therefore the consequent graphs generated also vary in size. We use the graphs to detect communities in the network structure over time.

Community Detection

For the graph models in each temporal bin, we detect communities. Community detection requires the partition of a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected. Our objective is to determine the structural characteristics in the behavior of nodes in the network over time.

For this, we apply the Louvain method described in (Newman, 2006b). This approach consists of two phases. First, it looks for "small" communities where each node forms a single-member or single-node community. Then for each node *i* we consider the neighbors *j* of *i.* Here a given node *i* can be moved into the community of another node *j*. The quality of the newly formed community detected in the network is evaluated by modularity of the partition. The modularity score ranges between -1 and 1 where -1 denotes sparsity within a community while 1 denotes highly dense intra-community structure. The objective is to attain a maximum modularity. If no positive and maximum gain is possible, node *i* stays in its original community. The second phase reduces the size of these communities and aggregates them iteratively in order to form larger communities as discussed in (Newman, 2006a). The overall process is illustrated in Figure 3.

Figure 3: Community Detection


b. First Pass to form Small Communities
c. Second Pass to form Larger Communities
Nodes
a. Full Network
d. Final Aggregated Communities

We apply this approach because its accuracy is known to be very promising when applied to ad-hoc networks such as peer to peer networks that have a known community structure. In studying retention, there are two known communities: those students who are retained, and those who are not.

In terms of graph theory, our assumption is that each unique MAP event is intuitively a community of nodes that forms a clique such that nodes associated to the same MAP event are all connected to each other. The size of each clique will differ depending on the attendance of an event, however, our objective is to look beyond each clique and determine the reachability of nodes through merging of cliques to detect larger communities that are formed due to a set of nodes that are highly central, and thus connect small communities to form larger ones.

For example, larger communities can be formed through several short paths where some nodes serve as junctions between smaller communities. These short paths can be used to determine the betweenness centrality (Freeman, 1979) of nodes on the network. Additionally, some nodes with high degree centrality (Freeman, 1979) are highly adjacent to other nodes in the network and thus serve as hubs for communities. Hence, we validate the communities detected by determining correlations to the centrality of nodes in the network. We also validate the communities detected by determining correlations to the growth of the network by evaluating its density (Kosorukoff, 2011) and diameter (Kosorukoff, 2011) over time. Overall, identifying such communities can be used to enhance information flow in the network. Next we discuss our experimental results.

RESULTS

This section presents a summary of the results from an analysis of the graphs of student participation in MAP events from the first year since the programs' inception.

Analysis of Student Participation

Figure 4:Event Count Over Time

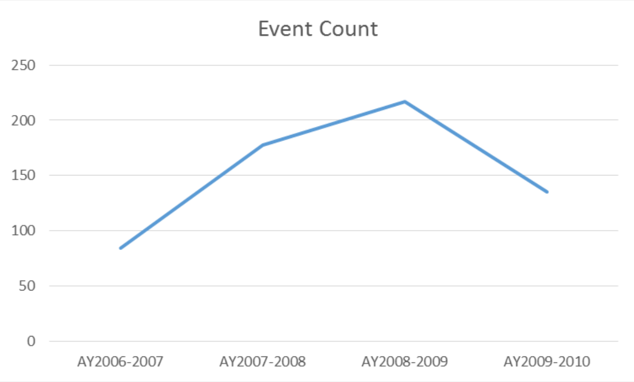


Figure 5: Student Count Over Time

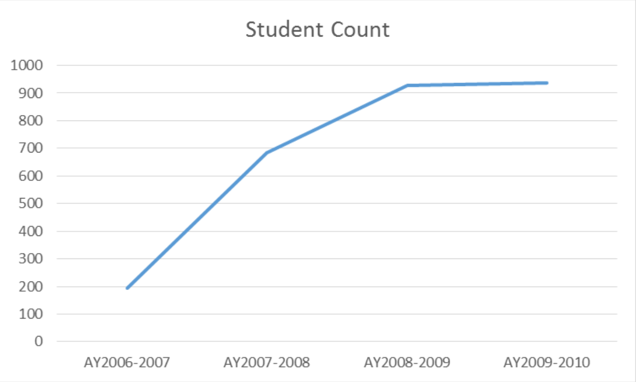
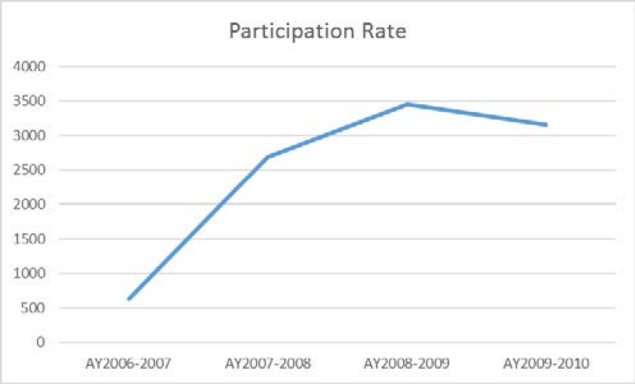


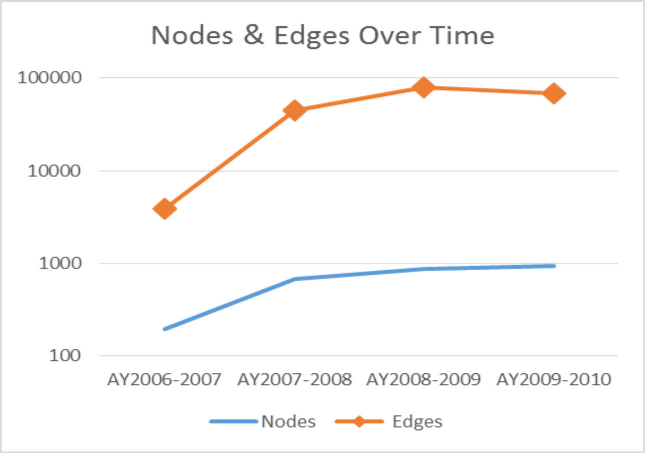
Figure 6: Participation Rate Over Time



Our findings in Figure 4 indicate an increase in the count of events over time except in AY2009-2010. Additionally, the student count based on enrollments in the program, increased over time as the number of students required to enroll in the MAP program also increased as shown in Figure 5. We also observe that the participation defined in terms of attendance records indicated in Figure 6 increases over time with an increase in the student enrollments. However, in AY2009-2010, we see a decline in the participation rate. This is characterized by the smaller number of MAP events during the same academic time period, which despite the high number of student count (enrollments), their participation was limited by having a smaller number of events.

Interestingly, we also see an exponential increase in the number of nodes and edges from AY2006-2007 to AY2008-2009 as shown in Figure 7. This is an indicator that the network densifies over time particularly in the first 3 time periods. However, the number of edges decreases slightly in AY2009-2010 which can be characterized by a decline in student participation during that time period.

Figure 7: Count of Nodes and Edges Over Time



Analysis of Communities Detected

Figure 8: Count of Communities Detected Over Time

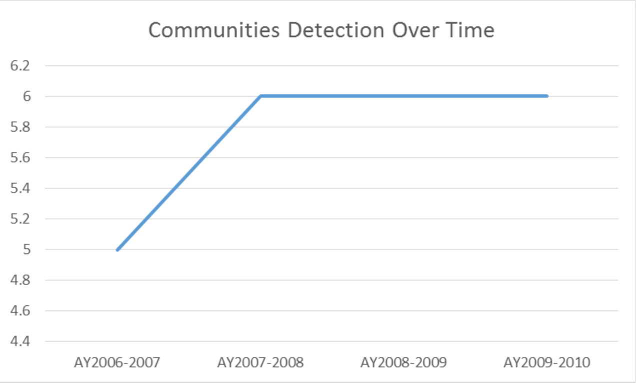
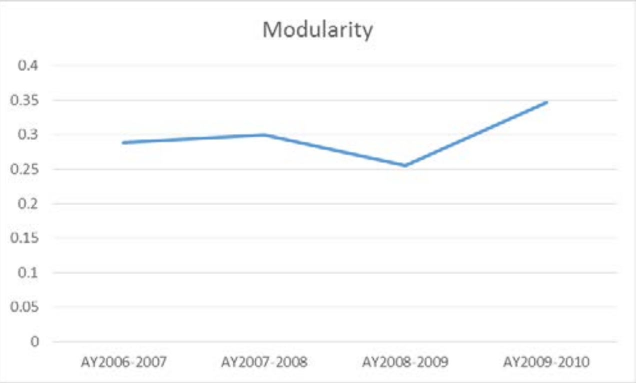


Figure 9: Modularity of Community Structures Over Time



Our findings in Figure 8 indicate that there are four communities detected in AY2006-2007 which also marks the inception of the MAP program. The number of communities increases to six following that and remain consistent over time. Given the count of events, student enrollments and participation relatively increases significantly following AY2006-2007, which can be used to characterize the slight increase in the number of communities following this time period. Nevertheless, the overall positive modularity as shown in Figure 9 portrays dense community structures over time which further indicates high levels of peer engagement amongst students.

Specifically, we observe denser communities between AY2006-2007 and AY2007-2008 which can be attributed to the increase in overall MAP events and potentially increments in student participation. On the other hand, we also observe sparse communities in AY2008-2009 as indicated by the lowest modularity score. Given that AY2008-2009 had the largest number of events and highest student participation overall, such community sparseness can be attributed to a high diversity of events, thus causing very low attendance for some specific events (isolated events), or groups/individual students restricting themselves to events. This potentially limits reachability to other events and potentially other students.

Interestingly, we also observe very dense communities in AY2009-2010 as indicated by the highest modularity score despite that its' number of events is significantly lower than AY2008-2009. However, given the highest student enrollment in that time period, the overall attendance in events is also high. As a result, such limitation on the number of events, potentially leads to more student interactions and thus enhances peer engagement.

Analysis of Node Level Behavior

Figure 10: Count of Shortest Paths Over Time

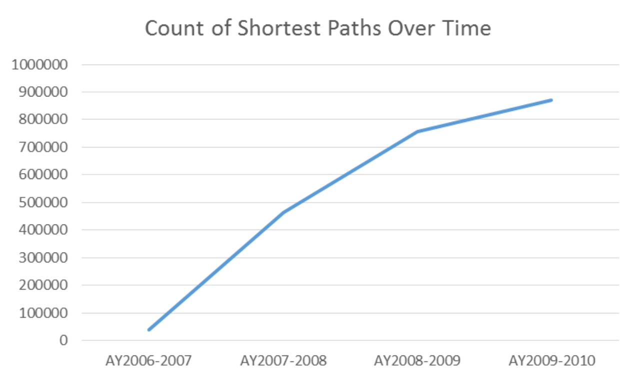
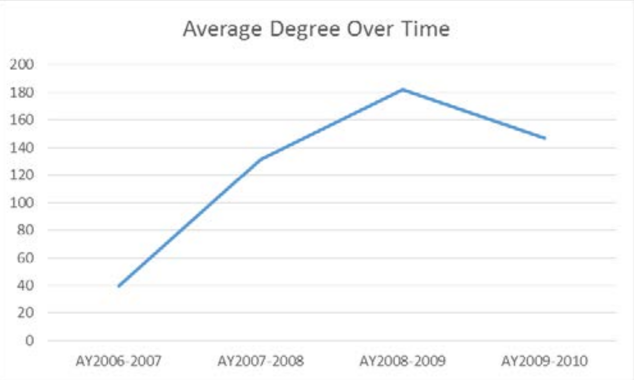


Figure 11:Average Degree of Nodes Over Time



Our findings in Figure 10 indicate an increase in the number of shortest paths over time. This further indicates high average betweenness centrality over time whereby several nodes that lie on these paths connect other nodes in the MAP network. Hence, such nodes that represent students, create junctions on the network for students in disparate circles to identify peers across the College despite not attending any similar events. This in turn can be used to target events to other groups of students through cross-disciplinary experiential learning by organizing cross-disciplinary events or further develop cross-disciplinary courses down the road. The increase in shortest paths is also supported by an increase in number of nodes and edges over time which indicates high cohesiveness in the communities detected.

Similarly, the average degree in the network increases over time especially in the first three time periods in Figure 11. Our findings support the view outlined by (Leskovec et al, 2005a; 2005b; 2007; Leskovec & Faloutsos, 2007; Leskovec et al., 2007; Leskovec, 2008) that the networks are becoming denser over time with the average degree increases over time as well. This indicates that the influence of key players in the network increases over time. This further shows that there are some students who are very central in the network and can thus be used as hubs to reach others directly by drilling down to identify the role of other factors driving similarity such as the major of study, demographic characteristics and more. Additionally, it could be used to identify those students who are very active in the MAP program and are thus classified as high achievers. For our future work, we plan to evaluate node behavior over time to identify critical and non-critical nodes on the network based on the map miles acquired by a student. For example, a student x can be defined as a critical node if they have low miles while a non-critical node is defined as otherwise. Essentially, this can be used identify students who need extra attention and are thus at risk of not being retained. More so, identifying such critical nodes can be used to identify curriculum needs to improve the program to better target different groups of students.

Analysis of Network Level Behavior

Figure 12:Diameter of Subgraphs Over Time

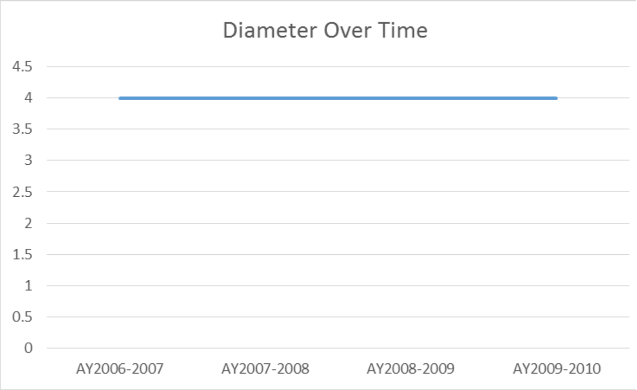
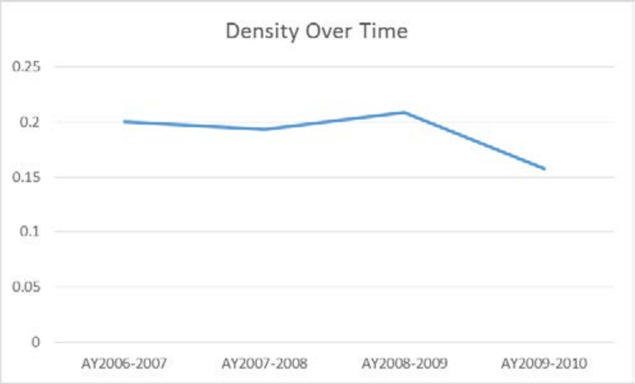


Figure 13:Density of Subgraphs Over Time



Our findings indicate that the increase in the number of shortest paths and average degree centrality of nodes on the network impact the network's overall connectivity as defined by consistent network diameter in Figure 12. Given that as the network size significantly increases over time, the average distance between points on the network does not increase. Thus it can also be argued that such a consistent diameter, essentially increases reachability across the student body as it grows over time.

On the other hand, we observe fluctuations in the density of the overall network over time as shown in Figure 13. The overall densities are low, which signifies that direct connectivity among students is low. This further indicates that measures have to be taken to increase or foster direct student engagement by considering the multiple dimensions governing the student body.

DISCUSSION AND CONCLUSIONS

Our approach is based only on data that represents each individual student as a node and their participation in an event with another student as an edge. Additionally, the data also captures whether each student ultimately either graduated or was not retained. Given the relatively low degree of connectivity we found by analyzing the network, we do not know the degree to which we may be able to differentiate between the two sub-groups of those students who were retained and those who were not. But we have developed metrics from graph theory to measure student engagement. Based on these and our initial results, we believe that we have a strong base to proceed and evaluate the degree and extent to which student engagement actually does influence student retention.

REFERENCES

ACT. (2010). What Works in Student Retention. Fourth National Survey. Iowa City, IA 2010.

Blake, R. & Gutierrez, O. (2011). A semantic analysis approach for assessing professionalism using free-form text entered online. Computers in Human Behavior. 2011.

Delen D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems. 2010; 49:498-506.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. Social Networks 1 (3): 215–239

Kosorukoff, A. (2011). Social Network Analysis: Theory and Applications. Publisher: Passmore, D. L, 2011

Leskovec, J. (2008). Dynamics of large networks.

Leskovec, J., Chakrabarti, D., Kleinberg, J., & Faloutsos, C. (2005a). Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD).

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005b). Graphs over time: Densification laws, shrinking diameters and possible explanations. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Leskovec, J. & Faloutsos, C. (2007). Scalable modeling of real graphs using kronecker multiplication. In International Conference on Machine Learning (ICML).

Leskovec, J., Kleinberg, J. & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. In ACM Transactions on Knowledge Discovery from Data (TKDD), volume 1.

Robert J. M. (1979). Cliques, clubs and clans. Quality and Quantity. International Journal of Methodology. 1979.

NCES. (2005). College persistence on the rise: Changes in 5-year degree completion and postsecondary persistence rates between 1994 and 2000. In: NCES, editor. Washington, DC. 2005.

Newman, M. E. J. (2006a). Detecting community structure in networks. 2006.

Newman, M. E. J. (2006b). Modularity and community structure in networks. Proceedings of National Academy of Science of United States of America

Takes, F. W. & Kosters, W. A. (2011). In Proc. of the 20th ACM International Conference on Information and Knowledge Management (CIKM)

Tinto V. (2006). Research and practice of student retention: what next? Journal of College Student Retention: Research, Theory and Practice. 2006; 8:1-19.

(Wang et al, 2015) Wang, M., Wang, C., Yu, X, J. & Zhang, J. Community Detection in Social Networks: An Indepth Benchmarking Study with a Procedure-Oriented Framework. Journal Proceedings of the VLDB Endowment. Volume 8 Issue 10, June 2015. Pages 998-1009.

(Wasserman & Faust, 1994) Wasserman, S. & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge: Cambridge University Press.