University of Glasgow | School of Computing Science

Analysing the interplay between students' preferences and sociological variables

Irina Preda (1102452p)

April, 2016

ABSTRACT

This papers presents the research conducted to explore the relations between prospective students' characteristics and their academic preferences. Career decisions are eminently important in determining ones socio-economic status and quality of life. The factors which influence people's decision to pursue a certain academic subject are mostly unclear, which motivated us to explore the relevance of factors such as gender and social status. Also, we decided to use the data set to examine another possible physical trace of academic intention, in-app behaviour. A mobile app was used to extract the relevant characteristics, by collecting data such as users' first names and postcodes, as well as the sequence of events representing their app behaviour. Their personal characteristics were used to examine their influence on academic choices. Their app behaviour was also used to study engagement. Our results suggest that these factors are meaningful and that they can helps us understand prospective students' decision-making. Research has yet to take advantage of the large number of users mobile devices attract. With this wealth of data, we can more easily explore the relations between physical traces of behaviour and internal human intentions. Also, the discreteness of a mobile application versus a more traditional data gathering method means it minimizes the risk of users feeling self-conscious or manipulating the results. By employing such data collection and analysis, we can take steps towards determining the unconscious factors involved in such career-defining choices.

1. INTRODUCTION

The growth in ubiquitous computing has led to major shifts in our daily lives, changing our relationship to data. Our mobile devices have become vast repositories of information, as we become more and more reliant on them. One of the biggest beneficiaries of this development could be the research community, which could take advantage of this easily-available data to advance the state of the art across disciplines, well beyond classical computer science. We look specifically at the potential for developing tools which can process sociological data and app-behaviour, and enable the critical analysis of it.

The examination of how users interact with applications has been shown to provide a wealth of information[5][15][4]. Mostly this has been applied for the purposes of improving devices and applications[39], however there has been increased interest in applying this data analysis for more diverse purposes. Through app usage analysis, we can also understand the users better and predict behaviour, by mapping their background and profile (gender, age, location, etc.) with their in-application behaviour. These approaches could provide interesting applications for sociologists, marketers and policy advisers. And it could also be used specifically by universities in their attempts to reach out to underrepresented communities.

This project targeted the development of automatic approaches for the sociological analysis of the prospective students that participate in the Open Days at the University of Glasgow. In particular, the project applied statistical analysis techniques to the corpora collected in collaboration with the Data Lab[14] and Bizvento[2]: registration data (Open Days participants' information, expressed interest about particular subjects), usage data extracted from the app developed by Bizvento, Scottish Index of Multiple Deprivation[11]. By using this data collected by mobile app, the experiment involved 4,885 participants, which far surpasses the number of subjects studies of this kind have attracted in the past (the largest one had 1,668 participants[25]).

This research was focused on examining the relationship between three elements: academic preferences (what subjects participants were interested in), sociological characteristics (gender, social status, etc.) and app behaviour (by categorising participants' interaction with the app).

The proposed approaches address the following sociological questions:

- What are the main characteristics of the people interested in a particular subject?
- Is it possible to cluster prospective students into groups of individuals characterized by particular features or behaviour?
- Is there a difference between people interested in different types of subjects?
- Is it possible to predict which students will finally get an offer from the University?

Our motivation is to try and understand academic decision-making at this early stage and the factors influencing it. The choice of university degree is a highly risky endeavour for prospective students, as this decision is very likely going to shape the next chapters in their lives. Of course, this decision is not taken randomly, in which case we would

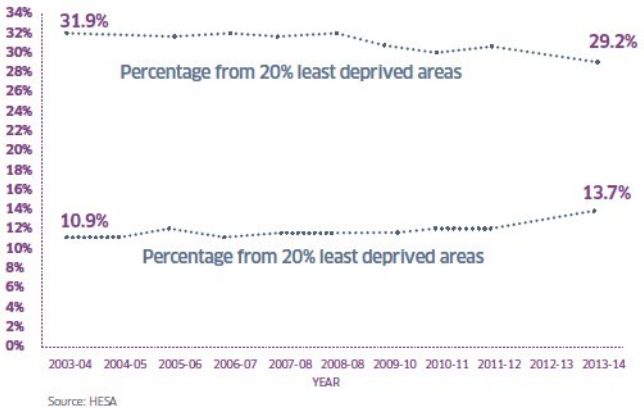

Figure 1: This graph is from the Scottish Government website (the bottom line has a mistake, it should be most deprived). There has been steady progress on widening access in Scottish Higher Education Institutions, with the percentage of full-time first degree entrants from Scotland's 20 per-cent most deprived areas increasing from 10.9 per-cent [11]

see an approximately even distribution among disciplines over the population, but is rather more likely the result of a number of factors. The question is what factors are the most decisive in determining these preferences? If academic preferences were objective assessments based on the return of investment, we should observe high-paying subjects in overwhelming demand across the board and find sociological factors to be negligible. But experimental surveys have indicated that when people make risky choices they often do not do so with objective reasoning in mind, but rather choose different paths "tend to follow regular patterns that can be described mathematically"[28][24].

This study set out to identify the relevant patterns. To do so we chose to use mobile devices to collect the necessary data through a mobile app. This was done to ensure its widespread use and, very importantly, a "natural setting" for data acquisition. Instead of using a questionnaire which could introduce certain biases associated with this technique[23], we offered the app as a timetabling service and used inference based on the information provided by the participants. This was done to avoid triggering any bias in the participants' responses relating to social norm and desirability.

Our usage of this data collection method reflects an attempt to obtain data about academic preferences that is supplied as honestly as possible, whilst ensuring high standards of privacy. No sensitive data was acquired and participants' identities remained obscured all throughout. The participants' data was combined with publicly-available information to extract societal trends. The methodology developed in this work can be applied in future research with similar approaches.

2. RELATED WORK

To conduct this research it was necessary to examine the literature surrounding the topic of app usage data collection and analysis.

In 2010 researchers looked at developing a new probabilistic framework that would allow automatic mining usage patterns of mobile phones [22]. They were interested in finding out whether they could extract patterns of usage and whether they could characterise users' behaviour based on these patterns. For the purposes of this research they developed a new way of representing data to summarize app usage and an author-topic model that probabilistically infers relation between usage patterns and users in an unsupervised way. The results show that the system was able to perform the detection of patterns as well as the analysis all on its own. Their framework was thoroughly evaluated with hundreds of thousands of hours of real-life app phone data and it showed that relevant patterns can be extracted. One downside to this kind of system is inevitably the need for a large set of training data. Also atypical patterns have not been proven to work by their methodology but could be explored in a subsequent study.

In 2011 a similar study [21] was conducted with the same researchers together with Jan Bloom to examine the relation between phone usage and two contextual cues, location and social context. For information on the contextual cue they designed the program so that it remembers visited places and requires human input for semantic labelling. The research questions they asked related to identifying the impact on app usage of these cues both individually and jointly. To perform this experiment they enlisted 77 participants and collected data over 9 months. Their findings indicate that patterns of phones usage rely on these contextual cues and they use these findings to suggest improvements to smartphones relating to user needs. One downside to this study was that it is not generalizable because of the small sample size and the type of phone used which was quite outdated (together with available applications). Also information while participants were moving was not collected. However the results are promising enough to encourage further study, however these limitations need to be resolved.

A really early study on data mining of usage patterns came from the year 2000 [36] when researchers explored these techniques on exploring usage patterns for web data. This paper presents a comprehensive although slightly outdated survey on web usage mining and the three necessary phases of preprocessing, discovery of patterns and pattern analysis. The most interesting aspect of this papers is the description of pattern discovery methods like statistical analysis, clustering, classification, sequential patterns, etc. Especially relevant is the sequential pattern discovery which is used to predict future visit behaviour based on existing data through trend analysis, similarity analysis or change point detection. Unfortunately the paper does not delve into the practical applications of these methods.

One really comprehensive study in how applications can be used to perform data collection and analysis is the paper from 2012 from Carnegie Mellon University [35] where they collected sensory mobile data to analyse the context of the

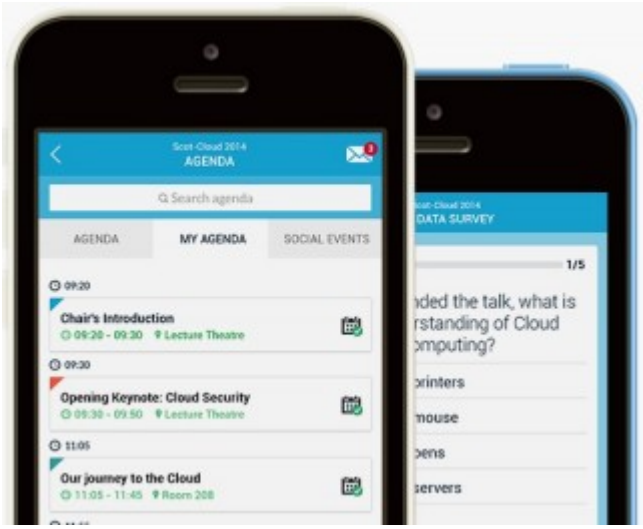

Figure 2: The Bizvento app[2] which prospective students could use to select events they wanted to participate in during the university's Open Day.

app usage. They then built models that would calculate the probabilities associate with an app being used in the present context. And they developed an application that would help users interact with the results (homescreen with suggestions) and at the same time evaluate their model. This study stands out in its development of a new model which performs data collection on a wide array of contextual cues. The prediction method they used was Naive Bayes and they found that the most important contextual influencers were the time of day, the last app used and cell ID. The results are really convincing (the app predictions outperforming previous research) but the researchers also express confidence in allowing their system to account for newly-installed or rarely used applications. But those elements will require some innovation in terms of methods.

Rosenfeld's paper from 2000 [33] informed my understanding of n-grams and influenced my decision to generate up to a trigram based on the size of the corpus available.

3. METHODOLOGY OVERVIEW

Our aim was to better understand prospective university students and the influences behind their subject choices. To do so, we decided to use the data set collected by Bizvento through their mobile application2. Rather than ask the students for information regarding their academic preferences and personal characteristics (gender, social status, etc), which could be vulnerable to biased self-reporting, we instead used the app as a tool for collection of data from the participants and then inferred the necessary information. The application was really successful with 4,885 participants, 2,721 accesses and 32,664 app events.

The app was offered as a service which allowed prospective students to pre-plan their open-day visit by choosing the subjects they were interested in, so as to join their respective introductory sessions. Alongside their academic preferences, their first name and postcode were also obtained. Their first name was useful as it could then be used by us as an unconscious trace of their gender. To convert the names to gender, we used an already existing Python library[8] that had a dictionary of names and their classification with values such as either male, female, mostly female, mostly male or ambiguous. The reasoning for this is that priming a social category can elicit stereotype-consistent behaviour and bias the results. For example the information on gender was extracted from the first names provided by the participants, as there is literature to suggest that the mention of gender can act as a primer leading to girls to underrate their own mathematical abilities or interests [37]. We also collected postcodes, similarly, to gauge participants' social status without expressly asking sensitive questions that might introduce bias. The postcode data was combined with open source data offered by the Scottish Government[11], the Scottish Index of Multiple Deprivation, which provides information regarding each postcode relating to educational attainment, income levels and other sociological factors. These factors were tested for significance using the Chi-Square test.

The second part of the experiment involved the extraction of the app behaviour from the mobile app's data set. User access sessions and their sequence of page views were isolated and then fed into an algorithm which processed their entropy. And then n-gram analysis was used to explore the relation between events. This was done for to observe the different types of behaviour users exhibited.

4. METHODOLOGY - THE PROGRAM

A program in Python was created to perform the unpacking of the data set, the n-gram sequence analysis and the generation of graphs. To create the graphs a Python library named Plotly[7] was used.

To understand and categorise the app behaviour, sequence analysis was performed. From the data set, the full access list (which contains all the sequences of events) was extracted and used to construct the lexicon (which contains all unique events). Each unique event was labeled with a letter from a to m, as there were 13 events in total. This was done to make processing easier, as each sequence of events would then be a string of unique letters and could be easily manipulated. Once the access list and the lexicon were assembled, the entropy of each sequence was calculated. A list of all entropies was built, ordered from highest to lowest and then displayed in a graph.

A python program was designed to allow for input arguments to be offered when run, so as to easily obtain the n-gram for any value n. For the program to be able to do so, the basic functionality was modified to construct the access list and lexicon differently, by allowing for the value n to influence the pattern of the unique event code. Thus, while for a unigram it would be single events represented by one letter, for a bigram it would be two events concatenated and represented by two letters (e.g.'ab'), and so on for larger values of n. One addition was also the dummy value 'f' which, when added to an event (e.g.'fa'), preceded the event code that was the first in the sequence. This was done

so as to separate starter events as distinct, which is important to recognise in sequential data analysis, because they are not preceded by others. The entropies were calculated in the same way as for the unigram.

This program is easy to use, requiring the user to only have Python installed, run the script with a properly formatted data set and specify the n value as an input argument. As a result, a graph of the entropy of each sequence is generated and displayed in the default browser.

An optional feature was added to display the graph with superimposed information for each sequence regarding the entropy contribution of the page view events in relation to the entropy of the full sequence of events. This was done because the page views represent a specific kind of user behaviour representing low activity.

For the purposes of this experiment, this program was used to obtain graphs of the unigram, bigram and trigram; also to produce superimposed information relating to the contribution of page view events.

5. METHODOLOGY - THE THEORY

5.1 App Behaviour

Sequence analysis was used for this research so as to study the in-app behaviour of users by categorizing the different behaviours according to their likelihood.

To do so, we had to calculate the entropy of each sequence of events. For information theory, entropy as described by Shannon[34] is the measure of the uncertainty associated with a random variable. Here, our random variable is the sequence of events performed by the user. The formula can be derived by calculating the mathematical expectation of the amount of information contained in a sequence of events.

To calculate the entropy, we first expressed the data mathematically: s is a sequence composed of M events and each event is an element of the lexicon.

$$\begin{aligned} s &= \{s_1, s_2, ..., s_M\} \\ where \quad s_i &\in \{e_1, e_2, ..., e_L\} \end{aligned} \tag{1}$$

Then we estimated the probability of the sequence occurring as the product of the probability of each event:

$$p(s) = \prod_{k=1}^{M} p(e_k) \tag{2}$$

Afterwards, the geometric mean of the probability is calculated to eliminate the influence of the size of the sequence:

$$[p(s)]^{1/M} = [\prod_{k}^{M} p(e_k)]^{1/M} \tag{3}$$

Once the mean has been obtained, its common logarithm can be calculated to simplify the equation:

$$log[p(s)]^{1/M} = \frac{1}{M} \sum_{k=1}^{M} \log p(e_k) \tag{4}$$

Then the equation can be expressed in terms of the events of the lexicon, where $n(e_k)$ is the number of times the event is observed in a sequence and $p(e_k)$ is the estimated probability of that event occurring in relation to the entire corpus:

$$log[p(s)]^{1/M} = \frac{1}{M} \sum_{k=1}^{L} n(e_k) \log p(e_k) \tag{5}$$

Thus, the expectation of the amount of information contained in each sequence of events can be calculated:

$$\mathbb{E}(p(s)) = \sum_{k=1}^{L} \hat{p}(e_k) \log p(e_k) \tag{6}$$

5.2 Gender & Social Status

For the analysis of the sociological background of the participants, the gender was inferred from users' first names and social factors from their postcode (education, income, housing, employment, crime, health and access levels).

The Chi-Square test was used to verify whether sampling is random. Chi-square is a way of measuring the statistical significance of observed means that deviate from expected random distribution. It is calculated as the sum for each individual value of a category (G is the number of categories, e.g. for gender we considered two) minus the expected mean times the number of values in a category, squared (so positive and negative values don't cancel out) and then divided over the expected value.

The following formula expresses that:

$$\chi^2 = \sum_{k=1}^{G} \frac{(O_x - E_x)^2}{E_x} \tag{7}$$

The p-value can then be calculated as such:

$$p(\chi^2) = \int_{\chi_0^2}^{\infty} \rho(\chi 2) d\chi^2 \tag{8}$$

The cut-off for statistical significance (p-value) was decided to be 0.05.

6. RESULTS

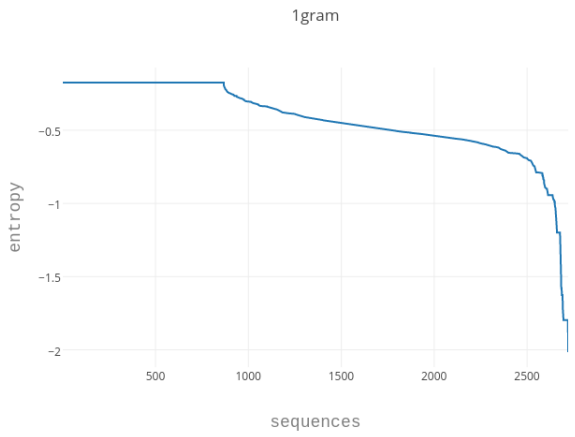

Figure 3: Unigram graph

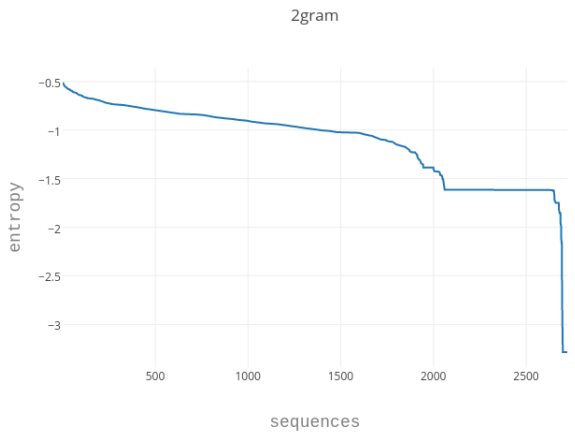

Figure 4: Bigram graph

6.1 App Behaviour

We can infer that the app was used properly, as 94.7 percent of the participants chose five subject preferences or less, which indicates that responses weren't random. The app allowed for users to enlist for all events, but the Open Day lasted between 9.30am and 3pm, so it was only possible to attend at most 5 events.

Graphs were generated to represent the distribution of sequences, from most likely to least likely (on the x axis) and their respective entropies (y axis). The ngram model was used to separate sequences into independent events (unigram), and then into pairs of events where each event affects the next (bigram), and then grouped in three. Deriving a larger n-gram would not be possible on this data set, as we would require a larger corpus.

Figure 3 shows the unigram, with a distribution of the most likely sequence of events to the least likely, from left to right. As we can see from the figure, the data can be divided into three groups. The group of sequences that are the most frequent which involves minimum user activity (looking at a page). A group with some variety in their actions (logging in, enlisting for an event, etc.). And a final group which consists of the outliers, sequences of events that consist of a wide variety of actions and which possibly suggest the user might have been confused about some aspect of the application.

The above figure 4 represents the bigram analysis. This shows the entropy of each sequence with events paired. This was done to examine the effects of changing the model to allow for the events to be statistically dependent on the previous event. This choice was made as it is expected for app events to be linked and performed in succession.

The resulting graph shows a change in distribution, reflective of the strong dependence between events. There is a sharper contrast between behaviours and the most likely behaviour has become one marked by medium levels of activity. The plateau suggests only page views and the falling slope intense activity.

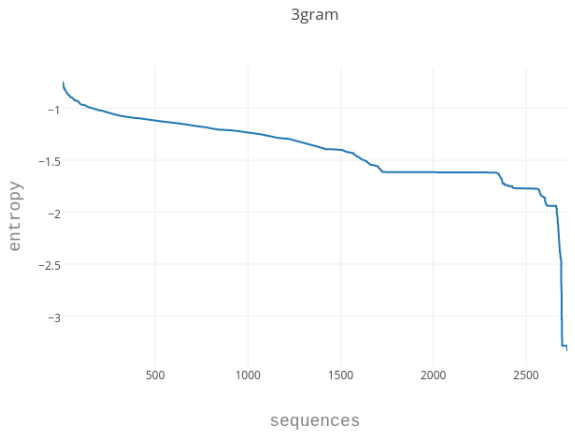

Figure 5: Trigram graph

This figure 5 represents the trigram analysis, which considers events as grouped in threes. This graph further consolidates the conclusions of the bigram.

To clarify the results of the previous n-gram figures, I chose to visually represent the contribution that the page

view events had towards the overall entropy in each figure.

For the unigram, we can see in the following figure 6 that the page view is highly prevalent (the green line) and defines the first category of behaviour understood as inactivity or low activity. The other two categories contain such events but show more active participation.

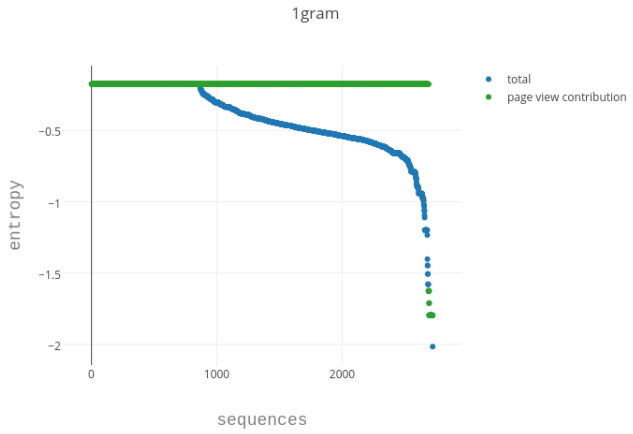

Figure 6: An example figure

The next figure 7 shows the contribution that the page view events had towards the overall entropy for the bigram. As we can see, for the first group (with high entropy) there was a combination of page views and actions suggesting a moderate level of activity, the second group a plateau of page views and the third exhibiting the outlier behaviour (with numerous and varying actions).

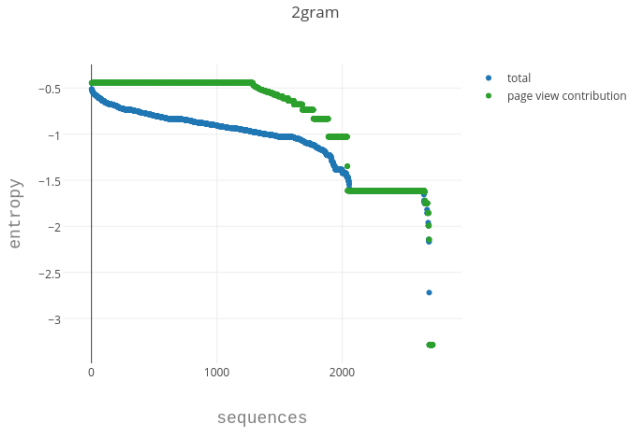

Figure 7: An example figure

And the final figure8 shows the contribution of the page view events for the trigram and fortifies the conclusions of the previous graph.

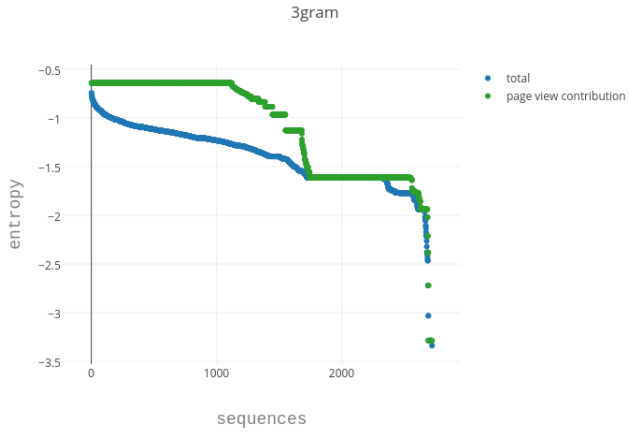

Figure 8: An example figure

The successful representation of these distinct forms of interaction with the application suggests that there is potential in mapping these behaviours to intention and engagement relating to educational attainment.

6.2 Gender and Social Status

To extract gender information we used participant's first names and processed them to remove those that are ambiguous. We found that we could unambiguously map 86 percent of first names to gender (4,112 out of the 4,885 subjects). And we found that women are significantly more frequent among participants in this study than men, a Chi-Square test revealing the p-value $\ll 0.01$. However we also observed that males submit more preferences than females, with a p-value of 0.0001 according to the Chi Square test.

In terms of choice of subjects, we found that men choose more often from the sciences whereas women from the social sciences. However among individual subjects the most female dominated fields were human biology, psychology and english literature. Whereas the largest male to female ratio was observed in accountancy, mathematics and sports science. The clearest gender differences was in the choice of school, where women where highly likely to choose the School of Arts and men the School of Science and Engineering.

Another aspect to the sociological analysis concerned social status. We looked at the following factors relating to areas of Scotland based on the SIMD[11]:

- Employment
- Income
- Health: amount of medical assistance
- Education, Skills, and Training
- Geographic Access: distance from basic services
- Crime: police records
- Housing: heating and people per household

We found that our participants' social background was distinctly affluent, with 56.5 percent of students coming from the 30 percent most affluent areas of Scotland. Our results also found that education levels were by far the most important factor with 15.5 being the ratio of top decile to bottom decile, followed by housing (9.4) and income (9.3).

7. CONCLUSIONS

My main work in this project represented the sequence analysis which contributed to our understanding of the in-app behaviour of the participants. By using an n-gram model, it was possible to not only explore the entropy of each sequence of events in a user session but also the relation between the events. It was observed that while in a unigram the most likely singular event is for the user to merely view a page and thus do nothing, the bigram and trigram show that, if we consider events as related, the most common behavior is that of alternating between a variety of actions (such as logging in, viewing an event page, joining an event, etc). Our results suggest three distinct behaviours with the most prevalent, characterised by relatively high engagement, being the desired one. Future data regarding prospective students' university applications and respective admissions would allow us to analyse the relations between these behaviour with academic attitudes.

While the results suggest there is a link between sociological factors and subject choice, this can serve as the base for a wide range of interpretations. On the topic of social status, we observed the highest influence is exerted by the levels of education within the area that the prospective student originates from. While the link between income and education is often assumed to be the most important [17], these results suggest that the picture is more complex and requires future research.

For the gender-based differences, one possible way of explaining them is by using the systemising/ empathising theory of Baron-Cohen and [26]. This model states that to systemize (as in to understand object-based interactions) and empathise (as in to grasp social human-based dimensions) - are separate functions which operate independently (an individual being able to be highly skilled in both, either or neither). In their paper the researchers find evidence of the gender disparity, similar to our findings. These functions strongly correlate with gender. And the explanation that they give for this is that these functions are influenced by a mixture of the biological differences among the sexes together with the environment's gender-based influence.

One study by Manson and Winterbottom found that there is an association between degree subject and Systemising and Empathising scores, but also that individuals' scores were better predictors than gender [29]. When they controlled for these scores, the relation between gender and subject choice was not significant. These findings suggest the importance of looking at individuals in terms of the discrepancies between these two functions, rather than simply in terms of their gender.

Another study with the same focus found that "self-assessments of task competence were found to influence career-relevant decisions, even when controlling for commonly accepted measures of ability" and that men are more likely than women "with the same math grades and test scores to perceive that they are mathematically competent" [20]. The researchers behind the study concluded that men do not pursue mathematical activities more than women necessarily because they are better at these activities, but because they rate themselves higher. And that people's choice of career is based on assessing their own competence.

These differences in career choice represent an important social problem, considering the fact that there is a considerable pay gap between men and women: the average woman will earn 19.7 per cent less than the average man per hour [1]. And also if we consider society's need for more workers in STEM [3][9].

Another conclusion as a result of this study are the benefits of using data collected through a mobile device. Firstly, it provides a platform that is highly accessible, especially among the university's target audience, with 90 percent of 16-24 years olds owning a smartphone [6]. Another benefit arises from the flawed nature of analysis based on self-assessments [23]. By avoiding explicit input from the user in this experiment and instead relying on the tool to provide a service to the user (timetabling) we minimise bias. And most importantly, the use of phones as part of large-scale statistical studies present enormous potential in using sensor and application data as an extension to traditional ways of collecting information.

7.1 Future Work

This project could be extended in the future in a multitude of ways:

Firstly, because the subject pool was restricted to parties interested in studying at the University of Glasgow, one could argue that it is biased towards a certain audience. Historically, far more students from affluent backgrounds have been able to study at this university compared with more deprived areas in Glasgow[30][13]. This socioeconomic restriction in our data could be rectified through a cross-university endeavour, where a sample is collected that is more representative of the country as a whole.

Secondly, an evaluation of the application would preferably be conducted to verify its accessibility. A form of bias could be introduced by the level of difficulty users face which can be influenced by digital exposure, income levels, disability etc.

Thirdly, the most interesting extension would be the production of a predictive machine, through the collection of Applicant Visit Day data, which could then be used in verifying the predictive ability of the app behaviour features on academic choices (e.g. is the lack of engagement with the app a predictor for likelihood of applying to the university). This could be used as a way to better isolate the factors that influence prospective students' decision-making. This information that can be useful for improving the university's ability of reaching out to under-represented communities. Also, if data regarding admissions could be obtained, we could see

the sociological features which influence access. Sociologists and policy advisers would benefit from having such a tool at their disposal, to enhance already existing knowledge regarding social factors and their influence on university access and academic diversity.

Acknowledgments. I would like to thank my supervisor, Alessandro Vinciarelli, for his advice and earnest support. Also my thanks to his research assistant, Walter Riviera, for his useful suggestions early in my project. It was a pleasure working with them.

8. REFERENCES

[1] Annual Survey of Hours and Earnings: 2015 Provisional Results. http://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2015provisionalresults. Accessed: 2016-04-10.

[2] Bizvento. https://bizvento.com/. Accessed: 2015-12-10.

[3] Failure to meet engineering skills demand to cost UK Âč27bn a year. http://www.engineeringuk.com/View/?con_id=490. Accessed: 2016-04-21.

[4] How studying mobile user behavior impacts app marketing. http://memeburn.com/2014/11/how-studying-mobile-user-behavior-impacts-app-marketing/. Accessed: 2016-03-20.

[5] Mobile Analytics: The Key to Understanding User Behavior. http://www.webanalyticsworld.net/2014/07/mobile-analytics-the-key-to-understanding-user-behavior.html. Accessed: 2016-03-20.

[6] Ofcom - The Communications Market 2015 (August). http://stakeholders.ofcom.org.uk/market-data-research/market-data/communications-market-reports/cmr15/. Accessed: 2016-03-20.

[7] Plotly for Python. https://plot.ly/python/. Accessed: 2015-12-10.

[8] Python - SexMachine 0.1.1. https://pypi.python.org/pypi/SexMachine/. Accessed: 2016-03-20.

[9] Report reveals scale of UKâĂŹs engineering skills shortage. https://www.theengineer.co.uk/report-reveals-scale-of-uks-engineering-skills-shortage/. Accessed: 2016-04-21.

[10] Scottish Government Equality Outcomes: Gender Evidence Review. http://www.gov.scot/resource/0042/00421042.pdf. Accessed: 2016-03-20.

[11] Scottish Index of Multiple Deprivation. http://www.gov.scot/simd. Accessed: 2016-03-20.

[12] Social Focus on Deprived Areas 2005 . http://www.gov.scot/Publications/2005/09/2792129/21376. Accessed: 2016-04-21.

[13] Students In Higher Education At Scottish Institutions 2008-09. A National Statistics Publication for Scotland ISBN 9780755978724. http://www.gov.scot/Resource/Doc/933/0122980.pdf. Accessed: 2016-03-29.

[14] The Data Lab. https://pypi.python.org/pypi/SexMachine/. Accessed: 2016-03-20.

[15] Understanding User Behavior via Visual In-App Analytics. http://online-behavior.com/analytics/in-app-analytics. Accessed: 2016-03-20.

[16] Y. Bengio. Markovian models for sequential data.

[17] J. Blanden and P. Gregg. Family income and educational attainment: a review of approaches and evidence for britain. *Oxford Review of Economic Policy*, 20(2):245–263, 2004.

[18] S. Briggs. An exploratory study of the factors influencing undergraduate student choice: The case of higher education in scotland. *Studies in Higher Education*, 31(6):705–722, 2006.

[19] F. Camastra, A. Vinciarelli, and J. Yu. Machine learning for audio, image and video analysis. *Journal of Electronic Imaging*, 18(2):029901–029901, 2009.

[20] S. J. Correll. Gender and the career choice process: The role of biased self-assessments1. *American journal of Sociology*, 106(6):1691–1730, 2001.

[21] T. M. T. Do, J. Blom, and D. Gatica-Perez. Smartphone usage in the wild: a large-scale analysis of applications and context. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 353–360. ACM, 2011.

[22] T.-M.-T. Do and D. Gatica-Perez. By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 27. ACM, 2010.

[23] D. Dunning, C. Heath, and J. M. Suls. Flawed self-assessment implications for health, education, and the workplace. *Psychological science in the public interest*, 5(3):69–106, 2004.

[24] D. Kahneman and A. Tversky. The psychology of preferences. *Scientific American*, 1982.

[25] E. Keskinen, J. Tiuraniemi, and A. Liimola. University selection in finland: how the decision is made. *International Journal of Educational Management*, 22(7):638–650, 2008.

[26] J. Lawson, S. Baron-Cohen, and S. Wheelwright. Empathising and systemising in adults with and without asperger syndrome. *Journal of autism and developmental disorders*, 34(3):301–310, 2004.

[27] N. R. Mabroukeh and C. I. Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys (CSUR)*, 43(1):3, 2010.

[28] R. MacMullen. *Why Do We Do What We Do?: Motivation in History and the Social Sciences*. Walter de Gruyter GmbH & Co KG, 2014.

[29] C. Manson and M. Winterbottom. Examining the association between empathising, systemising, degree subject and gender. *Educational Studies*, 38(1):73–88, 2012.

[30] A. McFadyen and Y. Ritzen. Glasgow: A city divided along learning lines. http://www.aljazeera.com/indepth/features/2016/03/glasgow-city-divided-learning-lines-160321140932335.html. Accessed: 2016-03-25.

[31] Y. J. Moogan and S. Baron. An analysis of student characteristics within the student decision making process. *Journal of Further and Higher Education*, 27(3):271–287, 2003.

[32] Y. J. Moogan, S. Baron, and K. Harris. Decision-making behaviour of potential higher education students. *Higher Education Quarterly*, 53(3):211–228, 1999.

[33] R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here? 2000.

[34] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[35] C. Shin, J.-H. Hong, and A. K. Dey. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 173–182. ACM, 2012.

[36] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.

[37] J. R. Steele and N. Ambady. âĂIJmath is hard!âĂİ the effect of gender priming on womenâĂŹs attitudes. *Journal of Experimental Social Psychology*, 42(4):428–436, 2006.

[38] D. Webbink and J. Hartog. Can students predict starting salaries? yes! *Economics of Education Review*, 23(2):103–113, 2004.

[39] D. Weir, S. Rogers, R. Murray-Smith, and M. Löchtefeld. A user-specific machine learning approach for improving touch accuracy on mobile devices. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 465–476. ACM, 2012.