Using Data Analytics to Predict Peer-Group Effects on Student Exam Results

Philip Scanlon & Alan F Smeaton
Insight Centre for Data Analytics,
Dublin City University,
Glasnevin, Dublin 9, Ireland
Alan.smeaton@DCU.ie

Abstract

There has been much research examining the influences on an individual student and his/her academic performance within a University environment and the impact of the heterogeneous social groups to which they become members. Manski [10] has addressed the concept as "Reflection" or the influences within group dynamics. Constructivism is a pedagogy theory that says knowledge is constructed and not acquired. Social Constructivism emphasises the importance of an individual's social and cultural environment within which they interact and learn. It considers how they are influenced by the past, their present interactions and ergo, their influence on their peer group members. We examine this hypothesis within a University environment and build on research which recognises the intricate nature of complex community structures.

Using anonymised campus WiFi access logs collected through use of the University's Eduroam system, we are able to identify locations where students congregate within academic and social environments and we have identified students who spend a higher proportion of their time together in comparison to with other class members, thus defining social groupings. As an illustration of our approach we randomly choose a mid-semester school day as representative of a University's activity. We mined the 7 million WiFi log events for that day and identified the activity of 4,700 students. On that day there was an average of 40 interactions or meetings between student pairs. From this we can determine which students collocate and those who interact less with other class members.

1. Introduction

Universities are unique micro-environments in which multiple individuals interact with the same overall objectives, to attend campus for classes, study sessions and in some cases social events. Through their interactions with the University on-line systems, students provide large quantities of data providing their individual unique digital footprints. This footprint is collected and stored by the University's IT systems. We believe this information could be usefully mined for knowledge using learning analytics approaches, which could ultimately benefit the student and inform theories of pedagogy.

During their University career most students will become part of one or more social groups. Our principal research question is as follows: is the academic achievement level of a student correlated with the levels of the friends and peer groups that they associate with and can we capture these associations automatically. Constructivism and specifically social constructivism believes that knowledge is constructed though the sharing of ideas based on life experiences and understandings. Constructivism at it's core is a theory that learning is a longitudinal collaborative process and that who we are learning from can influence what we learn and vise-versa. In our work we measure some of the collaborations and interactions each student has with others and determining this, we can examine our research question quantitatively. Ultimately gaining this level of insight could help with better planning of student activities, better organization of course curricula or to identify individual students who could benefit from additional assistance.

On entering their first year, University students are formed into exogenous units as dictated by the University administration, i.e. their class, study, work, labs or assignment groups. However the endogenous groups that form within the University community also have a bearing on a student's performance. These latter groups include self-formed study teams, social and sports groups. These groups overlap in composition, time and location and contain sub-units that exist within larger groups. We wish to examine the make-up and interactions of all groups and the effects or reflection they have on each member. Our research is undertaken within an environment where we consider it possible to identify groups as they successfully pass through

Tuckman's [15] stages of group development.

Our research focuses on the makeup of each group during the performing stage. Not every group completes all of Tuckman's stages and many disband at any stage for a variety of reasons. It is the group members' ability to interact that dictates the stability and longevity of the group from ephemeral and ad-hoc to long-term.

We will identify group dynamics and the influences of group on the individual and the effect on their academic achievements. The hypothesis is that individuals will become members of a number of emergent groups in the early stages of interaction which can lead to the forming of friendships that will become influential in a student's development. The strength of the friendship will determine the effects of the individual on the group. Individuals will effect and be affected on many levels within groups based on their own personality, academic achievements and social maturity (Constructivism). It is our intention to accurately profile students of interest with supplemental information including previous academic achievement, demographic data plus other interactions with the University's IT resources, thus broadening the definition of a students' digital footprint.

The novel element of our research is the data collection process. Data collection in the majority of the research in this area has been invasive. Typically it involves direct interaction and observational or census-gathering techniques. Such methods can introduce a bias into the data collection process due to interviewer or interviewee interpretations of questions. Interpretations can vary across subjects and team members, based on their own characteristics. It is difficult to estimate the bias effect of a subject's awareness that they are part of a research project. Carney, [4], carried out an extensive literature review in the domain of peer influence. As with much research in this area, it identifies that data in the main is collected through direct observation and/or the use of census. While the influence of the Hawthorne effect, prescribed by Elton Mayo [11] is an unavoidable bias in much quantitative research, our work is based on data collection which is unobtrusive and has minimal contact with the subjects providing the data. This is because we perform data analytics on ambiently-collected log files. These logs are effectively the students' digital footprints left through their interactions with the University WiFi system. This leaves the researcher removed from the subjects themselves.

2. Background

Our data set is derived from a University with a campus accommodating an academic staff of 440 and approximately 12,000 students each semester. All students, once registered are provided with a unique login identifier for accessing the University's Information Technology (IT) assets such as email, web browsing, Google apps and access to the University's virtual learning environment (VLE), Moodle. Access is either directly through network-enabled PCs or via mobile devices such as smart-phones, tablets and laptops. In addition to the main campus, there are five linked institutions that share University resources and all students registered in the University have access to all IT assets through the Eduroam system.

Eduroam is an international roaming service for those in Universities and other higher education institutions which allows seamless interconnected Internet access for University students, researchers and other staff. This access spans borders and provides access to all participating institutions using just their home institution's login credentials. It is based on IEEE 802.1X protocols and is currently deployed in almost 70 countries worldwide. It is the default network for students and staff using wireless devices on our University campus. When a student, or staff member, connects to the Eduroam WiFi network, a record of that connection is created in a log file and this is the raw data that we will use in our analysis.

3. Related Work

Much of the past research in the area of exploring group influence on academic performance involved the creation of an artificial environment from which analyses and hypothesis-testing could be performed. We interpret artificial to mean the environment is designed specifically for the experiment and/or the test subjects are reminded on a continuous bases that they are being observed. Carrell *et al*. [5] in their study at a US Air Force Academy monitored students exogenously assigned to groups. Their research reported a peer effect of "greater magnitude than previously found". The effect of experimental interventions within the research environment such as in this related work, causes bias which we believe could be avoided utilising a new and novel approach to data collection.

Brewe *et al*. [3] investigated the effect of a community within a purpose-built physics learning centre concluding that, social network analysis holds significant promise for the description and analyses of student learning.

In 2005, Eagle [6] used what was at the time, cutting edge technology to track subjects to infer contextual interactions between them. The approach was novel as it did not use any form of census data collection to develop their dataset, but used mobile phone location monitoring for temporal and

geolocation data.

In 2015, Rui Wang [16] used advances in mobile smart-phone technology to monitor a small cohort of students for the purpose of predicting "academic performance". He used regression analyses to develop a behavioral slope and behavioral breakpoints. These methods were used to identify changes in a student's behavior on a weekly basis. In both studies the subjects were fully aware of their role in the research. Recently, researchers such as Minaei-Bidgoli [12] now focus on data collection from web-based educational systems. The use of e-learning systems provides useful data based on a student interaction with on-line materials.

Our research uses Social Network Analyses (SNA) as a deterministic basis for the modularity of the domain and specifically social groups. It was the research of Wasserman [17] who first expressed the usefulness of SNA to identify patterns or regularities of inter-relationships among interacting units. He sub-divided networks into (1) one-mode (uni-partite) networks using students as the entities and (2) multi-mode networks also incorporating Lecturers, and University administrators. For our research we focused on the one-mode network with students as our entities. Grundspan [8] *et al*. utilised Social Network Analysis in the domain of Education Research and specifically in the analysis of groups formed within the classroom.

4. Data sources

The collection and assimilation of data is the first step of the usually long process of data mining. Data mining is a method of processing data with the specific aim to obtain useful knowledge from the data. Osmanbeovic *et al*. [13] compared numerous approaches to data mining when researching the impact of demographic variables, previous academic results and academic ambitions on their final exam results.

Hanneman [9] introduces the concept of Power and examines methods of measurement based on Freeman, Borgatti and Everstt who are the authors of the widely utilised social network analysis package, UCINET [2]. Longitudinal data collection examines how networks changed over time, giving rise to two questions, namely how has a network changed over time and how can the past predict the future ? These methods are usually labour-intensive and require extensive human interaction whether it is observational, recalling and recording events or interpreting questions and expressing their response in a manner that can be accurately recorded.

Our data is a set of WiFi access logs. These logs contain each access request from a WiFi-enabled device on the University campus, to the University's IT network. Each device uses the unique logon details of the registered student, or staff members, to verify their credentials and provide access to the University's online assets and records the date, time, the asset accessed through the address of the WiFi base station used.

The main campus occupies a 50-acre site with more than 30 buildings with universal Eduroam coverage. To support this there are 780 WiFi base stations distributed around the campus. This allows us to determine with fidelity or accuracy, where the student is at any point and therefore surmise their activity including their company. So, for example, we can determine whether a student is in the restaurant, or one of the cafes or library and who else is in the vicinity at that time i.e. potential group members.

Haven taken our sample day for our initial analyses we created a sub-set of data based on students for a single module. Figure 1 identifies 40 of the most frequently accessed stations. Not surprisingly the top locations are transit areas, areas students pass through or congregate in between classes. Within the top 10 are classrooms and sports areas.

For the quantitative stages of this research all user identities are anonymised. Their reasons for interacting with the Eduroam system are of no concern as we are interested only in identifying students' on-campus locations. Furthermore we focus on the subset of students in certain undergraduate degree modules. These were chosen as they span a broad section of disciplines and have a large number of registered students and student types such as full-time vs. part-time.

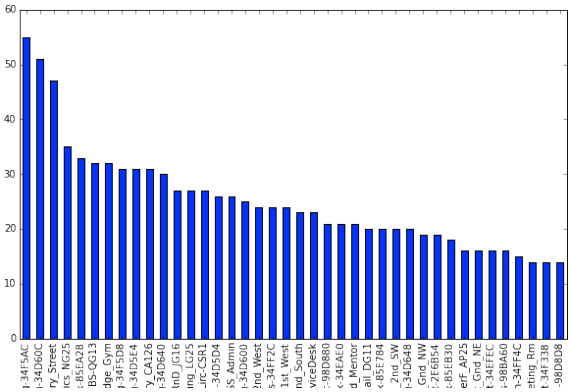

Figure 1: Location of students for sample day.

Our first hypothesis is that endogenous groups tend to form between people with similar characteristics, backgrounds and social demographics. To test this hypothesis we carried out SNA and identify groups with a high degree of centrality. Our second hypothesis will examine the correlation between the types of students and the number of groups they are members of, to determine a correlation, if any with academic performance.

This will be based on identifying a set of characteristics for each student and applying these to an analytical algorithm to build a profile for each. These characteristics include:

- Previous academic achievements
- Mature student y/n
- Gender
- Attendance at lectures
- Access patterns to the Moodle VLE
- Social-demographics
- Time spent in academic Vs. social areas
- Number of friends (Pairs)
- Friends' scores

Using the characteristics of group members, our hypothesis will test if students with similar characteristics form groups and that potential academic achievement can be derived from the group profile. Androushchak [1] examined the role of peers in student academic achievement in exogenously formed University groups. He found that the presence of high-ability classmates has a positive effect on individual grades in overall academic performance.

5. Methodology

Using a similar methodology to Manski [10] each student in the population is profiled using a set of characteristics derived from the categories academic, personal and social. Linear regression analyses will identify unique baseline scores for each student. Integrating the scores with those of their friends a total score for the student can be formulated.

This algorithm will be run each week to identify variance and re-calculate a Predictor based on the data collected from previous week's data. The Predictor identifies which groups predictive score has varied in a negative sense from the previous weeks scores, similar to Rui Wang's [16] behavioral slope. It is accepted that many students are members of more than one group. Palla [14] recognises the intricate nature of overlapping community structures and the sub-units from which they are comprised. He defines a community, or more specifically a K-clique community as a union of all K-cliques that can be reached from adjoining K-cliques. It will therefore be necessary to consider correlated effect. We refer to correlated effect as the impact of an individual being a member of a network of groups and their interactions within the University Campus community and its environs.

From these findings we can infer relationships or friendships, furthermore inferring the contexts in which they occur. The context, duration, number of participants and frequency of the interactions will be used to determine the type and strength of the friendships. There will be a separation between time spent with groups in various contexts. Time spent in an area that has been classified as "Social" (e.g. Cafe or Sports area) will have a different weight to time spent in an "Academic" (e.g. Library, study room) tagged location.

Using the DCU WiFi access logs collected by the Eduroam system for a complete academic year, we established the durations and locations that students spent their time on campus. We used the log data to identify which student groups interact on a continuous basis both within the confines of the class and also specifically in areas where gathering would be for more social or for shared studies. Using our previous mentioned data set representing the students of a single module on a single day we calculated the numbers of students co-located at any one time.

Figure 2 gives a breakdown in 10-minute segments of the maximum number of students congregating at anytime. From this we can see, for example, that at 9:40, 10:50 and 11:50, the largest groups of students were congregated together, which tallies with scheduling for lectures which commence on the hour. Observe that at 11:50, 18 students from one sample module were collocated.

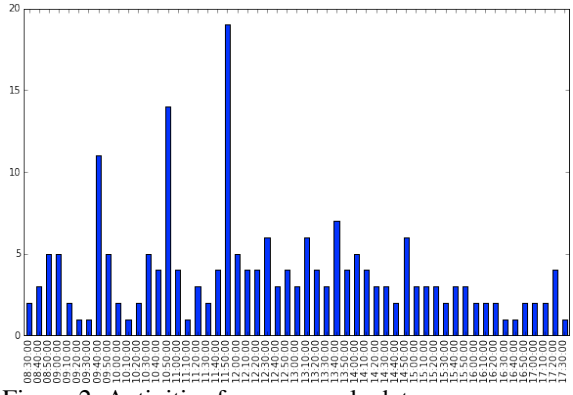

Figure 2: Activities from a sample date.

6. Conclusion

When commencing our research we posed a number of questions and developed a number of hypothesis. Our questions included:

1. Can we identify the make-up of student groups from the analysis of a university WiFi logs ?
2. Does the number of groups a student is a member of influence the academic performance of the student?
3. Does the make-up of a group dictate the academic performance of the students in the group?

We believe that we can identify groups from WiFi logs. From this analysis we can determine each groups constituent make-up. Our research has progressed to the point where an analysis of data has commenced. A preliminary analysis of WiFi traffic

has proven that we can identify clusters of students by location and calculate a score per student per day

Our approach is a non-invasive strategy using ubiquitous log data which can identify the movement of an individual through the University campus. As previously stated by Manski [10], inferring influence requires further information about the members of a group. We believe it will possible to collect and correlate the data we need from University sources preserving the integrity of our approach, which is a non-invasive study.

7. Future

At present, profiling of students is predominantly historical and based on static information. Our objective is to use our dataset for a more granular analyses culminating in the profiling of groups (communities) and membership of those communities. Once we have established a technique to do this our objective is recommend an optimum approach to the formation of study group course work to maximize academic achievements.

An interesting approach to the profiling of students is through their use of Social media. Greenhow, [7] identified the use of social media as a support mechanism for students. We believe there could be benefits to the examination to profiling individuals through their use of social media activities.

Acknowledgements: This research is part-funded by Science Foundation Ireland under grant 12/RC/2289

8. References

[1] O. Y. M. Androushchak, Gregory. Poldin. Role of peers in student academic achievement in exogenously formed university groups. Educational Studies, 39:568–581, 2013.

[2] S. P. Borgatti, M. G. Everett, and L. C. Freeman. Ucinet. Encyclopedia of Social Network Analysis and Mining, pages 2261–2267, 2014.

[3] E. Brewe, L. Kramer, and V. Sawtelle. Investigating student communities with network analysis of interactions in a physics learning center. Physical Review Special Topics-Physics Education Research, 8(1):105–108, 2012.

[4] M. H. Carney. Identifying Peer Effects: Thinking outside the Linear-in-Means Box. 2013.

[5] S. E. Carrell, R. L. Fullerton, and J. E. West. Does your cohort matter? Measuring peer effects in college achievement. Technical report, National Bureau of Economic Research, 2008.

[6] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. Personal and Ubiquitous Computing, 10(4):255–268, 2006.

[7] C. Greenhow. Online social networks and learning. On the Horizon, 19(1):4–12, 2011.

[8] Grunspan, D. Z. Wiggins. Understanding classrooms through social network analysis: A primer for social network analysis in education research. Life Science Education, 13:167–178, 2014.

[9] R. M. Hanneman, Robert A. Introduction to Social Network Methods. University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/), 2005.

[10] C. F. Manski. Identification of endogenous social effects: The Reflection Problem. The Review of Economic Studies, 60(3):531–542, 1993.

[11] E. Mayo. The Human Problems of Industrial Civilisation.

[12] B. Minaei-Bidgoli, D. Kashy, G. Kortemeyer, and W. Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In Frontiers in Education, 2003. FIE 2003 33rd Annual, volume 1, pages T2A–13, Nov 2003.

[13] E. Osmanbegovi´c and M. Sulji´c. Data mining approach for predicting student performance. Economic Review, 10(1), 2012.

[14] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society (article). Nature, 435, 2005.

[15] B. W. Tuckman. Developmental sequence in small groups. Psychological Bulletin, 63(6):384, 1965.

[16] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell. SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pages 295–306, New York, NY, USA, 2015. ACM.

[17] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications, volume 8. Cambridge University Press, 1994.