Self-organising maps and student retention: Understanding multi-faceted drivers

David Carroll Gibson
Curtin University

Matthew Ambrose
Curtin University

Matthew Gardner
Curtin University

Abstract: Student retention is an increasingly important yet complex issue facing universities. Improving retention performance is part of a multidimensional and deeply nested system of relationships with multiple hypothesised drivers of attrition at various sample sizes, population clusters and timescales. This paper reports on the use of a self-organising data technique, Kohonen's Self Organising Map, to explore the potential retention drivers in a large undergraduate student population in Western Australia over a six-year period. The study applied the self-organizing method to two point-in-time data sets separated by 18 months and was able to identify a number of distinct attrition behaviour profiles appropriate for creating new tailored intervention.

Keywords: Attrition, retention, predictive models, machine learning, educational data mining, learning analytics.

Introduction

The student retention rate is a broadly accepted and important measure of university performance, and is often considered as a proxy for the quality of education and support services provided (Crosling, Heagney, & Thomas, 2009; Olsen, 2007). Poor or declining retention is of concern for universities as it significantly affects financial performance and university reputation (Jensen, 2011), it is of little surprise that there has been significant research focused on understanding drivers of student retention and the development of models to predict student attrition (de Freitas et al., 2014).

In the experience of the authors there are number of challenges in the development and use of predictive models of student attrition.

- The rigorous experimental conditions that are desirable for the development of predictive models are difficult to achieve (many of the proposed drivers of attrition change simultaneously).
- There is a complex time consideration, it can be difficult to assess the exact time of attrition, and indeed a typical attrition scenario is identified only when students fail to re-enrol.
- The drivers of the attrition are broad and varied as are the demographic backgrounds and aspirations of students, consequently the functional dependencies of models on gathering and handling of data can be complex.
- Even when predictive models are available the outputs are not easily understood by support staff and planning staff, due to the applicability of predictions within a given timeframe, current institutional processes, and the role of increasing information in evolving the predictability characteristics of the modelling approach

Here we report on the use of the self-organising map technique, both its predictive ability and its utility in communicating potentially complex information about a student population to non-technical staff responsible for support and intervention planning services.

Problem Definition

In their interactions with the majority of higher education institutions, students typically access two types of services; academic (e.g. lectures, library materials and journals, tutorials, examinations, grading etc.) and supporting services (e. g. administration, counselling /advisory services, facilities, social services etc.). Additionally, each learner brings a number of demographic attributes (e. g. age, social economic status, prior aptitude for the subjects selected etc.). It is the goal of the education

provider to understand the dependencies between demographic attributes and the academic and support services they offer (or could potentially offer) and design interventions, actions and policy to optimise a desired outcome such as retention. One obstacle to optimising outcomes is a holistic understanding of the broad student population – also known as high dimensionality in the data – consisting of factors such as the variety of their sociocultural, psychological and historical characteristics and how these interact with their current intentions, daily patterns of private and social behaviour and academic performance. A well-established approach to understanding large high dimensional data sets is Kohonen's Self Organising Map (SOM) (Kohonen, 1990).

This section reviews the SOM technique before providing the specifics of our programme. A Kohonen model consists of input vectors $V = \{v_1, v_2, \dots, v_i, \dots, v_m\}$ with $v_i \in \mathbb{R}^n$ and a Self-Organised Map M ; a lattice of vectors $M = \{m_{i,j}\}\, with\, m_{i,j} \in \mathbb{R}^n$. M defines a mapping $f: V \to M : f(v) = m_{i,j}\ iff\ d(v, m_{i,j}) = min\{d(v,m), m \in M\}$ with d a metric function on $\mathbb{R}^n$, taken to be the Euclidean metric for our purposes here. M is calculated according to the algorithm below:

1. Randomise map M (a common heuristic is to evenly spread lattice vectors across the plan spanned by the first two principle components of V)
2. Randomly select input vector v_i and compare to each m to find the lattice point most similar to the input vector (i.e. $m_{i,j}$ such that $d(v, m_{i,j}) = min\{d(v,m), m \in M\}$).
3. Update lattice points in a neighbourhood of $m_{i,j}$ such to increase the similarity of the lattice points to v_i according to $\Delta m_{k,l} = n_0\, exp(^{-t}/_{\tau})\, exp(^{-S^2}/_{2\sigma(t)^2})$ where S is the distance between lattice sites and σ is a monotonically decreasing function usually taken to be $\sigma(t) = \sigma_0\, exp(^{-t}/_{\tau})$
4. If t is less than the maximum number of iterations increase t and return to step 2.

Applying the mapping f to the input vectors produces a 2 dimensional representation of the higher dimensional data set where similarity of vectors relates to lattice separation (with the most similar input vectors mapped to the same node). Colouring nodes according to a component $m_{i,j}$ produces a visually intuitive way to explore data.

The goal of the study was to generate profiles of students likely to attrite by combining a large amount of known data from a number of university systems and to engage the stakeholder community in exploring the data, understanding the systems of the university and apply their creativity to generating new interventions, actions and policy to improve retention.

Model

Parameters

The selection of 200+ fields from ten data systems in the university was prioritised based on the ease of data access and the perceived importance determined by interviewing a number of subject matter experts at the university. A consultation and engagement process with students, instructors and leaders from all areas of the university was undertaken to broaden the base of understanding of attrition and retention, surface the mental models of a wide range of stakeholders concerning their concepts and assumptions about potential drivers and leverage points in the system, and to ensure that the results of the project were visible to as wide as possible a group of concerned and active participants. Details of this process have been published in internal reports as well as briefly described in (de Freitas et al., 2014).

Based on the consultation process, over 200 hypotheses were created and evaluated (Gibson & de Freitas, 2015) which shaped the choice of factors based on fields in the data systems (Table 1) through a hybrid approach of *human shaped machine learning* in a series of cycles of consultation and data mining. Prior to applying the self-organizing map technique, the research team followed the typical processes of data mining to collect, clean, transform, and conduct exploratory analysis in an iterative process that resulted in the refinement of data models and algorithms before, during and after the SOM technique is applied and re-applied. We can think of the exploratory process as a series of mappings, refinements and re-mappings, from raw data to meaningful indicators for use in creating *M* as defined above. *M* is then optimized for stakeholder consumption, via visualizations, and

interpretive communications of findings and musings concerning a relevant subset of 50 hypotheses from the original 200+. Some hypotheses do not have indicators (yet) in the data systems and cannot be addressed by data mining, and some were superseded by a result from an earlier finding making further analysis pointless.

The SOM stage of the process is an example of *unsupervised machine learning* that is, once the data is made ready, computational resources explore and organize the data without human intervention until a data model 'settles' (converges to a solution in the form of a map representation). The map can then be further queried, manipulated and explored by stakeholders working alongside the data science team.

Table 1. Data sources

Data Source	Domains covered
Student Enrolment System	Student demographic information including: • Age • Country of birth • Gender Student University Performance • Unit and course enrolment, changes and cancellations • Unit performance • Graduation status Pre-university measures • Previous institutions attended • Admissions method (direct applicants, school leaving examinations, existing tertiary qualifications etc.)
Learning management systems	While the learning management system potentially contains a variety of pertinent domains, due limitations on time and complexities associated with extracting data, only log information (time of day) was included.
Library Computer Weblogs	Library web logs revealed indicate when a student accesses the library computer system and whether the access is from a university owned computer
Survey Data	Students take a number of surveys during their time at the university results from the following surveys are included †: • Unit satisfaction • University Facility Satisfaction • Course satisfaction.
High School Leavers Applications	High school students in the universities geography apply through a third party entity owned by public universities. Each university has visibility of all student applications in a given year and so it was possible to identify whether a student had a higher preference for a competing institution.
Card Access System	Students carry electronic cards which they can use to access facilities outside of normal hours. Logs of these cards can be used to track student usage of these facilities
Australian Bureau of Statistics	

After sourcing raw data from the above systems the authors combined the data into a single data set to take advantage of the SOM method to explore for trends in the high dimensional data set. For each domain it is not known *a priori* which features of a given domain are correlated with attrition and retention (e.g. no hypothesis is put in a privileged position) and so for each domain, multiple possible features are created by grouping, transformations, and other methods that combine business intelligence from the expert consultations with data and information expertise. For example from the learning management system weblogs, multiple features are possible based on which semester, the time of day of access and comparisons to the student's cohort (i.e. students in the same course with a similar proportion of the course completed). Examples include:

- In the first semester of their final course what was the most times in a day the student logged into blackboard
- In the first semester of their final course what was the average times in a day the student logged into blackboard
- In the first semester of their final course how many times did the student log into blackboard
- In the final semester of their final course what percentage of login attempts were made in the morning (7am – 12pm)
- In the second last semester of their final course, compared to their cohort, how does this students usage compare, on a directional scale, for login attempts
- In the second last semester of their final course, compared to their cohort, how does this students usage compare, on a directional scale, for login attempts in the afternoon 1pm – 6pm

Continuing in this manner 95 middle level features were generated from the learning management weblog data. Applying a similar approach the data from the 10 systems that were sourced for the single dataset, 1,273 attributes per student were derived. These features have been called *n-grams* and *motifs* when derived from dynamic, highly interactive digital learning experiences, and *meso-level* (the raw data are called *micro-level* features and the systems that encompass and act as exogenous influences on these features are call *macro-level* features or factors). See (Gibson & Jakl, 2013; Gibson & Webb, 2015; Shum, 2011).

Status Definition
Since there are multiple possibilities for defining when attrition occurs it worth commenting on the definitions used in the model presented here. In an ideal scenario, students wishing to leave a course would inform student services, formally withdraw and complete an exit survey. Practically few students at this university follow such a procedure, many simply stop interacting with university (i.e. stop attending classes or services). We opted to assign a status based on students with active units. A student is considered to attrite if they fail to take any units at the university for two semesters after they were last enrolled in a unit, excepting of students who graduate after their last semester. At any point in time then students can be assigned a status based on the last semester in which they were enrolled in units

- **Current:** the student has taken units in the most recent semester
- **Graduated:** The student has completed their course in the last semester that they interacted with the university. Students enrolled in two courses that complete one course in the last semester they interacted with the university are considered to have graduated for our purposes
- **Attrition:** The student is not current or graduated and two or more semesters have elapsed since they last interacted with the university.
- **Probable Attrition:** The student is not current or graduated and one semester has elapsed since they last interacted with the university.

When developing a SOM for exploratory analysis it is often useful to consider modify the definition of the metric function d so that the distance is invariant to certain parameters (so that the resulting map does not cluster on these parameters.). In this instance we do not cluster on the statuses above, to avoid having different behaviour profiles collapsed together because they result in attrition, a desirable outcome is to determine if there are different profiles associated with attrition.

Scope
Students analysed were undergraduate students that studied at least one unit on-site at the universities main campus between 2009 and 2014. Two data sourcing activities took place between one post semester 2014 and post semester 1 2013, in order to understand what movements across the map frequently occur.

Results

Map Overview
An underlying behavioral demographic map was generated using the commercial package Viscovery

to perform the SOM analyses, the resulting hexagonally packed map contains 1200 nodes (approximately square at 33x35 nodes). A modified Ward clustering algorithm (Batagelj, 1988; Murtagh & Legendre, 2011) takes into account the values of each input vector point as well as their positioning on the map and sets the distance between non-adjacent nodes to infinity (ensuring the clusters are connected regions in the lattice). We have broken the resulting map into 8 clusters (Figure 1) which can be thought of as representing 8 profiles of students.

The Ward algorithm can be used to divide the map into an arbitrary number of regions; eight regions were chosen to assist in socialising the map with users. With over a thousand parameters that can be viewed against the map, limiting the visualization to eight clusters assisted stakeholders in accessing information, creating meaning and developing insights from the map by generating an underlying easily-understood demographic profiles for non-technical users.

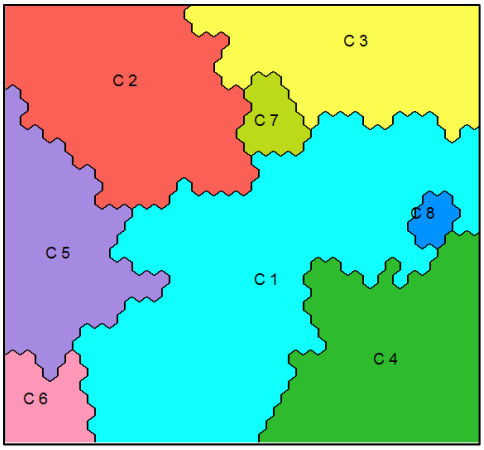

Figure 1. Eight clusters determined the Ward algorithm

When describing the clusters (or any subset of nodes) the mean value of parameters can be calculated and compared to the mean of the total map (or any other cluster) using a standard t-test. Categorical parameters such as country of birth are transformed into binary (0 or 1), in which case the mean on those parameters for any node or cluster is the proportion of students in that category; proportions are compared by considering the whether the Wilson intervals (Yan & Su, 2010) of the two values overlap within a given confidence. In this way regions can be described by parameters that make them 'most different' from the rest of the map. By way of an example some of the key demographic information for regions C1, C4 and C6 are given respectively in Tables 2, 3 and 4, along with examples of descriptions that were used in familiarising users with the map.

Table 2. Domestic near-graduation student cluster

Cluster Description C1 (n=14,995) Predominantly domestic students that have either graduated or are close to the end of their course in the most recent enrolled semester, slightly higher than average performance than other demographics.			
Parameter	**Mean / Proportion**	**Cluster mean difference from input mean (%)**	**Confidence** (mean is different from mean of entire set)
Citizenship is Australian	83.0%	15.0	>99.9%
Percentage of units taken in first semester at university are level 2	14.7%	-16.7	>99.9%
Percentage of units taken in first semester at university are level 3	5.7%	27.4	>99.9%
Percentage of course complete in final semester Curtin	66.6%	23.9	>99.9%
Students Graduated	46.9%	43.5	>99.9%

Age at Course Start	22.98	4.3	>99.9%
Course Weighted Average	64.06	9.0	>99.9%

Table 3. International near-graduation student cluster

Cluster Description C4 (n=8,434) International students that have either graduated or are close to the end of their course. They are distinct from C1 students in that they are typically taken a high number of level and level 3 units in their first semester of their course.			
Parameter	**Mean / Proportion**	**Cluster mean difference from input mean (%)**	**Confidence** (mean is different from mean of entire set)
Citizenship is Australian	5.4%	-92.5%	>99.9%
Percentage of units taken in first semester at university are level 2	50.4%	185.7%	>99.9%
Percentage of units taken in first semester at university are level 3	11.1%	149.7%	>99.9%
Percentage of course complete in final semester Curtin	62.9%	14.5	>99.9%
Students Graduated	56.8%	73.8%	>99.9%
Age at Course Start	21.8	-0.9%	>99.9%
Course Weighted Average	59.52	1.2%	>99.9%
Attendance mode External	0.02%	-90.3%	>99.9%

Table 4. Domestic external study mode student cluster

Cluster Description C6 (n=2,006) Domestic students that are significantly more likely to be taking an external study mode (to be in scope a student has to have taken at least one unit on campus, however the majority of external mode course have a small number of on campus components). On average students are older when commencing their course.			
Parameter	**Mean / Proportion**	**Cluster mean difference from input mean (%)**	**Confidence** (mean is different from mean of entire set)
Citizenship is Australian	94.9%	31.5	>99.9%
Percentage of units taken in first semester at university are level 2	19.9%	12.7%	99.5
Percentage of units taken in first semester at university are level 3	2.5%	-44.5%	>99.9%
Percentage of course complete in final semester Curtin	42.3%	21.3	>99.9%
Students Graduated	21.2%	-35.2	>99.9%
Age at Course Start	30.39	37.9	>99.9%
Course Weighted Average	52.27	-11.1%	>99.9%
Course: Attendance mode: External	50.5%	1,950.3%	>99.9%

For clarity we have compared only three of the eight clusters and selected a small number of parameters. In practice stakeholders are engaged in a series of workshops where considerable time is spent providing granular descriptions of each cluster, including areas of study, unit loads, past educational attempts, method of application and acceptance into courses, method of payment, and

other factors, in order to query the data model, test assumptions and understandings, and uncover or discover new relationships worthy of additional investigation or re-entering into the iterative model-building process.

Risk Profiles: Typical vs A-typical Risk

The SOM is not inherently a binary predictor (i.e. it doesn't assign likelihood of a particular outcome). Instead, in order to define an 'at risk' profile we consider areas of the map where there are a large proportion of students with the status 'attrition'. It is important to note that since a student can also either have the status 'current' or 'probable attrition' there are areas on the map where few students have status 'attrition' or 'graduation'. In the SOM these areas are largely concentrated in the top left of the map and overlap segment C2 and C5 (see Fig. 2 and Fig.3).

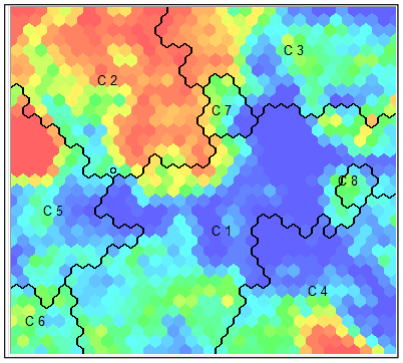

Fig 2. Current students: Colors represent the proportion of current students (blue represents 0% and red 100% of students) mapped to a node.

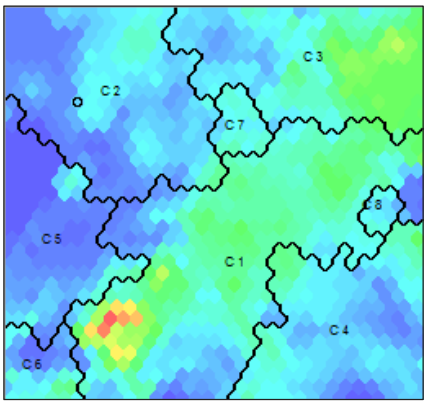

Figure 3. Semesters into course: Colors represent the proportion of current students (blue represents 0% and red 100% of students) mapped to a node.

Considering nodes where attrition is >40% identifies five connected regions larger than a single node, which we label R1 – R5, (Figure 4). It is reasonable to question whether occupying the same node as previous attrition students is indicative of likelihood of future attrition since by definition students that attrite are separated by two semesters from those that are current. To address this question we have taken two point-in-time data extracts (data slices or snapshots). We found that after 18 months the proportion of attrition for current students from these nodes is [32.01, 36.22] (99.9% CI) compared with [8.18, 8.81] (99.9% CI) for the entire map.

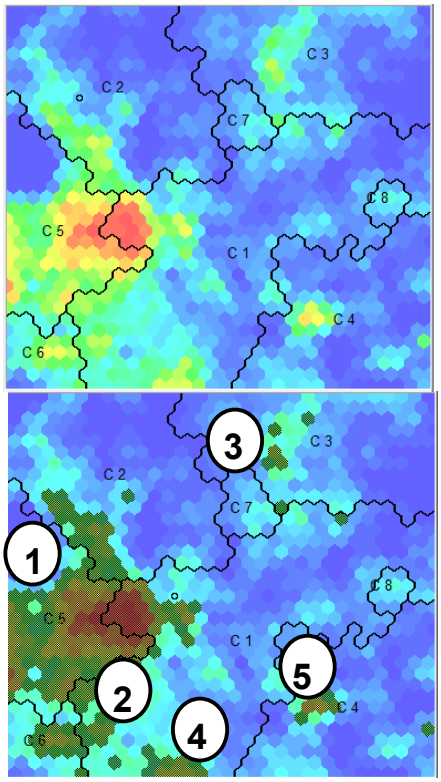

Figure 4. Attrition Rate: (Top) Colors represent the proportion of current students (blue represents 0% and red 100% of students) mapped to a node (Bottom) Five regions of the map with >= 40% attrition

Of the five regions we consider region 1 to be associated with what might be classed “typical attrition” as it aligns with common hypotheses of many subject mater experts. The students in this region are domestic students; males slightly over represented) studying full-time in on-campus courses, and generally taking between 3 and 4 units a semester, which is typical for the entire population. They live slightly further from the university than average and access library and learning management systems less often. They are significantly more likely to have failed units in their first and last semesters. Interestingly, while unit evaluation surveys response rates are lower than average, those students that do respond generally do so positively. When we compared region to 2 to region 1 we found those students to be generally older, more likely to be female and studying part time either externally or online. They access library systems almost exclusively outside of Curtin. Despite similar risk profiles; (Attrition Proportion: R1: [65.9, 70.0] (99.9% C.I.) and R2: [55.6, 68.3] (99.9% C.I.)) the proportion of units failed differs significantly in students first semester. (R1: 42.1% and R2:27.6% T = 8.19). This suggests that resilience to poor performance in part time students is potentially lower, this insight is important for designing targeted interventions; for example, the threshold for reaching out to a such a student will need to be lower.

Conclusions and Comments

We have demonstrated the use of the Kohonen self-organizing map (SOM) technique for approaching the multifaceted retention and attrition challenges in higher education. The approach outlined here is innovative for two reasons; the first is the utility of the visual element in communicating results to stakeholders and decisions makers. In this hybrid approach, an exhaustive set of hypotheses are collected from stakeholders, exploratory analysis takes place with appropriately sourced big data and the results are iterated with stakeholders as well as data scientists. The iterative exploratory analysis process investigates a large number of hypotheses by supplying evidence that clearly supports or challenges the stakeholder’s assumptions and understandings, making easier the often difficult process of translating untested qualitative and heuristic knowledge into testable quantitative models, and onward to the creation of interventions, actions and policy.

Secondly the approach is as broad as the sensor net of incoming and available data affords. Multiple

and varied domains of student behaviour can be analysed in a holistic manner. These behavioural domains range from a student's engagement with university systems, attitude towards the quality of the pedagogy received, academic engagement and performance and a number of external factors. . The SOM approach has been shown to successfully identify multiple profiles of student attrition, creating new more nuanced risk profiles by separating behaviours originally thought to belong to a single profile as well as creating whole new classes of profiles
SOM is not inherently a predictive technique in contrast with logistic models analysis and binary classifiers; but is effective for understanding the characteristics of a total population, identifying complex atypical clusters of behaviour and supplying other modelling approaches (e.g. linear regression, machine learning predictive techniques) with cohorts that have a high coherence among factors suitable further investigation. We have shown that SOM has potential to be combined with statistical and predictive analyses to form a complementary set of techniques for understanding the factors of retention and attrition for the purpose of developing new highly targeted interventions, actions and policy.

Future research is planned to test the impact of the definition of attrition to see if the historic at-risk status based on the 2 semesters missing (we waited three semesters to analyse the data) is truly at-risk and whether the factors can lead to predictive estimations before students leave.

References

Batagelj, V. (1988). Generalized Ward and Related Clustering Problems Ward clustering problem. In *Classification and Related Methods of Data Analysis* (pp. 67–74).

Crosling, G., Heagney, M., & Thomas, L. (2009). Improving student retention in higher education Improving Teaching and Learning. *Australian Universities Review*, *51*(2), 9–18.

De Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., … Arnab, S. (2014). Foundations of dynamic learning analytics: Using university student data to increase retention. *British Journal of Educational Technology*. doi:10.1111/bjet.12212

Gibson, D., & de Freitas, S. (2015). Exploratory Analysis in Learning Analytics. *Technology, Knowledge and Learning*, (March), 1–15. doi:10.1007/s10758-015-9249-5

Gibson, D., & Jakl, P. (2013). *Data challenges of leveraging a simulation to assess learning*. West Lake Village, CA. Retrieved from http://www.curveshift.com/images/Gibson_Jakl_data_challenges.pdf

Gibson, D., & Webb, M. E. (2015). Data science in educational assessment. *Education and Information Technologies*, *June*. doi:10.1007/s10639-015-9411-7

Jensen, U. (2011). *Factors influencing student retention in higher education*. *Research & Evaluation*. Retrieved from http://www.ksbe.edu/spi/pdfs/retention_brief.pdf

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*. doi:10.1109/5.58325

Murtagh, F., & Legendre, P. (2011). Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *arXiv Preprint arXiv:1111.6285*, (June), 20. Retrieved from http://arxiv.org/abs/1111.6285

Olsen, P. (2007). *Staying the course : Retention and attrition in Australian universities Findings*. Sydney. Retrieved from http://www.spre.com.au/download/AUIDFRetentionResultsFindings.pdf

Shum, S. B. (2011). *Learning Analytics*.

Yan, X., & Su, X. G. (2010). Stratified Wilson and Newcombe Confidence Intervals for Multiple Binomial Proportions. *Statistics in Biopharmaceutical Research*. doi:10.1198/sbr.2009.0049

Gibson, D.C., Ambrose, M., & Gardner, M. (2015). Self-organising maps and student retention: Understanding multi-faceted drivers. In T. Reiners, B.R. von Konsky, D. Gibson, V. Chang, L. Irving, & K. Clarke (Eds.), *Globally connected, digitally enabled*. Proceedings ascilite 2015 in Perth (pp. 112-120). https://doi.org/10.14742/apubs.2015.981

Note: All published papers are refereed, having undergone a double-blind peer-review process.