Draft Type: Complete

Revision Number: 1

Date Printed: September 12, 2018

Using Students' Digital Footprints to Identify Peer Influences on Academic Outcomes

Philip Scanlon B.Sc.

A Dissertation submitted in fulfilment of the

requirements for the award of

Doctor of Philosophy (PhD.)

to

Dublin City University

Faculty of Engineering and Computing, School of Computing

Supervisor: Prof. Alan F. Smeaton

15/07/2018

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed

Student ID

Date

Contents

Acknowledgements

I wish to thank Science Foundation Ireland for supporting the research which has lead to this thesis under grant number SFI/12/RC/2289, the Insight Centre for Data Analytics.

I also wish to thank DCU's Information and Systems Services, Registry and Faculty Administration Offices for their assistance during my research and the provision of the data needed to successfully conclude this research.

Thanks are also due to Dr. Wei Tai at SAP Ireland, for his assistance in the development of strategies for the use of SAP applications.

For their support, assistance and access to SAP applications, many thanks also to Ronan O'Connell, Paul O'Connor, Alan McShane and Colm Noonan at SAP Ireland.

I also wish to express a special word of thanks to Prof. Alan Smeaton for his help and guidance over the past four years.

"We approached the case, you remember, with an absolutely blank mind, which is always an advantage. We had formed no theories. We were simply there to observe and to draw inferences from our observations". Sherlock Holmes. The Adventure of the Cardboard Box.

Publications:

The following are the publications from this research:

- Philip Scanlon and Alan F. Smeaton. Identifying the impact of friends on their peers academic performance. *The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. August 2016, San Francisco CA, USA.

- Using WiFi Logs to Identify Student Activity in a Bounded Campus. Philip Scanlon and Alan F. Smeaton. *EC-TEL 2017, The Twelfth European Conference on Technology Enhanced Learning. Data Driven Approaches in Digital Education*, September 2017, Tallinn, Estonia.

- Innovative learning analytics research at a data-driven HEI. David Azcona, Owen Corrigan, Philip Scanlon and Alan F. Smeaton. *3rd International Conference on Higher Education Advances (HEAd17)*, June, 2017, Valencia Spain.

- Identifying student groups through their WiFi digital footprint. Philip Scanlon and Alan F. Smeaton. (2016) *Ireland International Conference on Education (IICE)*, 24-27 April 2016, Dublin, Ireland

- Philip Scanlon and Alan F. Smeaton. (2015) Predicting topographical and sociological information patterns from building access logs. In: *Event Insight Student Conference (INSIGHT-SC 2015)*, 30 Oct 2015, NUIG, Galway, Ireland.

- Philip Scanlon and Alan F. Smeaton. (2015) Predicting peer group effects on University exam results. In: *Insight Student Conference (INSIGHT-SC 2015)*, 30 Oct 2015, NUIG, Galway, Ireland.

Collaborations:

- The National Forum for the Enhancement of Teaching and Learning in Higher Education. Member of the Learning Analytics and Data Mining Advisory Group.

- Designing a best practice structure to assist institutions implement a Learning Analytics strategy. The National Forums Online Resource for Learning Analytics (ORLA). Launched 2nd November 2016.

List of Tables

Abstract

The ability of researchers to identify the type of activities and levels of interaction among students on a campus is important to research in Learning Analytics and in particular, anthropological studies which explore interactions among students. Historically the collection of base data in such studies has in the main been through observation, questionnaires or a combination of both. This work utilises the unique digital footprints created by student interactions with online systems within a University environment to measure student behaviour and correlate it with exam performance. The specific digital footprint we use is a students connections to the Eduroam WiFi platform within a campus. The advantage of this data-set is that it captures the personal interactions each student has with the IT systems. Datasets of this type are usually structured, complete and traceable. We will present findings that illustrate that the behaviour of students can be contextualised within the academic environment by mining this dataset. We achieve this through identifying student location and those who share that location with them and cross-referencing this with the scheduled University timetable. Our work uses the digital footprint to identify student location and thus co-location of students. From this co-location analyses we infer peer groupings and levels of interaction. This can be used for identifying peers in a University community and for identifying popular locations for different students and their peer groups to meet. This thesis examines the data collection process we followed and our data-mining process. Using the spatio-temporal data derived from the WiFi system we mined the data to produce actionable knowledge for use in the learning process. This research contains minimal Personal Data and no Sensitive Personal Data as defined by the DCU Data Protection Policy (Version 2.0). All data has been anonymised, stored and used in conformance with the Universitys Personal Data Security Schedule (PDSS).

Chapter 1

Introduction

Each student entering a University is unique even though the number of such students are quite high. Students' diversity is based on their life experiences, their social and home environments, their personalities as well as the educational system they encountered at both primary and secondary levels. Students in the main are gregarious and by their nature wish to form friendships with various degrees of comradeship. In attending University for the first time, some students may be moving out of their family home to take up temporary residence in rented accommodation. Others may be moving in with family relations or friends. There are many forms of such temporary accommodation from on-campus shared rooms, shared houses or apartments or "digs", where a student may be resident with a host family for five nights a week and receive room and board for the duration. While it is primarily the need for affordable and secure accommodation that shapes the decision of where to live, it will be the persons they interact with within this accommodation that could shape their formative years at university

While the attendance of students at secondary level institutions is predictable, based on the legal requirements in most countries to attend school for a specified time-frame, the same cannot be said about student attendance at formal education activities in third level institutions. At secondary level, the social groups to which a student is a member of are often formed in the early years and persevere through to their exit from second level education. When entering third level education, a young adult will be commencing a new phase of their development in a new environment, often without the supports previously relied upon in their earlier years in secondary level. Identifying the likely behaviour of students in a University environment and examining how that behaviour can contribute to their academic achievements been a source of research for many decades. The main reason for this is that, with such flexibility in their lifestyles and behaviour at third level, there is an opportunity to influence student behaviour in order to encourage students to do better, academically. Some of the seminal researchers working in this area include Astin, in [5], who was one of the first to study the impact of student interaction and the peer group effects on undergraduates at third level. The focus of much research is the identification of the influences on a student within the university such as their environment and those with whom they interact with, i.e their peers. Manski's, [40] "reflection problem" has been a central tenet to many peer influence research projects, since the term was coined in the early 1990's. Similarly Sacerdote, [51] Winston & Zimmerman [60], Hoel, *et al.* [32] and Carrell *et al.* [11] have all carried out extensive research into the effects of peer influences. This thesis will explore in greater detail the works of these researchers in the Background chapter.

Students entering University are initially gathered together into groups formed by University administration. These groupings are based on the degree programme they successfully applied for. Based on the random nature of applicants to a university program students will be placed in a group that they may be expected to remain in, for a whole year or possibly until the completion of their course. Lecturers in computer programming modules will often place students in small working groups for the dual purpose of helping students mix and also to work together on continuous assessment projects. These groups are also intended to mimic groups work in an industrial environment. Grundspan *et al.*[27] identified "the classroom as the principle domain wherein working relationships form between students" in a University environment, yet other work by Bruner [7] proposed that social interactions among students which take place outside the formal classroom or other teaching setting lies at the root of a student's ultimate academic achievements. We will be considering both these hypotheses during our research. Astin [5] found that the Academic Faculty also have a strong impact on students' development during their time at the institution and this is another aspect we may consider later.

1.1 Framing the Thesis Hypothesis

A university campus is comprised of a structure of numerous schools or Departments of Faculties, attended by students whose primary intention is to ultimately graduate with their desired academic qualification. From the moment a student applies to a university, then gains acceptance and starts to attend the campus they are creating a unique digital footprint of themselves within the University administrative systems. From the point of application and through registration, the university commences collecting data including demographic, prior academic performance and academic intentions through course preferences, for each student. Once a student is accepted and registered with the University for any degree programme, they are provided with a set of unique personal credentials that will allow them to access the University's IT assets and content, including the WiFi system. Like almost all Universities in Europe and further afield, the Dublin City University campuses have a WiFi platform provided by Eduroam with near total coverage throughout campus buildings.

As part of all students' day to day activities, the University will also gather information on activities like library attendance, book withdrawals, online activity on University Virtual Learning Environments (VLEs), use of ICT (Information Communication Technology) resources and access to WiFi networking system. In fact once registered, each student becomes a source of continuous streams of data relating to their activities on and even off the campus. Information regarding the modules students studied including examination achievements as well as some of their extra-curricular activities, like social club and society memberships, is also logged digitally. During their time attending the campus, students in varying numbers, will spend their time in academic, social and sports or recreational locations on the campus depending on their academic deadlines, their interests and the interests of the social and academic groups they are associated with. As each student's activities are unique to them, the activities and behaviour recorded about each student can be determined by their individual digital footprints. A student's digital footprint has many components, bound by a common thread of technology such that each component is captured and held in digital form. We will therefore refer to it from here on as the students' ***digital footprint***.

Students' digital footprints are a source of data that may be of interest to many groups including educators, analysts, administrators and policy makers in the education, sociology, and pedagogy domains, as well as being of interest to students themselves. Learning analytics (described later in Section 2.4.1) can offer the tools to mine this data producing knowledge databases usable by the various interest groups.

The ability of researchers to identify the type of activities carried out, and levels of interaction among and between students on campus, is important to research in the domains of academic and learning analytics and in particular, anthropological studies which explore interactions among students. Historically the collection of base data in such studies has in the main been through observation, questionnaires or a combination of both. The approach taken in this thesis is novel in that we do not depend on these traditional data capture methods and we gather data which is robust enough to be usable in decision-making.

As previously outlined, a student's digital footprint can comprise of a number of components many of which have been used in previous work in learning analytics including:

- Demographic information
- Previous academic history (prior to joining the University)
- Performance in assignments and exams
- Library usage and book withdrawals
- University affiliated Clubs and Societies memberships
- Online activity on University Virtual Learning Environments (VLEs)
- Usage levels of University ICT (Information Communication Technology) resources like programming environments
- . . . and more.

In an era where smartphones and WiFi use is widespread, this thesis will examine another source of research data which is of use to an anthropological study of the influences on student exam performance. We examine the use of WiFi-enabled devices within a bounded domain, i.e within a University campus.

We will examine this data source within the context of the physical educational space and learning analytics. Our approach uses the digital footprint that WiFi-enabled devices leave to identify student location and from this, the co-location of student groups. From this co-location analyses we will be in a position to infer peer groupings and group compositions and thus infer levels of interaction among students. While on one level this data can be used for the identification of peers amongst a student community, it can also be used for the identification of popular locations where different students and their peer groups spend time at various times of the day. In Chapter 4 there is further context given to student activities by subdividing the day into core academic and social hours. *Core hours* are defined as those hours within which academic classes, laboratory sessions and tutorials are scheduled i.e. 09:00 to 17:00 Monday to Friday. All times outside of this range are considered to be *social hours*.

1.2 Research Questions and Overall Hypothesis

As a major part of our research, we are interested in the data collection process and the role of technology in that process. We will use spatio-temporal data derived from the WiFi system access logs to determine on-campus location as a component of student digital footprints. Once this data is gathered and anonymised, as described in section 1.4, learning analytics tools will be employed to mine the data and to produce actionable knowledge for use by the various stakeholders in the learning process.

A prerequisite to the progress of our research is the ability to identify where various student activities take place including lecture attendance, laboratory usage, use of the library and other study locations, as well as places used for informal social gatherings. In our case we do this through the use of WiFi network access request logs and once we do this then this initiates a range of related questions including the following.

Research Question 1:

Can we identify a student's activities from the data held in WiFi access logs? This research question will be addressed more fully in Chapter 4. There are a number of avenues of access to the DCU Campus. On entering the university campus students' WiFi enabled device can be configured to automatically request access to the Eduroam network. The device will be identifiable as belonging to the student by the unique username that is making the request. The student's approximate location can now be identified based on the Eduroam access point that they connect to. We wish to determine if it is possible to continue to identify the location of the device throughout the day as they attend classes in various buildings as well as visit other campus locations.

Research Question 2:

Research question 2 will build on the research addressed in Question 1 and will ask whether it is possible to identify student friendships among student pairs and larger groups, through the analysis of WiFi logs? In this question we ask whether the digital footprints left behind by students can be correlated among pairs or larger groups and where we identify a strong temporal and locational correlation for a student pair/group then can this imply a friendship.

Research Question 3:

The third research question asks whether there is evidence from the analysis of WiFi logs which infer friendship and peer groups, of peer influence within student groups and more specifically, is whether peers can influence others' exam performance?

A number of interesting sub-questions will also be explored while examining the answers to our three main research questions. These include:

- Can we identify student groups such as the formally constituted groups like class attendees as well as informal groupings?
- Can we infer the activities of students from the location?
- Does the make-up of a group of students influence the academic performance of the

students in the group?

- Can students who are isolated and not engaging with others be identified early in a semester, especially in the case of first year undergraduates?

- Can we profile those students who are more likely to drop out of University, early in the semester, based on their group participation and perhaps intervene to offer additional assistance?

Our reason for exploring these questions is because central to much of the research on peer influence is the ability to identify those who spend time with others, and in what groupings, and for what purpose, from study group to social gatherings.

This line of thinking leads us to our overall hypothesis for the thesis:

> "That we can use students' digital footprints, especially those indicating their physical locations, which yields unbiased data, to identify academic collaborations and social friendships and from that we can quantify peer influences on exam performance at third level education"

1.3 Motivation

Data collection in educational research has been undertaken in an effort to answer many research questions regarding the effectiveness and efficiencies of academic systems and their many component parts. While endeavouring to answer these questions, researchers are interested in two broad categories of data collection. Depending on the format of the questions and expected answers, these are qualitative and quantitative data. Quantitative methods rely on structured observations and interviews often including pre-formatted tests. Qualitative collection methods are less formal and focus on the participants' narrative and use less structured observation and interviews. In both studies the subjects are fully aware of their role in the research.

With the ubiquitous use of digital technologies, other data sources are now available to researchers include access logs from University Virtual Learning Environments (VLEs), E-learning and Massively Open Online Courses (MOOCs) and logs from other knowledge management systems. The data from these systems can be used to identify a students behaviour on an ongoing basis. Recent work by researchers such as Minaei-Bidgoli *et al.* [44] focus on data collection from web-based educational systems. The use of e-learning systems provides useful data based on a student's interaction with on-line materials. This is another form of pure data which is of use in the Learning Analytics domain and in the future may be used to augment the research we carry out here. However this will be one of the few references to this kind of data in this thesis.

The advantage such automatically generated data-sets have over those generated through questioning or observation is that they grow naturally. Data-sets derived from the personal interactions a student has with the University IT systems, are a unique digital footprint of that student. Data-sets of this type are less susceptible to the inherent biases introduced through the intervention of human interpretation, they are usually structured, complete and traceable. Maslow [41] pointed out that when dealing with "the needs of human beings we talk about the essence of their lives" and questions how "this essence could be put to the test in some animal laboratory or some test tube situation?". We believe that he was questioning the ability to understand human behaviour when observed in an unnatural environment. Obviously, to gain a correct insight into a subject, the subject needs to be observed in a real life situation in their own natural social environment. Placing a research subject in an environment where s/he is aware they are part of a study or asking a subject to recall or put into context answers to a set of questions, is arguably a "test tube situation".

Initially our motivation for carrying out this research focused purely on the uniqueness of the data set we pursued and were able to use. We were unable to find any other research that had examined a student's digital footprint in the same manner as we intended and therefore there was no other work which could have addressed the research questions that we address in this thesis. During our research it also became apparent that a large portion of (from what was found) research in the domain of peer influence at third level educational institutions had been carried out in the USA. The majority of this work had been done at residential complexes i.e where students lived on campus. It can be reasonably assumed that living on, or within close proximity to a campus creates an environment where students share many of the resources and assets with the same groups of students and staff, both academically and socially. Thus, this posed an interesting perspective on our research as we could compare our work in a small to medium sized University against the majority of the literature which is based on largely residential Universities of much greater size.

When we consider our own data-set gathered and used in this thesis and its origin, coupled with the fact that it was collected from a partially residential and relatively small campus, this places this research as unique in the terms of educational peer influence research.

1.4 Ethics in Educational Analytics Research

Institutional IT and manual systems such as those in a University are a repository for large amounts of data specifically relating to individual students and their activities both current and historical. It is now legitimate to question the use of such data and the legitimacy of an institution to use this data for the betterment of the pedagogy approach and enhancement of the learning experience and ultimately improve the learning experience of present and future students. University administrators are the custodians of data which could potentially be of benefit to future students and therefore this data "should" be used for the betterment of those students or "should" it?

Willis *et al.*, [59] believe the availability of big data in higher education raises unique ethical questions for administrators. He states that "to know entails an obligation to act". Based on this premise we believe we should use available data, but with many obvious limitations and restrictions. Having made the decision to act we must consider what ethical obligations we are bound by and we will delve into this more deeply.

Big data analytics techniques which can be categorised under the heading of Learning Analytics (LA) which aims to develop actionable intelligence from this kind of data. As with the use of all data analytics there are legal and professional codes to be adhered too. As with all data mining on personal data, this research faces the ethical issues associated with educational inquiry. These obligations are defined by the ethical committees that bind us to advocate due diligence and have laid down comprehensive guidelines for us to work within to ensure there is a balance between risk, reward and community expectations.

We believe we should act for the betterment of future students but with due consideration given those whose data is being mined, to ensure they are not identifiable or relate-able to in any way. i.e. individual identities are protected. We recognise that institutions have an obligation to protect students' data while utilising it to better them and their peers. This research and subsequent research has been carried out in accordance with the University's ethical policies and values.

Another aspect of ethical debate surrounds the approach of whether the research is open or closed, that is if the research cohort are made aware of their participation. In the case of the research in this thesis we are examining historical data and therefore the relevance of this question is diminished.

As outlined in the Dataset and Data Used section (section 3.1) we have taken demographic and WiFi access log data which is personalised in its raw form. To ensure the anonymity of our student data-set we examined all the various datasets we intended to utilise in our research for personal identifiers. Utilising a Python implementation we have hashed any students identification data such as their personal id, username, fore and surnames and their student number. The result of the hashing algorithm is a 32 character string, for example: "0cj8sn5isbr4ojtna9ne678hg439nhed" Each student will therefore have a unique identifier, that is not traceable back to source. We believe that once we have secured the identity of our source cohort we could proceed with our analysis.

We had a requirement later in our research to carry out a Focus Group event with a representative sample of the student cohort. We will cover in greater detail the processes, participants and results in the section on **focus group** (see section 3.4) of the Data chapter. We will also examine the findings of this focus group from which we developed a questionnaire also covered in section 6.3. This questionnaire was sent to the complete audience of our research cohort and was designed to capture information on the normal pattern of student behaviour during an academic semester. In both the focus group and cohort survey we collected no personal information. Similarly to all data used in this research the identity of the participants and all data collected, could not be used in any way to identify individual students.

As this research is being carried out under the auspices of DCU Insight Centre for Data Analytics, we will be giving due regard to the Standard Operating Procedures (SOP) associated with the handling and controlling of personal data. Personal data is defined "as data relating to a living individual who is, or can be identified from, either the data itself or from the data in conjunction other available data". This research will be operating in compliance with the SOP, "Personal Data Security Schedule (PDSS)".

It is intended that the data secured for this project will be anonymised prior to any analysis being undertaken. No individual will be identifiable directly or indirectly from the results of the analysis. No data from this research project will be shared outside of the research group and all data on completion of the project will be deleted. While we would like to contribute to the philosophy of open access" to research data, because of the nature of this data, this will not be possible here All non-anonymised data will be stored securely within the confines of the DCU Insight research lab on a password protected PC. A Personal Data Security Schedule (PDSS) has been raised and returned to ensure complete compliance.

1.5 Contributions

There are a number of contributions from this thesis. One intended outcome of the study, on an impacting theoretical level, is the integration of some of the many data sources available within a university. We will examine students not just on a pedagogical level but also at sociological one. The approaches undertaken include the:

1. Development of a methodology to mine a unique set of data, which has not been utilised before.

2. Identification of friendships from a data set which has not been biased by human interventions.

3. Identification of popular locations for students to congregate in.

4. Differentiate between random, regular and strong friendships.

5. Examine the concept of peer influence over time in a bounded campus.

We believe that taking a unique approach to a well researched area, our contribution will stimulate a new conversation in this domain. Further conversations must include a number of relevant questions that would need to be answered into the future.

1. What are the capabilities and limitations, in a Learning Analytics domain, of the combinations of various data sources?

2. Examine potential biases of research results through choices of research cohorts chosen for research. i.e would this research be scale-able to other academic groups?

3. How relevant will WiFi be into the future with the advent of high speed (5G) cheap data packages for mobile devices?

4. How will this research methodologies keep pace with the advances in technology.

A central premise of this research is the use of unbiased data. It is important that it is understood that the references to biased and unbiased refer to the collection methods of the researches primary data source, i.e. the WiFi log data. It is accepted that once a research cohort is chosen from universities population we will be introducing an inherent bias based on the cohorts unique construction.

1.6 Overview of Thesis

This thesis consists of 6 more chapters, Background, Data and the three main chapters (Chapters 4 to 6) which are the main body of the work, presenting our approach, research techniques and findings. The final chapters will summarise the research and propose future research directions.

Background (Chapter 2) This chapter situates this current study in the relation to previous studies and the available literature. We explore the domain of "Educational Data Mining" which is concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational setting. We employ a number of these methods to better understand the students, and the settings in which they learn. We discuss the current practice of data collection relative to our approach and the use of technology in the processes. This includes a critical review of the historical context of educational research at third level and the domain of Learning Analytics.

Data (Chapter 3) We provide an in-depth exploration of the numerous data-sets required to develop this research. There is an examination of the processes and challenges that led to the development of each data-set and an explanation of the importance they play in every aspect of our research. It was necessary to interrogate multiple data-sets concurrently to glean the knowledge necessary to answer the research questions posed in this thesis. While a kernel of this research is the collection of non biased data through the use of technology, there was a consultation process carried out with the cohort of students being researched. This takes the form of a **focus group** which we will cover in-depth in this chapter and further in Appendix E. The cohorts comprises of two distinct student groups from the School of Computing, namely students who have registered for the Enterprise Computing (EC) and Computer Applications (CA) undergraduate degree programs.

Methodologies and Proof of Concepts: (Chapter 4) Our research is based on the premise that students on campus are engaged in either academic or social activities. We believe that a student's on-campus location can be used as an indicator of their activity. In this chapter we will examine the concept of mining WiFi log data to identify individual student activities on campus and from this we can infer activities and friendships. The development of these methodologies included the examination of the potential tools available to mine very large datasets. As part of the "Proof of Concept" we examined a sample dataset to determine the feasibility of differentiating individual students by time and location within the campus boundaries. As part of the proofing process we examined a number of potential tools that could be employed in our research. Presented here is the decision making process in choosing those tools we considered best suited for our purposes.

Experiment (Chapter 5) Having examined the concept of our research and determined that the it is possible to mine large sets of *log data* accurately, we then focused on the development of the methodologies to establish context from the raw WiFi data. The experimental stage is designed to develop a set of robust tests that could be used to address the “Research Questions” posed earlier. There were a number of techniques examined with an emphasis on various clustering techniques. The experimental stage examines each of the research cohorts (EC & CA) as two distinct groupings. The experiments are designed to identify distinct **strong friendships** and compare their examination marks over the course of three academic years. We carry out a number of micro and macro examinations of the cohorts and their individual members to establish what relationships they develop over time. We identified distinctive behaviour pattern differences between the two research cohorts with regard their daily activity patterns. These existance of these variances would have had to consider for each academic Programme regardless of School. Each programme will have a distinct personality bias inbuilt which would have to be considered if scaling the research.

Findings and Conclusion (Chapter 6) The concluding chapter will assess the question posed and outline the journey undertaken to answer those questions. This chapter will present the decision making process used in the development of the methodologies chosen in all aspects of the research. The research approach and the findings are summarised, appraised with the results presented as they relating them to our questions. The thesis will conclude with an evaluation of the research and suggest potential future research directions.

Chapter 2

Background

2.1 Background

In this chapter, we examine some of the concepts covered during the research process and the related work in the various domains we encountered on our journey. We will illustrate our research developments in terms of the tools in the fields of data collection methods, Learning Analytics and specifically peer influences in an Irish third level institute.

Data collection in research on people can be invasive. It is our contention that any interaction or involvement of a third entity in an environment changes the constituency of the environment and thus is in danger of introducing bias into the data collection process. It is however, difficult to estimate the bias effect of a subject's awareness of being part of a research project has on the final results. Early research by Mayo [42] referred to this as the Hawthorne effect. McCambridge *et al.* [43] investigated the "Hawthorne" effect in a heterogeneous group of studies, concluding that there is not one single form of bias, but multiple forms that can be introduced depending on the research methodology.

2.2 Data collection and the identification of groups

The identification of groups within a complex network of people is not an easy task. One method, identification by observation, would required extensive resources including multiple trained researchers. In our application area, training of researchers to ensure standardised understanding of student behaviour and the requirements to record, store and interpret this qualitative data is expensive and not feasible. Alternatively we could just ask, i.e. question each individual in a research cohort for a list of their friends, the groups they belong to and then try and measure the degree of friendship. This approach has been used extensively in research but has limitations, for example the work by Celant [12]. Celant asked students directly to recall who they spend time with, socially and academically for example in their social gatherings, jointly working on homework, or as part of formal study groups. He found that there was some **blurring**, that is an understanding of the level of a friendship among all those who have taken part. Students had different interpretations of meetings. While one may interpret an informal study session as just that, another may see it as a coming together of a group of friends i.e. "different students may have a different concept of preparing for an exam with a course mate". Other students who, for different reasons who may have a very narrow circle of friends and do not interact with peers at a level they would prefer, may claim friendships with popular peers who they actually had little interaction with. These students may not wish to admit to having few social or academic friendships and thus create links which would introduce a bias into the results.

Eagle, *et al.* [20] using technology, compared observational data from mobile phones with standard self-report survey data which identified variances between the two data-sets. Eagle attributed the variance to recall bias. Recall biases included *recency bias*, where memories are biased toward recent events and *salience bias*, where memories are biased toward more vivid events. Their findings inferred that recall can be affected in many ways that casts doubt over the accuracy of questioners and interviewing techniques for collecting data. Our method of collecting data will avoid the pitfalls of the questionnaires while also negating any Hawthorn effect [42]. While the influence of the Hawthorne effect seems to be an unavoidable bias in much quantitative research, our work is based on data collection which is unobtrusive and has minimal contact with the research subjects, eliminating many biases. This is achieved as we perform data analytics on ambiently-collected log files. This method ensures a degree of separation between the researcher and the subjects themselves.

Securing a data set is the first stage in the learning analytics process. Our research utilises a number of distinct data-sets, the first being a database of WiFi *access request* logs for the University's full academic years, 2014/2015 to 2016/2017 i.e. six academic semesters both Summer and Winter, including data for the intervening holiday periods. Our second data-set comprises the exam results of students registered for selected modules during this time. The exam data set also includes some categorical demographic data for the module participants. Thirdly we constructed the academic timetable of our student cohort's chosen program and finally we constructed a database of campus NAS locations and their associated MAC addresses. A greater level of detail will be provided on each of these data-sets in the data chapter 3.1

2.3 Dublin City University

Dublin City University (DCU) is a multi-campus university accommodating an academic staff of 440 and approximately 16,000 students each semester. It is a modern facility contained on a 50 acre campus, compromising of 27 separate buildings providing an approximate floor space of $180,000m^2$. On campus facilities include 1,400 residential apartments, 7 restaurants/cafes and numerous shops including convenience, book, print stores and a pharmacy.

The campus is located within a mainly residential area with good public transport system which links it directly to the city centre and other suburbs. There are a number of private coach companies which provide services to more distant towns within the commuter belt. In the section covering student demographics, Section 3.2.4, we will analyse the profile of students with respect of domicile and transport to and from college on a daily and weekly basis.

There are currently a number of constituent colleges within Dublin City University, including All Hallows College, Mater Dei Institute of Education and St Patrick's College of Education. These occupy separate self-contained facilities located externally to the main DCU campus in Glasnevin.

2.3.1 Program modules

In section 4.1.2 on **research approach** we examine the processes we undertook to determine the most appropriate school and academic programs to include in our research. We determined that our research cohort would be drawn from the Faculty of Engineering and Computing and specifically from the School of Computing. Within this School our research will focus on the students in two undergraduate degree programmes namely, Computer Applications (CA) and Enterprise Computing (EC).

The **Computer Applications (CA)** CA degree is a four years honours degree providing an emphasis on students developing an in-depth knowledge of software engineering. Students focus on software engineering, databases, multimedia, computer graphics, artificial intelligence and computer security. Applicants for this program are expected to have a very strong ability in mathematics. The entry level for this program, based on the Irish GPA system is approximately 400 CAO points.

Enterprise Computing (EC) EC students in contrast study a four years honours degree program covering the contexts within which software and information systems address real-world business problems. Students in this programme study how information technology and information systems are integral to business processes and they develop methods for improving and re-designing their use in organisations. The entry level for this program, based on the Irish GPA system is slightly lower at 370 CAO points and has less of an emphasis on mathematical ability.

These two programmes have been chosen as they share some modules and have a common degree format which attracts students with similar interests in the IT domain but within different spheres. Our cohort while in the main have taught classes in the Computing School building, have a large number of classes taking place across the campus in the Business, Nursing and Language Schools and this are marked in purple in Figure 2.1.

2.3.2 Eduroam

Eduroam (education roaming) is a secure, world-wide roaming access service for wireless access to internet resources, available across multiple sites worldwide. It is a cross-site

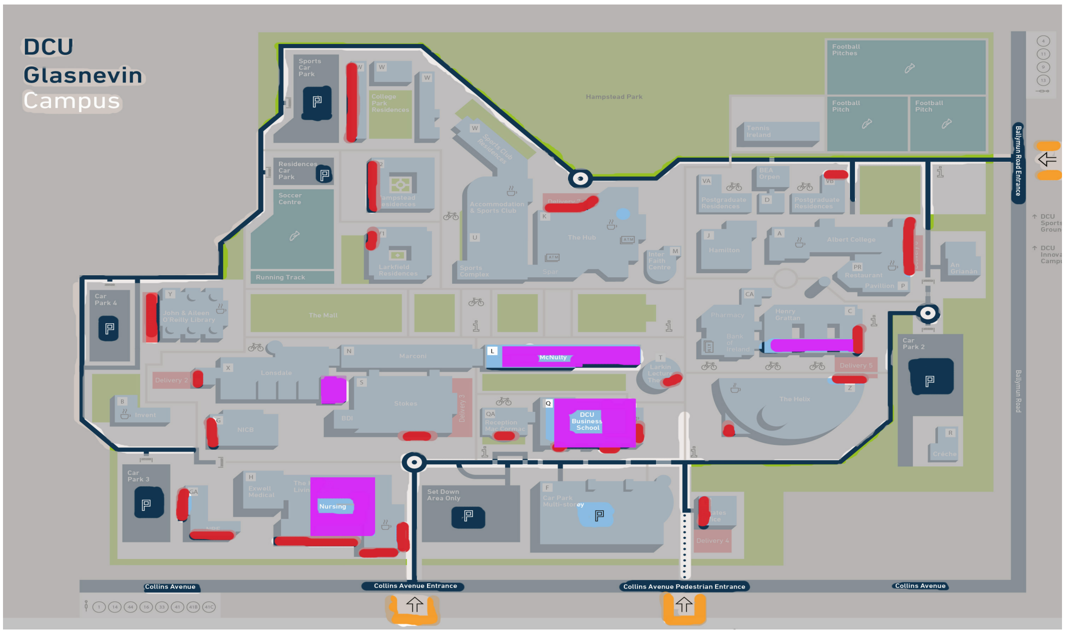


Figure 2.1: DCU access points and building identification.

infrastructure which allows users gain access to the internet in one location while being registered at another. This flexibility allows the use of the facilities at partner Eduroam sites. Access is provided through authentication via the user's home site, where their credentials are stored. Once verified and validated they are granted access to the host infrastructure. Network access at member sites is via WiFi, 802.1X protocols and utilises the Network Access Server (NAS) infrastructure at the institutions site. The campus at DCU comprises approximately 1000 individual NAS. These NASs are distributed across the complete campus ensuring continuous WiFi coverage for users regardless of their location.

When a user enters an Eduraom enabled site, if they possess a WiFi enabled device that is activated, it will automatically request access to the local wireless network. Figure 2.1 identifies in red, the perimeter points for the campus where a student is most likely to first request access to the Eduraom system, dependant on their mode of transport. Pedestrians will enter of any of the main routes accessing the first NAS they encounter. Car drivers once they exit the car park on foot will connect to the NAS they first encounter. Each of these requests, when it occurs, creates a log of the request and all the subsequent validation network traffic, but not the content that is actually delivered to the wireless device. It is these access logs that we will examine in greater detail later in this thesis, in Section 3.1, as it is one of the cornerstones of our research.

It is important to note that apart from the Eduroam system there are no WiFi platforms on the campus. Eduroam has two pathways on to it's platform, namely Eduroam and Guest Eduroam. This pathway offers basic web browsing SSLVPN(443) for guests of DCU. All Staff and Students are advised to configure their devices to connect automatically to the Eduroam system [2].

2.4 Co-location

In a University environment, some students may spend minimal time on the campus i.e. they attend formal classes only and then leave, while others may use the campus as a base for individual or group study and for socialising. The location of a student's domicile for the semester may influence the amount of time spent on campus. As described by Gonzalez *et al.* [26] there are patterns of locations visited by students at their University campuses and as a result there are some locations that students will tend to spend more time in, i.e. those that have a greater context to them and their peers. This context may be social or academic and may also be either shared or a private area for study. An important aspect of our research is the ability to place a person in a location and to determine the context of the visit to that location.

Co-location in the arena of our research can be interpreted as the location of two or more individuals in the same physical location at the same time. Individual incidences of a co-located pair cannot be simply interpreted as the individuals having a relationship. We will be examining the quantity of co-locations or meetings and the context of the meeting to determine if there is a relationship and if so what is the nature of that relationship. Li *et al.* [37] having mined subjects' GPS logs, used hierarchical clustering to develop a trajectory model that determines a semantic meaning to stay-points (points were time is spent) and inferred similarity between subjects based on this. In that work, the intention was to use the results as the basis of a recommender system. At the core of that research was the identification of subjects who visited locations at specific times and from this Li inferred context for the research subjects. Similarly in this thesis we will be inferring context based co-location and the nature of the location.

Meetings or gatherings of two or more people in the same place and at the same time, in the context of a University campus can be either formal or informal i.e. they can formally be scheduled to meet others through their attendance at a class or lab or they can be meetings in social locations. Social locations may be more important to them as they are the locations that their friends may also attend. These two classifications do not however categorise all the reasons that students visit a location. For example, peers who attend the library together do so for academic reasons and not social, and may not have any actual interaction in the library, though both are there at the same time. This is a very important distinction we will examine and clarify within this work. Gupta *et al.* [28] uses this distinction in his analysis of People-to-People geo-social recommendation system analysis. Their experiment examined a month of mobility traces collected from smartphones running Intels PlaceLab location app that identified user locations and their group meetings. One intention of his research was the identification of student's popular hangouts or study areas i.e applying context to groups.

It is an obvious development that technology is used in research to collect data to identify the activity of subjects. The use of technology will be covered in a later section but here we will briefly cover a prime example of the use of technology in the identification of co-location. Cranshaw's *et al.* [16] research project tracked the location of their 489 subjects using GPS and WiFi position applications installed on their mobile devices. Co-location was deemed to have occurred if subjects were within 30 meters of each other, within a 10 minute period. He supplemented his data with traces left by his cohort who used the Facebook location sharing app Locaccino plus. He believed this was necessary as co-location did not provide enough evidence to reliably establish a relationship between subjects. This study required all members of the cohort to be active members of Locaccino. This study ranged in duration from weeks to months, with the number of participants varying from week to week. The cohort was made up of 285 users who were recruited as part of a number of research projects using paper flyers around the campus and posting on the Universitys electronic message boards. The balance were invited by other participants or they found Locaccino through other means. A point to note and one that we are cognisant of, over 50% of the cohort were subjects who are serial participants in research projects. We consider it legitimate to ask whether the they are typical of the population we expect them to represent?

2.4.1 Learning Analytics (LA)

A precursor to Learning Analytics (LA) as an independent domain of research was Educational Data Mining (EDM). EDM was defined by Romero & Ventura, [50] as an “interdisciplinary research area that deals with the development of learners”. EDM emerged in the early part of the 21st century as a method to explore the big data originating in an educational context. It was a development whereby stakeholders in the Educational supply-chain wished to identify usable knowledge from the patterns in the data. They wish to identify useful knowledge to aid in the decision making process of setting policies, programmatic development, delivery to the students, monitoring and measuring the outcomes i.e success of both the process and the students.

LA evolved from EDM research as it focused on the *Learner* in the process of data mining. It focused on the impact of making decisions with the student as the central focus rather than the institutional processes that are there to ensure the effective running of the institute. EDM and LA cannot be separated and constitute a set of IT based techniques. In this thesis we will examine the domain of LA, while not ignoring the potential of EDM.

Siemens & Long, [52] cite one of the accepted definitions that encapsulates the meaning of Learning Analytics which arose from the 1st International Conference on Learning Analytics and Knowledge, held in Banff, Alberta, Canada in 2011. This definition is now accepted by SOLAR (Society for Learning Analytics Research) and is:

> “The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and

> the environments in which it occurs”

The underling steps in the learning analytics process can be generalised as follows

- Collect accurate data.
- Use this data to answer questions either *a priori* or *proposed*, based on the findings of other research on that data.
- Convert the data collected into knowledge.
- Act on the knowledge to improve the system being examined.

An extensive review of the development and challenges in the learning analytics field was carried out by Ferguson [21]. She believed that circa 2010, learning analytics emerged as a separate entity from the main stream data analytics domain with three main (overlapping) areas of interest, namely:

- Technical challenges, extracting knowledge from big sets of learning-related data.
- Learning analytics, educational challenges, optimising opportunities for elearning.
- Academic analytics focusing on the educational results at national levels.

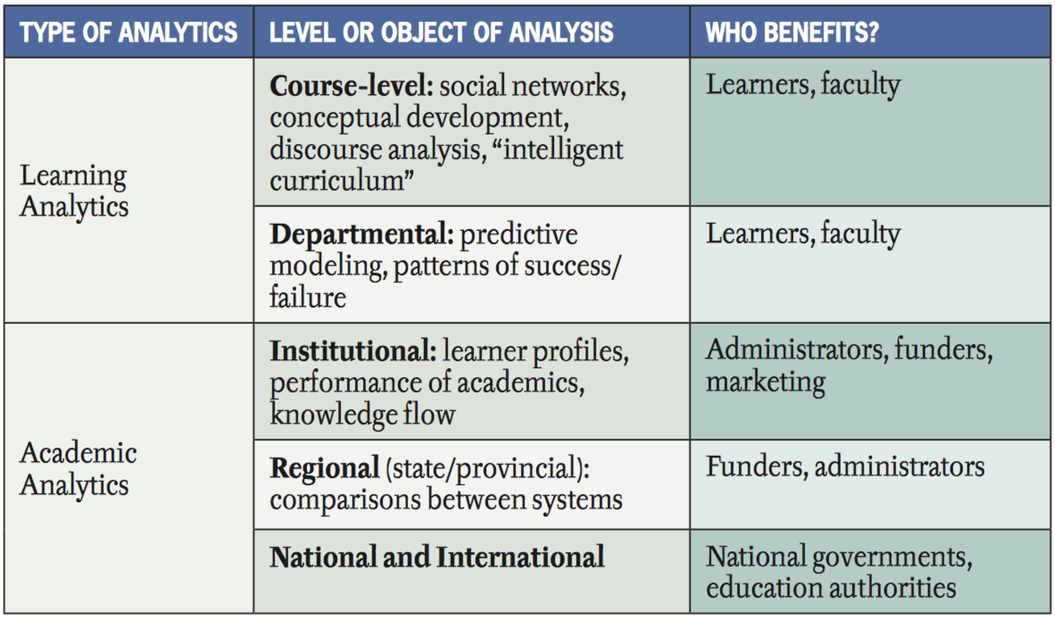
Our interest in the field of Learning Analytics is the technical challenges, i.e. the collection and analysis of large data-sets and the extraction of knowledge.

Learning Analytics has the potential to provide various levels of knowledge and actionable intelligence for Learners, Faculties, course administrators and decision takers at Departmental levels. Each of these groups interests include the provision of actionable intelligence for both learners and academics. Learners’ interests include course design and interaction with lecturers, students and University resources.

Academic analytics in contrast is a toolkit for administrators at institutional and funding authority level who dictate policies using more global data, for example, measuring the number of students retained year-on-year through to the conclusion of their programme.

This distinction between learning analytics and academic analytics was illustrated by Long and Siemens [52] and is shown in Table 2.1. This table distinguishes between the types, objects and beneficiaries of learning and academic analytics. Romero & Ventura’s.,

[50] examination of Education Data Mining (EDM) listed the stakeholders of the field which mirrored the distinctions made by Long and Siemens.



TYPE OF ANALYTICS	LEVEL OR OBJECT OF ANALYSIS	WHO BENEFITS?
Learning Analytics	**Course-level:** social networks, conceptual development, discourse analysis, "intelligent curriculum"	Learners, faculty
	Departmental: predictive modeling, patterns of success/ failure	Learners, faculty
Academic Analytics	**Institutional:** learner profiles, performance of academics, knowledge flow	Administrators, funders, marketing
	Regional (state/provincial): comparisons between systems	Funders, administrators
	National and International	National governments, education authorities

Table 2.1: Learning vs. Academic Analytics

As with many areas of analysis, Learning Analytics is based on historical data and thus retrospective, i.e. analysis is an examination of the past in an effort to be deterministic and have influence on, or predict the future. Often a students' measure of success is based on the outcome or a set of examinations at the end of a semester, and often on results alone which places a quantifiable measure on the previous semester's efforts vis-a-vis their own previous results and those of their peers.

Administratively, progress can also be measured against industry Key Performance Indicators (KPIs) such as retention levels or grades being achieved per programme module. While the extraction of knowledge is one challenge, this effort is immaterial if the educational process is not improved.

2.5 Related Work

While researching the domain of Third Level Educational institutions it was observed that a great deal of the available research carried out into student development was carried out in Universities in the United States of America. It is important and very relevant to this research to realise that Universities in the USA and specifically those most often cited in academic research are residential universities. Residential universities are those where a high proportion of students live on or within a very close proximity to the campus. Campuses are the centre of the University and are where most of the students' academic and social events occur. The campus to all intents and purposes is providing the kernal of the University community. Subsequently, students spend the majority of their time amongst the same community of people on a continuous basis. Therefore students are interacting with their own class colleagues, their room mates, dorm mates, and also on a regular basis, socially with students from other faculties.

Fraternities and Sororities are a common feature of US third level institutions that are central to the campus community. Fraternities are male social, professional groups whose members promote different combinations of interests plus social and academic achievement. Sororities are their female equivalents. Another common aspect of Universities in the USA is, "Dorm-life". Dorm-life is referenced to in many seminal studies such as by Zimmerman [60] and by Sacerdote [51]. Students who live on campus normally reside in dorm buildings and either have a single or a shared room. Rooms for first years are normally allocated on a random basis, with some Universities cognisant of not clustering athletes or racial minorities in Freshman dormitories. Renn [49] looked closely at this ecology of higher level institutes, examining the development of the student within the environment, while consideration is given to their previous influences.

In contrast, Irish Universities may have the majority of students living off-campus either at home or in rented houses or apartment accommodation, with a small percentage living in campus residences. Therefore, making direct comparison between these earlier studies based in the USA and our work would not be accurate. The students that make up our study cohort are a mixture of students residing on-site, locally or commuting from outside the campus neighbourhood. It is difficult to estimate the precise breakdown of their residence during the academic semester. As their semester residence is often temporary, many students will prefer to maintain their home address as their primary contact with the university. We have made some assumptions about the make up of our cohort's residence from the responses to our survey covering campus activity 3.1 which will be detailed in our chapter on Dataset and Data Used.

One commonality between Irish and US Universities and probably the wider academic world is the practice of streaming. Access to University is often down to a number of factors with the main ones being academic achievement (GPA/CAO) CAO and financial factors. Students may obtain the GPA or CAO points required to apply and be accepted onto the third level program that they prefer, but may be prevented from applying due to financial constraints or other circumstances. College applications may effectively be made based on financial or geographical considerations and this has an effect of streaming students based on these criteria. A seminal paper on the field of peer effect was carried out by Sacerdote [51] at Dartmount College, which at the time of the research was one of the most selective undergraduate schools in the USA based on incoming test scores and school class rank. Wang *et al.* [56] labelled the college as an Ivy League liberal arts college with its undergraduates being among the top high school performers. We recognised that the sample is skewed to high performers with good GPAs. Institutions that can afford to hire the brightest teachers are able to attract a higher standard of students and therefore will be in greater demand which effectively causes the entry level requirements to be raised. This could be considered streaming by financial stature.

Assuming all things being equal with regard to finances and geographical consideration, the Irish third level entry system has a requirement for students to obtain a minimum standard of CAO points. This can also be itself construed as a methodology of streaming students. Those who obtain the highest points get the first choice of programs and perceived popular courses attract high achieving students. This has the effect of bringing together students with a similar level of education and probability of similar economic and social demographics. We must now consider that if students in academic programs have been streamed and considered similarly, why is there a varied level of academic success between students in the same programs?

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 of. In this section we will compare and contrast some of this work to our own research.

Areas of research that have been examined and considered as impacting on a student's academic achievements include:

- Student profile entering the university, analysing demographics and families' previous educational background.
- The educational background from which they came from e.g. private, public, co-ed or single gender.
- The type of college and the program they undertook, i.e. diploma, degree, honours degree etc.
- Ranking of the College they are entering.
- Socio-economic profile of the student cohort.
- Student accommodation, private room, dormitory during academic semesters, or living at home.
- Who they socialise with.
- Academic programs undertaken.
- Peer influences.

Students entering a University are as unique as they are numerous. Each student will have been moulded and shaped based on their genetics and life experiences to that point. Influences include previous education experiences, socio-economic background, educational institutional supports by teachers and their home environment. In reality a students friendship with classmates are determined by a complex set of decisions and circumstances from their past through the influences of their parents, school administrators and teachers and often a great deal of coincidence and chance from interacting with previous classmates and friends. The impact of these influences on a student was highlighted in the Coleman report [14]. Coleman *et al.* carried out one of the largest educational surveys in the USA in the late 1950s. They sought to measure the features of school environment that led to differences in student academic attainment. A key finding of this study was that "... a pupils achievement is strongly related to the educational backgrounds and aspirations of the other students in the school". Peer characteristics were found to be notably more important than teacher characteristics or non-social aspects of the school. The survey covered 600,000 students and 60,000 teachers from 4,000 of the nation's public schools. While this survey focused on secondary level schools it's findings are relevant to tertiary level as it examined not just school activities but the students' backgrounds and the impact and influence of students on their peers.

It is with consideration of the findings of this research that we will give due consideration to the makeup of the peer groupings and friendships we identify. We do recognise that attempting to identify whom has the greatest influence on a student within that community network can be difficult as their social network can be extensive.

According to Bruner [7], "social interaction lies at the root of good learning". This school of thought is central to constructivism and based on this premise, the influence of an individual's groups on their learning has a large impact on their constructed knowledge and ergo their academic studies. Constructivism is the belief that knowledge is constructed and not acquired. Everyone has a different interpretation of events and constructs their knowledge based on their interpretation of their unique life experiences. Ferguson [22] compared the works of Piaget (1972) who was a cognitive constructivist and Vygotsky, a social constructivist [55] and Bruner [7] who emphasises the importance of an individual's social and cultural environment within which they are exposed to and the contexts within which they can accumulate knowledge. The process of learning is facilitated through individual participation in social interaction. Constructivism at it's core is a theory that learning is a collaborative process and therefor whom you are learning from can influence what you learn. Thus the groups to which a student belongs to, will influence their academic achievements.

2.5.0.1 Peer Influence

Peer influence can be understood as the modification of an individual's behaviour as they come into contact with one or more individuals, from the same or another similar peer group. In our context we are interested in the contact that takes place within the bounds of a University campus. While our interest is in identifying any academic Peer influence in our chosen student cohort, we recognise that students face consistent influences external to their university community. Peer influence sources and effects are examined in greater detail in this section.

One of the largest and possibly most cited third level research works of its time was a longitudinal study by Astin [5], who interviewed 25,000 Freshman students in 1985 and followed up four years later in 1989 using questionnaires designed to determine what impacted the most on each student during the four years of their University life. The results included 192 measures of the college environment. Three of the main items of influence were found to be;

- the environment created by the faculty and students;
- the type of colleges that produce favourable performance in standardised test and those that “enhance retention and other cognitive and effective outcomes”;
- the **single most important** influence on development of students is peer influence.

Carney, [10], carried out an extensive literature review of the models used in determining the peer influence in various domains. Many of these models try to measure the subtle and complex interaction between members of a group that can influence group members. Manski [40] examined this area and coined the phrase “reflection problem”, which identifies three different effects found within a group and the difficulties in separating and thus measuring their effects on the group members. These effects are:

1. Endogenous effect, wherein the propensity of an individual to behave in some way varies with the behaviour of the group i.e., the influence of the collective of peers on the individual’s behaviour;
2. Exogenous effect, wherein the propensity of an individual to behave in some way varies with the exogenous characteristics of the group;

3. Correlated effect, wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environments.

Each of these effects examines a different form of influence that the individual exerts on the other group members and conversely how they are being impacted by being part of the group. We examine this hypothesis within a University environment and build on research which recognises the intricate nature of complex community structures.

Due to the variate of known unknowns that these different effects have on a research cohort some studies try to mitigate the effect of variables by controlling for them. Some 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.* [11] in their study at a US Air Force Academy monitored students exogenously assigned to groups. At the time of their research they determined that there was little evidence of large positive peer effects in academic performance. He cited Sacerdote [51] as finding evidence of small peer effects for *roommates* at Dartmouth. Zimmerman [60] found small *roommate* contextual effects for students at Williams College. Foster (2006) [25] and Lyle (2007) [38] found no evidence of peer effects at Maryland and West Point respectively. Carrell's project wished to estimate peer effects in college achievement, utilising "a unique data-set" where groups of about 30 students had to spend the majority of their academic and social time together, effectively eliminating exogenous influences on other students. Their research reported a peer effect of "greater magnitude than previously found".

Alternatively Lyle [38] on the other hand found no evidence of academic influence amongst randomly assigned cadets at West Point Military Academy. He did however find that there were other influences amongst friends on the choice of major or the decision to remain in the army. While Lyle does refer to his cohort as being randomly assigned to each of 36 companies, West Point does control for some characteristics such as gender, race, recruited athlete and prior performance and behaviour, effectively reducing the claim of randomness. Lyle also recognises that environmental factors or **shocks** (biases) have an impact on the academic achievement of a group. These **shocks** can include quality of instruction, classroom environment such as seating arrangements and class timetabling. Lyle's research is one of the few projects that has considered these types of biases. Once again we would caution as to whether these environments represent what would be considered normal, when correlated against similar research in institutions where no external influences came to bear. Another attempt to mitigate for the external influences, this time outside of the USA, on students, included Androuschak *et al.* [4] at the National Research university — Higher School of Economics in Moscow, who chose exogenous groups formed by the school administrators as the research cohort. Their supposition was that because they had been formed into groups that would work together they also would avoid bias in peer effect findings. However students do interact with others outside of their study groups socially and students may be influenced by these interactions. We would also question the expectations of putting a group of students together and expecting them to act as if they had formed friendships in the normally organic way.

Examination of group formation will be discussed later with reference to work by Tuckerman [54]. Some studies were carried out in military schools where control can be implemented to try and prevent outside variables from influencing students and therefore drop them as possible influences of the students and their peers. However although students are placed together into units this does not mean that it is possible to determine if they influence each other either positively or negatively. Students may also interact outside of their Units and be influenced by this interaction. This may be difficult to observe as personas may mask their true feelings for each other.

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 automated approach to data collection.

The principle behind the research conducted without attempts to mitigate for exogenous effects included Zimmerman, [60] Hall & Willerman [29] and Socerdote [51]. Their research is based on students who are randomly assigned to dormitories and their roommates. These roommates are the peers that will be considered to be those that may, or may not, effect their academic attainments. Due to the random nature of the roommate allocation they are often from different backgrounds with different academic abilities, interests and attitudes. These researchers used student entry level SAT scores and measured the academic performance of these students and their academic attainment over time. This methodology seems to assume that students spend enough time with their roommates to influence each other and seems to ignore the possibility that these students develop alternate social networks. Foster [24] for example defined a peer group as "...all students residing in rooms that are on the same wing of a residence hall floor as the given student". Her investigation similar to others looked at the entry questionnaires of students at the start of the academic year and once again focused on the dormitory social networks.

Zimmerman [60] using the SAT scores of randomly assigned students and those of their roommates. His research set out to measure differences in grades of high, medium and their assigned roommates. He also examined students as part of clusters of rooms that constitute social units. He concluded that there is strong evidence of peer effect in third level institutions. Students of middle ability are usually more susceptible to peer influence from those at either end of the ability distribution.

At Dartmouth college where freshmen are assigned to rooms and dormitories randomly, Scaerdote [51] identified peer effects in GPA scores at room level and the decision on fraternity membership is identifiable at the room level and dorm level. While the allocation of freshman rooms is random there is a level of the allocation being controlled. New students are asked to answer four questions that are aimed to ensure some form of personal compatibility. The questions are

1. do you smoke (only 1% say yes to this) ?

2. do you like to listen to music while studying ?

3. do you keep late hours ?

4. are you neat or messsy ?

Overall, the findings of the research identified evidence of peer effects in student academic results.

Hoel *et al.* [32] carried out a 10-year longitudinal study which found strong effects of the academic quality of roommates and dormitory peers on a student's achievements, while finding no evidence of classmates peer effect. Hoel surmises that "better students enable a more challenging curriculum and, so the story goes, a better education through interaction with peers inside and outside the classroom". Interestingly the positive peer effects Hoel identifies appears among the dorm-mates and roommates and not classmates. He believed this is because dormitories that are strongly orientated toward academics, may actively discourage noisy and rowdy behaviour and therefore of its self is a peer effect on those who reside there.

We can conclude from the previous research that measuring peer effects is difficult. Student outcomes depend on multiple factors from their family and educational backgrounds to the quality of the college they attend and their engagement with the program they are studying. When students select a college and a program, they are unconsciously making a decision about their future peers with whom they will develop alongside, during their university life. The individuals they get a chance to share this life with have already been filtered by the university through their application and acceptance process.

2.5.0.2 Groups

In any effort to identify peer influence we must be able to identify groups. Our research will identify students who spend a higher proportion of their time together in comparison to other class members, thus defining social groupings. We will tackle the complex issue of the identification of various types of peer groups, namely social and academic. To further understand peer influence amongst a group we must understand the concept of social groups and why they form. A group as defined by Forsyth [24] is "a collection of two or more individuals who are connected to one another by social relationships". The concept of social relations is very broad. Forsyth explained that humans have a need to be part of groups and this is normal, and we choose to be part of numerous groups whether it is for working, learning, relaxing, playing or worshipping, we usually do this in groups. Researchers have been interested for centuries in the "Group Dynamics". They have examined the actions, processes and changes that happen within groups, often with conflicting opinions.

Groups are often defined depending on their contexts and their features. It is the degree of the relationship between the members that will differentiate the type of group. A family may not meet face to face on a regular basis where as a football team can meet regularly and can spend a great deal of time together. However there is an obvious disparity between these two group types. The former is based on kinship while the latter is on friendship and a collective interest in sport.

Recent work by Chen *et al.* (2016) [13] found that students, when given the choice, formed groups based on their own social networks and furthermore that self-forming groups performed better than exogenous organised groups. Chen believes group members perform better in a group if they had an input into its formation. He also determined that students create more connections if they are part of a self-formed group. Other researchers such as Oakley *et al.* [45] and Foster [25] believe the converse, that is that students do better in groups formed by instructors rather than self-formed groups. Foster's research is based on a comprehensive data set made available by the administrative panel from the University of Maryland, and she hypothesised that social peers do not impact on a student more than randomised peers would have. She analysed student campus accommodation requests as they entered their second year. Students can apply to share accommodation with other named students. Foster compared student requests against students who shared accommodation within the first year and inferred social connection through these requests. The results indicate that peers, based on these connections do not affect performance any more than random peer interactions would do. Oakley's area of interest was the development of student groups into effective teams and highlighted many of the different personality issues that arise during the development stages of a group. However Oakley's approach does require individual group management to ensure effective performance.

Benson [6] has identified a list of attributes referring to the reasons a set of people may engage in frequent interactions:

- They identify with one another.
- They are defined by others as a group.
- They share beliefs, values, and norms about areas of common interest.
- They define themselves as a group.
- They come together to work on common tasks and for agreed purposes.

He summarised that a group is organic, intentional and just a random experience, and come together for for some common need or interest. This infers that a group at a specific point recognises itself as a collective and is also seen by others in this vain. They come together "for a common purpose" if in an exogenously formed group but also because they share beliefs and interests with others within the collective. Tuckman [54] accepts that the principles that groups do form for numerous reasons, these groups may not remain intact over time.

Tuckman identified that groups pass through a number of stages namely Forming, Storming, Norming and Performing. When a group is formed the personalities within the group will pass thorough a storming stage when individuals develop an affinity with the individuals within the group. In this, the concept of the influence of peers and the context within which that influence occurs will surface continuously through this thesis, but it is the ability to identify accurately who is influencing whom within the group, that is central to this research. Our approach is the employment of technology as the resource of choice to capture this.

2.5.1 Technology and Data Collection

In one of our earlier sections on Learning Analytics and Educational Data Mining (section 2.4.1) we discussed the collection and analysis of large data-sets and the extraction of knowledge. We have referenced the digital footprint created by student and the technologies that we intend to use to capture that data-set. The use of technology in geospatial research has been ongoing for in various guises from the early part of this century.

An early example of the use of technology to collect geospatial data from student activities was in 2004. Data was collected over a two year period by two researchers using a hand-held GPS device as part of Project Lachesis carried out by Hariharan & Toyama, [31]. The aim was "…extracting stays and destinations from location histories in a pure, data-driven manner". Hariharan went on to define Location History as "…a record of an entity's location in geographical space over an interval of time."

Technology has advanced in the intervening years, but with the same requirements, structured, pure data. One of the most popular technologies utilised is the modern mobile devices, which has the same capabilities as the Project Lachesis equipment.

Other researchers, whose work we will examine in detail in this chapter, who have used technologies to collect data on the interaction between parties could be categorised as:

1. Using geospatial data collected through the use of GPS and GMS location data. Examples of this are the work by Li *et al.* in 2008 [37], Cranshaw *et al.* [16] and by Hariharan & Toyama in 2004 [31].

2. Specifically adapted smartphones utilising bespoke data collection applications. Examples of this are the work by Eagle & Pentland, in 2006 [19], Gupta *et al.* in 2007 [28], Wang *et al.* in 2015 [56] and by Harari *et al.* in 2017 [30].

3. Badges that collect data relevant to the wearer, including work by Pentland in 2012 [47] and Watanabe *et al.* in 2013 [58].

4. Smartphones which are used to collect WiFi base station data. This includes work by Rekimoto *et al.* in 2007 [48] and by Gonzalez *et al.*in 2008 [26].

The research data gathered in these papers ranged in scope from two users over a one year period, to 48 users for a ten-week period, through to 100 users for a period of nine months. Our research is based on two academic years (24 months) of data and a cohort of 174 students and we believe that this compares very favourably with similar research.

Eagle *et al.* [19] carried out a longitudinal research study using data collected over a nine-month period from 100 mobile phones, utilising what he termed "the ubiquitous infrastructure of mobile devices". He stated that it was the very nature of the mobile phone that made it a perfect device to capture data. He employed a mobile phone application to determine subject position based on proximity, time, location and date. The purpose was to demonstrate the ability to utilise Bluetooth technology to collect user behaviour and activity information. The intention was to recognise social interaction patterns and to cluster subjects at locations, therefore modelling activities and inferring relationships through the monitoring of temporal and geo-location data. This is an approach used in many research projects and will be utilised in our research.

The interaction between WiFi-enabled devices and the WiFi network can provide the capability to identify information previously impossible to gather on ad-hoc and formal groupings of people. Gupta *et al.* [28] presented a clustering algorithm based on co-presence that identifies both groups of people and the locations where they congregated. This algorithm used data collected from smartphones running Intel's PlaceLab location engine which identified groups that met over a one month period.

Outside of educational applications, other research projects infer similarities among people thorough shared locations and as such recommending products or services based on the purchasing patterns of other visitors to similar locations. Li *et al.* *et al.* [37] using spatio-temporal data collected from a cohort of 65 subjects over a six month period, proposed a system to model individual locations and infer *similarities from shared visited locations*. Li uses this technology to identify *trajectories*, paths or expected paths through a network and *stay points*, locations where they spend a greater portion of their time. These locations were considered significant for an individual subject, i.e. subjects visit many locations but visit some more often as they hold a particular interest for them. Work by Li *et al.* in [37] used location tracking from public WiFi access in a shopping mall in Australia over a one year period in order to determine the location habits of shoppers, a technique which underpins applications such as retail recommendation systems among others.

Gonzalez *et al.* [26] studied the behaviour of 100,000 anonymised individual mobile phone users and *"identified that human trajectories show a high degree of temporal and spatial regularity, each individual being characterised by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations."*. In our work we will endeavour to identify if similar patterns can be detected in a cohort of University students in a campus environment.

Further research by Wang *et al.* [56] in 2015 carried out a *SmartGPA* study using students' direct reporting and passive sensing data about their locations collected from their smartphones. The research aim was to understand students' study and behavioural patterns. The passively sensed data collected included GPS coordinates, movement and audio data. The audio data recorded from the phone is assessed to identify the context of the meeting. Quiet backgrounds may indicate academic locations and study whereas loud music would indicate a social gathering. This data was collected from a cohort of students of different academic abilities. Through the understanding of different student behaviours they sought to determine if academic achievement can be predicted through student behaviour. The project concluded that there is correlation between GPA and sensed behaviours and a significant correlation between GPA and behaviours inferred from that data. In similar work, Harari *et al.* [30] carried out continuous tracking of behaviours, again using mobile sensors including accelerometers, microphones, light sensors and phone logs. This study examined the physical activity and social activity of the 48 volunteers over 10 weeks.

Minaei-Bidgoli *et al.* [44] and Fire *et al.* [23] studied the logs and databases of web-based systems as the sources of raw data for their research. Fire *et al.* analysed the implicit and explicit cooperation among students while doing homework assignments both individually and in groups. Analysis of computer usage logs identified those who used the same machine to submit their individual assignments, to infer implicit friendships. Explicit friendships were identified through group assignment submissions.

Research by Watanabe *et al.* [58] employed wearable sensor badges that demonstrated a strong correlation between students physical behaviours and their scholastic performance. The research was carried out in two schools which utilised badges that could identify a subject's activities (body movement), locations and face to face interactions between other research subjects. Their results showed that face-to-face communication between students during break times correlated well with performance.

Outside of his work with Eagle at third level, Pentland [47] also worked in industry where he expanded his research in to the area of effective team building, which had findings relevant to the third level projects. He identified that face to face communications are more prevalent in successful teams. He concluded that patterns of communication are a better predictor of a team's success and that social time is critical to team performance. Pentland measured the groups' energy levels through the use of badges that collected data on subject behaviour including the participants, the tone of voice of the conversation and their body language. Based on this premise we believe that students who come together regularly, face to face will work well together and influence each other.

Rekimoto *et al.* [48], developed a life logging system called *LifeTag* and a platform service called "PlaceEngine" utilising a smartphone and a keychain-like device using custom hardware. Their work combines the use of WiFi-base signal strength and GPS base station identifier of the nearest base station. This form of model requires an extensive database of pre-determined station information. While this work can monitor activity both indoors and outdoors, it combines a number of the issues identified earlier, namely expensive setup and the Hawthorn effect on the research cohort, and there are biases introduced into their work, something we intend to avoid in this research.

2.6 Conclusions

In this chapter we discussed the concepts of and processes employed in the use of data within a University environment. Having presented the role of technology in the data collection and the systems used to fulfil that role such as Eduroam, we presented our research environment, i.e Dublin City University. inferring friendships. The discussion regarding the current practices of data collection and how it relates to our approach, included a review of the domain of Learning Analytics that provided a background to this research. In this chapter we examined a small portion of the research domain associated with peer influence situates this current study in the relation to previous studies and literature. We explored the domain of "Educational Data Mining" which is concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings.

In the next Chapter we will present in detail the sources, collection processes and preparation stages of the numerous data-sets that are the cornerstone of this thesis. We will examine the sources and type of data including the challenges in the development of the various systems used to create usable knowledge from the raw sources.

Chapter 3

Dataset Used for Learning Analytics

3.1 Dataset and Data Used

The collection and assimilation of data into an integrated and clean dataset is the first step of the usually long process of data mining. Data mining is a method of processing data with a specific aim i.e. to obtain useful knowledge from the data source. The aim of data mining within this thesis is to obtain knowledge about student behaviour within a bounded campus from the mining of WiFi access data logs. As outlined when framing our hypothesis in Section 1.1 our intention is to mine a dataset of WiFi logs for information, to augment them with additional student demographic and academic achievements and to do this with the intention of identifying implicit and explicit trends of influence from that data. We will be utilising a combination of numerous datasets of various sizes to develop usable knowledge to obtain this end.

Romero *et al.*, [50] considered educational environments to have many different data sources which are specific to the educational domain. They believed that such data contained semantic information and when considered in relation with other data sources, it can provide multiple levels of exploitable information and knowledge. As we defined in the Learning Analytics (LA) section earlier in Section 2.4.1, LA focuses on the process of decision-making by making the student the central focus rather than the institutional administration which is normally the case. A definition set out at the first international Conference on Learning Analytics and Knowledge (LAK 2011) and adopted by the Society for Learning Analytics Research (SoLAR) said that "Learning Analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs." It is the intention of this thesis to illustrate that through data mining, our analysis can determine the context of student activity within a University campus. Having mapped activities to contexts, we will then determine the effects that student friendships have on each other, based on the variations in academic achievements.

3.2 Survey of Datasets

In this chapter, I outline the datasets gathered for this work and which are the basis of our research and I provide a descriptor of each and its contribution to the research. To develop context for a dataset it should not be examined in isolation and requires the additional combination of other data to enhance the understanding of what each contributes. We have previously outlined the source of our main dataset i.e. the WiFi-logs generated by the Eduraom system and the interactions with that system by the cohort of students. We will also describe those datasets that will provide context and further interpretation to our log data. These include student timetables, Eduroam infrastructure assets plus student demographic and academic information. The following sections will provide an outline of each dataset, its origins and its role in our research.

3.2.1 Eduroam WiFi Access Logs

Following discussions with the University's Information Systems Services Information and Systems Services (ISS) Department and DCU's Institutional Research & Analysis Officer we received approval to use data collected by the University IT systems relating to student WiFi connections on campus under strict security and anonymisation control, as described earlier in Section 1.4. This approval also conformed to the Universities ethics guidelines. Approximately 1,000 Network Access Server (NAS) units located across the main Glasnevin campus capture and control all wireless traffic active on the WiFi network. Eduroam log data contains the captured connections from all devices connecting to the WiFi network on the campus. The specific log our research is interested in is called the “auth-details” log. This log captures the request made by a WiFi-enabled device such as smartphone, laptop or tablet, to access the network. For network efficiency reasons these files are saved to one of two separate servers maintained by the University’s IT Department. These servers Orcus2 and Senda2 have been configured to record the network activity on a continuous basis.

A sample of the data contained within a single request packet or log entry is shown below:

- Tue May 12 00:00:21 2015
- Packet-Type = Access-Request
- User-Name = ”0cj8sn5isbr4ojtna9ne678hg439nhed”
- Acct-Session-Id = ”1ADDD1B4-F437B7DFAC71-0000003805”
- Calling-Station-Id = ”F4-37-B7-DF-AC-71”
- Called-Station-Id = ”FC-0A-81-DE-1C-F1:eduroam”
- Vendor-388-Attr-2 = 0x656475726f616d
- NAS-Port = 6
- NAS-Port-Type = Wireless-802.11
- Framed-MTU = 1400
- Service-Type = Framed-User
- NAS-Identifier = ”Res_CP_W514”
- NAS-Port-Id = ”radio1”
- Vendor-388-Attr-17 = 0x5265735f43505f57353134
- Vendor-388-Attr-19 = 0x46432d30412d38312d44442d44312d4234
- Vendor-388-Attr-23 = 0x5265735f4873655f31372d3139
- Connect-Info = ”CONNECT 72.2Mbps 802.11bgn”
- EAP-Message = 0x0201000a01636f786337
- NAS-IP-Address = 136.206.23.253
- Proxy-State = 0x000100051addd1b4001a

- Message-Authenticator = 0x8e4c386a5adcaafe0faaef82e614cf7d

This packet contains a mixture of data fields of which some will be identified as useful. As part of our process definition we identified the minimum information from a request packet required to identify student location. We identified the following fields as being of important to our research. Using Python we developed a model to differentiate between the useful data,

- Date and Time of the access request
- User Name
- Session Id, assigns for the duration of that user's visit (session)
- Calling-Station-Id, is the devices unique id making the request
- Called-Station-Id, the id of the NAS recording the request
- NAS-Identifier, a descriptor provided by the network administrator
- NAS-IP-Address, unique physical address of NAS

and the superfluous fields that would not add value to our analysis, such as ...

- Vendor-information
- Connect-Info
- EAP-Message
- Proxy-State
- message-Authenticator

The fields we identified as containing the minimal data upon which we could base our research and necessary to be able to identify individuals by date, time and location are outlined here:

When:

- Mon Jun 2 16:04:35 2014

Who:

- User-Name = ”0cj8sn5isbr4ojtna9ne678hg439nhed”
- Acct-Session-Id = ”0B34FF44-28E02C402CD2-0000073359”
- Calling-Station-Id = ”28-E0-2C-40-2C-D2”

Where

- Called-Station-Id = ”5C-0E-8B-26-2A-50:eduroam”
- NAS-Port-Type = Wireless-802.11
- NAS-Identifier = ”Sports_Gallery”
- Connect-Info = ”CONNECT 65Mbps 802.11bgn”
- NAS-IP-Address = 136.206.23.253

The data collected per log can be categorised as *When* the connection event occurred, *Who* was making the request and *Where* was the request being made from. The quantity of data made available for the research was extensive as Table 3.1 outlines. The year 2014 contains WiFi log data from September 2014 to December 2014 inclusive, which is the first semester of the academic year 2015 (2014/2015), both 2015 and 2016 contain two semesters of WiFi data and 2017 contains the second semester (Jan - May) of the 2017 academic year.

Year	Size (Gigabytes)
2014	44 GB
2015	99 GB
2016	96 GB
2017	24 GB

Table 3.1: Volume of raw log file data used in experiment

All of the data cleansing, preparation and transformation stages were carried out using a bespoke Python system that reads in the data log files containing the log data, removes all unnecessary data and stores a single day’s log file per server in .csv format.

We classify the raw input file as containing semi-structured data comprising of non-linear fields which vary in order from record. Having extracted our required fields, we needed to standardise the data within each field, for example the ”Called-Station-Id” which collects the NAS identifier or MAC address, may be formatted as:

“5C-0E-8B-26-2A-50:eduroam”

or

“5C-0E-8B-26-2A-50”.

We therefore standardised these addresses to “5C-0E-8B-26-2A-50”.

Utilising our Python filters we built a standardised log for each day found on the two different servers (Orcus2 & Senda2), from these files we created a single file representing the WiFi activity for each day.

Access requests for access to Eduroam generate a number of logs per request and this occurs due to the handshaking or *conversation* between the device and the network as the authentication process is completed. Each packet in this conversation contains the same session id (Acct-Session-Id) allowing both parties to the conversation to differentiate it from other conversations on the network. We will be focusing our attention on the first of these tagged entries as it indicates the first request by the device for access to the NAS and therefore the most likely arrival time for the student at that location. We remove all other duplicate records from our dataset to avoid any form of double counting.

In addition to the access requests made by those associated with the University, the network will receive requests for access to the wireless network from University visitors who have Eduroam credentials from other institutions. As part of our data preparatory process we removed these non DCU based requests from the dataset. Once removed we believe we have a clean, robust dataset on which we can carry out our analysis. The following section 3.2.2 give a brief illustration of the ability of this data to be used to identify activity

3.2.2 Student Activity

To illustrate the activity of WiFi activity on the Eduroam network, we have plotted a sample day during a semester in Figure 3.1. This figure illustrates the activity between the hours of 10:00 and 19:00. This activity represents CA student attendance for a typical in-semester Tuesday. There is a peak of activity prior to the class commencement at 11:00 and again at around 11:45 and at around 12:45. There are a number of reasons for the increase in activity at these times, for example classes finish in the last quarter of the hour and once finished class, students will commence using their phones for social media, checking for messages or other reasons. Students will also commence movement to their next class which may be in a separate building. As they transit from one location to another their device will switch between the various NAS' as they drop out of range of one NAS and into the range of another. Every time a device enters the range of a NAS and requests access to the network it is generating traffic. The drop in traffic in between the peak occurs as students' devices becomes dormant during classes. We also observed that there is a drop-off in traffic later in the day illustrating that the majority of students leave the campus after 14:00 and the remaining traffic is probably generated by those studying in the Library, meeting socially or who are resident on campus.

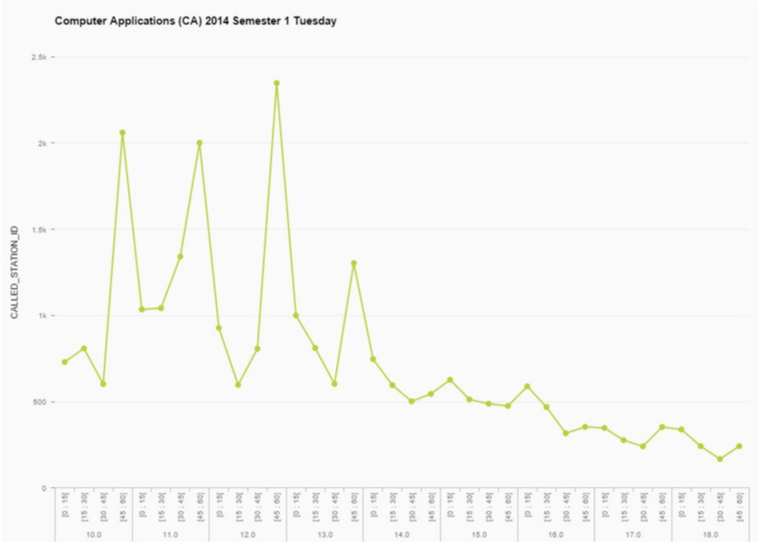


Figure 3.1: Daily sample activity shown by hour

Figure 3.2 illustrates the level of WiFi activity occurring throughout a semester. This figure demonstrates that activity commences in the middle of September and develops a recurring weekly pattern that remains relatively similar until the end of the semester in December. We have annotated Figure 3.2 showing a variation in the normal weekly patterns at the end of October, when there is a Bank Holiday weekend. Over this weekend there are three days with low activity illustrating that fewer students are on campus that weekend. In our next chapter in we will apply context to this activity through the examination of activity during academic hours (9:00 to 17:00) and importantly outside of these hours. This analysis will be the initial step in applying meaning to our student activity and the various student meetings generated.

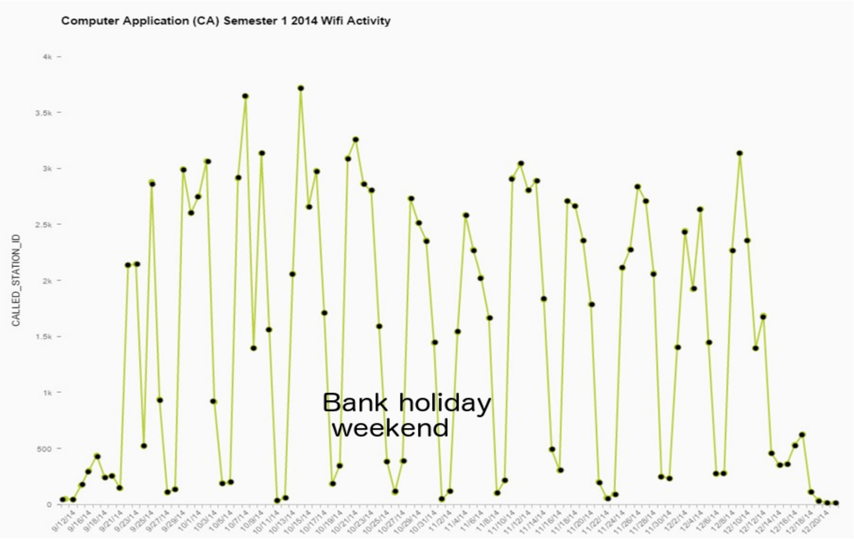


Figure 3.2: WiFi Connections for Semester 1

3.2.3 Dublin City University NAS Locations

Access to the Eduroam WiFi platform by users is provided through the provision and authentication of a request's credentials by Radius servers at their home institution. Network access at member sites is via NAS configured 802.1X protocols. The DCU network comprises approximately 1,000 individual Network Access Servers (NAS). These NAS's are distributed across the whole campus ensuring almost continuous WiFi coverage to users. Users do expect to avail of a continuous connection to the WiFi at all times and locations. For this reason there is a least one NAS reachable from each location on campus. In areas of higher traffic requirements, there may be two or often more in larger meeting locations such as the main restaurant or library.

In our initial research we identified the "NAS-Identifier" field as a source of location that would be the basis of identifying the physical location within the DCU Glasnevin campus. However our analysis showed that logs often had a number of different NAS units in a single room and these NASs had different location identifier descriptors and therefore would not be a reliable source of identification. Following discussions with the ISS Department it became apparent that there is no listing in the form that we required. All previous records were held for maintenance and servicing purposes only and were used by experts who had experience with the infrastructure and understood the standardised naming convention. We therefore had a need to create a usable database of NAS by locations.

Using the list of all NAS-identifiers contained within the log files and a consultation process with the ISS Information Systems Services network technicians, we developed a comprehensive database of the campus network system including MAC addresses for devices and their physical locations. Each NAS has two distinct channels namely, 2.4 GHz and 5 GHz, each of which provides multiple sub-channels, effectively providing thousands of access points across the campus. Figure 3.3 is an illustration of the DCU campus with buildings being represented in blue, pedestrian areas in white and the green areas as grass or gardens. All buildings have multiple NASs fitted, ensuring that not only the building has adequate WiFi coverage, but the extremities of the buildings and the green areas are also covered.

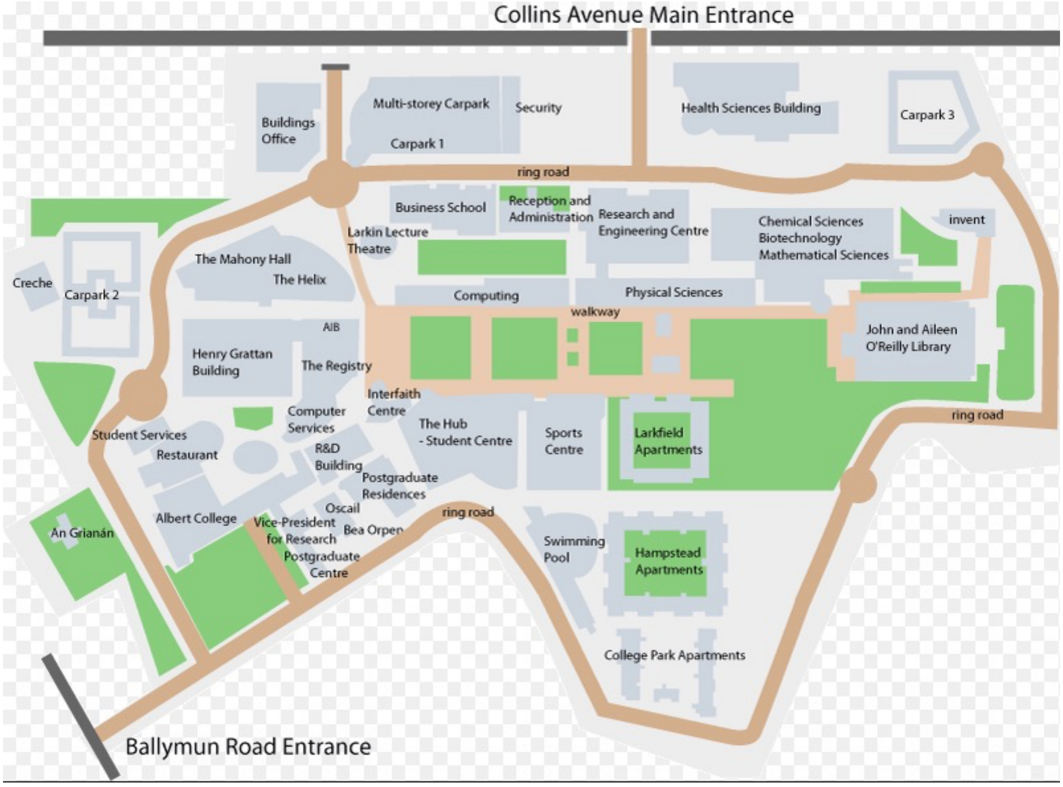


Figure 3.3: DCU Glasnevin campus

We took the opportunity during the development of the NAS identifiers table to enhance the dataset with additional information. This additional information included a classifying of the NAS as being either in an *Academic* or *Non-Academic* (social) location. We further categorised each location to apply some context e.g. academic areas will include classrooms and laboratories and non-academic includes hang_outs and residence. Table 3.2 contains the information on, and a breakdown of each NAS's details.

Name	**Description**
Academic	Area Category
Ref_index	Zone Type
Room/Zone	Room descriptor
X-Axis Y-Axis	DCU Campus Map reference
Type	Room type (Class, Lab, Library, Transit, Hang_out or Residence)
Map	Estates, map reference
2.4GhzMac_Address	NAS Mac_Addresses
5.0GhzMac_Address	NAS Mac_Addresses
Mac Address Location	Descriptor
Location1_Comment	Additional descriptor

Table 3.2: Eduroam WiFi log Fields (Example)

The completed NAS database has provided us with a comprehensive list of identifiable locations with additional contextual information.

3.2.4 Student Demographics

Following a number of meeting with representatives of the University's management department we were supplied with a data set of student demographics[1] for our student cohort. The dataset included the following. The data provided falls into a number of sub-categories

- Demographic - age, gender, nationality, country of birth and Domicile;
- Registration - year, program, CAO points;
- Program - modules, results, precision mark.

In this list, the "program" refers to the degree program a student is registered for. CAO points is a reference to the Central Applications Office (CAO), a centralised application process for all students to apply to all University and Institute of Technology courses through-

[1] All identification markers such as student number, names or usernames, were anonymised in advance to ensure no individual could be identified from our analysis

out Ireland. As we describe later in Section 3.3, students apply for courses in ranked order and are allocated a place on a course based on their performance in the national Leaving Certificate examination, similar to SAT tests in the USA. Depending on their performance in this exam they "earn" a certain number of CAO points and those with higher CAO points get first refusal on their first preference course. In summary, higher CAO points indicates better exam performance in the precursor to University entrance.

"Program modules" in the list above refers to the set of modules each student takes in each year of their course. Typically a student would take 12 modules in an academic year, 6 in each semester, with each module being rated as 5 "credits", totaling 60 credits per semester.

We will be using the demographic data to investigate the correlation between students who become friends and their academic marks on a longitudinal basis.

Academic performance is measured on a yearly basis using the *precision mark* . This measure is a weighted score of a students individual module marks for that academic year. To present a good understanding of our student cohort we have carried out an examination of our demographic data.

All data sources that could be used either separately or in conjunction with other datasets have been anonymised. Examples data fields anonymised include username, first and surname. These fields were anonymised using a Python application applying a hashing algorithm which generated a md5, 32 character code. This hashing algorithm was applied to all demographic and WiFi data that contained any form of personal identifiers.

3.2.5 Registered Students

In Chapter 4 we will present our research which is based on a longitudinal study of a cohort of students registered in 2014 for the year 2015 (i.e. 2014/15 academic year). We will use these students as our baseline cohort and then fold in those additional students who join or leave the class through the period of our analysis.

In 2015 a total of 223 students registered for the two School of Computing degree programs we studied namely the B.Sc. in Computer Applications (CA) and the B.Sc. in

Enterprise Computing (EC). As with all third level program classes the composition of the group is dynamic. Students may either leave of join the group at various times during the program's intended duration, typically four years for an honours degree. Those who leave the program do so for numerous reasons such as they feel incompatible with the program, they decide to take a year out of education, they have transferred to another 3rd level course or they may have failed examinations and intend to repeat the year or drop out completely.

Our cohort are those who have registered for either of the two degree programs at the start of the three years of our work. Students who join our cohort could be doing so as they transfer from other courses or they are repeating students. As these students are now part of the class network and interact with our initial cohort of students, they potentially could influence those students, academically and socially and are therefore of interest to our analysis.

3.2.5.1 Registered Profiles

Our research will focus on two distinct cohorts i.e. students from the CA and EC degree programs. There are differences between these courses in terms of course content and they will each be attractive to a different student type. We believe that each program will attract individuals with different competencies, characteristics and interests. We will therefore examine each group individually from this point forward.

In Table 3.3 we present the age demographics of the CA and EC program participants for the years 2015 to 2017 inclusive. We identified that there is an uneven gender balance in both of our chosen programs. There are approximately 14% of the CA registrants Female. While this imbalance is significant it is not peculiar to DCU but is in line with that found by the Irish Higher Educational Authority [8] across all Information and Communication Technologies (ICT) program entrants to Irish institutions, for the year 2015. By the third year the balance changes slightly, as the proportion of females has grows to 22%. In the EC program there are 20% female in the first year which drops to 17% in third year.

Table 3.4 is a more granular examination of the gender mix and age profiles of the cohort per year from the first to third. As expected, the majority of students entering their

Program	Gender	2015		2016		2017	
CA	F	19	14%	21	19%	17	22%
	M	118	86%	92	81%	59	78%
EC	F	17	20%	12	18%	10	17%
	M	69	80%	54	82%	49	83%
Total:		223	-	179	-	135	-

Table 3.3: Gender & Demographic profiles per academic program per year

respective programs are of school-leaving age i.e. age range 17 to 18 years old. Within the CA program for male entrants, there are proportionally more mature students than in the EC program. We also identified that the Female CA student group had less of a drop-out rate (19 to 17) than their EC counterparts (17 to 10).

CA	Age Range	1st year (2015)	2nd year (2016)	3rd year (2017)
Female	17-18	11	1	0
	19-20	8	18	9
	>20	0	3	8
	Total	19	21	17
Male	17-18	66	14	0
	19-20	42	64	29
	>20	10	14	30
	Total	118	92	59

Table 3.4: Computer Applications cohort count by age range, per academic year

EC	Age Range	1st year (2015)	2nd year (2016)	3rd year (2017)
Female	17-18	11	1	0
	19-20	6	9	6
	>20	0	2	4
	Total	17	12	10
Male	17-18	36	0	0
	19-20	21	43	31
	>20	2	3	18
	Total	69	54	49

Table 3.5: Enterprise Computing cohort count by age range, per academic year

The majority of entrants to the programs as expected are Irish born. The CA program had 102 of 137 (74%) and EC had 73 of 86 (85%) as profiled in Tables 3.6 and 3.7 respectively. The next largest grouping was from the *European countries* and the balance from the

Rest of the world. We have sub-divided our cohort in this manner as it will be interesting to identify trends based on basic demographics.

CA	1st year (2015)	2nd year (2016)	3rd year (2017)	Drop-out
Europe	19	17	13	32 %
Ireland	102	84	57	44 %
Rest of World	16	12	6	62 %
Total	137	113	76	45 %

Table 3.6: Computer Application count by Country of Birth, per year

EC	1st year (2015)	2nd year (2016)	3rd year (2017)	Drop-out
Europe	8	6	6	25%
Ireland	73	58	51	30%
Rest of World	5	2	2	66%
Total	86	66	59	31%

Table 3.7: Enterprise Computing count by Country of Birth, per year

The data tells us that there are a number of drop-outs occurring per year. These drop-out can be categorised in two ways. Those students that started the program but did not complete all examinations during the year and had a precision mark precision mark *less than 5*. These students will be deemed to have dropped out during the year. We use the value "5" as a student may in the early part of the semester participate in an in-class continuous assessment test, receiving a mark which contributes to their yearly overall precision mark. The second group are the students who did complete a number of examinations and gained a precision mark, but did not register for the following year. The latter group can be made up of those who did not get an overall pass mark for the year and may have either dropped out completely or will repeat the year in a new class in a subsequent year. Either way, they are removed from our research cohort. We will be using the precision mark as an indicator throughout. A precision mark is a weighted average of a student's academic examination and continuous assessment marks for the modules undertaken in an academic year.

Programme	Gender	2015	2016	2017
CA	M	5	7	1
	F	0	1	0
EC	M	0	0	3
	F	0	1	0

Table 3.8: Count of registered students who did not complete the academic year

3.3 Entry to University

There are a number of different routes into the third level educational system in Ireland. The standard route is through the CAO points system where students must obtain a minimum number of points from their Leaving Certificate exams in order to be accepted into a third level degree program. This is the most common route for progression from secondary to tertiary education in Ireland. There are a number of other routes into tertiary education for varying groups of students, these include mature students and students who have completed a non-university third level program and who wish to progress to higher levels in education, as shown in Table 3.9 which provides a breakdown of those by program:

CAO code	Path	2015	2016	2017
878	Mature	8	4	3
968 & 969	DARE	6	6	6
976	Access	5	4	3
978 & 979	HEAR	28	20	12
999	Deferrals	4	4	2
666 & 669	Others	13	12	8

Table 3.9: Count of students by alternate entry routes into Third Level Education

Mature student entry is available to individuals aged above 23 and who may have 'other' experience, apart from performance in examinations which may be taken into consideration when being considered for acceptance to a program.

DARE (Disability Access to Education) [1] is another third level alternative admissions scheme to assist students whose disabilities negatively impacted on their second level education.

The Access program is unique to DCU in Ireland in that it is the largest, and the oldest. It was formed to address the low numbers of students entering third-level from one of its closest neighbourhoods in North Dublin. Funded through corporate donations and philantrophy, it provides a host of financial, academic, professional & personal development supports to over 1,300 Access Scholars studying at undergraduate and postgraduate levels, comprising approx 8% of the student population.

The HEAR (Higher Education Access Route) [9] applicants are students who have availed of a scheme that offers places on reduced points to school leavers who are considered to come from socio-economically disadvantaged backgrounds.

Deferrals may be provided to students who have been offered a place on a program but are unable, or do not wish to take up the position in the year it was offered. These students will enter the program at a later date.

Other: There are no additional or specific information on the previous academic history of these students in the registration system.

3.3.1 Student Timetable

To apply any form of context to a physical location and the reason students visit it is central to this work. There is therefore a need to distinguish between the reasons for locations being visited at a particular time of the day. Students will be on campus to attend class and may remain on campus in classrooms or a laboratories to study or to work on group projects. One of the reasons a student or group of students are in a location is because they are scheduled to be there i.e. there is a scheduled lecture. To confirm the location of students at any time and to identify if they are attending a scheduled class is important. Equally as relevant is to identify those who are not in a class when scheduled, as their activity may indicate co-location that has a strong indication of friendship.

To do this we need to utilise the official timetable applicable to our two cohorts (CA & EC). Table 3.10 is a sample from the CA timetable indicating that in 2015 **CA** students were scheduled to attend a class on a Monday at **10:00**. This class was module **MS121**, which is the program code for Mathematics and it was scheduled to take place in classroom **QG13**, which is in the Business School building.

A sample of some of the 2014/2015 academic activities is provided in Table 3.10:

YEAR	PROG	DAY	MODULE	SEM	TYPE	START	ROOM	Building
2015	CA	MON	MS121	1	L	10:00	QG13	Business
2015	EC	MON	MS121	1	L	10:00	QG13	Business
2015	CA	MON	CA106	1	L	11:00	HG23	Nursing
2015	EC	MON	CA106	1	L	11:00	HG23	Nursing
2015	EC	MON	CA106	1	L	12:00	HG23	Nursing
2015	CA	MON	CA106	1	L	12:00	HG23	Nursing
2015	CA	MON	CA172	1	L	14:00	L114	Computing

Table 3.10: Sample Academic Timetable

The timetable information is a compilation of data collected from numerous sources including the associated program lecturers, school offices and students. It was necessary to collate these sources as the timetable can fluid in the early weeks of the semester and a number variants can be in existence.

3.4 Focus Group

One of the main tenants underpinning our research questions is the ability to identify friendships though the analysis of student co-location. We hypothesised that not all co-locations are of equal importance i.e. co-location during a formal scheduled class is not as relevant or important to a friendship as a social co-location outside University core hours.

We therefore convened a focus group to gain an understanding of student behaviour on campus and the relevance students put on the locations where they in general and *friends* in particular, meet. This focus groups' primary role was to collect a representative set of activity types for students during their time on campus. We were specifically interested in the locations they prefer to spend time, either alone or with their friends. We also wished to identify the context of the locations of their meetings, i.e. academic or social reasons. The intention was to use the outcome of the focus group to construct a questionnaire which would be put to a wider cohort of students from the two programs.

There were two parts to this Focus Group. After a general mood setting discussion, we focused the discussion through the use of a prepared questionnaire. Following that we extended the discussion into a general ranking of locations students meet with their friends.

We have outlined the questionnaire in Appendix B and the table used in the discussion on ranking of these locations is in Appendix A.

3.4.1 Participants

The participants for the focus group were selected from the cohort of 3rd year students from 2015. The focus group attendees were all from the same program and it was accepted that their habits would be unique to the them but would represent the areas that students can go to outside of their core (formal) class hours. The intention was to identify these areas and use them as the basis of a student survey that would be circulated to all students in the 3rd year i.e my longitudinal study cohort or subjects. All attendees were voluntary willing participants.

3.4.2 Locations

The environment for the focus group meeting was a technical lab at Dublin City University. It had seating for all participants in an informal arrangement, a semi circular layout. All attendees seemed comfortable with the arrangement as there was minimal to no rearrangement of the chairs upon arrival. Other equipment was a whiteboard with a diagram outlining my research process.

On arrival there was a number of minutes of general "chat" to set the relaxed mood of the meeting. The moderator (me) outlined the research topic and the research questions. I used the outline diagram on the whiteboard to explain the work being undertaken and their role within the research. I further outlined what I was aiming to achieve from the group and what the data collected on the day would be used for. I outlined the importance to my research of my ability to identify locations within the campus which are used to identify locations that students would attend with friends.

The meeting took place on a Wednesday (28/Sept/2017) mid-day prior to the attendees going to class, they had not been to a class earlier in the day and had come in early (2 hours) before their class. This was important as it ensured that the attendees were fresh and could focus on the topic in hand rather than any distractions occurring in the previous class. The attendees represented a group of friends from 3rd year in the EC program. The participants of the group comprised of 4 female and 2 male students.

3.4.3 Opening

Once the group was settled and they had a good understanding of the objective of the focus group meeting I commenced the discussion. I played the role of facilitator and did not give my opinions but led the discussion while ensuring all students contributed. They discussed our understanding of the concept of *friendships* as identifiable through co-locations. Once I was satisfied that all understood the intention of the question, I introduced a prepared questionnaire with ten questions, this is outlined in Appendix B. I ensured all students understood what their role was and it became apparent that all did as their discussions were well focused. The group worked through the questionnaire and engaged in a discussion on each question to ensure understanding, each student marked their questionnaire independently. At the end of the session each student returned their questionnaire anonymously.

3.4.4 Ranking and Rating

As part of the group discussion the concept of the **ranking** of the various locations within the campus and the separation within these locations by time zone (core vs. non-core) was raised for discussion.

Our participants distinguished three distinct groups of locations and those areas where it is most probably a meeting of friends will occur, while also considering location and time. The discussion ranked the locations in this order but believe there is very little difference in the weights that could be applied, within each:

- Non-core Hang-out
- Non-core Residence
- Core Residence

The following locations were identified

- Core Hang-out
- Non-core Labs

- Non-core Classroom
- Non-core Transit
- Non-core Library
- Core Labs

It was considered that if students are in the following locations that they are indistinguishable from other groups in the same location or they could most probably be there on their own.

- Core - Transit
- Core - Classroom
- Core - Library

While there was a great deal of discussion on the ranking of the various locations, there was a consensus of opinions. Further discussions were used to grade the different locations relative to each other.

3.4.5 Summary of the Focus Group findings:

As part of the qualitative research all comments were recorded. Based on the comments made and the quantitative data collected in the questionnaire, we developed the survey questionnaire, which is covered now.

Student comments:

- Students, unless they live on campus only attend on the days they have classes that interest them or that are ”Not Boring”.
- The majority of students do not have a regular arrival time but prefer to arrive on campus for the first class of the day. They will also leave the campus after the last class has concluded.
- For lunch some prefer the canteen, but will go to a friend’s apartment on campus if they are around.
- While on campus they will attend classes and then head to the Canteen, Nubar, Residence (friend/self).

- When studying the majority prefer, if studying on campus, to study alone in the library, alternatively if there is group work they will meet in a Lab in the Computer building or on occasion in the canteen or in the common area of the library.
- From a social aspect those who attended the Sports complex did so with friends or as part of a team.

During the focus group conversation a number of comments were made which included some regarding group dynamics within the class:

- "Females do tend to form quicker friendships with other females due to the limited number of females in the wider class." Our research did not necessarily confirm this statement. While it may be true that females do form friendships with other females in the early part of their college career, they may not remain sole friends with other female students.
- As part of the process of assisting students develop friendships in the first couple of weeks of their undergraduate course, lecturers place students into groups for group projects. The intention is that the students are provided a reference group that they can identify with. Our focus groups "would consider two years later, those people are not part of their friendship group but a class acquaintance i.e. few of those who had been put into groups in the early part of 1st year remained as friends into later years."
- These EC students who believed that "there are a small number of groups within the class corpus, with a couple of people who are not tightly tied to any one group."
- Most students leave the campus after the formal class finishes and only attend college on their off days to study or meet for group work. This would add weight to the occurrence of a pair of students are on site regularly together on days they do not have classes are friends.
- "Meeting places are limited to the Canteen and Bar. There are few alternatives. Students prefer to arrive for a class and leave afterwards and there are few reasons to attract or keep students on site outside class hours.

From the Focus Group discussion and the questionnaire a ten-question survey was pre-

Location	Sub-Category	Computer Applications	Rank	Enterprise Computing	Rank
Canteen	Social	3.2	6	2.5	6
Nubar	Social	3.8	4	3.2	5
Class	Academic	5.0	2	4.9	2
Labs	Academic	2.1	7	2.4	7
Library	Academic	3.4	5	4.2	4
Sport	Social	4.5	3	4.3	3
Residence	Social	5.4	1	5.9	1

Table 3.11: Survey location ranking by student group

pared using Survey Monkey. The survey was designed to compare the discussion points and location ranking of the Focus Group against that of the students in two programs and their findings. The survey was sent to all third year students in our research programs. Further details are in Appendix B.

3.4.6 Survey results

A full analysis of the results of the questionnaire can be found in Appendix B. We surmised that the opinions aired in the Focus Group were indeed representative of the general student cohort. From the survey the respondents ranked the campus locations where they spend time outside *formal classes*. It is the result of these rankings that will be used in our analysis as detailed in Chapter 5

A complete summary table of the weighting that were determined from the analysis of the Student Survey can be seen in Section 5.3.1 Table 5.2. These weights are a representation of the locations that students consider the most likely to frequent with their friends outside of the formal class timetable locations.

3.4.7 Conclusion and contribution

The collection and correlation of data for the use in any research is often a long and enduring process with the majority of a projects time taken up by this preparation. This thesis faced many of the same challenges, with the collection process having numerous challenges from the identification of data sources to the collection and preparation of that data. In this chapter we presented a survey of the various datasets required to be collected and mined, to extract the knowledge required to approach our stated research questions. Some of those sources identified included the Eduraom WiFi logs, student timetables and campus NAS locations. As this research is predicated on understanding the behaviour of students within a bounded campus and with their peers, we stepped outside the technological sphere of data collection, to talk to the students. A comprehensive examination of student activity on campus was undertaken using a focus group with a sample of the research cohort. The contribution from the focus group had a enormous input to how the next phases of the project were approached. It also identified that although the EC and CA students were interested in different aspects of ICT their opinions on the locations they share with their friends is very similar.

In the following Chapter there is an examination of the research tools available to the project and those that were ultimately employed. Using a dedicated test dataset of students from a collection of the different schools across the DCU campus. The chapter will outline the experiments that were undertaken to test the feasibility of the proposed proof of concept.

Chapter 4

Developing Research Methodologies: Proof of Concept

4.1 Preliminary Research

In this chapter the focus is on the development of the concepts and methodologies developed during the research projects' life cycle. The chapter outlines the approach taken to scope the project objectives and to examine the requirements needed to answer the research questions posed earlier in Section 1.2. The intention here is to present the foundations for the experimental process, the decision making process that was followed, the experimental results and the conclusions which will be further detailed in the following chapter.

4.1.1 Software Tools Used

During the initial scoping and objective definition stages of this project we looked at a number of potential strategies with various approaches which have been used in previous research involving the examination of third level education and the impact of peer influence. At a macro level, the research approaches that are considered appropriate include

1. Social Network Analysis (SNA)
2. Summary statistics

3. Clustering techniques

SNA was identified as a useful tool to "identify patterns or regularities in relationships among interacting units" by Wasserman [57]. He examined the concept of *Relational Ties*, that is the type of relationship which links units together. He specifically identifies "Behavioural interaction" and "Association" which can be used to identify the types of links and their strengths. These tools were also considered by Dawson [17] who explored the relationship between a student's position in a classroom social network and their reported level of *sense of community*. SNA was therefore considered a legitimate starting point for this research project.

Due to the large dataset that required analysis there was an obvious requirement for a software specialist platform to support the analysis, something that was not an issue for other SNA research. Desk research identified a number of potential systems that are designed for the mining of large datasets. Two of the leading "off the shelf" industrial platforms were identified as having the ability to handle large datasets namely:

1. SPLUNK, a platform for analysing machine-generated big data.
2. Knime, a data analytics, reporting and integration platform.

SPLUNK [53] is a platform developed for monitoring, and analyses of machine-generated big data, and it is controlled via a web-style interface. This product was considered because one of its unique selling points is its ability to process large datasets of digitally generated log files. SPLUNK has a number of the features which were deemed advantageous to our research project and one such feature of interest is the ability to accept as input, zipped files, i.e. the format in which the WiFi logs were presented to the project.

A large number of experiments were carried out to assess SPLUNK's ability against measures of speed, agility, accuracy and flexibility. These experiments were carried out on subset of the main WiFi dataset. The results indicated that this platform did not perform well under a number of these requirements and would not be suitable for this project, and these can be summarised as follows:

- The platform was found to be slow at processing filtering queries developed for the sample data-sets. The comparisons were gauged against similar data-sets using bespoke functions written in Python. The Python functions were found to be considerably faster, albeit that they did require each python program to be developed and tested before use.
- The licence secured from SPLUNK Corp, was a temporary “Researchers” licence which was granted for a period of six months only. As there was no guarantee that the licence would be renewed, it was believed that this time period did not provide the security needed to cover the duration of this project.
- SPLUNK’s inability to handle zipped files of semi-structured data as previously detailed in Section 3.2.1 on data, negated the expected benefits original anticipated.
- One of the previously identified benefits of SPLUNK was the ability to handle zipped files as inputs. This proved impracticable as this projects zipped datasets contained semi-structured data which proved problematic and could not be handled effectively.

Because of these disadvantages we abandoned the idea of using SPLUNK for data processing.

The second platform identified and similarly assessed in terms of its potential for data processing in this research was the open-source data analytics and reporting platform KNIME [35]. This platform has a number of additional plug-ins with the one of most interesting being a Social Network Analysis plug-in. As with the testing process undertaken with SPLUNK, using a sub-set of our data, there was an examination for speed and flexibility. These tests included the development of the queries required to establish a network with identifiable dyads based on co-location. The queries, while easily constructed using the platform’s GUI, proved to be complex with processing times being excessive. As both the SPLUNK and KNIME platforms had varying difficulties dealing with unstructured data, it was considered necessary to develop bespoke applications to import, cleanse, analyse and present the data using a combination of Python and R.

During the early stages of the development of our bespoke application this project was provided the opportunity to work with various application platforms provided by an industrial partner, SAP. SAP software corporation is one of the largest enterprise software solution providers in the world. SAP Ireland made available their application suite and provided additional training and technical support to this project. Initial testing of the platform, using the same principles as those used with the aforementioned platforms proved positive. Initial testing was carried out using SAP's HANA featuring In-Memory Database structure, Parallelism, Column store, Dynamic aggregation and SQL scripting. The intention was to utilise HANA to perform the computationally expensive analyses which heretofore had proven prohibitive. The other SAP application that was evaluated was SAP Lumira, a data visualisation application which was identified as providing opportunities for use later in the research analysis process.

Having carried out initial feasibility into the software tools, how these different computer languages and the platforms were subsequently used during our research will be examined and explained as we continue to explore our individual research questions.

4.1.2 Research Approach

To ensure trace-ability of data at each stage of the analyses process, analysis results need to be saved and securely stored to ensure the quality of data and the ability to return to any point of the process and revisit the results developed at each stage. This data management allowed for the ability to fork the research in various directions without compromising the original data validity.

Having prepared and formalised various datasets, the initial analysis focused on the WiFi logs as introduced earlier in Section 2.2 and the examination of these for quantitative traits and data patterns. This involved the testing of the dataset to determine the scope of useful knowledge that could be extracted. The testing commenced by determining that the Eduroam network could accurately track and record user devices as they traversed the campus. To test this hypothesis a small sample of volunteers were canvased to allow the mining of their WiFi log data. The five volunteers had different daily routines affording the

DATE	Time	NAS_identifier	Location	Room
12/12/14	08:52:27	HG-86E5DB	Henry Grattan building	CG12
	08:52:44	HG-86E758	Henry Grattan building	C105
	08:53:08	HELIX_STUDIO_AP34	Helix Theatre	AP34
	08:53:28	HELIX_BLUEROOM_AP37	Helix Theatre	AP37
	08:53:48	Computing-34F5AC	School of Computing	L101
	08:54:27	Computing-34F5D8	School of Computing	L208
	08:57:07	Computing_LG25	School of Computing	LG25
	08:57:15	DCUBS_QG13	Business school	QG13
	08:57:23	Computing-34D57C	School of Computing	L121
	08:58:07	Computing-34D648	School of Computing	L125
	08:58:11	Computing-34D644	School of Computing	L128

Table 4.1: Sample WiFi activity for a volunteer's trip

opportunity to sample activity patterns for various disparate WiFi Logs. The logs of these volunteers' WiFi activity was isolated in the log files, accessed and analysed by filtering the dataset by their unique **username_ids**. Having shown the ability to identify students by their **id's**, their logs for a number of randomly chosen dates was chosen and the actual activity was compared against the logs. A sample log extract from one of our volunteers is shown in Table 4.1 which tracked and logged their morning journey through the campus to their lab. The first row in this table shows their first connection to the Eduroam system and the location of that connection. Record 1 identifies this NAS location being room CG12 which is located in the Henry Grattan building, see Figure 4.1 . As they passed out of range of this NAS, they connect to the NAS in room C105. This is a typical journey for this volunteer and one taken most mornings. They park their car in the car park at the end of the campus and walk to the School of Computing through the access road between the Henry Grattan building and the Helix Theatre. They emerge from the access road and cross over to the Computing building, where they connect to the NAS in room L101. Having passed other NAS in the computing building (L208 & LG25) the student takes the stairs up to the first floor, at this point the nearest NAS that they connect to is in the Business School. As soon as they pass back into the spine of the Computing School building, they connect to various NAS' on the 1st floor on their way to their Lab.

To ensure the integrity of the data of the volunteers, an examination of their activity for

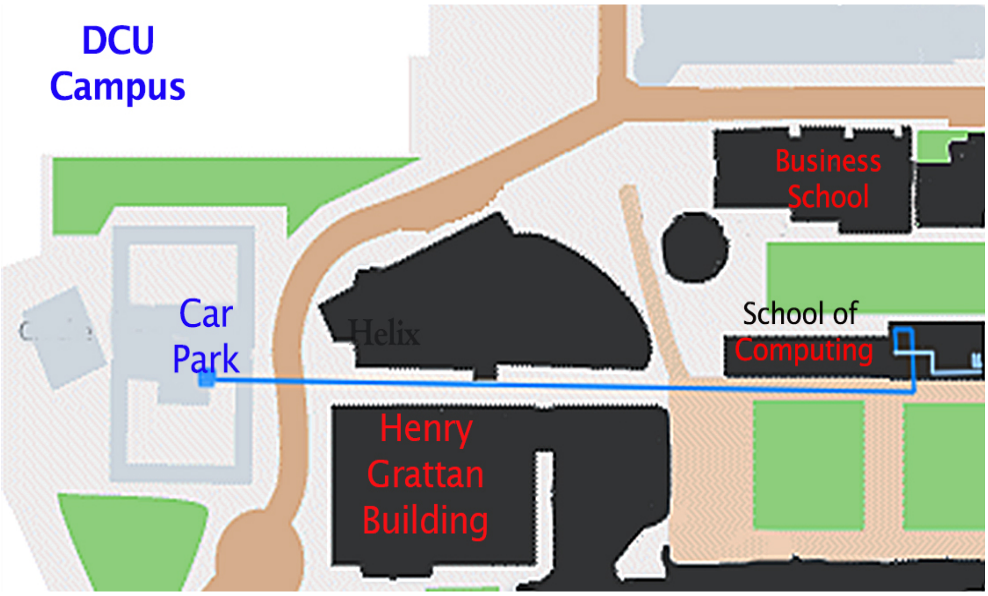


Figure 4.1: Volunteer daily route to Lab.

a sample month was carried out to determine if their recorded activity matched their actual behaviour. The findings indicated no anomalies or unexpected activities being present in the data. One finding which required further investigation was the presence of a number of entries at the same time in the same place but with different *Session_ids* for the same student_id. It was identified that volunteers sometimes access the Eduroam WiFi system using up to three separate devices. Three devices were identified by their distinct unique "calling id" and the volunteers' "Username". Examination of the "calling id" and comparing it to the MAC addresses of the volunteers' Laptop, Phone and iPad identified the source of the entries. This would prove to be an important consideration during the programme development stage of the project.

During a more granular examination of a sample month, one volunteer generated 2,282 events and accessed 20 distinctive "NAS Identifiers". As would have been expected the majority (1,978) of events occurred in the confines of the building they frequent the most, in this case the Computing School building. A smaller number occurred in many of the locations previously mentioned such as the routes taken to and from the car park and those involving other daily routines such as visiting the on-campus shop.

The data logs of a number of other volunteers were similarly mined and these also confirmed their normalised routines to be accurately reflected by the WiFi log data. They demonstrated that each individual automatically logged onto Eduroam at the nearest NAS point as they enter the campus boundary. Depending on a student's circumstances, they may arrive at the campus on foot, by bicycle, via public transport or in private transport such as a car or a motorbike. They could arrive at various entrance points onto the campus and their WiFi device or devices will connect at the campus perimeter.

4.1.3 Sample Network

Having shown that we can track individual volunteers on-campus whereabouts from the WiFi access logs it was considered a logical progression from identifying singular volunteers to larger groups and their collective locations. It was important to identify a cohort of students who would be expected to have a level of interaction within the group that could be identifiable as a network and thus could be measurable. The intention was to chose a group of students registered to the same degree programme, i.e. one of the many exogenous formed groups. These groups would therefore be expected to co-locate at known times and locations identifiable through their shared academic calendar.

To identify such a useful dataset, an examination of other work carried out by the Insight Centre for Data Analytics on comparable datasets was made available. PredectED was an analysis of students' digital footprints created in the DCU Virtual Learning Environment carried out by Corrigan and Smeaton [15]. A subset of their data that was of value to our research was the access to a subset of the data set comprising the **username** of the students who had opted into taking part in the PredectED research project, their module code and exam results. This test dataset comprised first year participant in ten varied modules with a total of 2,028 students registered from a number of different schools in the university. This was considered to be a representative sample of the student population at the University with each programme requiring different entry levels (CAO). The different entry level requirements and programme content ensured a mix of students with varying levels of academic interest and topic interests which makes analysing their digital footprints for possible group formation, quite valid and interesting.

Table 4.2 and Figure 4.2 illustrates the percentage of students who enrolled in each of the ten modules and their CAO points achieved in their Leaving Certificate examination[1]. An important aspect of this graph is that it shows the range of these points. For example the majority of students taking the module ES125, was in the 300-325 points range, in BE101 the greatest proportion of students scored in the 480-500 points range. An examination of the range of points scored by the majority of students in each program demonstrates the diversity across the 10 modules chosen.

Range	BE101	CA103	CA168	ES125	HR101	LG101	LG116	LG127	MS136	SS103
300-325	0.0	4.3	2.7	37.9	0.9	0.0	7.3	0.4	1.0	0.0
330-350	0.4	5.7	5.5	20.7	0.0	0.0	5.5	2.7	1.0	0.0
355-375	1.2	2.9	13.7	13.8	0.0	2.2	12.7	5.1	0.0	2.4
380-400	4.1	5.7	19.2	20.7	0.9	3.3	23.6	3.5	4.8	4.8
405-425	5.8	30.0	26.0	3.4	4.6	5.5	23.6	10.1	18.1	0.0
430-450	12.4	21.4	12.3	3.4	9.3	1.1	12.7	32.7	5.7	4.8
455-475	24.5	17.1	13.7	0.0	37.0	13.2	7.3	30.4	27.6	9.6
480-500	30.3	2.9	2.7	0.0	15.7	34.1	1.8	10.1	21.0	33.7
505-525	13.7	5.7	2.7	0.0	15.7	19.8	0.9	2.7	12.4	27.7
530-550	6.6	4.3	1.4	0.0	12.0	13.2	3.6	1.6	6.7	12.0
555-575	0.8	0.0	0.0	0.0	3.7	7.7	0.9	0.8	1.9	4.8

Table 4.2: Percentage CAO points by sample Academic Module

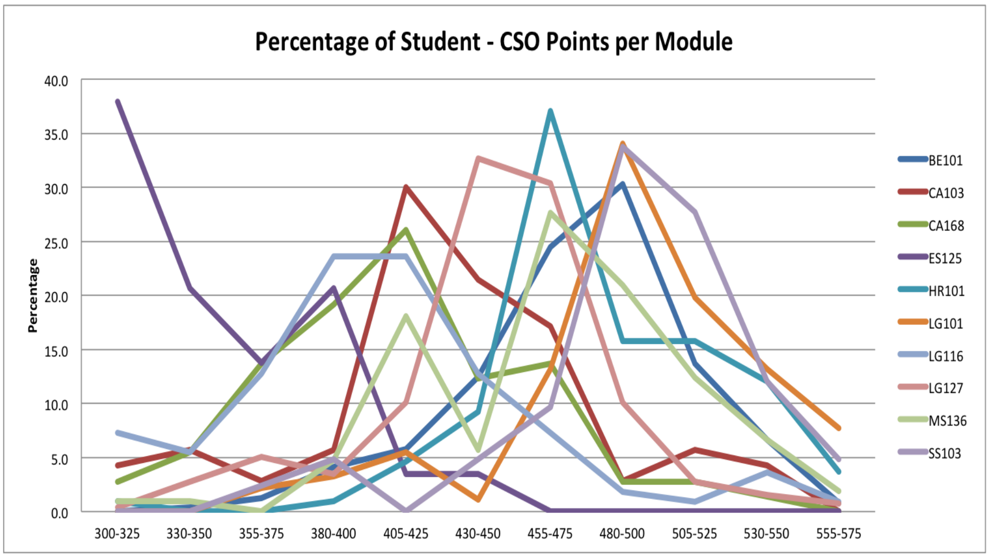


Figure 4.2: % CAO Points per student per module.

[1] The Leaving Certificate is a nationwide state examination which is used to determine academic ability and is thus used as an assessment tool for entry into courses in Universities in Ireland

4.1.3.1 Data Summarising

The data examination process commenced by identifying the location of students and correlating this activity with their academic class timetable i.e. identifying the time and location of the class plus the expected number of students. The examination were positive, as a large number of students were identified as co-located in locations and at times that corresponded with an academic class timetable. These findings confirmed that through simple data summations it is possible to identify class member interactions through physical meetings and the context of such meetings. The context most readily identifiable was the academic meeting of those attending the same classes. To identify other meeting contexts and friendships, there was a requirement to identify peers at a more granular level i.e. dyads.

4.1.4 Dyad Identification

Dyads as define by Wasserman [57] are "a linkage or relationship establishing a tie between two actors". For the purposes of this research, Dyads will be defined as a unique meeting between two personal WiFi-enabled devices that are co-located during a specified time window. The time window in the initial research is defined as being 20 minutes in duration. For example if two devices are connected to the same NAS point within a 20 minute period, they are deemed to be co-located and classified as a **meeting**. This is irrespective of how many other devices are co-located with the pair in question. The 20 minute interval was considered indicative of the time span within which students can come into contact with each other e.g. Student A enters a location at 11:05 all other devices that enter that location within the subsequent 20 minutes will be deemed to have met or interacted with Student A. We developed a Python module to apply these criteria to the dataset of students taking generated 22,800 pairwise dyads (meetings) in the first semester (2014/2015).

The expectation that should emerge from analysis of the data is that students enrolled in the same programme and taking the same modules will be on campus at similar times, as dictated by their scheduled class timetable. As students will be co-located for scheduled classes, study sessions or project group meetings, there is an expectation of a large number of dyads in academic locations i.e. their academic count will be high. When students

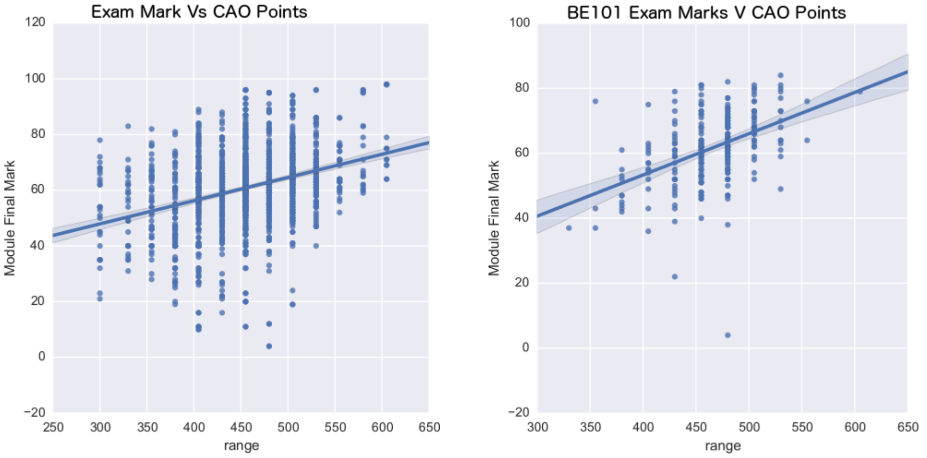


Figure 4.3: Left: Total CAO V Precision Mark. Right: BE101 CAO V Precision Mark

vacate academic areas for lunch or other social reasons the social dyad count will similarly increase. Figure 4.4 is a representation of the number of pairwise meetings between students in our test dataset. The dyad count in the first week of the semester was 3,300 in academic areas and 1,622 in social locations. By week 5 this figure had increased to 7,058 academic and 4,333 social dyads from the same number of registered students. Figure 4.4 illustrates the growth in dyads over the semester with an additional two trend lines. There are two notable variances in the chart i.e. weeks six and nine. Week six is a "reading week" with many schools not holding formal class and the students carrying out research or preparing for mid semester examinations off campus. Similarly in week nine, students are preparing for semester tests and are finalising projects for submission off campus.

For demonstration purposes, the results of the investigations will be illustrated using the findings of a single module i.e. BE101, Introduction to Cell Biology. This is a module taught in first semester of a number of the first year science degree programmes. As it is being taught to variety of programmes it was considered to represent a varied cohort of student types and personalities. Figure 4.3 has two scatter grams with the left pane plotting the Precision marks of the students against the entry CAO points they had achieved. The right pane demonstrates similarly the BE101 module students CAO points on entry and their semester exam marks. These panes illustrate similar characteristics between BE101 and that of the total cohort.

Summarising the dyad meetings is shown in Figure 4.5 which illustrates that in week 1 of the semester there were 118 students from BE101 who frequented different academic locations and 111 students visiting social locations. At these locations there was an average meeting degree of 28.4 (academic) and 14.61 (social) meetings respectively. That is, during that week, students on average meet 28 other students in academic and 15 other students in social settings. These figures increase to 44 and 29 respectively in week 5, and the average for the semester as a whole is 34 academic and 19 social. We interpret this to infer that there is a growth in contact among students, which peaks in week 5 after which students have commenced an adaption to their new environs and began to have favourite places to visit.

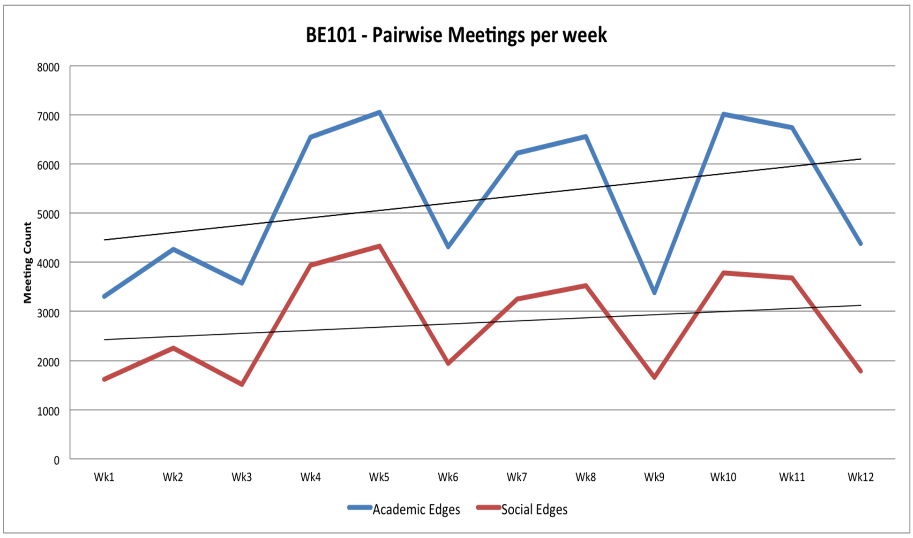


Figure 4.4: Weekly Pairwise Meetings (Academic & Social).

These results also indicate a continuous growth of student meetings through the early stages of the semester as we would expect. This is interpreted as demonstrating the initial stages of development of group formation and that students are taking some time to form groups, passing through the development phases prescribed by Tuckman [54] i.e. forming, storming, norming and performing. Tuckman's principle identified that the coming together of a group of individuals does not form a cohesive stable group immediately. The group dynamic remains fluid and can lead to friction between members before stabilising into a functional group. This research did not identify any empirical evidence that can point to the optimum time taken for the “coming together of individuals for a purposeful experience or to satisfy some common goal” Benson [6] i.e. the formation of robust groups or friendships. He summarised that a group is organic and intentional and not just a random experience, they come together for some common need or interest.

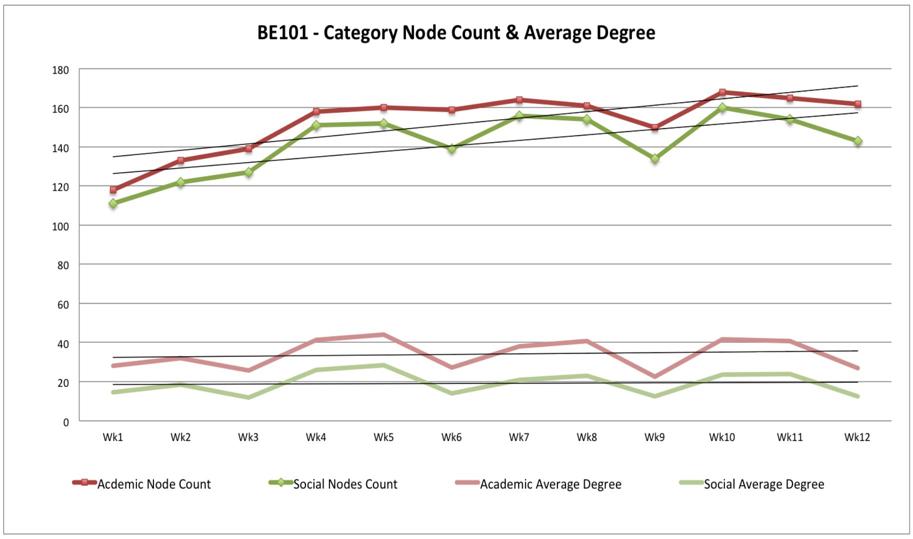


Figure 4.5: Average Weekly Pairwise Meetings.

The process undertaken has demonstrated the practicalities of identifying group meetings from a cohort of students from the same programmes. To identify closer friendships a more granular examination of these meeting and a summary analysis of a dyad’s meetings was undertaken. Using the “Location Identifier” 3.2 from the WiFi logs, a summary analysis representing the number of occasions the students met in each location was developed. Table 4.3 demonstrate a pair’s (dyad) count of two randomly chosen students 1 & 2 from this summary. The table lists the anonymised user names, the locations where they met and the number of times they met. Each location is tagged and identified as whether it is an “academic” location or not, and the sub-type of location (e.g. t = transit, m = meeting, r = residence). The location sub-types illustrated in this table are an early subdivision of the campus and were redefined at a later stage in the analysis process. This later classification was previously listed in Chapter 3, Section 3.4.4.

This preliminary research has shown that it is possible to accurately identify the movements of students while on campus and has demonstrated the ability to accurately identify

Student 1	Student 2	LOCATION	Academic	TYPE	OCCUR’
00824df0ec4a’	014aedf2b8e’	Foyer 1st Floor	n	t	3
00824df0ec4a’	014aedf2b8e’	Gym Gallery	n	m	16
00824df0ec4a’	014aedf2b8e’	Hub_Venue_Stage	n	m	1
00824df0ec4a’	014aedf2b8e’	ISS ServiceDesk Area	n	o	1
00824df0ec4a’	014aedf2b8e’	LarkinC8-2343F0	n	c	7
00824df0ec4a’	014aedf2b8e’	Mezz_POS	n	d	111
00824df0ec4a’	014aedf2b8e’	Mezz_Upstairs	n	d	2
00824df0ec4a’	014aedf2b8e’	Office Landing 1st Floor	n	s	2
00824df0ec4a’	014aedf2b8e’	Reception Area	n	m	28
00824df0ec4a’	014aedf2b8e’	Res_Lark_V156	n	r	1
00824df0ec4a’	014aedf2b8e’	The Street	n	t	2
00824df0ec4a’	014aedf2b8e’	Beside 1st Floor Lift	y	l	5
00824df0ec4a’	014aedf2b8e’	CA126	y	o	1
00824df0ec4a’	014aedf2b8e’	DCUBS-Q111	y	o	1
00824df0ec4a’	014aedf2b8e’	DCUBS-QG13	y	c	5
00824df0ec4a’	014aedf2b8e’	HG19	y	c	6
00824df0ec4a’	014aedf2b8e’	L101	y	l	1
00824df0ec4a’	014aedf2b8e’	L114	y	l	1
00824df0ec4a’	014aedf2b8e’	LG25	y	l	2

Table 4.3: Summarised count of dyad co-location meetings. For TYPE, t=transit, m=meeting, r=residence

pairs and groups of students co-locating. The next phase of the research is to identify those that are more than just class colleagues but actual friends. Thus can we answer the final question in this thesis, can we identify those that may have an influence on behaviour as an individual or group who are friends. Having undertaken a process of discovery this project has determined that the principal concepts of the research is achievable, using the test data set. It was therefore considered that the analysis of the main cohort of students could be undertaken with confidence.

4.2 Main Research Cohort Analysis

The remainder of this thesis will focus on answering the Research Questions posed earlier in Section 1.2 utilising the previously identified Computer Applications (CA) and Enterprise Computing (EC) program cohorts of students, which are described in Section 3.2. The primary analysis began with the examination of the Computer Applications (CA) and Enterprise Computing (EC) programs for the academic year 2015 (2014/2015).

Commencing with the WiFi logs for a randomly chosen sample day, a set of summary statistics of activity was created focusing on location visited, the frequency of the visits and an examination of a sample of the activity of students. The primary location visited by the students was the School of Computing, shown in the left pane of Table 4.5. Other buildings adjacent to the School also featured prominently in the top ten locations visited on the day by the cohort. An additional list of the top ten students, by activity count (right pane) is also included. This student activity table illustrates the contrast in activity between students, Student 1 is over twice as active in terms of visiting locations and interaction with others, as that of the 10th (Student 10) most active student.

Locations	**Count**	**%**
Computing-34F5E0	1,403	8.764%
S-Block-35064C	930	5.81%
Tech-Hse-1679DC	675	4.217%
HG-85EC54	517	3.23%
Tech-Hse-E883A4	432	2.699%
Computing-34F594	426	2.661%
Sports_Bridge_Gym	422	2.636%
Hub_Landing_SU	396	2.474%
Hub_Venue_Control	350	2.186%
Registry_Street	342	2.136%

Table 4.4: Top 10 visited locations

Student	**Count**
Student 1	1,738
Student 2	1,554
Student 3	1,308
Student 4	1,043
Student 5	999
Student 6	934
Student 7	922
Student 8	875
Student 9	812
Student 10	763

Table 4.5: Top 10 Students on 10th April'15

The activity count for the most active student listed him/her as generating 1,738 unique session id's on the Eduraom network. This student's activity mirrored the total cohorts' activity with the majority of their time being spent in the School of Computing and local environs. The top locations for Computing students are illustrated in Figure 4.6 and high-

Top 10 Locations	Count	%
Computing-34F5E0	1,364	77.588%
AC-Ext-34F330	51	2.901%
Computing-34D61C	39	2.218%
Computing-34D600	25	1.422%
Computing-34F5AC	22	1.251%
SPD-234878	22	1.251%
Computing-34D5E4	21	1.194%
Nursing_H116	21	1.194%
S-Block-34ED54	21	1.194%
AC-Ext-34EFEC	17	0.967%

Table 4.6: Student 1 location activity for a sample day

lighted in red. This graphic identifies that not surprisingly the Computing building is central to all locations.

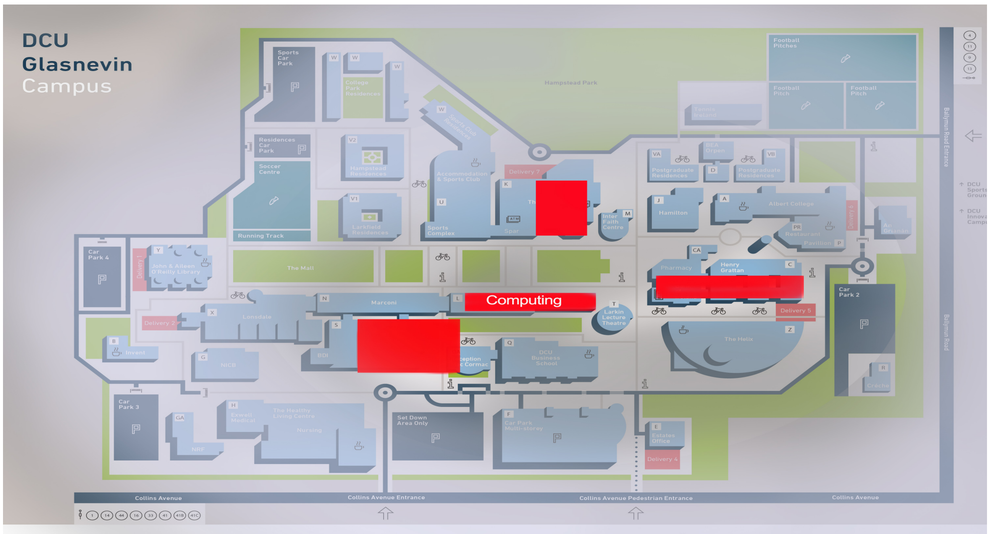


Figure 4.6: Top location of Computing Students (CA & EC programs).

The ability to summarise student activity and the identification of those with the greatest activity levels does not however identify the students with the greatest influence, which is the primary interest of this research. Summary statistics alone could not provide this information. To determine the identity of key students, or students of influence, the **Pagerank algorithm** [46] synonymous with the Google search engine, was applied to the cohort network, with students being the nodes and the number of their meetings being the edges. This algorithm was originally developed to determine the "importance" of a web page to aid the ranking of search engine results to a searcher. Applying the basic principle of the algorithm, a weighting is calculated based on the meetings (edges) between the student (nodes) to the cohort it was expected to determine the **important** members of the cohort. The algorithm was applied to a sample of both Academic and Social meetings. The results of this test is presented in Table 4.7. On the Table's left-hand side is the list of the top twenty "Pagerank" values for social location meetings and on the right hand the equivalent for academic meetings. Examination of the lists highlights the top-ranked student in the Social and Academic arenas are one in the same, with the second and third ranked students in the Social listing in the top ten of the Academic areas. There are fifteen of the top socially ranked students listing in the top twenty Academic listings.

These results introduced a number of issues that needed to be examined before considering applying the algorithm to the complete dataset. While these scores indicate who is **important**, based on interactions within the network, they may only indicate that a high scoring node has a wide circle of friends. However, it may not necessarily identify a sphere of influence. These students may not have any influence on other individual students. They may be friends with everyone i.e. akin to a social butterfly, that meets many, but does no remain for any length of time with any. They may have a high ranking as they are are always on campus and hang out where many of their class mates hang out, but not necessarily with anyone in particular. Alternatively students with a low score may spend a great deal of time with a small number of individuals away from the network. Based on these considerations the research approach now changes focus from the larger network focused at a more granular level. Our next strand of research will focus on the identification of dyads within each academic program.

4.3 Meeting analysis

As previously demonstrated, the statistical summary model has the capability to identify pairs of students based on a "meeting count". Table 4.8 lists the number of students and dyads who had meetings, i.e. student pairs from the same degree program that interact or

Social		Academic	
Student #	Pagerank	Student #	Pagerank
Student 1	0.02049	Student 1	0.01560
Student 2	0.01811	Student 5	0.01437
Student 3	0.01658	Student 20	0.01271
Student 4	0.01568	Student 6	0.01253
Student 5	0.01447	Student 10	0.01226
Student 6	0.01390	Student 2	0.01221
Student 7	0.01390	Student 3	0.01209
Student 8	0.01352	Student 4	0.01208
Student 9	0.01331	Student 14	0.01166
Student 10	0.01310	Student 10	0.01160
Student 11	0.01220	Student 23	0.01141
Student 12	0.01214	Student 16	0.01136
Student 13	0.01199	Student 7	0.01127
Student 14	0.01184	Student 8	0.01115
Student 15	0.01179	Student 11	0.01107
Student 16	0.01160	Student 25	0.01096
Student 17	0.01127	Student 26	0.01092
Student 18	0.01126	Student 40	0.01061
Student 19	0.01103	Student 18	0.01042
Student 20	0.01078	Student 29	0.01038

Table 4.7: Comparing Social and Academic Pagerank scores

"met" during the semester and the number of meetings during the semester among those dyads.

The Table shows there were are larger number of *academic* meetings in comparison to *social* meetings between students in both programs. Anecdotally, this is expected as students spend time in classrooms and laboratories not just for formal classes and tutorials but also to meet in project groups or to study. As previously discussed, Wang's [56] approach to better understand the academic (e.g., study duration) and social (e.g., partying) behaviour of undergraduate students required the classification of each on-campus building with "semantically meaningful labels" such as study areas i.e. classrooms, or social areas i.e cafes. To achieve a greater understanding of this thesis's research student activity, the DCU campus was similarly divided into two categories and three sub-categories in each. The academic locations are: Classrooms, Laboratories and the University Library, while the social locations are Hang_out, Residences, and Transits. The premise is that *friends* spend a lot of time together in the same location at the same time, with the location type providing the context for their meetings.

The summary analysis of students collocation in Table 4.9 outlines the number of meetings in social locations and Table 4.10 the Academic meetings by sub-category. These Tables identified a variance between the programme groups that had not been previously considered. It can be seen that socially 70% of the the CA students met in *Hang-out* locations compared to 78% of EC students. However there is a larger variance in the academic summary with 67% of CA students meeting in the Classroom to 33% for EC students and conversely 62% of EC academic meeting occurred in the laboratories.

Program	No. Dyads	No. Meetings	Avg. Meetings per Dyad
CA	5,523	335,465	60.7
EC	1,230	106,500	86.6

Table 4.8: Student numbers, Dyads and Meetings

There are also indications from this summary that the EC students spend a greater percentage of time in the Library than their CA counterparts. Socially there are also slight variances between the two student groups e.g. there is a greater level of activity at the

Program	Social	Hang Out	Transit	Residence
CA	36,543	25,386 (70%)	8,409	2,865
EC	16,962	13,268 (78%)	3,282	261

Table 4.9: Student Social meeting count by location

Program	Academic	Class	Labs	Library
CA	298,922	199,968 (67%)	97,803 (33%)	1,027
EC	89,538	32,284(36%)	55,619 (62%)	1,786

Table 4.10: Student Academic meeting count by location

Residences by the CA students than the EC students.

In the Academic domain, Table 4.10 shows the largest number of meetings took place in the *classroom*. A large portion of *classroom* meetings occur within the formal environment of lectures with a smaller amount where students have study groups in classroom locations. It is common for students to congregate in the Labs to study or work on group projects.

This was an important finding as it identified that the students from both programmes have different routines and practices. These routines could be based on personality types of the cohort or more likely the structure of their academic timetable and the resultant project work. Regardless of the reasons for these variances, it has identified the need to analyse the groups independently.

4.3.1 Correlation between Meetings and Exam Result Delta

Having identified the requirement to analyse our programs separately the next phase of the project was to identify the clusters within each group and develop a methodology that would identify peer influences, if they existed. Peer influence will be a measure of the Precision Mark (PM) achieved per year by each student and comparing that PM against the PM of the peers. This comparison between PMs will be refereed to as a **Delta**.

Using each students' Precision Mark 3.2.4 as a feature, we grouped pairs of Students by the number of "Academic Meetings" they had with other students in their program. Table 4.11 illustrates the number of meetings as groupings of 200 and lists the Average Delta between students who fall into that group. For example where a student meets another student on 300 occasions the pair are placed into the group range 200-300 and the differences between these two students Precision Mark is averaged with all other pairs in that range.

It could be interpreted from this table that as the number of meetings between a dyad increases in Academic settings, there is a decrease in the Delta score between these dyads. Additionally the Max variance between the student pairs per range is included, which once again could be interpreted that as the number of meetings between students increases the Delta decreases. The variance in the results of this experiment between the CA and EC programs is wide and similarly to the previous Summary Analysis, there is a wide variance in the results between the two programmes. In Table 4.12, the Social meetings analysis, the average delta decreases as the number of meetings between pairs increases. However it can be seen that the range of differences in deltas between the various range groups and also between the two programmes.

Both tables indicate that the more meetings a pair of students have with each other, the closer their exam grades. Whether this is an indication of peer influence or that students of similar ability naturally group together will require further study. This study commenced with a granular analysis of each of the main location categories, i.e. Academic and Social.

	CA		EC	
Academic Meetings	Avg. Delta	Max	Avg Delta	Max
0 : 200	**23.82**	54.84	**6.85**	25.00
200 : 400	11.90	44.59	5.79	19.17
400 : 600	12.22	37.84	7.76	19 .00
600 : 800	11.02	32.75	5.58	15.17
800 : 1000	10.72	28.92	5.22	12.17
1000 : 1200	9.79	15.25	3.71	5.09
1200 : 1400	**1.59**	1.59	**0.84**	0.84

Table 4.11: Average Delta for Academic Meetings, Grouped

Examining the number of meetings and the specific locations of those meetings, yielded similar results, indicating that the more meetings a pair of students have, the more exam their mark delta reduces. These findings are laid out in Table 4.14. This table demonstrates that student pairs have different meeting patterns, for example the highlighted pair on line

	CA		EC	
Social Meeting	Avg. Delta	Max	Avg Delta	Max
0 : 30	**13.75**	54.84	**6.86**	25
30 : 60	12.39	42.34	6.27	19.17
60 : 90	11.51	32.09	6.67	18.17
90 : 120	9.95	30.42	5.56	11.33
120 : 150	14.56	37.84	5.84	15.17
150 : 180	9.38	15.34	3.02	5.83
180 : 210	6.05	10.25	5.23	10.09
210 : 240	9.38	17.42	5.30	11.84
240 : 270	**10.96**	11.25	**2.39**	4.25

Table 4.12: Average Delta for Social Meetings, Grouped

Program	Activity Location	Max	min
CA	Academic	23.82	1.59
CA	Social	13.75	10.56
EC	Academic	6.85	0.84
EC	Social	6.86	2.39

Table 4.13: Program Delta Summary

four have a large number of both academic and social meetings, while others may have a large number academic but minimal social meetings. This distortion is probably due to the fact that students are attending their formal timetabled classes but not mixing with their peers socially. The meetings that take place in the Labs, Library or residence may be a greater indicator of friendship. It was therefore necessary as identified in Chapter 3.1 and Section 3.4 to develop a weighting system that could be utilised to identify locations that *friends* frequent together rather than just *acquaintances*. The development of the weightings is explored in the following Chapter, Section 5.3.1

		Stu_1	Stu_2					Hang					
Student 1	Student 2	Mark	Mark	Delta	Meeting	Academic	Social	Out	Transit	Class	Labs	Library	Residence
a7c220a2391204a	cf0caade99e1055	68.75	39.67	29.08	999	927	72	68	4	588	331	8	0
9eadb93b4f3cc9a	f6d9768a17dbda8	47.83	48	0.17	996	944	52	48	4	372	572	0	0
ad7bdcd6d4f2189	d70601092a0d6ba	49.25	51.92	2.67	996	988	8	8	0	568	420	0	0
482873727621bd6	**7a88e267dba3090**	67.33	81.08	13.75	992	460	532	516	4	156	292	12	12
3440aee39ee0855	b37f2ca45ca9a6e	51.42	53.42	2	986	864	122	102	20	552	308	4	0
482873727621bd6	53b1026ce2c5750	67.33	60.5	6.83	985	920	65	45	20	480	420	20	0
482873727621bd6	68a7768adb80f61	67.33	58.17	9.16	984	888	96	92	4	688	188	12	0
bfe26b4a4bb3c39	cf0caade99e1055	65	39.67	25.33	984	960	24	20	4	568	391	1	0
2b264df9656da40	dad9bbed96d8908	74.42	78.08	3.66	980	853	127	99	28	437	416	0	0
ae2273e5dee794e	d027424da77ef85	62.83	66.42	3.59	977	844	133	121	12	544	300	0	0
482873727621bd6	65d48f06eb25377	67.33	59.42	7.91	977	828	149	149	4	480	340	4	0
bfe26b4a4bb3c39	d70601092a0d6ba	65	51.92	13.08	976	904	72	72	0	568	336	0	0
3440aee39ee0855	65d48f06eb25377	51.42	59.42	8	974	831	143	107	36	552	276	3	0
482873727621bd6	6724d3974f33050	67.33	51.17	16.16	974	920	54	54	0	472	408	40	0
26decb0adcb2f06	9eadb93b4f3cc9a	53.08	47.83	5.25	971	732	239	231	4	224	498	10	4
093ba16377c1206	5a2278f80443782	63	66.83	3.83	970	888	82	70	12	648	240	0	0
2b264df9656da40	82ae4adeefe603e	74.42	66.67	7.75	970	825	145	137	8	373	452	0	0
206fe5037d94a09	e5dd7fcd00cd313	70.92	48.92	22	969	944	25	25	0	672	268	4	0
1da7bb37a8404bf	2499abe788a2b9f	52.25	50.75	1.5	968	760	208	204	4	488	272	0	0
093ba16377c1206	65d48f06eb25377	63	59.42	3.58	966	916	50	30	20	468	448	0	0
071e7661015ea15	e5dd7fcd00cd313	79	48.92	30.08	964	744	220	220	0	240	480	24	0
2499abe788a2b9f	2b264df9656da40	50.75	74.42	23.67	963	756	207	195	12	500	240	16	0
6724d3974f33050	f6d9768a17dbda8	51.17	48	3.17	961	937	24	24	0	421	500	16	0
2499abe788a2b9f	f6d9768a17dbda8	50.75	48	2.75	960	936	24	20	4	464	464	8	0
b37f2ca45ca9a6e	cf0caade99e1055	53.42	39.67	13.75	958	874	84	60	24	292	582	0	0

Table 4.14: Example of Dyad Precision marks, deltas and meeting count by location

4.4 Clustering

Having developed a weighting system for the campus locations in terms of their relative importance for student meetings, the next process was to apply the weightings to locations and correlate them to the frequencies of student meetings within their network. A common research method used in network analysis and one tested here for suitability is **Clustering**. As part of an in-depth overview of clustering methods, Madhulatha [39] defined clustering as "a process of grouping similar objects into different groups, or as the process of partitioning a data set into subsets, so that the data in placed in a subset according to some defined distance measure". The resultant subsets are a collection of objects which are "similar based on some features and "dissimilar to the objects in other subsets. Clustering methods and their usefulness in this research will be examined in detail in the next chapter (Chapter 5)

4.5 Other Developments

4.5.1 Additional Analysis

It was important to the main project to understand if there were influences that needed to be considered and could impact on the overall results. A small number of these strands will be mentioned here and detailed in the accompanying Appendices, these include:

1. The influence of Student drop-out on our clustering algorithms;
2. The use of multiple WiFi enabled devices by students;
3. Project group progression over time;
4. New students joining the research cohort.

4.5.2 Drop out

While this thesis does not focus on the **why** a student does not continue with their undergraduate degree program, it is necessary to examine whether a student that retires should be included in the longitudinal analysis, if they drop out during the academic year. Appendix 6.3 is a detailed examination of the students registered for the academic year 2015 and those who progressed to 2016. It is determined that a student who registers for a program, but at the end of the year has no or a minimal Precision mark for that year's examination, has dropped out during the year. If a student does not register for a year having completed the previous year, they are also considered to have dropped out of the program but not during the year.

The conclusion of the analysis is that students who drop out during the year have had a minimal attendance, therefore minimal contact with their peers and thus minimal influence on them. These student will not be included in the analysis for that year, but will be removed. Those who do not re-register will be removed from the analysis from that point forward as they will have no input, i.e activity or output i.e. Precision Mark.

4.5.3 Module group progression

As part of student introduction to third level education, lecturers may assign project work that requires students to work as part of a project team. This can have the added benefit of aiding the integration process and introduces students to others and helps them blend in. This has many benefits including the avoidance of student isolation and is an aid to student retention. It was considered worthwhile testing the hypothesis that these students become friends with the other members and develop a friendship that endures for the rest of their academic career. Analysis as presented in Appendix 6.3 indicates that project teams formed in first year do not remain as friends into their second or third year.

This question was asked informally of students who took part in the Focus group 3.4 and they concurred that in their experience, that students may not remain "friends" with the people they were grouped with in their first year.

4.5.4 New student joining cohort

Students can join a programme at a particular point in time for a number of reasons, including those repeating a year due to failure of previous exams, those transferring in from other programmes or other Universities or others returning to the program after taking a year or more away. Once a student joins a programme they are interacting with others within the cohort and can have influence through various forms of interaction and friendships. Any student that joins the research cohort and remains in the programme for a year will be considered in all experiments for that year.

4.6 Conclusion

As with many research projects, the investigation process and the methods employed to pursue answers to a **research questions** is often unknown, or requires proper definition. This Chapter outlined the approach taken to scope the project objectives and develop the processes and methodologies used to answer **our** research question. The intention here was to present the foundations for the experimental process, the decision making process that was followed, the experimental results.

A large portion of this section was given over to the examination process and endeavoured to identify student location from their WiFi logs and therefore the co-location of multiple students. There was an examination of a number industry tools appraised for their suitability for this thesis. Adopting a test data set that was compatible to our research cohort we examined a number of research approaches before a decision was taken on a specific course of action.

By the Chapter end we had adopted methods that had proven the concept of being able to identify students from their WiFi logs and provide context to their meetings with their peers. In the next chapter these concepts and methodologies are further refined and applied to the research cohort.

Chapter 5

Experiments

5.1 Introduction

This chapter will outline the process and the methodologies used to prepare and analyse data and from that to identify the unique clusters of students within the various cohorts on which we can explore the research questions and thesis hypothesis. Further explanation of the rationale for choosing these methodologies and their development through a summary of the experiments undertaken will also provide detail of the approaches undertaken. For the upcoming analysis stage of the research the software language package “R” was employed. This package was chosen for its statistical analysis features and high quality visualisations capabilities. As the work progressed it was determined that R was better suited than Python for the analysis and presentation of results both visually and tabular hence the switch to R.

This phase of the research work is an examination of the experiments designed in the main to test the second of the thesis’ research question formulated earlier in Section 1.2 and to lay the foundations for the subsequent questions. This question queried the ability to identify singular friendship groups from within a network of students based on collocation on campus. Having previously employed a number of summation techniques including a PageRank algorithm, it was determined that it is possible to identify the interactions between individuals and students within their network. It was determined that an individual’s degree within the network (number of meetings) did not identify the closest friendship of an individual student, i.e. the one(s) that they spend the greatest amount of time with and may have an influence on them. To identify a close friendship there is a need to examine other methods such as clustering techniques that are often used in statistical analysis. Clustering algorithms can fall into a number of categories including partitioning, hierarchical and density-based. This Chapter commences with an examination of Partitioning and Hierarchical clustering models as they were considered suitable for this research.

5.2 Techniques

5.2.1 K-Means Clustering

There are a number of partitioning techniques tested including the K-Means Algorithm and the less commonly used medoid Algorithm both of which would be very popular in similar research projects. Jain et al's, [33] examination of the Landscape of Clustering observed that "K-means can give reasonable clustering results that are not far away from other algorithms, and consistent with the general perception of the K-means approach." K-means however is a static model requiring a manual input i.e. the choosing of the value of K, the number of clusters, which has a significant impact on the model's results. This was supported by Karypis [34] who observed that K-means algorithms can break down if the choice of parameters in the static model is incorrect with respect to the data set being clustered, or if the model is not adequate to capture the characteristics of clusters. Due to the nature of this thesis' research, the number of groups that students will naturally form within their programme cannot be estimated in advance. Therefore it would be incorrect to make a judgment call and pre-define the number of clusters, that **may** form, based on a best guess.

To confirm these assertions a number of tests were undertaken, utilising R, to analyse the results of K-means clustering methodologies using a test sample of a semester's data. These investigations attempted to identify if it was possible to determine the optimal number of clusters (K) formed in a semester. The investigation undertook a number of approaches, from a chronological methodology whereby choosing a small K value and iteratively increasing that value of K and monitoring the number of students allocated to each cluster after each iteration. Clusters are generated numerous times using the same K value and examined for individual membership. Clusters were additionally compared to how a single student's cluster partnerships changes as the number of groups varies. Having examined multiple "K" partitioning approaches, it was identified that the final cluster taxonomy was dependant on the initial random seeding (starting points) and the number of Clusters defined (the value of K). Although the clusters' composition did not vary dramatically, there were slight variances within the same sub-set of data. Results indicated that the effect of the randomness of the starting point in the process, was considered untenable as the constituents of a cluster can change based on this algorithm's starting point.

5.2.2 Hierarchical Clustering

An alternative investigation into Hierarchical Algorithms (HAs) was carried out using the same data subsets as in the partitioning algorithm testing. Hierarchical clustering involves creating clusters that can be pre-determined and ordered from top to bottom or vice-verse. A HA can be visualised as a tree structure or *dendrogram*. The dendrogram plots are a useful visualisation and summary of a dataset by illustrating the arrangement of the clusters produced. The dendrogram displays a list of objects along the x-axis, and the distance at which these are clustered on the y-axis. As can be seen from the Dendrogram in Figure 5.1, in the next section, the separation of branches can be relatively small i.e. the comparable height at which there is a split in a major branch is indicative of the closeness of clusters. It is this closeness that will be a challenge the identification of distinct clusters. Taking any height along the y-axis from which a line is drawn parallel to the x axis, will cause that line to cut through a number of branches, each of these lines representing a cluster. Identifying these branches is referred to as branch or tree cutting or dendrogram pruning. A common tree cutting method is the fixed height branch cut. A fixed height on the dendrogram is defined and each branch below that height is considered a separate cluster and with each containing the members of that cluster. Figure 5.1 is a typical example of a dendrogram which will be examined in greater detail. Hierarchical clustering methods differ primarily on how the dissimilarity between clusters is calculated given the dissimilarities between their constituents. There are a number of theses methods and these will provide different solutions. Section 5.3 examines those methods in greater detail.

5.2.3 Distance Matrix

Clustering processes are based on the identification of the similarity between objects i.e. a distance measure between the nodes of the network. This thesis is examining the similarities based on co-location among students on campus and requires the output to be presented in a manner that can be utilised by a clustering algorithm. A common form of presentation of data is as a distance matrix.

A distance matrix (Dis-similarity Matrix) is a two-dimensional array containing a set of calculated distances between the elements of a set. For this research it is a measured distance between students based on the the number of times they co-locate on campus. This process is demonstrated with the use of Table 5.1, this matrix lists the location of a student within a pre-defined time-frame. From this data it can be determined who was co-locating, where, when and how often. This table illustrates an example of the activity of three students at four sequential times.

Student 1 and Student 2 share a location at Time-1, but as this is a scheduled class time slot, the relevance is not as important as Student2 and Student3 sharing two time-slots at Time-2 and Time-3. These student are in a Lab at the same time and later at Time-4 at the campus cafe. Attendances in a Lab can be for either formal classes, group study or individual study reasons and it cannot be determined if the students are together deliberately or by coincidence. The other meetings at the cafe may be significant in determining their friendship.

Student	**Time-1**	**Time-2**	**Time-3**	**Time-4**	**Time-n**
Student1	Class1	Class1	Class1	Lab2	Lab2
Student2	Lab1	Lab2	Cafe	Cafe	Cafe
Student3	Class1	Lab2	Cafe	-	-

Table 5.1: Student, Time and Location – sample table

As presented in the discussion on dyad identification earlier in Section 4.1.4, the earlier exploratory experiments used a time window of 20 minutes to test theories and develop experimental models. Consideration was give to the appropriateness of this time-frame and what interval would best represent student meetings. Students outside of formal class times are transient and may visit a number of locations in any given period. Those students who are friendly with other students will often travel together or visit the same same location within a fairly short time period. It was therefore considered appropriate that the time-frame used was reduced from 20 to a 10 minute interval. Therefore students who arrive at a location at the same time and whose WiFi device logs on the Eduroam within a ten minute window of each other will be deemed to be co-locating. Any other device that logs in within an overlapping 10 minute period will also be deemed to be co-locating.

It was also considered that a student may visit a number of locations in any chosen time-frame and it was necessary to determine where a student spent the greatest length of time, in that period. As part of the matrix generation function, the location at which that a student spends the majority of time, within the 10 minute time-frame, is the location (NAS identifier) recorded for that student in the time period. Taking these points into consideration and applying this process to the total WiFi activity logs a single location matrix would be generated per day.

These daily co-location matrices were the basis from which the student-student distance matrix would be generated and provide the input to the clustering algorithm. To generate the dissimilarity matrix required a methodology that could interpret mixed data-types such as categorical variables, i.e. those found in the NAS identifiers or location identifiers. Following desk research, the method chosen was Gower's General Similarity Coefficient a popular measures of proximity for mixed data types, to develop the required distance matrix.

5.3 Hierarchical Clustering Experiment

This research examined a number of agglomeration methods available to carry out hierarchical clustering, such as Single linkage, Complete linkage, Average linkage, and Ward linkage. Each method was tested using a subset of the overall data through a number of experiments to determine the best fit for the complete data-sets in this research. The determination process involved the development of a number of dendrograms using the various methods and the analysis of the number of clusters generated. The experiments included the examination of the dendrogram and the individual clusters, both by quantity and their composition. Both the Complete Figure 5.1and Ward Figure 5.2 methods produced the most interesting visual dendrograms and distinctive clusters. The only discernible difference between both methods is the visual representations of each, i.e. their dendrograms.

Clusters are identified from dendrograms through the process of "Tree cutting" where branches are identified as distinct clusters. While the clusters in a dendrogram can be identified visually they are not as easily separated computationally. This is a problem that Langfelder [36] addressed with a novel cutting approach by developing a Dynamic Tree Cut package for R. This package uses the shape of the branches of a dendrogram to aid in the identification of closely bound clusters.

There are two variants of the package of which the Dynamic Hybrid cut method, a bottom-up algorithm, is employed here. Using a previously generated dendrogram and applying a number of criteria, preliminary clusters are identified. The process involved:

- Firstly setting a minimum number of cluster members,
- Identification and removal of outliers within branches
- Each cluster must be distinct from others based on minimum distances, i.e the **Cut-Height**.

The clusters produced using the Ward Hierarchical method, and the Dynamic Hybrid cut method are very similar in terms of the number of clusters and their composition compared to clusters produced using the complete method.

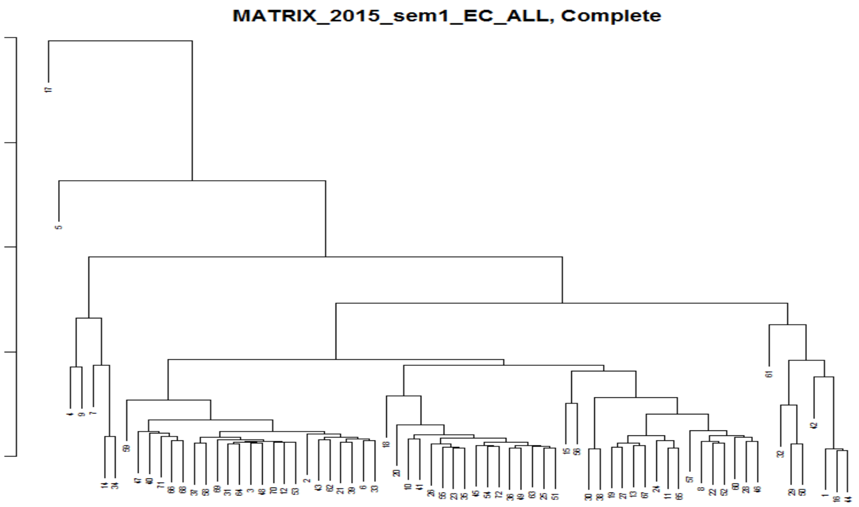


Figure 5.1: Dendrogram using Complete Linkage Calculation.

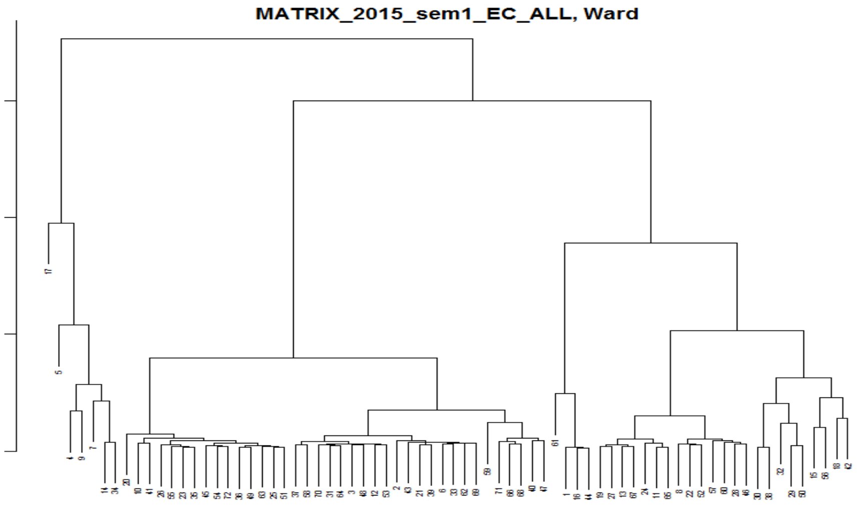


Figure 5.2: Dendrogram using Ward Linkage Calculation.

The results of these early experiments presented an issue that had not previously been considered. From an examination of the dendrograms, there were a number of students who appeared to continually be separated from the main cohort. Examples can be seen in both Figure 5.1 and Figure 5.2 where it can be observed that on the left of **each tree** i.e. students 17, 5, 4,9 and 7 do not seem to be clustering as part of the main tree. Further examination of the dendrogram provides other examples of students not readily fitting into any one cluster. It is accepted that this can occur naturally as some students are not gregarious in terms of their interaction with classmates. They may be socially shy and do not interact with fellow students or possibly they visit the campus only for formal classes. Whatever the reason, they will generate a smaller number of WiFi access log entries and minimal "co-locations" with other students and subsequently prevent the model from fitting them into any one cluster.

Utilising the anonymised list of registered students and the WiFi logs, the "isolated" students were identified. Further examination identified that these students had a lower number of WiFi interactions during the academic semester and their academic performance for the year indicated that they may have actually dropped out. As previously identified, not all students remain in the programme and some drop out for various reasons. A detailed examination of the drop-outs is presented later in Appendix C. As demonstrated there, the activity of students who ultimately drop out is sporadic and generally occurs in the early parts of, or near the end of the first semester of an academic year. Consideration had to be given to the inclusion or exclusion of the logs of such students from distance matrix generation.

While it is accepted that these students will have had some interaction with the cohort's network, at best, it is minimal. One determining factor was the consideration of these students, who had no, or minimal (less than 10 marks) overall precision mark, a measure that is central to the majority of this thesis calculations. Testing this hypothesis identified that including students with a precision mark of 0 or less than 10 marks, distorted the results of clustering algorithms. To prevent any bias being introduced through their inclusion required their removal from semesters of the year that they dropped out. Using the aforementioned R-Console platform, we identified each student of these students had their WiFi records removed from the dataset prior to the commencement of the clustering process.

Another cohort of students identified at this juncture, are those who did not commence their studies with the original cohort in the year 2014/2015, but joined the test programmes in later years. These students were either repeating a year or returning from a sabbatical and once integrated into their programme interacted extensively with their new class colleagues. These interactions may have influenced others and therefore had to be considered. Any student that joins the initial research cohort and achieved a precision mark of greater than

10 marks were thus included in the clustering analysis.

5.3.1 Weighted locations

In an earlier Chapter and Section 3.4, the findings of a focus group were presented, and one of the outcomes from that group was that not all co-locations are considered to be of equal importance e.g. sharing a classroom during a formal scheduled class is not as important to a friendship as a social meeting outside academic core hours. To ensure we capture and use the differences between each area in a friendship context, a method of weighting areas based on students' opinion of where they spend the most time with their **friends** was required. A focus group and student survey was carried out with the aim of identifying the variance in importance between the campus locations. The output of the focus group was used in the preparation of a questionnaire with the resulting responses being used in the development of a weighting system. The full explanation of the focus group process and findings are presented in section 3.4 and Appendix A and B.

The weighting system devised is based on the locations where students were **less** likely to encounter a random meeting with others from their programme cohort. Table 5.2 for example illustrates that the top ranked area is the Hang_out areas, outside of the academic core hours. Hang_out areas would be the bars and cafes, i.e. meetings in the campus bar in the evening will most probably be predetermined and not random. The most likely location and time to encounter random meetings are the Labs or classrooms, during core hours. These location weightings will be utilised within the next phase of the project, the clustering process where they are applied to a Distance Matrix and subsequently used in the clustering algorithm.

As stated previously (Section 3.4.4), the weightings are a correlation of the opinions of the members of the Focus Group and the responses to a Student Survey, specifically asking where they spend the majority of time outside of their academic timetable. Based on the opinions of those who completed the survey and specifically questions 5 through 8, we have made the **assumption** that these weights are representative of the research cohort.

Reference_id	Location	Period	Ranking	Weight
m_n	Hang_out	Non-Core	1	0.03
li_n	Library	Non-Core	2	0.06
c_n	Classroom	Non-Core	3	0.06
l_n	Labs	Non-Core	4	0.07
r_n	Residence	Non-Core	5	0.08
t_n	Transit	Non-Core	6	0.08
r_c	Residence	Core	7	0.085
m_c	Hang_out	Core	8	0.085
li_c	Library	Core	9	0.09
t_c	Transit	Core	10	0.09
c_c	Classroom	Core	11	0.13
l_c	Labs	Core	12	0.14

Table 5.2: Location weightings per place and time

5.3.2 Clustering Summary

A synopsis of the clustering process is outlined below:

1. Generate a co-location table by location category and per day of the monitoring time;
2. Generate a distance matrix per co-location table;
3. Apply a weighting to each distance matrix;
4. Merge all distance matrices into one per semester;
5. Generate a dendrogram;
6. Apply the Dynamic Hybrid cut method;
7. Identify Clusters and their memberships.

There are two distinct stages in the clustering process. Steps 1 to 4 and 5 to 7. From the WiFi data logs a co-location table is generated per day, per location and per period classification, i.e. location and either during core or non-core hours. Each table is the basis for the creation of a distance matrix which includes the application of the location weightings. The resultant daily distance matrices are combined into a single overall distance matrix. This combined matrix contains the calculated distances between all students for that semester.

In the second phase of the process, a dendrogram representing the hierarchical clustering of the the distance matrix is generated. The resultant dendrogram is further processed utilising the Dynamic Hybrid function. The result of this dual process are a collection of clusters. The number and composition of the clusters are dependant on the configuration set up of the Dynamic Hybrid function. This method provides greater control over the cutting of a dendrogram and the resultant clusters.

The output of this process is a list of student ids and the cluster number they belong too. Clusters are based on the weighted distance matrix and the cut-height variable defined in the algorithm, specifically the cut height. An example of how students are clustered, depending on cut-height is outlined in Table 5.3. This table is an extract from the analysis of a single programme, EC year 1, semester 2 activity. Each row lists the student number and their cluster (number) based on various cut-heights. Closer examination identifies that the first two records are students who have been clustered together regardless of the cut-height chosen, i.e. at the various cut-heights both have been clustered in the same cluster i.e. 31, 27, 18, 11, 10 and 10. Further analysis identified that other students do not become part of the same group on a consistent bases until the cut-height increase significantly e.g. the last two students in the table become paired and remained so from a cut height of 40.

Examination of other groups such as 8, 9 & 10 identified that they formed their groups at an early stage and remained independent from others regardless of the cut-height.

Group 7 had two consistent members from an early cut with other members joining up to a cut-height of 30, after which no new members joined.

Studenty Group — EC 1 Sem 2	Cut Height						
	1	**5**	**10**	**20**	**30**	**40**	**50**
Student	Assigned group number						
51dcbedccd9addff357f6944bc7b196e	0	31	27	18	11	10	10
7e7620fb584d8147d784be811f7ac5a2	0	31	27	18	11	10	10
4beab22fa7b4aefb12dad6f13d57e04e	0	13	14	15	10	9	9
73358d59a7a8eb7e0cb5c4b6dceb7697	0	13	14	15	10	9	9
b98f5cc754ef75567aae4412c467ad6b	0	13	14	15	10	9	9
52e38904a624a2f7cee567e7f6be53de	0	5	9	11	9	8	8
5ccdf3bfba68447225c9b78e8cc4eb05	0	5	9	11	9	8	8
776a71857e7a6070b67c6418c455c5f9	20	5	9	11	9	8	8
b22474774a018433574d322d3b7876d0	20	5	9	11	9	8	8
4fabce8b1c971fc98a28644ea43947e0	0	26	22	16	8	7	7
5091741b48dfe74b121e569bc8595d45	0	28	24	13	8	7	7
7f7bb3bdaf27c2a353cce9010e02f868	0	26	22	16	8	7	7
d678ae5829848f4be5bb8fa57aae619e	21	23	19	13	8	7	7
f0381e3ba5f7f6e7c3cbf6ed0f3e3888	0	28	24	13	8	7	7
f93952869e1d5cbd54a426a4d8baaa58	21	23	19	13	8	7	7
33a98797cd1b5befdab5b4f80e076a50	0	29	25	9	6	6	6
355efff071a2c3ddbc91937b02bd438d	0	29	25	9	6	6	6
4c59e0d414556ecab43187ac7ed7930e	0	12	13	9	6	6	6
c7afeba75b0274da8657336df2f8cd0d	0	30	26	17	6	6	6
e6cd9e7c7b66ff8fa71b0e559c0f2a2d	0	12	13	9	6	6	6
fcd00bacc7327dc417e56d1ebdb67189	0	12	13	9	6	6	6
feff8db2f6e8ecf428c7da9408012dda	0	30	26	17	6	6	6

Table 5.3: Group membership by cut height

5.3.3 Cluster results

In general for both program, CA (Computer Applications) and EC (Enterprise Computing) the number of clusters increases up to a cut-height of approximately 20 after which the number of clusters decreases. The number of clusters stabilises around cut-height 40 for the EC program and at 60 for the CA program. These findings, which will be the subject of greater scrutiny later in this chapter, illustrates that the methodology for the identification of clusters of students is robust and Research Question 2 has been addressed positively, at this point.

As previously identified, in chapter 4, section 4.3,there is a uniqueness of the activity and behaviours of the student groups found in the CA and EC programs. It is therefore appropriate that each program should be independently subject to the same vigorous analyses and their result are presented individually. Testing was carried out in the areas of:

- Clusters identifiable at various cut heights per program per semester;
- Cluster analysis for range of marks and differences in marks between the members of the clusters;
- Longitudinal analysis of the divergence or convergence of exam results within clusters;
- Relationship of student marks to the members of their cluster;
- Ranking of students changes within their cohort over time.

The number of clusters identified at various cut heights per program per semester are presented for the CA program in Table 5.4 and for the EC program in Table 5.5. Examination of the clusters' composition at these various cut-heights identified that at the smaller cut heights (less than 5) there were no discernible clusters. At the height were the number of clusters peaked (10 & 20), the clusters comprised mainly of singular dyads. At the larger heights ($>$ 40), there are a smaller number of clusters with larger membership. Cluster examination was carried out on the various clusters formed at the various cut heights. The object was to identify those groups or pairings that demonstrated some consistency in their composition across a number of semesters. For a cluster to be considered robust and not formed randomly, it was considered necessary for the membership of the group to, at a minimum, be traceable from one year to another and preferably span at least three semesters. An examination of the dyads generated at the mid level of cut height revealed that those pairings in the majority of cases did not meet these criteria. Testing of the clusters that occurred at higher cut levels did demonstrate some repetition. These findings are further explored later in this Chapter as it was the case that the most **robust** of clusters were found to contain sub-clusters that could be traced through a number of semesters. These **sub-clusters** have been identified as containing longitudinal relationships that could be considered as **friendships** . These friendships are considered to be the groupings that hold the key in identifying any effects of peer influence on their performance.

Cut height	**1**	**5**	**10**	**20**	**30**	**40**	**50**	**60**
CA1_Sem 1	37	49	41	22	11	10	8	7
CA1_sem 2	43	52	45	18	10	7	7	7
CA2_Sem 1	36	39	29	13	8	6	5	5
CA2_Sem 2	36	43	31	12	7	7	6	5
CA3_Sem 1	27	33	26	11	7	6	5	5
CA3_Sem 2	15	29	29	17	11	7	5	5

Table 5.4: Number of CA - Clusters by Cut height per semester

Cut height	**1**	**5**	**10**	**20**	**30**	**40**	**50**	**60**
EC1_Sem 1	20	31	31	23	14	12	11	9
EC1_Sem 2	22	31	27	18	11	10	10	9
EC2_Sem 1	19	20	16	10	7	7	7	7
EC2_Sem 2	15	22	21	13	9	9	9	9
EC3_Sem 1	15	18	13	8	7	7	7	7
EC3_Sem 2	5	17	15	9	8	7	7	7

Table 5.5: Number of EC - Clusters by Cut height per semester

We next approached an examination of the cluster composition for the CA and EC cohorts that were generated by the clustering algorithms.

5.4 Cluster Composition and Examination Marks of the CA (Computer Applications) Cohort

As identified the cut level of “60” in the CA program and of “40” in the EC program generated a consistent and stable data-set of clusters (groups). It was these groups that would be the bases of future experiments. The granular examinations of groups commenced with the generation of a Box-plot per semester of the groups generated at the chosen cut heights. Figure 5.3 to Figure 5.5 are box plots that visually represent each group generated and the range of marks within each group.

A large proportion of the analysis that will be presented in the remainder of this thesis involves the comparison of relationships between students, their groupings through semester analysis carried out across the six semesters of this research. Semester one of year one will be considered the **baseline** with all subsequent semesters are referenced to it. To ensure a legitimate comparisons is undertaken, it was necessary to take account for any variances in the program design and semester variances, that may affect comparisons. This is particular in reference to semester two of the third year of the CA & EC programs. During this semester a large portion of students do not attend the campus but are placed in industry for internships.

Examination of a singular group (# 3) identified from the CA year one semester one dataset, as illustrated in Figure 5.3. This group comprises of eighteen students in it's first semester and had a Precision mark range of, 13.25 to 88.58 with a median of 58.25. An examination of the member composition of this group over a three year period identified that a number of the students remained as part of the same group while others either dropping out of the program completely, or joined other groups. Due to these composition changes, the initial group reduced to seven students in the third year, see Table 5.6. In semester one of the third year the group had sub-divided into two distinct groups i.e. Students 1 to 5 in group 1 and 6 & 7 in group 4. In the second semester (CA3_2), two members (1 and 3) were not placed in groups with their original groups. There are two possible reasons for this occurring, these students potentially did not continue to co-locate with the rest of their group or due to the Programme structure which involves the placement of students in industry during this semester, thereby are not being on campus to develop meetings, with their peers. Students in the CA programme spend the first three weeks of the semester on campus working on projects before they depart for their six to seven month work placement. It is therefore considered that the first semester of year three is more indicative of true friend groupings than those created in the second semester. In the second semester the students that do remain on campus to carry out research or conclude other projects will spend a greater length of time together which they may otherwise may not have. These meetings will create a number of new groups while they are reflective of the meetings for that semester, but may not be truly representative of previous semesters friendships. Other relevant factors include the similarity of the class and exam scheduling in all other semesters of the programme.

A granular examination of the individual groups per semester was undertaken to identify the spread of the marks within the group and also the closeness of the the individuals marks from the Median and Mean values of their group. Using the first years figures as a baseline and analysis was undertaken to identify how the behaviour of the student Precision Marks change in relation to each other from year one to year three.

An examination of each plot in the Figure 5.3 to 5.5 presents a group of CA students and illustrates the range of marks achieved by the members of that group.

		CA 1		CA 2		CA3	
id	**Username**	**Sem'1**	**Sem'2**	**Sem'1**	**Sem'2**	**Sem'1**	**Sem'2**
1	ec9673d21f129	3	2	3	2	1	2
2	d4e688852304b	3	3	5	1	1	1
3	f7920468276fd	3	3	7	6	1	5
4	0ec5d39d8e8eb	3	4	8	8	1	1
5	9addfb2295e7	3	5	2	5	1	1
6	3dc2ba84ad58c	3	5	3	4	4	4
7	de215da0edf40	3	6	1	2	4	4

Table 5.6: Example of a group from Year 1 to Year 3

The first charts illustrates that in year one (CA1_1 and CA1_2) there were 7 groups with a range of precision marks from 12.25 to 88.58. Of the seven groups, five had low outliers that dropped below the 40% precision mark. As it is a requirement for students to pass all subjects before progressing to the next year of their programme, the thirteen students who fell below the 40% did not progress into the second year, at that time. A similar examination of the second year two, identified that each of the groups also had a tail of the box plot dropping well below the 40% mark. This examination revealed that a approximately half of the forty three students that dropped out after the second year, had a precision mark of less than 40%. It also revealed that 39 of the 43 dropouts had a precision mark below 50% before retiring. By the third year CA3_1 the marks range had narrowed to 31.1 to 80.6 with a reduced number of student struggling to achieve the pass mark.

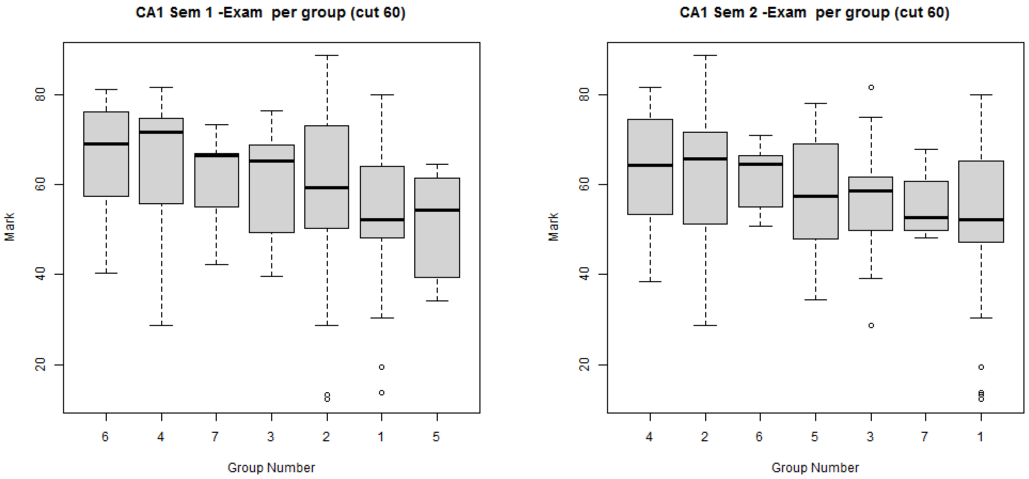


Figure 5.3: CA1 Semester 1 and Semester 2 Exam Mark range.

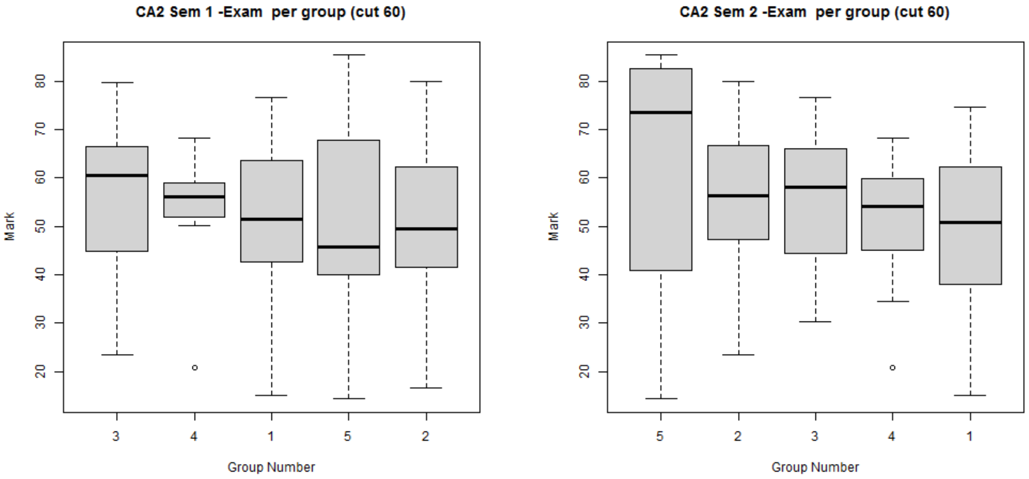


Figure 5.4: CA2 Semester 1 and Semester 2 Exam Mark range.

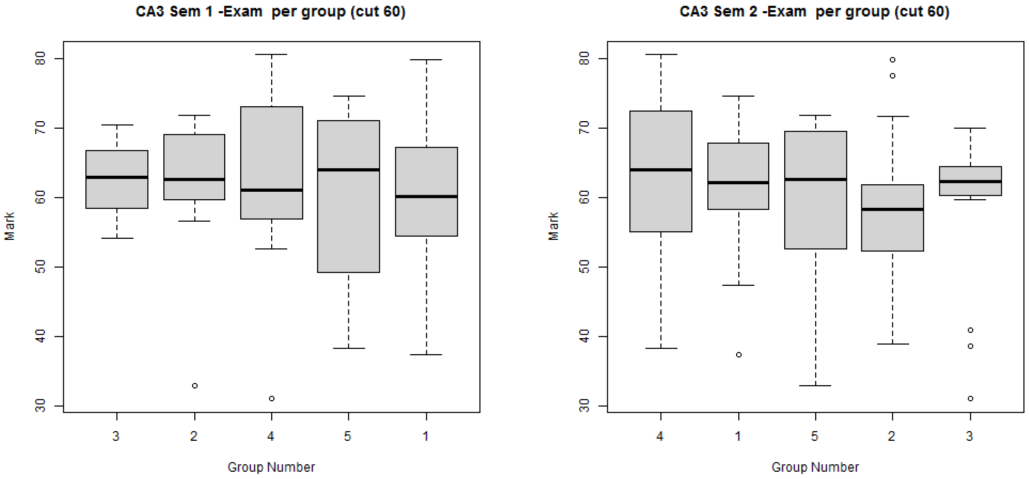


Figure 5.5: CA3 Semester 1 and Semester 2 Exam Mark range.

Due to the variance in the group composition year on year direct comparison of the median and mean of a groups exam results could not be carried out. However an examination of the collective Median and Mean values (see Table 5.7) per semester was undertaken to identify any potential trends. The median values from each group are analysed to identify the min and max value per semester, the mean values are similarly analysed. The **max Median** value in CA1_1 was 72 marks and the min 52.3 giving a spread of 19.7 marks. In the third year the range changed to a max of 64 and a min of 60.2 marks, which is a variance of 3.8.

There was also a corresponding change in the mean values, from a max of 66.2 and of min 51.5 marks in year changing to 62.6 to 59.9 in year three. This movement of the min and max values for both mean and median values, could potentially be demonstrating that the marks range within groups are narrowing. This narrowing of ranges could be an indication of groups influencing the precision marks of other students or that students with similar ability come together in the same group over time. Conversely, and more likely reason is that it may be simply that the min is increasing as less able students drop out and the remaining students have adopted university life and are working harder. This will be explored in later sections.

	Median			**Mean**		
	Max	**Min**	**Var**	**Max**	**Min**	**Var**
CA1_1	**72**	**52.3**	19.7	**66.2**	**51.5**	14.7
CA1_2	65.7	52.3	13.4	62.4	53.2	9.2
CA2_1	60.5	46	14.5	56.7	50.81	5.89
CA2_2	73.5	50.8	22.7	55.5	52.8	2.7
CA3_1	**64**	**60.2**	**3.8**	**62.6**	**59.9**	**2.7**
CA3_2	64	58.4	5.6	63.4	58.7	4.7

Table 5.7: Semester Mark CA - Median and Mean measures per group

5.4.1 CA - Group Delta Mark

A separate examination of the groups was undertaken with the analysis of the Average Delta Mark per group. The Delta measure, calculates the difference between an individuals precision mark and the average precision mark for their group. The group deltas are presented graphically in the Figures 5.6 to 5.8 and summarised in Table 5.8.

From this table it can be identified that there is a minimal difference in the Median and Mean Marks between year one and three. The max average (Mean) of the delta groups in CA1_1 is 13.2 (highlighted) and this figure increases in the second year but returns to 11.8 in the reference year, year three. This variance between these year was just 1.4 marks, the Min during the same period goes from 8.8 to 4.7 a variance of 4.1. As the values are decreasing over time it is indicating that student deltas are getting closer to their groups average mark and a narrowing of the ranges of marks between students. These findings may also be an indication that students marks are converging within their groups. The Median values are not varying by any degree between year 1 and year 3, in year 2 they do increase, in line with the mean values for that year.

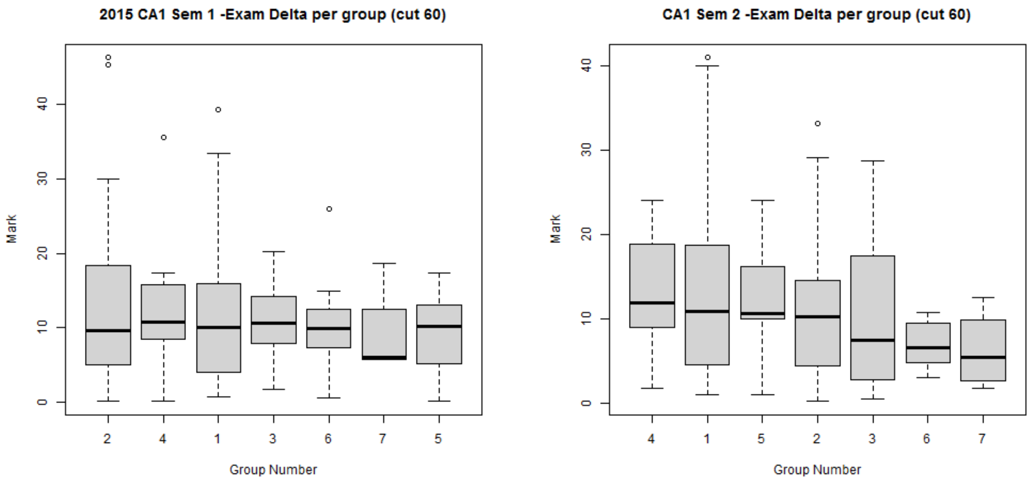


Figure 5.6: CA1 Semester 1 and Semester 2 Group Delta

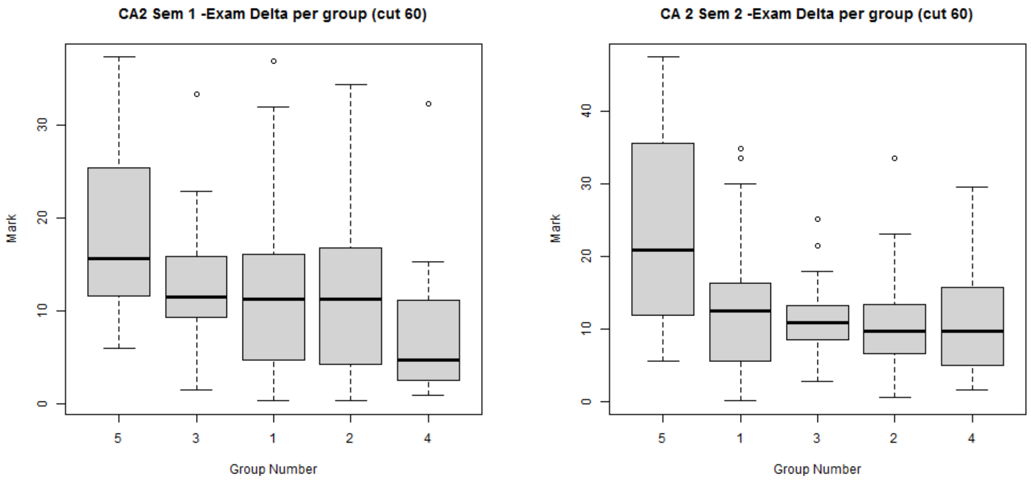


Figure 5.7: CA2 Semester 1 and Semester 2 Group Delta

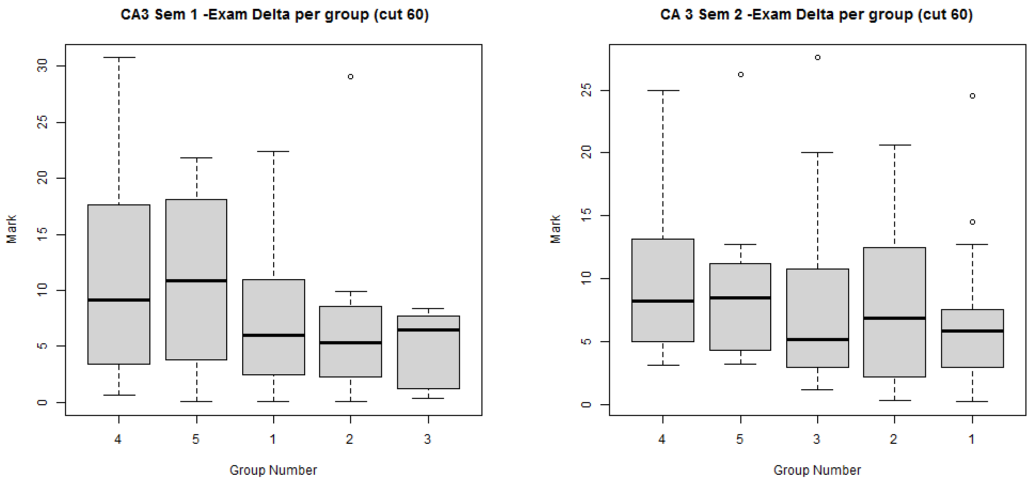


Figure 5.8: CA3 Semester 1 and Semester 2 Group Delta

Delta	Median			Mean		
	Max	Min	Var	Max	Min	Var
CA1_1	10.8	6.1	4.7	13.2	8.8	4.4
CA1_2	11.9	5.4	6.5	13	6.3	6.7
CA2_1	15.7	4.7	11	18.9	8.7	10.2
CA2_2	20.9	9.7	11.2	23.6	10.8	12.8
CA3_1	11	5.3	5.7	11.8	4.7	7.1
CA3_2	8.5	5.1	3.4	9.5	6.7	2.8

Table 5.8: Semester delta - Median and Mean measures per group

5.4.2 Program CA - Area Under the Curve

To place the narrowing of group delta marks in context, the delta values per student were plotted i.e. the distance a students mark is from their group average, and presented in Figure 5.9. These graphs present the complete distribution of variances between each student in their group. A comparative measure was developed that could be used for trend analysis. Using a threshold value of 10 marks or less from their groups mean a density measure or an Area Under the Curve (AUC) value is calculated, per semester. That is a value was calculated that represent all the students that had a delta of 10 marks or less from their group average precision mark

The individual semester AUC values are presented in Table 5.9. In the first year semester one (CA1_1) the AUC the value is 0.489. This figure indicates that approximately 49% of all pairs have a precision mark variance of 10 marks or less. This value varied an insignificant amount between year one and year two, but had a significant increase in year three, to 0.694, i.e. approx 69% of student pairs had a variance of 10 marks or less with their group mean. This indicates that the variance between group members precision marks decreased over this time, supporting the previous indications that group marks tend to converge. A further test was carried out using a narrower boundary of 5 marks. The results in CA1_1 is 0.255 or approximately 25% have a variance of 5 marks from the mean, this value increased to 0.403 (40%) by the third year. The results of this new boundary are illustrated as the shaded parts of year 1 semester 1 and year 3 semester 1 graphics. An additional comparison test was carried out to identify what percentage of the 10 mark density measure is made up of the 5 mark density measure. In year one the 10 mark AUC was 0.489 and 5 mark AUC was 0.255, that equates to 52 % of the 10 mark being made up of the 5 mark AUC. Over the intervening period that value increased to 62%. This indicates that there is a greater improvement of marks in the 5 mark range than the 10 mark range. Further analysis of the relationship between the student precision marks and their variances was required to determine which student and which groups, marks varied and in what manner.

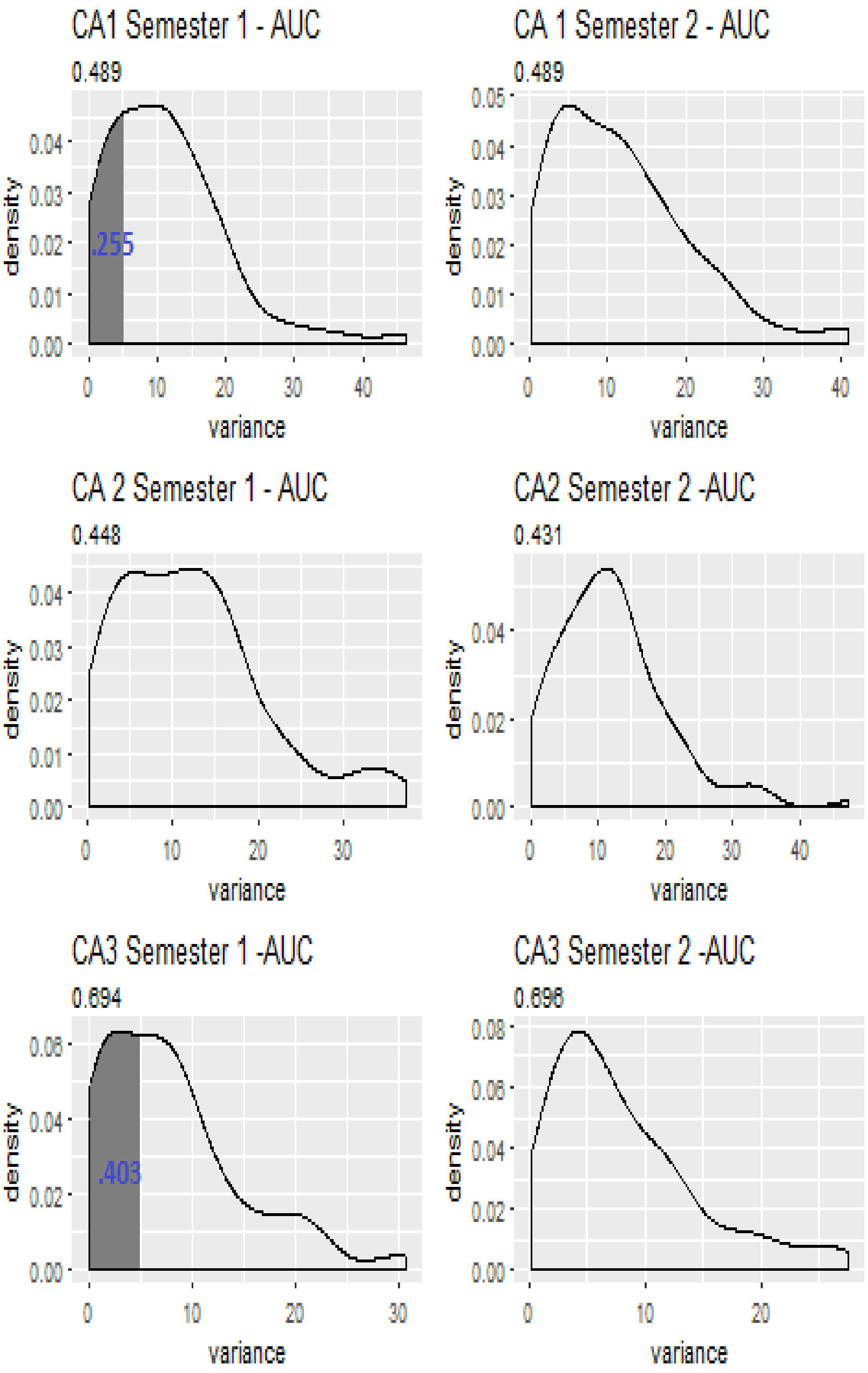


Figure 5.9: CA - Area Under the Curve - Variance from Group Mean.

Boundary	YR1		YR2		YR3	
	Sem'1	Sem'2	Sem'1	Sem'2	Sem'1	Sem'2
5	0.255	0.258	0.232	0.197	0.403	0.427
10	0.489	0.489	0.448	0.431	0.694	0.689
Ratio	52%	53%	52%	46%	58%	62%

Table 5.9: Area under the curve for boundaries of 10 and 5 marks.

5.4.3 CA Friendship Groups

The following experiment was carried out on the previously identified friendships (5.3.3) sub-set. This subset comprises of those students who had a longitudinal relationship with others and with whom the clustering algorithm has identified them with. This experiment examined how each student's precision marks varied within their individual groups. The objective of the experiment was to identify if the variance, over time, was either positive or negative (Converging or Diverging). The experiment examined the variance from year one and examined how those variances changed with each subsequent year. For example if the variance between two student precision marks reduced between year one and year two, it is considered to have converged in that period. Conversely if the variance widened, they will be deemed to have diverged. If the total number of hours for each category is calculated it can identify if the movement in that period is substantial or minimal. Minimal movements would indicate that there is only small changes in the precision marks indicating that they were initially close together or the marks obtained by the individual students remains consistent year on year. There was a substantial variance between the total convergence hours in both "Year 1 to Year 3", 208.39 marks. The average convergence figures are also relatively large at 10.97. Examination of the "Year 1 to Year 2" and "Year 2 to Year 3", infers that a large portion of the improvement occurred in the latter period. These indicators could be interpreted that the pairs that converged were further apart initially and the diverging pairs did not separate excessively.

As an illustrative example, a subset of the friendship pairs and their relationship is presented in Table 5.12. Examination of Pair "#1" from the table, illustrates the point that in year one the pair had a precision mark variance of 11.58, in year two, the variance reduced to 5.25 and in year three, the variance was 5.2. Therefore as the values decreased each year, the pair are considered to have converged, over time. In contrast pair "#6" are deem to have converged between years one and two, with variances of 6.41 and 0.34, but diverged between year two and year three, with values of 0.34 and 2.7.

Using these criterion, a summary of findings of the Converging and Diverging pairs (see Table 5.10) highlights that between year one and year three there were 17 Diverging pairs and 19 Converging pairs. While the number of pairs in each category are similar, the summed total number of hours associated with the converging pairs was almost double that of the diverging pairs. When summed the total hours was 110.79, averaging 6.52 hours. For the converging pairs, the average was 10.97. This indicates that when pairs are diverging there is less of a movement than if they were converging.

In a comparison between year one to two, there is a substantial difference in the number of diverging (25) and converging (12) pairs. As these results were at odds with the other periods a closer examination of these totals was warranted. The results identified eight diverging scores that consisted of less than two marks. A two mark variance could be considered as an indication of consistency between the pairs rather than divergence. These findings were not repeated in the subsequent period. The average diverging score was 5.52 marks. It was also noted that the Converging pairs had a total summed hours of 104.9 averaging with an average of 8.74, indicating that the precision marks that converged during this period had relatively similar starting precision marks.

	Count	Year	Avg	Count	Year	Avg	Count	Year	Avg
	1 - 3			1 - 2			2 - 3		
Divg	17	110.79	6.52	25	137.96	5.52	18	103.23	5.74
Conv	19	208.39	10.97	12	104.9	8.74	25	246.55	9.86

Table 5.10: CA: Summary of Converging and Diverging pairs

Follow on experiments examined in detail the hours by which specific pairs either converged and diverged from year to years. The following example in Table 5.11 is an extract from of seven pairs of friends.

Pair "#1", it is observed that in the period year one to year two, there was a convergence of the pair precision marks by 6.33, in the following period, there was a further convergence of .05 marks. This pair therefore converged continuously, to a total of 6.38 in the period one through to three. Pair“#4” in contrast, demonstrated a divergence between year one and two, but they converged between two and three.The overall effect between one and three was for the pair to converge by 1.27 marks. These findings indicate that the behaviour of a pair in a single period does not give a true reflection of their related behaviour over a longer period of time.

	Year 1	**Year 2**	**Year 3**	**Year 1 - 2**		**Year 2 - 3**		**Year 1 - 3**	
1	11.58	5.25	5.2	6.33	Conv	0.05	Conv	6.38	Conv
2	25.83	4.33	3	21.5	Conv	1.33	Conv	22.83	Conv
3	16.83	11.34	9.7	5.49	Conv	1.64	Conv	7.13	Conv
4	4.17	8.33	2.9	4.16	Divg	5.43	Conv	1.27	Conv
5	9.5	7.42	9.4	2.08	Conv	1.98	Divg	0.1	Conv
6	6.41	0.34	2.7	6.07	Conv	2.36	Divg	3.71	Conv
7	3.83	1.83	6.9	2	Conv	5.07	Divg	3.07	Divg

Table 5.11: Example of CA pairs variance year on year

5.4.4 Ranking within CA Cohort

The previous experiments had focused on the identification of groups of students and a precision marks comparison with their peers. The following experiment is an examination of how student positioning (Ranking) within their cohort changes when they are compared based on specific measures. Students were initially ranked based on the students CAO points value and subsequently on their precision mark. It was previously identified that not all students enter their programs based on their CAO points, but through other routes. Those students who access the programme through non-standard routes are classified by codes. Students who enter in this way are given a CAO code, which will have a value of 550 or greater and as as such will be ranked by this value. It was identified that all non-standard entrants would be placed in the deciles 7 and above which would also skew the overall results. It was therefore believed that the year one deciles will give the first real indication of the students position within the cohort, based on their precision mark. The focus of this experiment is the subset of students who completed the third year and received a precision mark. The majority of these students in this cohort had commenced their studies in the academic year 2014/2015 with 10 additional students joining in the second year. Student #24 (Table 5.12) is an indication of being a late joiner with the consequence that they are not easily clustered and identifying a trend in their decile ranking is also stunted.

Analysing the decile precision mark ranking of a students position, it was apparent that slight variations in precision marks can lead to a ranking which could be misinterpreted. For example from the sample Table 5.12 student number #5 had a mark of 61.42 in year one, placing them in "Decile #6". They had a slight improvement of 1.9 in the next year they are ranked as "decile #8", however a further improvement of 4.57 saw the students remaining in the 8th decile. To assist in the development of a system for comparing ranks from year to year, students were grouped into one of three categories based on their decile score. The categories are either **Top decile, 8 to 10**, **Middle, 4 to 7** and **Bottom, 1 to 3**.

Using the "Friendship" sub-set and applying this grouping system, 56% of the data-set had started in year one within the top group, labelled as "T", 38% in the Middle grouping (”M") and 7% in the bottom ("B") grouping. Due to the small numbers of students in the bottom groupings an analysis of the remainder of students who started in that grouping with some interesting findings. Approximately 66% of those who dropped out after first year, were ranked in the bottom 3 deciles in year one i.e. "B". A further 37 % who dropped out after the second year were also found in the bottom deciles of that year. This would indicate that those whose precision marks are at the lower end of the cohorts range from the first year have a higher possibility of dropping out in the first two years. The implications for this thesis is the role these students may have had when interacting with other students. Examination of the groupings these students were members of indicated that they had been part of various groups, i.e. there was no identifiable trends.

Num	Group CA1S1	Group CA1S2	Group CA2S1	Group CA2S2	Group CA3S1	Group CA3S2		CAO Points	Mark CA1	Mark CA2	Mark CA3		Decile CAO	Decile CA1	Decile CA2	Decile CA3	Rank (T-M-B)		
1	1	1	1	1	1	1		878	68.92	67.75	68.8		8	8	9	8	T	T	T
2	1	1	1	1	1	3		425	72.42	64	59.8		3	9	8	5	T	T	M
3	1	1	1	1	1	4		505	34.42	47.25	57.9		6	1	4	4	B	M	M
4	1	1	1	1	1	5		450	60.25	51.58	54.9		5	6	5	3	M	M	B
5	2	1	1	1	1	1		425	61.42	63.33	67.9		3	6	8	8	M	T	T
6	2	1	1	1	1	2		405	58.42	51.25	49.6		1	5	5	2	M	M	B
7	4	1	1	1	1	2		510	78.25	74.67	77.6		6	10	10	10	T	T	T
8	1	2	1	1	1	2		410	63	50	57		2	6	5	3	M	M	B
9	2	2	1	1	1	1		979	74.58	55.25	62.2		10	9	6	6	T	M	M
10	4	3	1	1	1	1		520	74.92	68.08	67.2		7	9	9	7	T	T	M
11	2	4	1	1	1	2		420	74.33	63.42	71.2		3	9	8	9	T	T	T
12	6	4	1	1	1	1		430	81.08	70.83	74.7		4	10	10	10	T	T	T
13	1	5	1	1	1	2		420	47.83	35.83	60.8		3	3	2	5	B	B	M
14	3	5	1	1	1	1		460	68.83	67.58	59.1		6	8	9	4	T	T	M
15	3	6	1	1	1	3		515	71	60.25	62.6		7	8	7	6	T	M	M
16	4	4	4	4	1	1		440	74.42	68.17	67.5		5	9	9	8	T	T	T
17	2	4	5	5	1	4		666	71.58	67.25	53.9		8	9	9	2	T	T	B
18	2	2	1	1	2	2		969	78	62.42	59.5		9	10	7	4	T	M	M
19	1	3	1	1	2	3		530	74.75	62.42	68.7		7	9	7	8	T	M	T
20	6	5	1	1	2	5		666	76.08	65.25	71.9		7	10	8	10	T	T	T
21	7	6	1	1	2	2		415	55	37.92	56.7		2	5	2	3	M	B	B
22	3	7	2	1	2	3		530	67.83	63.08	66.7		7	8	8	7	T	T	M
23	1	1	3	1	2	2		979	63.75	72	71.67		10	7	10	9	M	T	T
24	-	-	3	3	2	3		415	-	42.08	62		3	-	3	6		B	M
25	5	6	2	4	2	2		400	50.75	48.58	60.1		1	4	5	5	M	M	M
26	1	5	4	4	2	3		425	48.25	57.08	62.67		4	3	6	6	B	M	M
27	1	1	3	5	2	5		969	79.83	79.75	69.6		9	10	10	9	T	T	T

Table 5.12: Example of CA Summary of Student groups and rankings

5.5 Cluster Composition and Examination Marks of the EC Cohort

The following analysis will mirror much of the same experiments carried out in the previous section on the Computer Application (CA) cohort. Some comparisons may be made during the presentation of results, with the main comparison and conclusions being made in the following chapter. Figure 5.10 to Figure 5.12 are box plots that visually represent the groupings generated by the clustering algorithm. Each box representing one of the six semesters, from year one to three. Tables 5.15 and 5.16 summarise the graphical content and provide some context to the findings.

Similar to the CA cohort analysis, there were a number of experiments undertaken to identify the relationships between peers, precision marks, examine the clusters formed and an analysis to identify strong groupings. A set of guidelines had been previously defined what constituted a strong group i.e. a collection of students whose membership composition does not change for a minimum of two consecutive semesters spanning two separate academic years. It is preferable that these groups can be traced for a minimum of three consecutive semesters as the longer the relationship is in place the more reliable their behaviour traits can be determined. Another criterion used when considering the relationship between the EC and CA cohort is the third year programme structure. Similarly to the CA programme EC students do not attend the university but work in industry in the second semester. The same consideration has been given to the groups generated in this semester as the CA analysis, i.e. that groups formed in the second semester may not be an accurate reflection of the longitudinal friendships leading up to then.

Upon commencing the examination of the EC year one, semester one, it was observed that there were very few cohesive or strong groups developed in that semester. A granular examination of the semesters clusters could only identified a small number of clusters. For example in Table 5.13 we present cluster one and two of the twelve clusters formed in the semester. Records (students) # 1 to #12 are the members of cluster one and #13 to #22, are the members of cluster two. From this table it is seen that students # 1 and #2 are grouped in Semester 1, again in the second semester but they were not clustered again. The first identifiable student group that shown any sign of group strength, is the pair numbered #14 and #15. There grouping commenced in this first semester and continued through four consecutive semesters. Examination of the remaining ten clusters identified similar weak groupings, with minimal strong groups identified. Other observations taken from this table are the number of students who did not return in year two. Students #5 & #6 and #9 to #11 were shown signs of being part of a strong group in year one, but did not return in year two. The significance of the fact that these potential strong groups dropped out, was not pursued further as it was outside the scope of this project.

		EC 1		EC 2		EC 3	
	Username	**Sem’1**	**Sem’2**	**Sem’1**	**Sem’2**	**Sem’1**	**Sem’2**
1	**134041d965dffbv3**	1	1	1	1	1	2
2	**d42efa1ae87c9b97**	1	1	5	5	4	1
3	**53f4fe4a9d759074**	1	1	6	1	4	6
4	**6dd53e10d2d39748**	1	1	7	9	-	-
5	**9b1c28c65fa9cc7dd**	1	1	-	-	-	-
6	**beaae290d55eebw**	1	1	-	-	-	-
7	**7316e742c9947df064**	1	2	1	4	6	1
8	**f7fb4a52b3bc08ada7**	1	2	5	5	4	3
9	**0e03d7ec0d75eadew**	1	2	-	-	-	-
10	**91ed59444633548r7**	1	2	-	-	-	-
11	**f13947bb2ecfd41ae5g**	1	2	-	-	-	-
12	**33a98797cd1b5befda**	1	6	2	4	1	2
13	**82515a20d37a24bda**	**2**	**1**	**3**	**5**	-	-
14	**e8e6181ad45a046251**	**2**	**1**	**3**	**5**	1	1
15	**abc47010f81fbff71ed**	2	1	5	3	1	1
16	**edfd30e5bb2025cb53**	2	4	4	3	3	5
17	**4f9b3a9974e038ca5db**	2	4	6	4	3	7
18	**e6cd9e7c7b66ff8fa71**	2	6	2	8	5	1
19	**4c59e0d414556ecab4**	2	6	3	6	1	2
20	**fcd00bacc7327dc4sfds**	2	6	-	-	-	-
21	**d678ae5829848f4be5**	2	7	4	3	1	3
22	**f0381e3ba5f7f6ghdgf**	2	7	-	-	-	-

Table 5.13: EC Year 1 Sem’1 groups 1 and 2

Changing focus to the second semester of year one, an examination of the clusters formed did present groupings that conformed to the guidelines of *strong groupings*. Table 5.14 is an example of group of twelve students that have been clustered in the second semester. Student 1 and 2 were clustered together in this semester and remained in the same group for each semester through to the EC 3, semester two. Another pair in the grouping was students #4 and #5 who displayed strong friendship traits. This pair had been identified in the previous section as one of the few grouping formed in Semester one.

A granular examination of all groupings formed in both the first and second semesters demonstrated similar patterns, with semester two being the time of the creation of the strongest groupings. These results may be an indication that students from the EC cohort are slower to make strong friendships until the second semester in year one or later. Alternatively other reasons may be that a large section of the cohort remains as a homogeneous group through out the first semester and do not start to fragment into smaller distinctive groups until later. As previously identified the non-return students may have an impact on the groups they become members of before leaving the programme. Similarly to the CA analysis the identification of strong groupings and friendships will be determined through the analysis of the those groupings in semester one of year three, which can be traced back through earlier semesters.

		EC 1		**EC 2**		**EC 3**	
	Username	**Sem'1**	**Sem'2**	**Sem'1**	**Sem'2**	**Sem'1**	**Sem'2**
1	**134041d96531c985364e**	1	1	1	1	1	2
2	**65c3236efd2fc384426e**	6	1	1	1	1	2
3	**6ccbcb030e54c7bf1dba**	5	1	1	2	2	7
4	**e8e6181ad45a04625515**	2	1	3	5	1	1
5	**82515a20d37a24bd2f65**	2	1	3	5	-	-
6	**6273417eaa3b2cec1604**	3	1	3	5	-	-
7	**757e5e772c06bace680t**	5	1	3	6	1	1
8	**abc47010f81fbff71eb3**	2	1	5	3	1	1
9	**d42efa1ae87c9b97cf96**	1	1	5	5	4	1
10	**5b80c76df8134e496de**	5	1	5	6	3	5
11	**53f4fe4a9d759074c2**	1	1	6	1	4	6

Table 5.14: Example of a EC Cluster formed in Year 1 through to Year 3

Using the EC groups as presented in Figures 5.10 to 5.12, the graphics provide a representation of the grouping and their development over time. It is observed that the number of clusters in the first semester of year is twelve, and as presented previously it contains very few *strong* groups. In the first year first semester, the precision marks range is from from 24.50 to 70.33 with seven of the clusters having **tails** dropping below the 40 mark. In year three, the comparison year, this changed to a minimum of 38.67 and a max of 77.33 marks. In the CA analysis it was speculated that the minimum value improves as the non-performing students drop-out, these findings are presented in the following sections.

Comparing the cohort's median and mean values by year, as presented in Table5.15, it indicates that there are slight variances. The mean min values improved year on year from 45.6 to 49 .6 which is a slight variance of three marks. The max mean value increased from 60.75 to 62.2 a variance of 1.45 marks. The median's minimum increasing from 45.3 (EC1_1) to 51.9 (EC3_1) for the same period.

These early indications were that the precision marks for the EC cohort, do not vary significantly within any of the clusters leading to a stable set of figures. Follow on analysis was carried out on the Delta measures to determine any trends could be determined from those results.

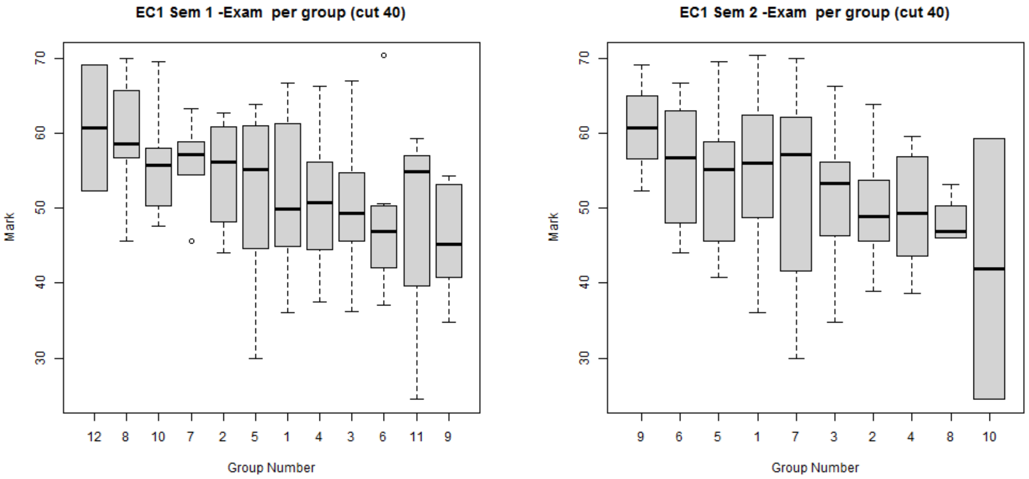


Figure 5.10: EC1 Sem'1 - Group mark range. -EC1 Seme'2 - Group mark range

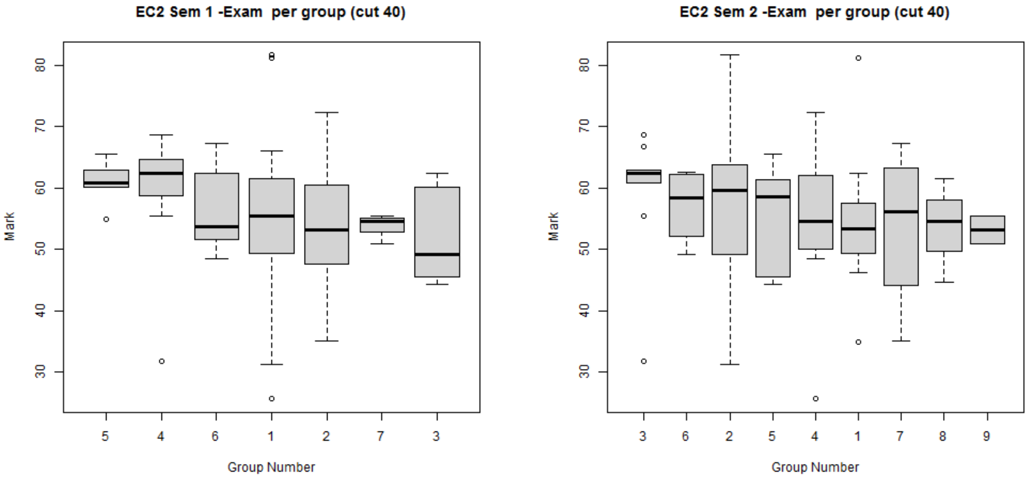


Figure 5.11: EC2 Sem'1 Group mark range. EC2 Sem'2 Group mark range

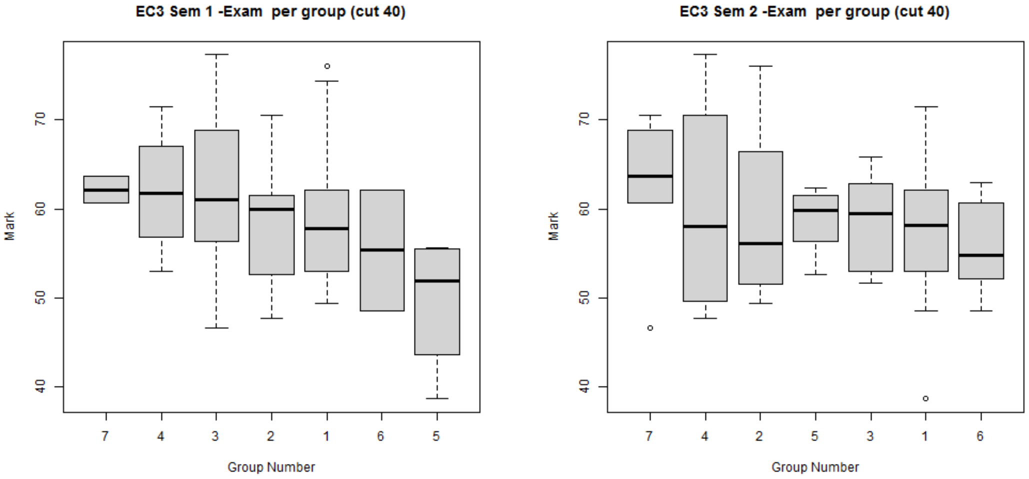


Figure 5.12: EC3 Sem'1 Group mark range. EC3 Sem'2 Group mark range

Mark	**Median**			**Mean**		
	Max	**Min**	**Var**	**Max**	**Min**	**Var**
EC1_1	60.8	45.3	15.5	60.75	45.6	15.15
EC1_2	60.7	41.9	18.8	60.75	41.9	18.8
EC2_1	62.3	49.0	13.3	60.9	52.5	8.4
EC2_2	62.3	53.2	9.1	59.3	53.2	6.1
EC3_1	62.2	51.9	10.3	62.2	49.6	12.6
EC3_2	63.7	55.0	8.7	62	55.6	6.4

Table 5.15: Semester Mark - Median and Mean measures per group

5.5.1 Group Precision Mark Delta Analysis

Graphs illustrating the average **deltas** between the members of each cluster per semester are presented in tables 5.13 to 5.15. A summary table of the outlining the movement of marks within their groups per semester is presented in Table 5.16. The *max mean* value is that value representing the student with the value furthest from their groups *average* precision mark. Conversely the *min value* is the precision mark closest to its groups average mark. There were two clusters, group 12 in EC1 Sem'1 and Group 6 in EC3 Sem'2 that are examples of a group with only two members, with each being equidistant from the mean value of their group, they are represented by single line, as there is no min or max values to be represented. In the majority of cases the groups are made up of three or more members, with each box plot illustrating the min and max values of their group.

Through out the period the minimum mean and median values are consistently low with the max values reducing over time. The max median value drops in EC1_1 to EC3_1, 13.1 to 6.8 respectively and max mean from 14.5 to 6.8 in the same period. These reductions would seem to be indicative of the group members marks converging between these years. This summation is tested through the plotting of the AUC calculation, as previously undertaken with the CA cohort.

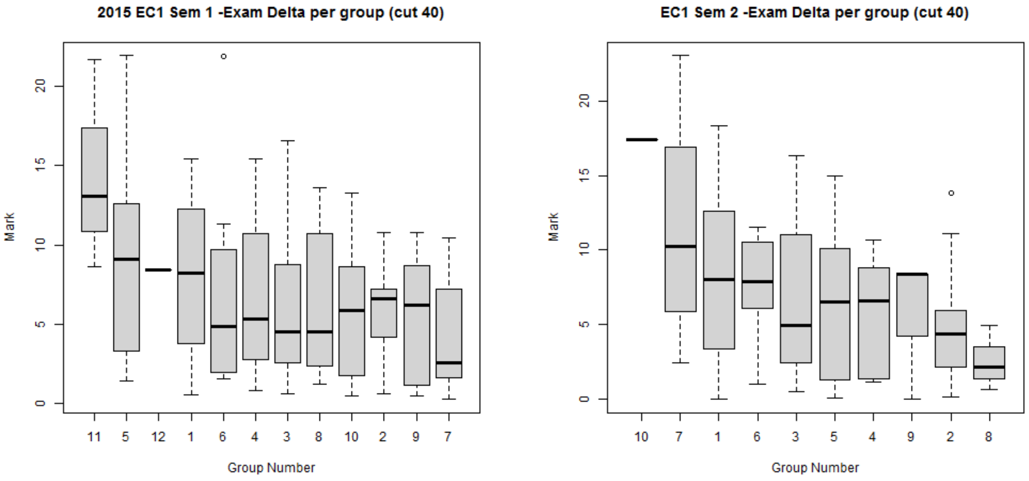


Figure 5.13: EC1 Sem'1 group delta. EC1 Sem'2 group delta

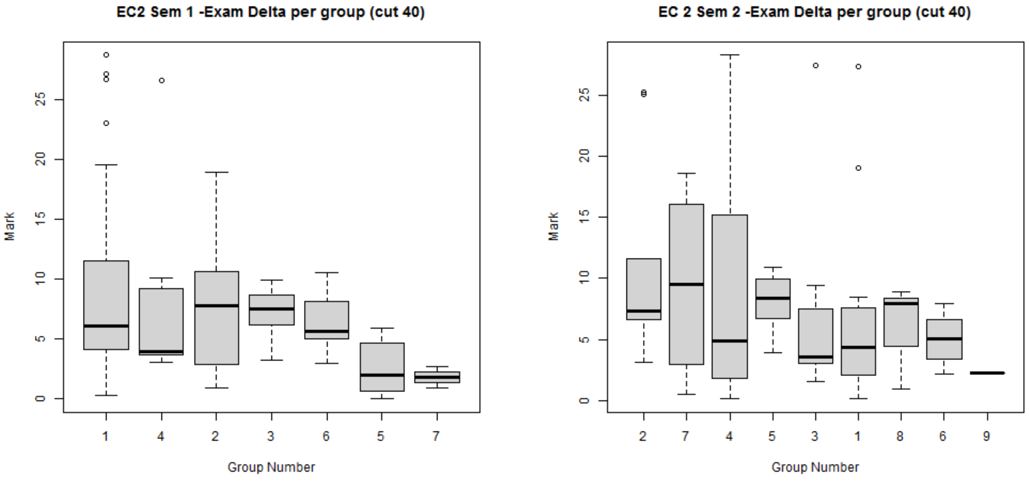


Figure 5.14: EC2 Sem'1 group delta. EC2 Sem'2 group delta

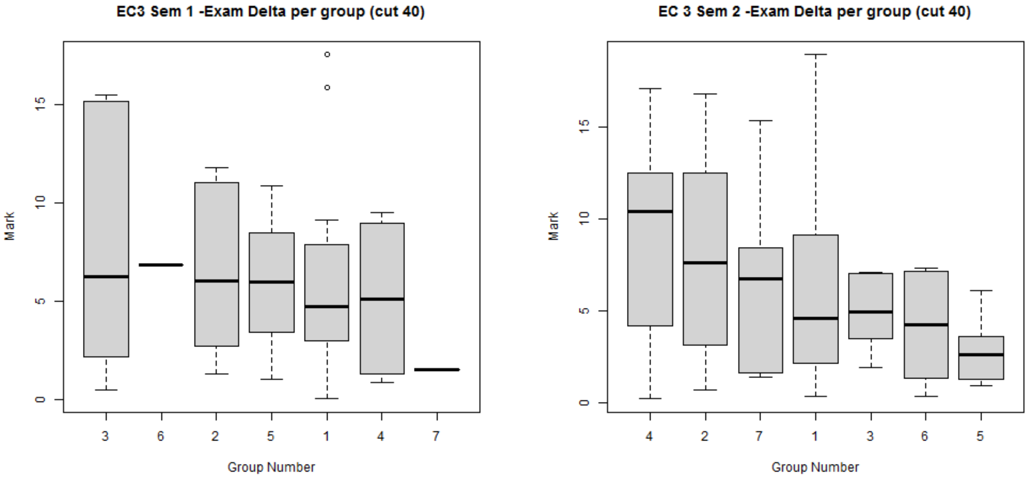


Figure 5.15: EC3 Sem'1 group deltas. EC3 Sem'2 group delta

Delta	Median			Mean		
	Max	**Min**	**Var**	**Max**	**Min**	**Var**
EC1_1	13.1	2.6	10.5	14.5	4.1	10.4
EC1_2	17.4	2.1	15.2	11.5	2.5	9
EC2_1	7.8	1.8	6.0	9.8	1.8	8
EC2_2	9.5	2.3	7.3	11.3	2.3	9
EC3_1	6.8	1.5	5.3	6.8	1.4	5.4
EC3_2	10.4	2.6	7.8	9.1	2.8	6.3

Table 5.16: Semester Delta - Median and Mean measures per group

5.5.2 Program EC - AUC

The EC delta marks have been plotted and presented in Figure 5.16 with an accompanying summary in Table 5.17. A generic AUC value represents the total value of the total area under the plot line. As with the corresponding experiment in the CA section, a upper boundary,(10 marks) was set and, the AUC calculated i.e.the sum of all delta values between "0" and "10". The AUC for year one, semester one, within this range, is 0.726. This approximates to 73% of deltas that are within 10 marks of their groups average mark. The AUC value increased through each semester to a max value of 0.818 (approx 82%) in year three semester one. This is an indication that the number of students who have a delta mark less than or equal to 10 marks is increasing over this period. These findings are consistent with the CA albeit they had started at a much lower point of 0.489 and increased to 0.694.

By adjusting the upper boundary to 5 marks, the AUC value of 0.406 (41%) in calculated for the baseline year and 0.473 (47%) in the referencing semester (EC3_1). Both these findings support the previous premise that the members of EC groups tend to have similar precision marks within their group.

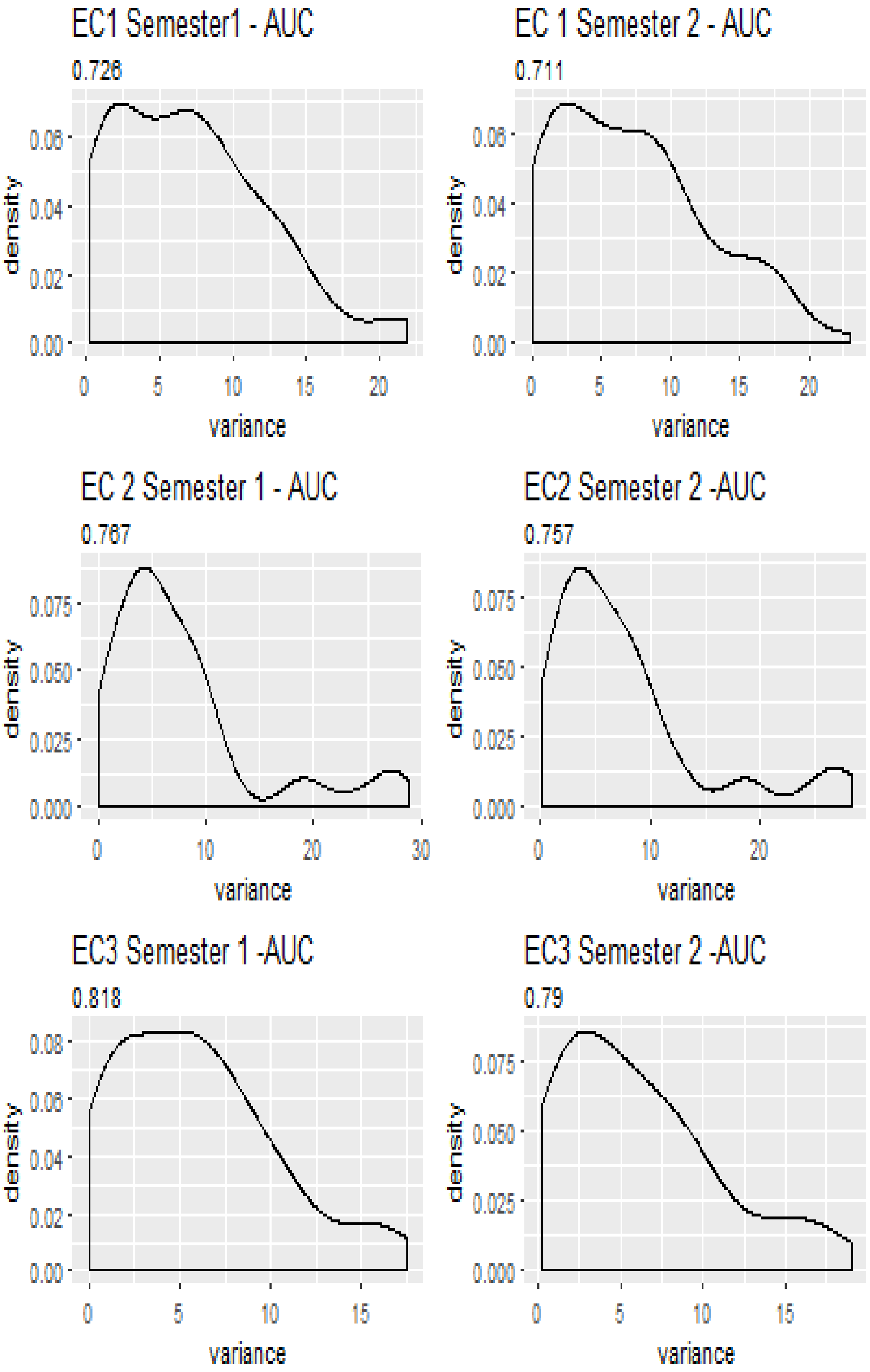


Figure 5.16: EC - Area Under the Curve - Variance from Group Mean.

Bounadry	YR1		YR2		YR3	
	Sem'1	Sem'2	Sem'1	Sem'2	Sem'1	Sem'2
5	0.406	0.415	0.426	0.431	0.473	0.482
10	0.726	0.711	0.767	0.757	0.818	0.790
Ratio	56%	58%	56%	57%	58%	61%

Table 5.17: Area under the curve for boundaries of 5 and 10 marks

5.5.3 EC Friendship Groups

This friendships (5.3.3) groups are a sub-set comprises of students previously identified through the clustering algorithm, as having a *friendship* with certain other students, with whom they have been clustered i.e. groups found within clusters. This experiment examined how these students precision marks varied within their groups. As with the previous CA analysis, the objective of the experiment was to identify positive or negative variations (Converging or Diverging) from year one and examined how those variances changed each subsequent year.

An example of the EC friendship pairs is presented in Table 5.18. Each row represents a pair with the variance between that pair presented per year (years 1 to 3). The students in pair #1 began with similar precision marks in year 1 with a 0.25 mark variance. The variance in marks expanded to 6.58 marks in year 2 and diverged further to 13.67 in year 3. The overall pattern for this pair between their first year and third year was a divergence of 13.42 marks. An examination of this pair's marks shows that as students one marks increased, student two marks decreased.

Pair #2 were similarly examined and it was found that students one's marks remained static (58.92 V 59) the others students marks decreased (60.75 V 49.83). The examination of all other groups did not identify any particular trend that could be attributed to the group. The indications are that within the EC cohort there is, over the three year period, a minimal variance in marks between students that make up a **friendship**.

The summary Table 5.19 identified that there are a relatively small number of strong pairs present in the cohort. The variation in the count from year 1 of 9 diverging and 6 converging pairs increased in year 2 as new students joined the cohort and formed strong

	Year 1	**Year 2**	**Year 3**		**Year 1 - 2**		**Year 2 - 3**		**Year 1 - 3**	
1	0.25	6.58	13.67		6.33	Divg	7.09	Divg	13.42	Divg
2	1.83	12.17	9.17		10.34	Divg	3	Conv	7.34	Divg
3	21.58	23.67	20.5		2.09	Divg	3.17	Conv	1.08	Conv
4	9.34	13.58	9.66		4.24	Divg	3.92	Conv	0.32	Divg
5	3.67	4.42	6.33		0.75	Divg	1.91	Divg	2.66	Divg
6	9.5	2	4.33		7.5	Conv	2.33	Divg	5.17	Conv

Table 5.18: Example of EC pairs variance year on year

	Count	**Year 1-2**	**Avg**	**Count**	**Year 2-3**	**Avg**	**Count**	**Year 1-3**	**Avg**
Divg	9	37.9	4.25	9	31.84	3.54	7	42.09	6.01
Conv	6	19.92	4.36	9	50.68	5.63	8	24.01	3.98

Table 5.19: EC Summary of Converging and Diverging pairs

enough friendships that they were identifiable in the third year. The average variance between student pairs between year 1 and year 3 is 6.01 for those diverging and 3.98 for the converging pairs. These indications point to a minimal movements between pairs, possibly indicating that they were initially close together or the marks obtained by the individual students remains consistent year on year.

5.5.4 Ranking within EC Cohort

The final experiment is an examination of a students ranking within the cohort and identifying any change in their ranking and the rankings of their group over time. In the Section 5.4.4 on ranking, it was laid out how the ranking process would be applied to a cohort. Using this process an examination of the EC cohort was carried out on the dataset which is demonstrated in Table 5.20. Additionally an examination of the students rankings and how they fit into a Top (T), Middle (M) or Bottom (B) category as per the CA of the overall rankings.

An examination of *strong friendships* was carried out to identify if there are any discernible trends amongst friends. What is present here are three example of how their precision marks varied by year and relative to their peers.

Firstly the group containing students 1 to 3 a strong group which can be traced from year 1 to year 3 for both students # 1 and # 2. Additionally student # 3 joined the group in year 2 and remained part of the group until year 3. Both student #1 and #2 precision marks increased from year 1 to year 2 and reduced again in year 3. The third year reduction was less than the previous improvement, therefore their marks were overall up from year 1 to 3. The third students, who joined the cohort in the second year, had an improvement from year 2 to year 3. From a cursory examination it seemed the group improved but an examination of their rankings indicates a different interpretation. Student 2 was ranked in the top decile from the commencement o the conclusion of the analysis period and therefore was placed as in the Top (T) categorised. Student 1, while their marks improved overall, relative to the cohort they dropped into the bottom category by the third year. Similarly student #3 was in the bottom category in both year 2 and 3.

Secondly, students numbered 7 to 9, a group that can be traced from the start of year 2 to year 3. Student 8 and 9 were also clustered in the second semester of year 3. An examination of the students categories, it can be seen that each of the students having started in the Top (T), dropped to either the Middle or Bottom by the third year.

Thirdly, students 14 and 15 who are similarly clustered from the start of year 2 to year 3. Both students improved their precision marks over the study period enough to maintain their position in the Top (T) category.

The examination of these groups would indicate that there are no distinctive trends amongst the groups. Although all the students in both the second and third groups were ranked in the top category in year one, the ultimate outcome for the students in year three can not be predicted.

What was determined was that students who are in the Bottom category after year one have a very high probability of dropping out before the third year. The majority of students who are in the Bottom group in the second year will remain there in the third year.

Num	Group EC1S1	Group EC1S2	Group EC2S1	Group EC2S2	Group EC3S1	Group EC3S2		CAO Points	Mark EC1	Mark EC2	Mark EC3		Decile CAO	Decile EC1	Decile EC2	Decile EC3	Rank (T-M-B)		
1	1	1	1	1	1	2		380	48.75	57.5	53.83		3	4	6	3	M	M	B
2	6	1	1	1	1	2		450	70.33	81.17	74.33		8	10	10	10	T	T	T
3	-	-	1	1	1	2		485	-	34.83	53.67		9	-	1	3		B	B
4	4	8	1	1	1	3		370	46.08	47.08	51.67		1	3	2	2	B	B	B
5	9	5	1	1	1	1		375	54.25	50.25	55.5		2	6	4	4	M	M	M
6	5	4	2	1	1	2		385	50.5	46.25	49.33		4	5	2	1	M	B	B
7	5	4	1	2	1	6		405	59.58	59.5	54.33		6	8	7	4	T	M	M
8	2	9	1	2	1	1		440	60.75	49.75	49.83		8	8	3	2	T	B	B
9	3	7	1	2	1	1		979	58.92	61.92	59		10	8	8	6	T	T	M
10	4	3	4	3	1	2		380	56.25	55.42	58.33		3	7	5	5	M	M	M
11	10	3	4	3	1	3		370	50.33	62.33	62.17		1	5	8	8	M	T	T
12	8	6	5	3	1	3		410	65.67	60.83	53		6	10	7	3	T	M	B
13	2	1	5	3	1	1		390	56.17	62.83	57.33		4	7	9	5	M	T	M
14	1	6	2	4	1	2		420	66.67	65.92	76		7	10	9	10	T	T	T
15	7	5	2	4	1	1		435	63.33	72.25	66.33		8	9	10	9	T	T	T
16	2	1	3	5	1	1		370	62.75	58.5	61.83		1	9	6	7	T	M	M
17	5	1	3	6	1	1		420	56	61.75	60.67		7	7	7	7	M	M	M
18	11	3	1	1	2	4		425	54.83	49.58	56		7	6	3	4	M	B	M
19	7	5	1	1	2	4		365	58.5	54	49.67		1	8	5	2	T	M	B
20	-	-	1	1	2	4		425	-	49.25	47.67		8	-	3	1	-	B	B
21	8	6	1	1	2	5		415	60.42	54.75	60		6	8	5	6	T	M	M
22	10	3	3	5	4	3		370	58	62.42	62.83		1	8	8	8	T	T	T
23	1	2	5	5	4	3		425	60.58	60.17	56.83		7	8	7	5	T	M	M
24	1	1	5	5	4	1		979	63.5	65.5	67		10	9	9	9	T	T	T
25	1	2	1	4	6	1		978	54.42	58.25	62.17		10	6	6	7	M	M	M
26	3	1	6	4	6	1		375	54.67	51.67	48.5		2	6	4	1	M	M	B
27	5	1	1	2	2	7		400	62.42	66	70.5		6	9	9	10	T	T	T

Table 5.20: Example of EC Summary of Student groups and rankings

5.5.5 Conclusion

Having proven that the research approach is viable in the previous chapter, this chapter took the concepts a stage further. There were two distinct parts to this chapter, the algorithmic development and the analysis of the results of the processed data. Having tested a number clustering techniques a set of experiments were carried out using a combination of Hierarchical clustering and "Dynamic Tree Cut" algorithm.

The two cohorts (CA & EC) were considered independently and tested as such. The decision to treat both coorts independently was vindicated with the early experiments. EC and CA cohorts produced different sets of results from the outset. EC analysis identified that there were not as many groups in year 1, semester 1 as CA had formed. The marks when compared using the AUC calculation identified that the EC group precision marks tended to be a good deal closer within their respective groups from year 1 through to year 3. This may be explained by the technical content of the topic in the CA programmes whereby the CA students had varying degrees of sucess based on their technical capabilities. Alternatively it may be the various types of students attracted to the different programs, this is an area for future examination.

A common trend with both cohorts is that students who finish their first year ranked in the bottom 3 deciles have a much higher probability of dropping out of the course completely by year 3. The precision mark average per group which is used as an indicator, found little movement in the EC group while for the CA group this was not as obvious.

A complete set of findings and conclusions will be presented in the next Chapter.

Chapter 6

Findings and Conclusions

6.1 Findings

Upon the commencement of this thesis journey a number of questions had been laid out to frame the domain of research we wished to address. Framing of the questions led to the setting out of our hypothesis, and the journey began. This research was bounded in the realms of Learning Analytics and the collection, measurement, analysis and reporting about undergraduate University learners in an educational context, for the purpose of understanding and optimising their environment. Romero [50] considered the educational environment as a boundless source of exploitable knowledge. While our research sits in the Learning Analytics (LA) domain and amongst a multiple other similar research projects, we believe our approach to be unique.

In the introduction chapter we presented our motivation for wishing to undertake the research we investigate. Having framed our hypothesis we presented several research questions, that would be the bases of the thesis. To summarise, the research questions were:

1. **Research Question 1:** Can we identify a student's activities from the data held in WiFi access logs?

2. **Research Question 2:** Is it possible to identify student friendships among student pairs and larger groups, through the analysis of WiFi logs.

3. **Research Question 3:** Is there evidence from the analysis of WiFi logs of peer influence within student groups and more specifically, exam performance?

A number of supplementary questions that arose during the early research stages included the following:

- Can we identify student groups such as the formally constituted groups like class attendees as well as informal social groupings?
- Can we infer the activities of students from their locations?
- Does the make-up of a group of students influence the academic performance of the students in the group?
- Can students who are isolated and not engaging in groups with others be identified early in a semester, especially in the case of first year undergraduates ?
- Can we profile those students who are more likely to drop out of University, early in the semester, based on their group participation and perhaps intervene to support them?

We believe that following extensive research and a comprehensive set of experiments that were completed, allowed us to present answers to these questions. We restate here our hypothesis and we will summarise how we arrived at the answers that supported this hypothesis;

> "That we can use students' digital footprints, especially those indicating their physical locations, which yields unbiased data, to identify academic collaborations and social friendships and from that we can quantify peer influences on exam performance at third level education"

We believe that successfully researching and answering these question will contribute to the large body of work already in the domain of Learning Analytics (LA). By employing a unique approach to a well researched arena, we believe our contribution will stimulate a new conversation in this domain. One of the uniquenesses of this thesis' approach is the separation between data collection and the subjects providing the data. There can often be a contention in research that data collection methods can be open to bias if the subjects are aware that they are part of an experimental environment. We believe our methodologies for the collection of data negate many of these biases while accepting that bias can be introduced by virtue of the selection process of research cohort. In the opening paragraph of this thesis we theorised the uniqueness of each student entering third level education. In Section 4.1.3 we presented the different CAO achievements of various student groups entering a number of different DCU academic programmes. We therefore have accepted that by choosing a particular set of programmes (CA & EC) will introduce a particular bias to our research that must be considered when scaling the methodologies into other academic schools or different third level institutions.

Challenges

Taking the unconventional and unique approaches can also present challenges that had not heretofore been encountered and had to be overcome. The identification of the data requirements and sources did not guarantee the availability or robustness of the data. With the commencement of this particular new research approach came the challenge of identifying the sources of data that may never have been sourced before. The collection and storage process established as part of the system design my not be conducive to easily accessing the data, but additionally the archived data may be purged periodically as part of the University's normal operating procedures for capturing, storing, and deleting this data. It was the intention of this research to delve into the archives in an effort to mine a minimum of one complete academic programme period i.e. four years. However this was not possible as a lesser amount of data was available. For this reason the first set of complete data was only available from the academic year 2014/2105. Similar issues arose when efforts where made to obtain academic timetable data for the specific period and programmes being measured. It was necessary to approach a number of sources to collect various data formats and build a complete robust academic timetable.

Findings

In the background chapter we identified that the use of technology in the role of identifying peoples' location and placing context on those locations is not new, but the methods we employed had not been used previously. In undertaking this research there were a number of assumptions made about the Campus WiFi infrastructure. It was necessary to assume that in the majority of cases that student who were co-locating were connected to the same Network Access Point (NAS). While we believe this assumption to be accurate, there may be situations where a member of a co-locating group connected to a NAS before entering the shared location while the other members only accessed a NAS only after entering the shared space. While accepting there may be instances where co-locating groups are connected to different NAS, this occurs in a small minority of cases.

In that chapter we presented a review of the various datasets employed in this thesis. To answer the **Research Questions** posed required the ability to interpret the WiFi logs collected from the Eduroam system. The interpretation of the logs required the development of a bespoke applications that could manage a very large set of data plus to clean and format that data into a usable form. Other datasets required included academic timetables, NAS identification and location databases, student demographic and student academic results datasets.

A comprehensive set of experiments were carried out to determine if it is possible at a macro level to determine the activity of students within the bounds of a University campus. We firstly examined yearly, semester, monthly and daily activities to determine the completeness of the data. The results indicated that the expected activity trends were present, such as activity growth over the semester with peaks of activity before and after scheduled class times and variances in activity which mirrored class timetables. Using a users unique username and the unique mac address of their wifi enabled devices, we mined the wifi logs of a number of test subjects by time and location. These tests confirmed that the WiFi log data was complete and robust and could positively answer **Research Question 1**, i.e. **"it is possible to identify a student's activity by their WiFi logs"**.

With the processes confirmed as effective on the test cohort, the analysis changed to an examination of the primary research cohort i.e. undergraduate students of two degree programmes namely Computing Applications (CA) and Enterprise Computing (EC). Testing of the activities of both groups was carried out to determine if they could be analysed as one homogeneous group or should be treated independently. Analysing the activity and specifically their *meetings* determined that they act in a different manner to each other.

Following further informal discussions with members from the programmes that made up the research cohort, it was determined that students did not consider that all locations were considered the same by all students. There was a realisation that students shared different routines with their friends and classmates. To gain a full understanding of the differences, we held a focus group and followed up with a survey of the complete cohort of CA and EC students.

Having answered research question one we commenced our determination to investigate our ability to identify groups of co-locating students and to distinguish between random meetings and co-ordinated collection of students, i.e. *friendships*. Analysing our cohort of students and correlating their activity with their programmes' academic calender, the groups were identifiable by the density of students at expected location at specific times. We therefore consider that *Research Question 2* regarding the identification of formally constituted student groups has been answered in the affirmative. We showed that context can be applied to student co-location i.e. when there is meeting of student groups, depending on the location type, context can be inferred. Location types were classified in general as Academic or Social and more specifically by activity e.g. the library, the restaurant or the classrooms or labs.

Other supplementary questions that were addressed during the research included the ability to identify the make-up of a group of students. To confidently answer this question it would have been necessary to obtain a greater level of additional detail about each student such as:

1. Midweek residence i.e home or away from home

2. Part time work requirements (financial situation)

3. Club and societies membership and activity

4. Family academic history

As this information was not available, we were unable to pursue this course of inquiry any further.

A question that was addressed during the research was whether it was possible to identify students who do not engage with others early in a semester, especially in the case of first year undergraduates, and who potentially have a greater chance of dropping out of the course altogether ? By analysing WiFi logs it was possible to identify students with minimal activity and it was shown it is these students who are more likely to drop out. There were a number of other indicators of students who were more likely to drop out, but these required an analysis of their end of year exam results.

Peer influence

We carried out an examination of peer influence through the analysis of student examination results of those who have been classified as friends. Friendships were positively identified through the use of Hierarchical Clustering algorithms that identified the clusters within which there were strong groups. This hierarchical model used a distance matrix generated through a combination of meeting occurring at weighted locations and within categorised time-frames at the University. Once groupings and friendships were identified a number of experiments were employed to identify the presence of strong friendships and hence peer influence in student exam results. These experiments included:

1. Precision mark, range per group, per semester.

2. Cohort mean and median analysis.

3. Group delta mark trends per group

4. Area Under the Curve, per cohort per semester

5. Ranking by student per semester, relative to their group.

6. Exam mark Divergence and Convergence per semester per group

The strong friendships which were identified demonstrated no apparent correlation between changes in individual precision marks relative to their friends. We could not identify any trend whereby one category or grouping of students performed in a similar manner. Based on the results of these experiments the exam performance of one or more students could NOT be predicted by the type of student they were friends with.

Experiments (AUC experiment) did indicate that individuals' exam marks do narrow in relation to their group average mark. However indications are that as the less capable students drop out, removing the low outlying marks does improve the group average marks.

An examination of this narrowing of group average marks identified a distinct difference between academic program cohorts. It is therefore acknowledged that Peer effects within different academic cohorts should be considered in any future similar research.

6.2 Conclusions

We determined that using the student digital footprint available to this research, we can not identify with any certainty, that peer influence on exam performance appears in our research cohort. We have a number of suppositions that can explain this.

- The number of strong friendships identified within the programmes is relatively small . . . we identified only thirteen in the EC and sixteen in the CA programmes. These quantities are not considered large enough to state definitive findings

- Due to the size of the main DCU campus there are only a limited number of locations for students to congregate for socialising when not in class, library or labs. This gives rise to a high probability of students having random co-location meetings with non-friends and this clouds whatever data might be present to indicate genuine friendships re-enforced through co-location at social locations on campus.

- The DCU campus has universal coverage of WiFi through Eduroam, and in order to achieve this the NAS base stations are powerful and each has a good geographic range. This means that there is not the density of NAS stations on the DCU camps and the precision needed to locate individual WiFi devices, within the larger locations such as the restaurant, that would allow for differentiation of groups of students visiting the same location at the same time.

- The WiFi infrastructure design can in certain circumstances allow students with mobile devices who are in close proximity to each other, be connected to different NAS stations. These NAS could be either within the same room or near by locations. Consequently there may also be others from the same peer group be connected to a NAS station but are not part of the same group.
- There are indications that group composition is continuously evolving as some students drop out and other students join the cohort each year. This raises difficulties in determining the true friendships

6.3 Future Research and Recommendations

If we were to give due consideration to extending and building on this research and to identify peer influence there are a number of methodologies that could be considered:

1. Use traditional research methods such as question and/or observation techniques.
2. Use technology such as mobile applications on student phones to measure activity.

Each of these options were discussed earlier in this thesis as methods whereby data collection bias could be introduced, as the subjects are aware they are part of a research environment.

Alternative options would include:

3. Re-run this research at a larger facility where students have a greater diversity of locations to congregate with friends and avoid the random meetings and where there are a larger number of NAS base stations so as to give more precise student location from their WiFi-enabled devices.

4. Use the present Eduroam infrastructure to collect and record the *probe requests* from WiFi enabled devices. These requests are a consistent stream of communications between the network and mobile devices checking for the nearest and or strongest gateway to the network. By measuring signal strengths and triangulating probed NASs, accurate coordinates can be calculated for each device. Analysis of these coordinates can provide a level of accuracy to ensure accurate identification of collocated devices.

5. Tools such as those developed by Kitto [3] for data extraction from social media can provide the ability (with approval) for the analysis of social media activity and identification of friendships.

6. Fire [23] Used the Implicit (timestamp) and Explicit (IP address) of student who upload assignments to identify co-operation between students and potential friendships.

7. Alternatively with the advent of new wireless technologies and in particular the development of 5G, more precise location-determination may become available.

Dublin City University sits on a 43 acre site, which in comparison to European and USA campuses is relatively small. Having the opportunity to repeat this research on a large campus would reduce the probability of random meetings and increase the chances of identifying a larger number of strong friendships. Using the same processes as those applied in this thesis, there is a greater probability of being able to predict if peer influence does occur at third level educational institutions.

Appendices

Appendix A

Focus Group Questionnaire

These are the questions that were used in the discussions with the focus group.

What program are you in? CA/EC

What Year are you attending? 1/2/3/4

Do you tend to attend college every day? Y/N

If N, can you comment why?

Do you prefer to arrive at the same time everyday or is it dependent on timetable? *Comment?*

What is a typical arrival time? 8,9,10 – 1 hour before – 1st class or for 1st class

When on Campus: Where do you spend the majority of your time?

Excluding Formal Classes, where do you spend the majority of your time?

Do you prefer to study: Alone, in a group

Where do you prefer to study:

Library, Lab, Classroom, Other. *Comment?*

Do you attend the Gym/Pool? If Yes, alone or with friends? Typical time: 8-17:00 or after 17:00 *Comment?*

Do you leave the campus after the last class of the day? *Comment?*

Do you have the majority of your meals at the restaurant or from the Shop? Restaurant, Shop.

If you bring your lunch or purchase from a shop where do you eat your food? Lab, Class, NuBar

Other /Comment

Location Ranking Questionnaire

This is the questionnaire that was distributed to participants.

Rank location 1 12:
Rate:

Location	**Example**	**Core hour**	**Non-core**	**Comment**
Social	Caf/Bar/Sport			
Classrooms	All			
Labs	L101-L25 etc			
Library	All floors not Caf			
Residence	All			
Transit	Public areas			
Core:8:00 to 17:00				
Non Core:00:00 to 7:59 + 17:01 to 24:00				

Table 1: List of DCU location - Ranking by time

Appendix B

Survey questionnaire

Based on the discussions and findings of the Focus Group 6.3 the following survey was prepared for distribution to all students in our cohort.

The survey comprised of 10 questions designed to harvest information on the individuals behaviour while on campus.

Survey questions

1. Program CA or EC
   (a) Computer Application:
   (b) Enterprise Computing:
2. Gender
   (a) Female:
   (b) Male:
3. Where do you live during the week?
   (a) Home:
   (b) DCU Campus residence:
   (c) Other Rental:
   (d) Other:
4. Do you attend DCU on days you have no formal classes?
   (a) Yes:
   (b) No:
   (c) Comments:
5. Please rank where you spend most time OUTSIDE of formal classes? 1-7

   (a) Canteen

   (b) NuBar

   (c) Classroom

   (d) Labs

   (e) Library

   (f) Sports complex

   (g) Residence apartments

6. Do you remain on campus after your last class of the day?

   (a) No, I leave after the last class of the day.

   (b) I remain on the campus to study

   (c) I remain on campus to hang out with my friends

   (d) I remain on campus to go to the Gym/Pools

   (e) Additional comments:

7. Do you attend the Gym/Pool?

   (a) No.

   (b) Yes - Alone

   (c) Yes - With Friends

   (d) Yes - As part of a team

8. Do you bring / buy lunch on campus?

   (a) Bring and eat in Canteen/NuBar

   (b) Bring and eat in Lab/Class

   (c) Buy onsite and eat in Canteen/NuBar

   (d) Buy onsite and eat in Lab/Class

   (e) Other (please specify)

9. Your trip to DCU?

   (a) Live on campus

   (b) Local, short walk/cycle

   (c) Bus

   (d) Drive

   (e) Other comments

10. Do you study on campus

   (a) Library alone

   (b) Library with friends

   (c) Labs alone

   (d) Labs with friends

   (e) Other comments

Survey Results

The following table outlines the results of question 5 which ranked the locations where students went outside the official class times.

The higher the number, the more likely people are going to attend that location with a friend. Therefore the lower the score, the more likely people who are there together are friends. Therefore they should be given a lower score (distance matrix, smaller scores mean closer relationship)

Breaking down the survey questioners by program we developed a profile of the different program student cohorts 3.2.5 activity patterns.

The following table outlines the average ranking of each location. Although there is a number of difference such as location of living during the week to the location at which they eat their lunch Lab for CA and Canteen/Nubar for EC students.

Quiz results for the EC program:

Q1	CA	Respondents	23		
Q2	M	F			
	19	4			23 CA students 19 Male and 4 female , how is this relative to the demographic breakdown of the course
Q3	Home	Rent			
	12	11			There is an almost 50/50 breakdown of students who rent or live at home and similarly a split within walking distance or travel by bus.
Q4	N	Y	some		
	14	8	1		61 % of students do not attend the college on the days they do not have formal classes. For this program there is no classes on a Friday. The focus group confirm that this is true and as the semester progresses toward the examinations, students tend to spend more time on campus.
Q5	See Table				
Q6	N	Y	Y gym	Y friends	
	9	9	2	3	60% remain on the campus after their last class of the day, with the balance (40%) leaving at the end of the last class of the day. Of the 60% who remain on site the main reason presented was to study, go to the gym or meet friends
Q7	N	Y friends	Y team	Y alone	
	15	5	1	2	The Gym question was to understand the number of students who attended the college gym as either alone or with friends individually or as a memebr of a team. The focus group identified the gym as a location where close friends can share time.
Q8	bring	buy	both		
	12	9	2		61% o students will bring in their lunch with the balance purchasing their lunch onsite. CA students in the majority of cases reported they consumed their lunch in the computer labs.
Q9	Walk	Drive	Bus		
	9	2	12		All drivers leave after the last class, 4 who walk and 4 who bus.
Q10	Labs friends	Labs alone	Lib-Alone		
	16	5	2		70 % of student have reported they studying with friends I the Labs and 30% studying alone with 22% doing so in the labs and the remainder studying in the Library

Q1	EC	Respondents	20		
Q2	M	F			
	15	5			Need to check the demographic profile of the program
Q3	Home	Rent			
	13	7			65% are resident at home while the remaining 35% rent 0f which 70% locally and 30% within a bus journey.
Q4	N	Y	some		
	11	6	3		75% of EC students do not go to the gym, with the majority of those who do go alone
Q5	see sheet				
Q6	N	Y	Y study	Y societies	
	9	7	3	1	
Q7	N	Y Friend	Y alone		
	14	2	3		75% of EC students do not go to the gym, with the majoity of those who do go alone
Q8	bring	buy	both		
	6	12	1		A majority buy their lunch and eat it in either the canteen or Nubar. EC students reported to eating their lunch in the Nubar or Resturants
Q9	Walk	Drive	bus		
	10	4	5		A large proportion of those who get the bus (4) leave after class.
Q10	Labs friends	Lib alone	lab alone	Lib friend	
	11	5	2	1	

Table 3: EC Questioner results

Location	Sub-Category	Computer Applications	Rank	Enterprise Computing	Rank
Canteen	Social	3.2	6	2.5	6
Nubar	Social	3.8	4	3.2	5
Class	Academic	5.0	2	4.9	2
Labs	Academic	2.1	7	2.4	7
Library	Academic	3.4	5	4.2	4
Sport	Social	4.5	3	4.3	3
Residence	Social	5.4	1	5.9	1

Table 4: Question 5 results

Appendix C

Drop-outs

By the time a student has commenced third level education they have completed thirteen or fourteen years at primary and secondary level education with the final three of those years focused primarily on attaining the necessary points to gain access to their desired third level program. For many the progression to third level is a positive and exciting experience that is looked forward too. For others it can be a challenging and stressful experience, as they adjust to new social and academic norms. For some, it could be the first time living away from home which includes adapting to new environments, learning to survive independently whilst at the same time coping with the academic demands of third-level. Some students may also struggle to adapt to a different teaching and learning format. The multiplicity of factors effecting a student transition in an Irish sense were examined by the National Forum for the Enhancement of Teaching & Learning. [18] which was based on a survey of 1580 students at four Irish institutes. Analysis focused on student experiences during their transitions including all influencing factors. Focus groups also looked at the challenges faced, duration of the transition and any impacts on academic performance. They summarised that older students as well as those commuting struggled the greatest with the transition. Student commuting longer distances struggling more than those living closer.

In summary student may drop out from their chosen program for multiple reasons:

- Program change - moving to another program within the same institution.
- Change institution - continue third level education at a separate institution
- Deferred to a later date - drop out with the intention of returning at a later date
- Leave 3rd level education altogether

Any one of these events could have been initiated for many reasons:

- Program incompatibility

- Finance or personal reasons
- Academically struggling
- Move to full time employment

Drop-out 2015

Preliminary analysis commenced with the 2015 Retirees, that is, the first year students of our CA and EC cohorts. Retirees are identified as the students who at the end of their first academic year, had a precision mark of less that one. A precision mark is a weighted average of all the modules examination marks for the academic year. Based on this criteria there were four students who we deemed to have retired during the year. These students are of interest as it is necessary to be able to determine the impact, if any, of their actions on the remainder of the cohort. A determination will be required as to include or exclude them from the analysis. Of the four students there was an equal division of retirees from each program. The WiFi logs were examined for the engagement activity of the students for an academic year, and the analysis showed a limited level of activity in the first semester, September to December. Two student, one each from EC and CA did have a greater activity in the second semester between April and May, however this activity was minimal, approximately 400 and 120 log-ins respectively. When compared against the average activity per student in the CA cohort of 4,577 and 3,800 in the EC program. It can bee seen from Figure 1 that of the four students only three recorded any activity in the WiFi logs. The fourth can be assumed to have registered but never commenced the program. While they may have converted to another program we can assume that if they did, it was not at DCU as there is no activity on Eduraom.

The activity of student 01a3885268f31717c55079b54523da62 was examined in greater depth as they had recorded WiFi activity in both semesters. From Figure 2 it can be seen that the greatest level of activity was during the month of September. The activity dropped dramatically over the following months before recovering during the second semester.

A granular examination of the wifi activity as illustrated in Figure 3 showing the top

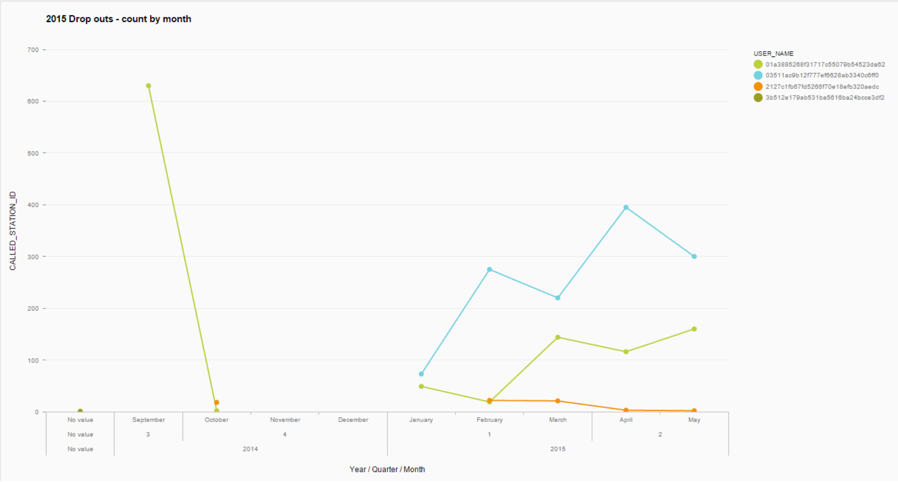
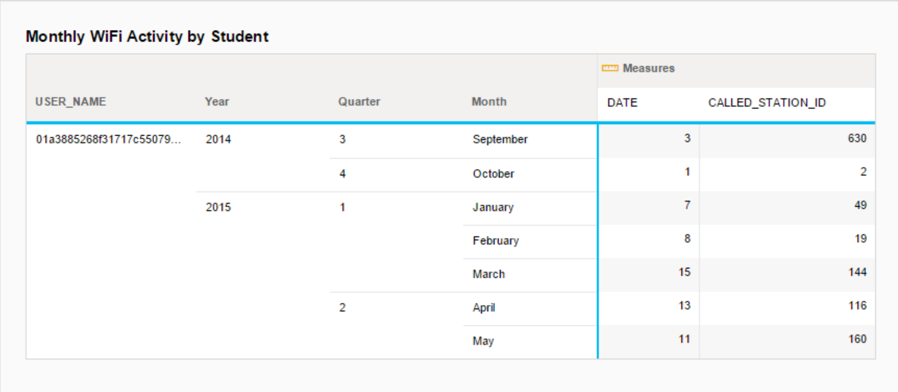


Figure 1: Monthly WiFi Activity - Retirees

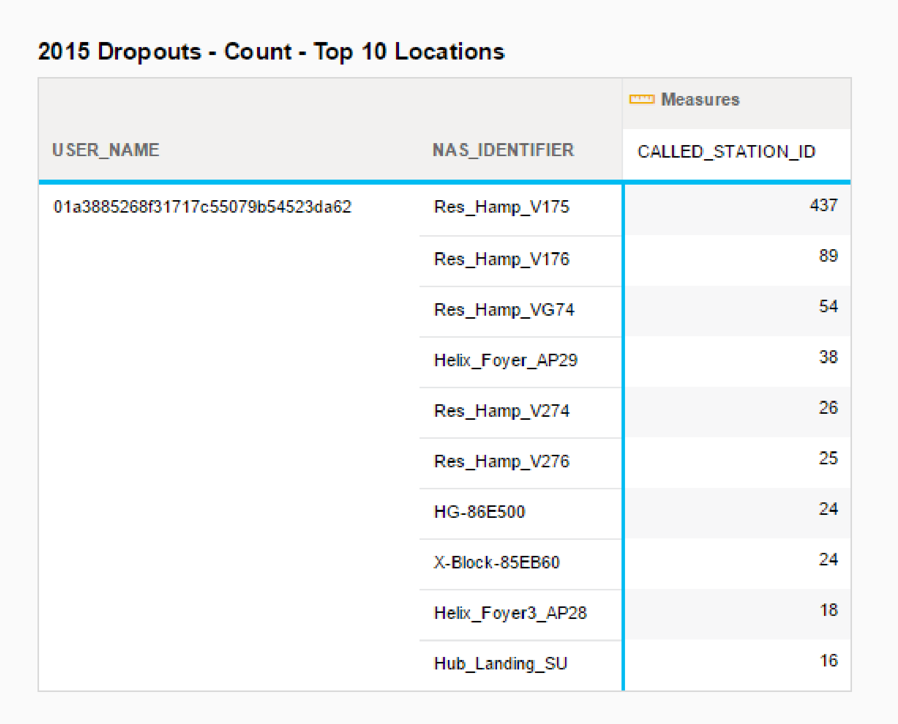

Monthly WiFi Activity by Student

				Measures	
USER_NAME	Year	Quarter	Month	DATE	CALLED_STATION_ID
01a3885268f31717c55079...	2014	3	September	3	630
		4	October	1	2
	2015	1	January	7	49
			February	8	19
			March	15	144
		2	April	13	116
			May	11	160

Figure 2: Monthly WiFi Activity

ten locations visited by the student identified as 01a3885268f31717c55079b54523da62 was carried out. The results indicated that this student spent little time in academic areas but the majority of their time on campus in the residential rooms, eg: Room 175 in the Hampton Block (Res_Hamp_V175).

As the research progressed, and the development of methodologies for the identification of Peer groups continued, it became apparent that it was necessary to determine how to deal with these **retirees**. As the level of activity is low, if they remain as part of the analysis, the Average and Delta calculations for the total cohort will be influenced. There may also be some influence on the clustering model employed. The decision therefore was to exclude those that retired during the year from further analysis.



2015 Dropouts - Count - Top 10 Locations

		Measures
USER_NAME	NAS_IDENTIFIER	CALLED_STATION_ID
01a3885268f31717c55079b54523da62	Res_Hamp_V175	437
	Res_Hamp_V176	89
	Res_Hamp_VG74	54
	Helix_Foyer_AP29	38
	Res_Hamp_V274	26
	Res_Hamp_V276	25
	HG-86E500	24
	X-Block-85EB60	24
	Helix_Foyer3_AP28	18
	Hub_Landing_SU	16

Figure 3: Monthly WiFi Activity

Retirees 2016

Similar analysis was carried out on the remaining cohort of students who completed the 2015 year i.e. all students who at the end of the academic year received a precision score greater than 1. Of the remaining students per program i.e. 111 in CA and 70 in EC, not all students returned for the following (2016) academic year. Using the same criteria as in our research of the 2015 retirees 6.3 the data was filtered for those students with 2016 precision mark of less than one (¡1). This filter will produce a list of those who did not complete the 2016 examinations. We identified 19 students from the Computer Applications (CA) and 28 from the Enterprise Computing (EC) who had a Precision mark less than 1.

Table 5 provides a brief synopsis of a number of CA students who completed the 2015 academic year but subsequently retired i.e did not complete the academic year through to the completion of 2016 exams. The table is sub divided into the students who entered via the standard route of the CAO as outlined in section 3.3 and listed in Table 3.9 In the case of those who entered via the Non-CAO route we can see that there were two students who performed the best in 2015 had a relatively high activity with 3887 and 3465 wifi log entries while the remaining five students had a very low level of activity. The average activity for CA students for the year was approximately 4,500 unique log entries. Figure 4 illustrates the point that the student, name commencing with a79f7c7d', attended the campus consistently through out the year. Examination of their log activity reveals that the majority of their time on campus was spent in the School of Computing computer labs. Student bfc88ea', with a slightly lower activity count in contrast spent the majority of their time on campus during the months of September and October. This student did not attend the college for the remainder of the academic year. Closer analysis of this students WiFi logs indicates that they spent the majority of their time between the on-campus residence and the Student Union complex.

Other observations included that 13 of the 19 students who retired had passed the previous years exams with the another of the remainder six failing by the narrowest of margins and the final five failing by a considerable margin

CA

Student	CAO Points	Precision'15	Domicile	WiFI Activity	Result
a79f7c7db6e3c99e6706118df414551f	999	73.17	08	3887	P
eb74802bc5753d514be49de6af8a70fc	979	38.42	02	491	F
09951fc006e51d8f8d5f9ada03be9835	979	48	04	6	P
845d4aa9ec355ace268023dbd458c70b	978	12.25	05	286	F
bfc88ea9d3bc5b3d04599edba9bd43ad	976	66.33	08	3465	P
6a4f447475b7da05814b3b94568541ee	878	58.92	01	11	P
a1eb95761a955dfa9c830d6670d94b9e	878	19.5	60	67	F
cf6a6cd79c9f8c17622d0bbeca870afd	535	73.75	14	401	P
8086342d42076256d768225656458f3d	485	50.17	65	2132	P
26decb0adcb2f06af6102b99a8c5b604	480	53.08	09	3201	P
58bb7b88e3dc9dcfc2154d4e1525c32c	460	37.42	20	19	F
49019e46daec956d9a3c9d6dcced476f	445	13.67	01	1	F
68a7768adb80f61f06c3a74795611d2e	440	58.17	20	3326	P
def6cfbdfc6835b08df7249bbce722ad	435	13.25	18	1	F
acfc9d6aa47054c87ae543231ccf791c	425	45.08	02	262	P
5730a19087aa01388a1cc97917bbae89	420	46.75	15	1	P
9b70113754a905a57adc7eb15396e0ab	415	50.67	10	1037	P
fdac6594ab2c91f06c6f2c1948565a3b	415	56.92	66	1853	P
1f5440434281 1cdc9de45317e98d39fe	405	60.08 T	02	11007	P

Table 5: Computer Application (CA) Retirees 2016

When we examine the CAO entries using the same criterion we identified one student which has an activity level which exceeds the remainder of the cohort. Student 1f54404' activity commences at the start of the year slightly above all other students and drops off

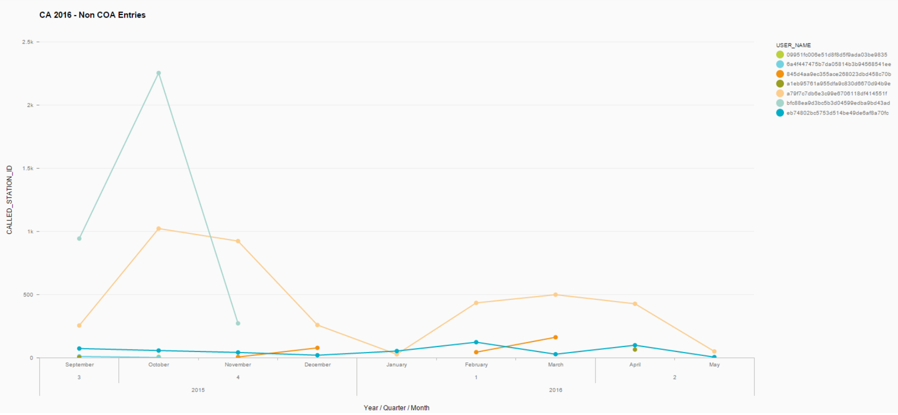


Figure 4: CA non-CAO Entry Activity

in line with the remainder of the students, but had a dramatic rise in the month of April. Examination of the April activity placed the student for the majority of there time between the School of Computing's four computer labs. Why this student did not complete their examinations is unknown. The student did appear in the wifi log records from September 2016, this would indicate that they did continue with their studies.

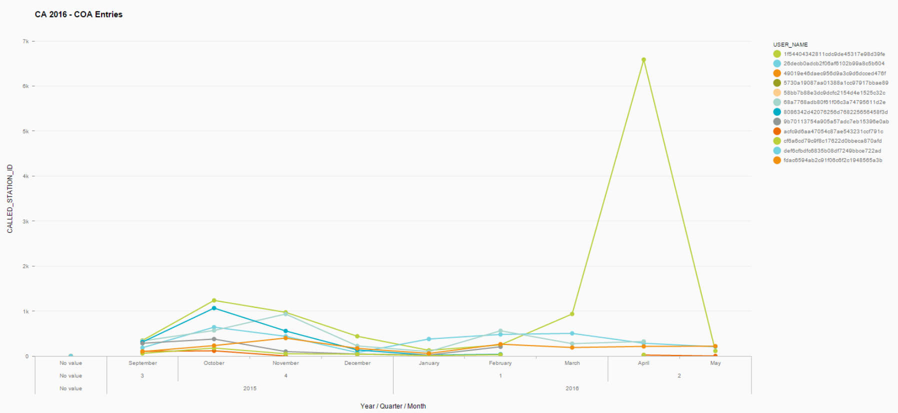


Figure 5: CA, CAO Entry Activity

We concluded from the analysis of the CA student retirees that those who did attend the campus had an activity level, with the exception of 1f54404', below the average activity of the CA cohort. In the case of a79f7c7d' there activity level was also below that of the cohort during the academic year but rose in April bring there activity level upwards. Our research is designed to examine Peer influence, we deem that the students on this list had minimal attendance at the college and therefore minimal impact on the cohort for the academic year. We therefore deem it appropriate to exclude them from the 2016 analysis.

We carried out the same analysis on the EC students to determine if any of the students warrens their inclusion in the 2016 analysis. Of the eleven students identified as retirees, there is one student, namely 9b1c28c6', we consider requires further analysis. All others have a very low activity count. The student of interest is easily identifiable in Figure 6. We can see that the student maintained a consistent level of attendance through out the year before increasing during the months of February and March.

EC

Student	CAO Points	Precision'15	WiFI Activity	Result
48251818e5f0a03a28c4ef61e9d736fb	430	56	118	P
fcd00bacc7327dc417e56d1ebdb67189	420	48	1	P
b22474774a018433574d322d3b7876d0	400	46	1	P
bd03060cb17cbb600612a585c53030df	395	36	716	F
a188c3ae93fa49162e157527546b26e6	395	37	1185	F
653d6ec441f92e10c56dad2c0d630d4f	385	40	1	P
3507575dff2ad165ad87a0f2f12ce6b4	385	37	1	F
91ed5944463354813ceae37906241aaa	375	38	547	F
9b1c28c65fa9cc7a2c877c9168b5ea80	375	36	5070	F
7e7620fb584d8147d784be811f7ac5a2	375	24	37	F
a1d1abe16748a25a787a0493afe4e72e	370	54	330	P

Table 6: Enterprise Computing (EC) Retirees 2016

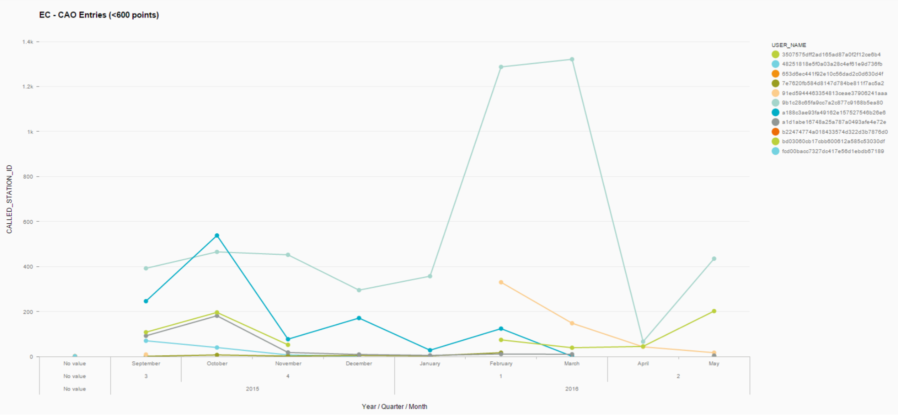


Figure 6: EC Non-CAO Entry Activity

Observations

We have observed that there can be retirees at any stage of an academic year, with some students retiring later in the year. Those who retire late in the year seem to be in a minority, with the majority dropping out at the end of an academic year and do not return at the start of the next year. Retirees occur in every year and cannot be easily predicted who will drop put as some achieve marks well above the pass grade and others would be classified in the lower deciles of the cohort.

Appendix D

Multiple devices

We believed that it is conceivable that a student may have access to a number of devices that they have registered to their account. That is they have accessed the wifi in DCU using their unique credentials to access the system. For example many student have a mobile smart-phone, a laptop, a tablet or any combination and in a semester may have more than one of a particular type of device. To ensure a thorough understanding of the systems we are analysing and avoid ambiguity with our data collection and analyses we examined the number of devices that a student may use in an academic year. The analysis shown that students on average registered on four distinct devices, with some students having more. In some cases we identified students that *seem* to have a number of devices that they only use for a day or two. Anecdotally this may be accounted for with students sharing their credentials with friends who visit them on campus. It was apparent that some students use their devices a great deal more than others.

For our analysis we identified the first and last date of wifi interaction per device and a summation of the occurrences during that period, from this we can see a definite pattern of student use, It was apparent that students have one main device that they retain for the academic year. This device demonstrated that the majority of the students wifi traffic occurred on it between September to May. The other devices which had a considerably less traffic over a shorter time period. Where students have two devices with significant traffic levels there is either no or a short overlap in times between the two devices being used. This could indicate that the student is replacing one device with another and once the new device is commissioned the original device was retired and would not be active on the WiFi system.

Usage - a large number of events may indicate that the student is logging on and off the system or is transiting around the campus connecting to the various NAS in different locations

Duration of life of a device is calculated as the earliest and latest dates the device connected to eduroam, see Table 7. It is conceivable that the device only was used very sparingly at the start and end of the period in question.

Table 7 illustrates the variance in the number of appliances a student can use during a semester. In the case of student f30cb7 who had numerous devices logging onto the wifi during the year either individually or potentially at the same time.

The conclusion from this analysis is that due and careful consideration is given to the summation of a students activity as collected in the Eduroam logs. We need to ensure that there is no possibility of duplicate devices being used in simultaneously, causing a double count.

This table is a sample of students listing the CALLING_STATION_ID, which is the unique identifier of the device requesting permission to connect to the wifi system and registering in the logs. The MIN_DATE and the MAX_DATE, specify the earliest date a log for that devise is registered and also the last date found. The count column is a summation of the unique log-on requests from that device and the final column is a count of the days between the MIN_DATE and MAX_DATEs.

Appendix E

It is observed that students who are placed together in project groups have different levels of interaction in their first year, over time in the majority of cases the interaction reduces. This could indicate that these groups do not maintain a degree of interaction that could be construed as friendship.

Group 1		MEETINGS		
USER1	USER2	2015	2016	2017
5784a44df24cc8f41cbefbfb8a7f17fb	ec9673d21f129a3e41c4a08b75dbf2c4	2	12	1
a69c985c94fca6d2ca77d91b93720e63	ec9673d21f129a3e41c4a08b75dbf2c4	3	16	4
2b264df9656da407d35d7337e0169170	ec9673d21f129a3e41c4a08b75dbf2c4	41	17	36
2b264df9656da407d35d7337e0169170	5784a44df24cc8f41cbefbfb8a7f17fb	73	68	32
2b264df9656da407d35d7337e0169170	a69c985c94fca6d2ca77d91b93720e63	120	98	82
5784a44df24cc8f41cbefbfb8a7f17fb	a69c985c94fca6d2ca77d91b93720e63	141	78	5

Group 2		MEETINGS		
USER1	USER2	2015	2016	2017
65d48f06eb253773a471344d95c6a5a1	d3dc0557965ec7ccc85d71157aeea9bf	275	55	4
65d48f06eb253773a471344d95c6a5a1	c27ff4936ee7f5a16c1b86ec9409b6be	674	53	13
c27ff4936ee7f5a16c1b86ec9409b6be	d3dc0557965ec7ccc85d71157aeea9bf	527	146	49

Group 3		MEETINGS		
USER1	USER2	2015	2016	2017
07d3d55ac4fa5aa3f765e78dea12300c	d900d84ffe30b6a68c667af243e67eb4	3		
b37f2ca45ca9a6e3fcc5c56fc455904f	d900d84ffe30b6a68c667af243e67eb4	31	43	2
33207637bba0ab58ac282b95403ff667	b37f2ca45ca9a6e3fcc5c56fc455904f	37	58	2
33207637bba0ab58ac282b95403ff667	d900d84ffe30b6a68c667af243e67eb4	161	89	
07d3d55ac4fa5aa3f765e78dea12300c	b37f2ca45ca9a6e3fcc5c56fc455904f		1	
07d3d55ac4fa5aa3f765e78dea12300c	33207637bba0ab58ac282b95403ff667		2	

Table 8: Formal- Project group meeting count over three years

2015

Student	CALLING_STATION_ID	MIN_DATE	MAX_DATE	COUNT	Days
f30cb7-	1C-AB-A7-A8-0F-20	24-Sep-14	03-Feb-15	13	132
	60-8F-5C-B5-89-A1	25-Sep-14	25-Sep-14	39	1
	9C-D2-1E-2C-27-B9	29-Sep-14	13-May-15	118	226
	D0-DF-9A-9B-86-1A	09-Oct-14	21-May-15	71	224
	D0-DF-9A-9B-97-80	10-Oct-14	10-Oct-14	7	1
	34-A3-95-CC-4E-C9	04-Nov-14	12-Nov-14	846	8
	00-0C-E7-B6-03-C0	10-Nov-14	04-Mar-15	56	114
	00-C1-40-51-0A-CC	14-Nov-14	14-Nov-14	11	1
	28-37-37-19-5B-36	17-Nov-14	17-Nov-14	2	1
	B0-AA-33-88-88-88	10-Feb-15	10-Feb-15	4	1
	C0-EE-FB-25-9E-2B	10-Feb-15	21-May-15	1,823	100
	2C-BE-08-ED-FF-5A	23-Feb-15	23-Feb-15	3	1
	E8-3E-B6-A3-2A-B5	25-Mar-15	25-Mar-15	1	1
1997c2-	18-1E-B0-05-2F-74	22-Sep-14	02-Apr-15	3,082	192
	D0-DF-9A-9B-86-1A	07-Oct-14	21-May-15	198	226
	78-DD-08-FE-AF-D9	09-Oct-14	02-May-15	94	205
	48-74-6E-97-62-F1	22-Oct-14	04-Dec-14	245	43
	C8-B5-B7-7B-E9-F0	22-Oct-14	04-Dec-14	208	43
	F4-F9-51-85-A4-CD	14-Nov-14	19-May-15	360	186
	30-75-12-A9-5D-AA	02-Jan-15	02-Jan-15	6	1
	2C-BE-08-ED-FF-5A	10-Mar-15	10-Mar-15	5	1
9a5ae2-	78-A3-E4-4D-DC-FD	15-Sep-14	09-May-15	5,940	236
	68-17-29-B2-F5-02	16-Sep-14	24-May-15	1,042	250
	14-1A-A3-28-DE-D1	12-Jan-15	23-May-15	2,151	131
	00-0F-55-A9-2D-22	04-Feb-15	04-Feb-15	8	1
	D0-E1-40-69-3F-94	18-May-15	22-May-15	259	4
	14-10-9F-D0-F3-C1	19-May-15	22-May-15	53	3
48c0f0-	1C-3E-84-C5-81-CD	22-Sep-14	17-May-15	85	237
	00-22-41-6B-C3-0C	26-Sep-14	26-Sep-14	12	1
	CC-89-FD-A9-75-5C	14-Oct-14	16-Oct-14	36	2
	08-70-45-09-BA-7D	05-Nov-14	19-May-15	766	195
	84-8E-0C-8C-46-BE	11-Feb-15	24-Apr-15	80	72 1
4e72b1-	58-55-CA-6D-6B-B8	17-Sep-14	12-Dec-14	401	86
	24-EC-99-48-79-C3	06-Oct-14	12-May-15	57	218
	18-AF-61-23-57-AD	14-Nov-14	14-Nov-14	86	1
	64-76-BA-C6-3F-F3	11-Jan-15	06-May-15	643	115
de076a-	CC-08-E0-43-73-3D	24-Sep-14	07-Mar-15	1,996	164
	84-38-35-85-74-38	05-Mar-15	21-May-15	1,474	77
	74-E5-43-58-E2-61	12-Mar-15	12-Mar-15	2	1
	40-E2-30-25-4E-B7	06-May-15	20-May-15	20	14
ab78c3-	60-03-08-6E-19-11	11-Sep-14	24-Sep-14	100	13
	00-61-71-C0-CB-99	25-Sep-14	31-May-15	7,780	248
	5C-96-9D-82-88-95	11-Jan-15	11-Jan-15	1	1
80a12a-	0C-14-20-55-79-87	24-Sep-14	12-Dec-14	526	79
	D0-E7-82-A9-77-3D	25-Sep-14	28-Oct-14	4	33
	5C-0A-5B-78-2A-1A	14-Jan-15	21-May-15	752	127

Table 7: Multiple Devices Sample

Bibliography

[1] The disability access route to education (DARE). http://accesscollege.ie/dare/. Accessed: 2018-Apr-05.

[2] Eduroam wifi system configuration recommendations. https://www.dcu.ie/iss/networks/eduroam/wireless-network.shtml. Accessed: 2018-Sept-11.

[3] *Learning analytics beyond the LMS*, 2015.

[4] Oleg. Yudkevich Maria Androushchak, Gregory. Poldin. Role of peers in student academic achievement in exogenously formed university groups. *Educational Studies*, 39:568–581, 2013.

[5] Alexander W Astin. Four critical years. effects of college on beliefs, attitudes, and knowledge. 1977.

[6] Jarlath Benson. *Working More Creatively With Groups 3//e*. Routledge, 3rd edition, 2009.

[7] Jerome S Bruner. *Acts of meaning*, volume 3. Harvard University Press, 1990.

[8] CAO. Full-time Undergraduate New Entrants in All HEA-Funded Institutions by field of study (ISCED) at 1 March 2015. http://hea.ie/statistics-archive/.

[9] CAO. Higher Education Access Route (HEAR). http://accesscollege.ie/hear/about-hear/what-is-hear/. Accessed: 2018-04-05.

[10] Monica Harber Carney. Identifying peer effects: Thinking outside the linear-in-means box. 2013.

[11] Scott E Carrell, Richard L Fullerton, and James E West. Does your cohort matter? Measuring peer effects in college achievement. Technical report, National Bureau of Economic Research, 2008.

[12] Simone Celant. The analysis of students academic achievement: the evaluation of peer effects through relational links. *Quality & Quantity*, pages 1–17, 2013.

[13] Roy Chen and Jie Gong. Group formation and performance: Field experimental evidence. 2016.

[14] James S Coleman et al. Equality of educational opportunity. 1966.

[15] Owen Corrigan, Alan F Smeaton, Mark Glynn, and Sinéad Smyth. Using educational analytics to improve test performance. pages 42–55, 2015.

[16] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

[17] Shane Dawson. A study of the relationship between student social networks and sense of community. *Educational Technology & Society*, 11(3):224 to 238, 2008.

[18] Eleanor Denny. Transition from second level and further education to higher education. *Focus Research Report #6*, 2015.

[19] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[20] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[21] Rebecca Ferguson. The state of learning analytics in 2012: a review and future challenges. *Technical Report KMI-12-01*, 4:1 to 18, 2012.

[22] Robert L Ferguson. Constructivism and social constructivism. *Theoretical frameworks for research in chemistry/science education*, pages 28–49, 2007.

[23] Michael Fire, Gilad Katz, Yuval Elovici, Bracha Shapira, and Lior Rokach. Predicting student exams scores by analyzing social network data. In *Active Media Technology*, pages 584–595. Springer, 2012.

[24] Donelson R Forsyth. *Group dynamics*. Cengage Learning, 2018.

[25] Gigi Foster. Its not your peers, and its not your friends: Some progress toward understanding the educational peer effect mechanismb. *Journal of Public Economics*, 90:1455–1475, 2006.

[26] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[27] D. Z. Grunspan, B. L. Wiggins, and S. M. Goodreau. Understanding classrooms through social network analysis: A primer for social network analysis in education research. *Life Science Education*, 13:167–178, 2014.

[28] Ankur Gupta, Sanil Paul, Quentin Jones, and Cristian Borcea. Automatic identification of informal social groups and places for geo-social recommendations. *International Journal of Mobile Network Design and Innovation*, 2(3-4):159–171, 2007.

[29] Robert L Hall and Ben Willerman. The educational influence of dormitory roommates. *Sociometry*, pages 294–318, 1963.

[30] Gabriella M Harari, Samuel D Gosling, Rui Wang, Fanglin Chen, Zhenyu Chen, and Andrew T Campbell. Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior*, 67:129–138, 2017.

[31] Ramaswamy Hariharan and Kentaro Toyama. Project lachesis: parsing and modeling location histories. In *International Conference on Geographic Information Science*, pages 106–124, 2004.

[32] Jessica Hoel, Jeffrey Parker, and Jon Rivenburg. Peer effects: do first-year classmates, roommates, and dormmates affect students academic success. In *Higher Education Data Sharing Consortium Winter Conference, Santa Fe, NM*, 2005.

[33] Anil K Jain, Alexander Topchy, Martin HC Law, and Joachim M Buhmann. Landscape of clustering algorithms. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 260–263. IEEE, 2004.

[34] George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[35] knime.com. Knime analytics platform. https://www.knime.com/about, 2015 (accessed June 12, 2015).

[36] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 2007.

[37] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

[38] David S Lyle. Estimating and interpreting peer and role model effects from randomly assigned social groups at west point. *The Review of Economics and Statistics*, 89(2):289–299, 2007.

[39] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

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

[41] Abraham H Maslow. Personality and motivation. *Harlow, England: Longman*, 1:987, 1954.

[42] Elton Mayo. *The Human Problems of Industrial Civilisation*, volume VI. Routledge, Taylor & Francis Group, 2003.

[43] Witton J. & Elbourne D. R. McCambridge, J. Systematic review of the hawthorne effect: New concepts are needed to study research participation effects. journal of clinical epidemiology. *Journal of Clinical Epidemiology*, 67:267277, 2014.

[44] B. Minaei-Bidgoli, D.A. Kashy, G. Kortemeyer, and W.F. 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.

[45] Barbara Oakley, Richard M Felder, Rebecca Brent, and Imad Elhajj. Turning student groups into effective teams. *Journal of Student Centered Learning*, 2(1):9–34, 2004.

[46] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[47] Alex Pentland. The new science of building great teams. *Harvard Business Review*, 90(4):60–69, 2012.

[48] Jun Rekimoto, Takashi Miyaki, and Takaaki Ishizawa. Lifetag: Wifi-based continuous location logging for life pattern analysis. In *LoCA*, volume 2007, pages 35–49, 2007.

[49] Kristen A Renn and Karen D Arnold. Reconceptualizing research on college student peer culture. *The Journal of Higher Education*, 74(3):261–291, 2003.

[50] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[51] Bruce Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704, 2001.

[52] George Siemens and Phil Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5):30, 2011.

[53] Splunk.com. Splunk business analytics. https://www.splunk.com/en_us/solutions/solution-areas/business-analytics.html, 2015 (accessed June 7, 2015).

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

[55] Lev S Vygotsky. Mind in society: The development of higher mental process, 1978.

[56] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew 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.

[57] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, 1994.

[58] Jun-ichiro Watanabe, Saki Matsuda, and Kazuo Yano. Using wearable sensor badges to improve scholastic performance. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 139–142. ACM, 2013.

[59] James E Willis III. Ethics, big data, and analytics: A model for application. *Educause Review Online*, 2013.

[60] David J Zimmerman. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, 85(1):9–23, 2003.

Glossary

a-l Average linkage: Average represents a natural compromise, but depends on the scale of the similarities. Applying a monotone transformation to the similarities can change the results. Average linkage. 110, 183

c-l Complete linkage: Complete linkage has the opposite problem to chaining. It might not merge close groups because of out-lier members that are far apart. Complete linkage. 110, 183

ca Computer applications program, emphasis on the development of computer applications CA. 35, 183

CAO Central Applications Office, The Central Applications Office processes applications for undergraduate courses in Irish Higher Education Institutions (HEIs) CAO. 43, 183

ec Enterprise Computing program, combination of business and ICT principles EC. 35, 183

Information Systems Services Providing services and support to DCU staff and students through the application and use of information related technologies and processes. Information Systems Services. 59, 66, 183

precision mark A weighted average mark for a students academic results in an academic year precision mark. 71, 183

single linkage Single linkage: The single-linkage method produced asymmetric-looking clusters; this is the so called chaining effect, which refers to the tendency of the method to incorporate intermediate points between clusters into an existing cluster rather than initiating a new one. Chaining can occur in the early stage of cluster development and impact on the final cluster shape Single linkage. 110, 184

w-l Ward linkage: Ward looks at cluster analysis as an analysis of variance, instead of using distance metrics or measures of association. The Ward error sum of squares hierarchical clustering method involves an agglomerative clustering algorithm. Ward linkage. 110, 184