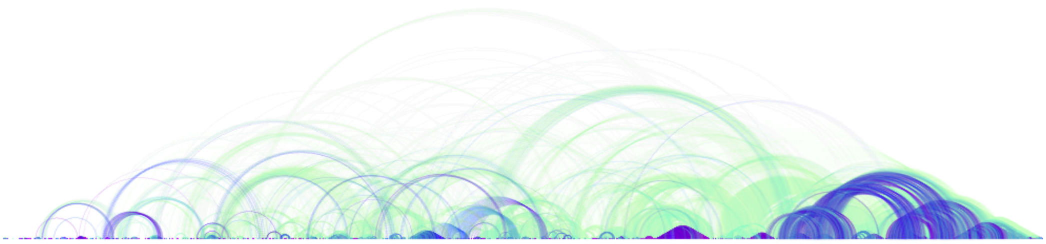
COMP5048 Information Visualization

Final Report

University of Sydney Course Explorer



Hanley Weng and Adam Gawronski

Semester 2, 2011

Table of Contents

Table of Contents

Abstract

The University of Sydney offers a diverse range of courses from over 20 different faculties and schools which cover a very broad range of academic disciplines. Because of the vast range of subjects on offer, it can often be difficult to choose subjects for study that not only hold a personal interest but also are almost most relevant to a person's study aims. We use interactive matrix and multilevel radial drawings in order to expose relationships between subjects and courses to investigate areas of mutual academic interest and interdisciplinary links, and help simplify the complex and large dataset used.

Introduction

Our visualization system is designed to facilitate future potential University of Sydney students with exploring, and deciding upon, available courses and degrees. It also serves to provide academics and other interested parties with a broad overview of inter-disciplinary links at the university based on the courses offered and their subject makeup.

This paper begins by exploring the data domain and exploring its unique characteristics. The core dataset is then analyse to discover the core relationships that will be focused in the development of the visualisation.

The final visualisation is then presented and critiqued. The domain is quite intricate and complex, so not all possible avenues for visualising the dataset could be investigated within the scope of the report; therefore the final section is devoted to possible ideas for future exploration and extension of the visualisations presented.

Domain

The University of Sydney offers a diverse range of courses from over 20 different faculties and schools that cover a very broad range of academic disciplines. Because of the vast range of subjects on offer, it can often be difficult to choose subjects for study that not only hold a personal interest but also are almost most relevant to a person's study aims.

The increasing popularity of students undertaking double degrees covering two separate disciplines, and the ability for students to undertake electives units has led to an increase in the diversity of subjects a student studies during their time at university and increasing cooperation between different faculties in the university. As more of these subject combinations become available, this added complexity. has also increased inter-department relationships. There is also a secondary need to visualize these for interdisciplinary purposes, which is a prominent area of interest in academia.

The main aims of this project however are to visualise the connections between different courses and subject areas. This would aim to reveal the strength of links between different parts of the university, both in terms of relevance between subject areas and administrative cooperation. The relative strength of these links could be used, for example, by students when deciding which double degree combinations to consider, and also which elective subjects could prove most relevant to their existing degree.

Data Set

Our core data-set is a database containing various information about subjects and courses available at the University of Sydney.

The dataset used originated from a university database containing information about the various subjects and courses the university has on offer. The database described which subjects were available to students of particular courses, as well as which faculties courses were administered by and other administrative information. The dataset was rather large - the SQL dump file representing the database was 147MB in size. The data is currently being used in a web page format by the university to display subject and course details in a relatively plain, web page format. Although such a format gives a good level of detail about a subject, it is difficult for the user to easily navigate and comprehend both the scale and variety of subjects on offer.

The full database contains information about subjects, courses, faculties, subject sessions, UAC courses, campuses, and areas of interest. It does not explicitly contain information about subject pre-requisites; although some pre-requisite information was contained in subject and course descriptions, it would have been very difficult to accurately convert this information into a useable format, so it was not used in the final visualisation.

Our visualization system will utilize a subset of this data. Namely, descriptions of Areas of Interest, Subject codes with titles and descriptions, and listings of the courses subjects are in (as one subject can belong to many courses, and one course can contain many subjects).

When considering both subjects and their parent courses as nodes, this subset of data has a scale free network structure. However our data set does not contain certain relationships, such as explicit information regarding subject pre-requisites, though such information can be text mined from subject descriptions in the database.

The data that was specific to our visualisation consisted of 1214 courses, 11,544 subjects and 1,478,061 links between subjects and courses. The database contained duplicate entries for certain subject and course combinations, as subjects were offered in multiple semesters over multiple years. For the purposes of our visualization, we wanted to ignore the duplicates, resulting in 268,320 subject/course links remaining.

Other dimensions in the data included additional subject information such as subject alpha code, descriptions, the faculty and department they belong to, as well as the number of credit points required for completion and other such ancillary information.

Clustering

Because of the large number of courses and subjects, several static methods of clustering were devised to assist in categorising the data.

As mentioned above, there were several dimensions in the data set that were suitable for clustering. For example, each course had a total credit point value that indicated the duration of the course and this could have been used as a clustering method (grouping courses together which require 48, 96, 144 credit points and so on). However if such a clustering mechanism were used it would likely group together courses from the full range of academic disciplines; this would confuse the visualisation and would no longer clearly show connections between academic disciplines.

Subjects consist of an 8-character unique identifier in the format of four alphabetical characters followed by four numerals. The four letters define a particular category of subjects, for example the “ACCT” prefix defines subjects related to the study of accounting, “INFO” defines subjects related to information technology and so on. Considering one of the goals of this visualisation project was to identify similarities in subject areas between different courses, it would be logical to assume that course that share subjects would share a number of subjects with the same prefix. Hence, clustering subjects by prefix could simplify the visualisation of relationships between courses, reducing the number of edges displayed at any one time (with a consequent reduction in edge crossings).

Courses have a similar natural grouping mechanism to subjects. Each course belongs is of a certain level (Bachelor, Masters and so on), belongs to a particular faculty or school, and requires a certain number of credit points for completion. The issue with these is that they do not group like courses together - Bachelor degrees for example span almost the entire array of subjects available at university; likewise, two courses which both require 144 credit points for completion could involve two completely different disciplines. Clustering such courses together would be non-sensical, even if it did produce a visualisation with good quality metrics, as it would not be convey much useful information to the viewer.

Courses do, however, have a 5-character identifier in the form of two alphabetical characters followed by three numerals. The first two letters group course by specific schools and disciplines within faculties; there are 178 different combinations of these at the University of Sydney. Hence, grouping courses by the first two letters of this attribute permits a reduction in the number of course nodes in the graph and consequent increase in the visualisation’s quality metrics without affecting readability or usefulness of the information presented.

Aims

The main target audience for this visualization system are future University of Sydney students, predominantly freshmen. We aim to support both potential students who already know what areas of study they are interested in, and those who are still exploring their options.

The initial aims of the visualisation were to create an interface that would facilitate the free exploration of courses and subjects alongside a hierarchical representation of course structures. Other aims encompass the expansion and collapsing of course and subject nodes to reveal more in-depth information; this zooming metaphor will also be extended to node clusters. The zooming metaphor was later changed to a simpler

Additional features include the weighted bookmarking and export of such nodes, as well as user-inputted-keyword nodes that are connected to existing nodes by similarity. Such features can act as information filters.

One of the original aims was to visualise subject prerequisites and hence try to expose potential inter-faculty or inter-departmental relationships between subjects in courses. However this information turned out not to be available in the dataset, so we couldn't pursue this direction.

- Purpose of relationship, what benefits do users gain from being shown these

Architecture

The system was comprised of two main components, namely the database which stored all the data for the visualisation, and the application that processed this data.

The database was a standard relational database, hosted in MySQL. Basic SQL queries were then used to extract the required data and export it in CSV form that the main application could then accept.

Originally, we intended the application be written with a data access layer so that it could dynamically query the database while it was running. A mini-framework was developed which was designed to convert the relational tables into Java objects that could be directly used and manipulated by the Processing framework. The problem with this, however, was due to the large amount of data being processed, keeping all the necessary objects instantiated proved to be too resource intensive to handle any sort of real time interaction.

At this point it was then decided that a more feasible method would be to export the required, formatted data in CSV format from the database ready for importation in the visualisation application. Such an approach also made sense from a rapid prototyping perspective as the same data could be easily imported into various graphing tools such as Gephi and Tulip for quick analysis.

This visualisation application itself was developed in Java, heavily utilising the Processing library for graphics rendering and manipulation. The Processing library is ideal for building rapid prototypes of possible visualisations as well to handle the drawing capabilities of larger scale applications.

Code Structure

After loading all the data, joining them to one another in the program.

Data iterated through and analysed to bring up certain values like predominant Subject Clusters for individual courses and course clusters. To generate visualisations on the fly, pre-rendering is done at program initialization. Both real time, on the fly visualisation as well as pre-rendered images with interactivity built on top.

Visualizations were compartmentalized into separate 'modules' so as to be able to handle touch points properly - in this way, the values of one could change without detrimentally effecting the other, yet still sending useful information to other visualizations.

The application allows for user interactivity, as it is able to track the mouse movements and actions of the user.

Tools

A wide variety of graphing programs, programming frameworks and other tools were used to facilitate the creation of the visualisation. Using appropriate tools allowed for rapid development of prototypes that were then evaluated and incorporated into the process of creating the final visualisation.

Analytics

Several quantitative and qualitative methods of analysis were used in the early prototyping stages to explore the data set and determine which subsets of the data would be most useful to fulfilling the aims of the visualisation.

The freely available interactive visualization and exploration platform, Gephi, was used to investigate the data set and computer simple quantitative visualisation metrics. Initial experiments included transforming different parts of the data set into a traditional node graph and running standard force directed algorithms to gleam some initial information about relationships. For example, all courses and subjects were imported as nodes, with edges linking courses and subjects. One of the better results from this process is shown below.

It was apparent that the traditional node graph did not satisfy many quality requirements so alternative visualisation metaphors would have to be investigated. The effectiveness of various clustering methods was also analysed using basic single-level arc drawings such as in Figure 2.

Several quantitative analyses were performed over different parts of the dataset to determine which could be used most effectively to fulfil the aims of the visualisation. Average degrees, graph density and clustering coefficients were calculated for variations of the data set to quantify quality metrics and measure effectiveness of various clustering techniques. Graph modularity was also investigated (Newman, 2006) however it was not particularly attributable to the data set.

Data

The relational database made it easy to perform certain basic data analytics, and also made it easy to export pre-processed data into CSV format for further analysis using Excel.

Excel was used for simple joins, ordering, sorting, and hence could quickly provide visuals to show to benefits and detriments of using different variables for different sorting's and groupings.

Rendering

Once a particular implementation was programmed it had to be rendered. Rendering operations could often a long time to show any results, and this initially proved to be a bottleneck during rapid prototyping.

Because of the portability of frameworks such as Processing, we were able to package possible renderings and distribute them across the various computers we had at our disposal.

These many renderings resulted in a collection of static screenshot images that we were then able to compare against each other, judging the relative qualities of the different images, then choose to pursue those which looked the most promising. We were essentially creating our own small visualisations of our dataset and it's attributes, visually judging the disparity and dispersion of nodes, edges, and or, pixels, to decide the parameters on which to optimize our own final visualisation.

Development of Visualisation

Initial Node Network

Node networks were initially chosen as the predominant metaphor for this visualization. However, further examination revealed that this was an inefficient choice.

The initial base metaphor for this visualization was a simple and free node graph. Due to the predominantly static data, fixed mapping metaphors such as terrain or space were considered as they could be effectively utilized to extend the theme of this visualization (Emden et. al., 2010). These same spatial algorithms would restrict the topology of the nodes.

In exploring this initial metaphor, we implemented our own force directed methods (Barnes, 1986), with various variations of Hooke's law, to do the initial mapping of the network (of courses and related subjects). We also used readily available software such as Gephi to apply different algorithms to the network, such as Yifan Hu's Proportional and Multilevel, and Fruchterman Reingold, methods.

Due to the size of the dataset, we had begun and evaluated the practicality of this metaphor at much smaller scales, and had intended to solve the scaling problem with interactivity. However, parsing the total size of the dataset was impractical in terms of complexity and resolution. Also a visualisation shouldn't be using vast amounts of interaction to compensate for such resolutions when alternatives would suffice.

As the use of node networks for our data set were too impractical, other metaphors were explored, along with ways of simplifying the data. Throughout the following processes, we continued to evaluate the practicality of the node network metaphor for different clusters and sets of data to little success due to their density, and other metaphors were more appropriate.

Matrices and Data Exploration

In discarding the node network as the predominant metaphor, the data was explored for other relationships. Gephi also contributed to this process, allowing the quick visualization and evaluation of graph attributes (such as density), and relations. The effective import of the data into a standard relational database allowed for complex querying and export of data. Exported data (commonly in CSV format) could also be processed in spreadsheet software such as Numbers, and Excel to perform more detailed functions, such as mathematical functions, sorting, grouping and ordering. This software also allowed for quick visualizations and statistics on the current relationships and data-sub-sets we were assessing (such as, subject-to-course relationships grouped by different clusters).

Not only were spreadsheet and visualization software utilized in assessing, parsing, and analysing our data, our own java implementations were created for specific purposes. One significant tool that we utilized was a matrix of pixels ('matrices') comparing one set of data to itself based on some external metric (such as, subject-to-subject relationships based on shared courses).

Such a tool utilizes the idea of 'mapping' to better retain users' mental models of a visualisation, ideals that were carried on from our initial exploration of our node network.

Matrices aided us in quickly assessing the visual complexity, disparity, and clusters of different data sets formed by different comparison-metrics and orderings. They were a form of information visualization that allowed us to improve our own information visualization that we were developing.

Eventually, the matrices became such a useful tool that it was decided that one of them would be used in the final visualization.

User Testing

User tests were important to evaluating and developing our visualization, similar to other visualisations (Plaisant, 2004). Many frequent, iterative informal user tests were conducted to assess our evaluation, with specific regards to mental mapping and comprehension. The feedback received helped us to improve our developing visualization, and fed back into itself.

Certain features that were significantly contributed by user testing were interaction metaphors, and the choice of data sets and metrics to represent in order to optimize the visualisation's usefulness for users. Other features included general aesthetics, visual comprehension, and the development of keys. Arc drawings were one of the visualizations that were decided on by this, in terms of initial visual attraction to users.

Level Drawings

In addition to visualising data between data in the same set (e.g. course-to-course, cluster-to-cluster), a dissimilar relationship also required visualising (e.g. course-to-subject). As such, level drawings were also generated.

Combined with matrices, these drawings helped simultaneously evaluate the appropriateness of different datasets and their attributes (such as ordering).

Cross Minimization was utilized in minimizing edge crossings between sets of data, with the BaryCentre method being used. The final sorting (after some initial ordering) is based on applying this method to clusters, than grouping the courses by course clusters, then applying this to subject clusters. This happens a few times, and finally, subjects are sorted by their clusters and courses by their clusters.

Through these various evaluations, outlined here and above, the final clustering decide on for grouping clusters were via 'Faculty', and subjects were clustered by their 'Alpha' codes. Visualisations were centred on representing course-to-subject relationships, with such clusterings.

Eventually a level drawing was decided on to be used for the final visualisation, with a course_cluster-to-subject_cluster relationship, depicting individual courses and subjects per node). The final level drawing was transformed to a concentric level drawing to more aesthetically represent the data, and to, with the aid of interaction, better visually represent edges.

Interaction

User tests were a major asset in contributing to the establishment of natural interaction metaphors, and consequently, the successful augmentation of mental models.

Interaction metaphors were initially primarily created for each visualization (arc, matrxi, line representations), to allow diving into information. Through iterative user tests, these evolved into interaction design elements that were connected across visualisations (for example, the cluster selected in a matrix would be simultaneously represented on a radial drawing). This evolution arose from observations of users augmenting their mental models of the total visualization through simultaneous understanding of the different visualizations (multiplied by the feedback loop provided by testing).

The presence of such cross-visualization interaction elements allow for users to establish mental models of the different visualizations, their individual relationships, and the relationships formed by combining visualizations (for example, a course-to-course-similarity-comparison in one visualization, could be mentally mapped to a cluster-to-course relationship in another).

As with all good information visualizations, this visualization allows the telling of a narrative. Having trouble representing sufficient data for users to form their own links with the visualization, we summarized values of the data based off user interaction (e.g. mouse hover over specific pixels), in the form of message, whoms visual change, in correspondance with user interaction, allows the user to relate the visualization with it's data. The message represents statistics, and lists of the content of information interacted with, directly reflecting the visualization.

Final Visualisation

The final visualisation depicts four different types of data; Faculties, Courses, Subject Alphas, and Subjects. Faculties help cluster Courses, whilst Subject Alphas cluster subjects. Three sub-visualisations are used, these are a two level concentric-level drawing, a matrix, and an arc drawing. Interaction and text are also used to tie all the visualisations together nicely.

The radial drawing depicts Faculties in it's inner circle and Subject Alphas on it's outer. Each edge in between these go from a Course inside the inner circle (faculty), to a Subject on the outer circle (subject alpha). The clusters determine unique colours of each data set. The inner-circle's ring of colours represent individual faculties, and the same applies for any edges originating from the inner-circle. The outer-circle's ring of colours represent individual subject alphas. Due to the higher quantity of subjects than courses (similarly with their clusters), subjects were placed on the outer-ring, allowing for better resolution. (There are 320 iterated subject alphas, compared to 20 faculties. Similarly, there are 11544 iterated subjects compared to 1048 courses).

The Matrix represents a course-to-course relationship, with comparisons based off a similarity metric that is determined by the number of shared subjects as a ratio to their total existing subjects. Examples of matrix representations can be seen in Figure 3 and Figure 4 in the appendix.

Each square (or 'pixel') represents a "Course A" to "Course B" comparison. "Course A" is determined by the pixel's x-value, and vice versa for course B (for example, as "Course A" and "Course B" equal to each other along the diagonal, the diagonal is quite evident). Due to the way the pixels are coloured, and how the courses are ordered (which is the same to the radial drawing), clusters of courses sharing subjects are quite evident. The Course A to Course B comparison also produces an 'almost' mirrored diagram, which makes identifying differences much easier (such as the university-wide study abroad program, whom relationships can be seen along the horizontal roughly 80% down the vertical of the matrix).

An arc drawing is also attached at the end of this diagram, visualising an average of this comparison metric between courses. It serves as a nice attractive visual addition, representing existing information in a condensed (average), and in a different manner. An alpha channel is also mapped to the metrics values here to reduce visual complexity.

Two keys are utilized for this total visualisation. One set of colours are mapped to a cluster (as seen in the radial drawing's inner circle and at the sides of the matrix). These colours have high saturation, brightness, and differ by hue, allowing for easier recognition. Another key set of colours are the similarity metric representing number of shared subjects. These are utilized in the arc drawing and the within the pixels of the matrix. Hue is limited to just under half of the spectrum as this is metric is a linear relationship (ranging from 0.0f to 1.0f), saturation similarly is mapped from a low to high value, and inversely with brightness. The selection of such colours allow for more of a user's focus to be lead to higher values of the metric (dark purple), from lower levels (light green), in any sub-visualisation.

Text is utilized to convey a message depicting the information a user has selected. In this instance, it tells the user how a course compares to another in terms of similarity. It shows the courses' faculties and the percentage shared and it displays their shared alpha codes and shared subjects. Like the other sub-visualisations, this text is not intended to exist on it's own, but rather as a whole with it's effective data representation mapped to what the user is observing in the other visualisations.

Interaction is a large part of this visualisation. Visualisations are connected, in that, viewing any element in one, also represents its viewing in the other. The radial drawing can have clusters selected by hovering over them in relation to the it's centre, simultaneously, this will register the radial' graphs' hovering course, which will be represented in the matrix, and it's information displayed in text. The same happens vice versa. The cross-interaction of multiple forms of visualising information effectively allows the user to establish mental maps across visualisations representing different types of data and relationships.

As with all good information visualizations, this visualization allows the telling of a narrative. Having trouble representing sufficient data for users to form their own links with the visualization, we summarized values of the data based off user interaction (e.g. mouse hover over specific pixels), in the form of message, whom visual change combined with user interaction allows the user to relate the visualization with it's data. The message represents statistics, and lists of the content of information interacted with, directly reflecting the visualization.

Evaluation

Throughout our development and tests, we evaluated our system on functionality, effectiveness, efficiency, usability and usefulness.

Functionally, the visualisation allows users to quickly get an overview of the connections between different courses and subjects, then further investigate particular areas of interest by hovering over the relevant section.

This visualisation improves previous methods of accessing and evaluating this data. Previously, the data was presented on an individual subject and course basis, thus interested parties could only recognise connections through tedious exploration. This way, multiple dimensions of data can be seen quickly with further information derived just as fast.

The visualisation is quite efficient, with most users comprehending the data quickly upon interaction. The connections between sets of data (faculties, courses, and subjects), have clusters visualised instantaneously whereas additional information can be acquired from the selective 'diving in' of data. The end result is that academic disciplines with stronger links appear more prominently.

The system has good usability. A very simple hovering mechanism, shows contextual information based on what section user is investigating. Text display shows detailed information about particular selection. Extremely approachable, and on testing, quite quickly comprehendible in terms of interaction metaphors and content.

The visualisation is useful for both students and academics alike. Students can use it to find subject areas that share a strong connection with those they are currently studying or considering studying, better informing their choices when considering which elective subjects they wish to pursue or when considering. Members of academia can explore other areas that may be similar to theirs and act accordingly in collaboration and promotes further research.

Further exploration

There are several ways the visualisation could be further improved.

Due to the large size of the data-set, additional filtering out of parts of the graph would reduce the complexity of the on screen image improving a user's understanding. The visualisation could take advantage of greater user customisation to enhance usefulness. For example, ideally the system could have access to the student's current university course enrolment, or have some knowledge of their interests (from the user's online social networking data for example). The system could then use such information to pre-filter results.

Many of the quality metrics of the image could be improved if the data on-screen could be further filtered based on what information would be determined to be useful to the user. Dynamic clustering could be implemented which would use heuristic algorithms to determine which subjects / courses to cluster as opposed to the static-type clustering currently implemented. This could work in tandem with filtering in order to optimise the resulting image based on the remaining information, rather than still being clustered by predetermined algorithms which were optimised based on the entire data-set.

The detailed text information that is displayed when a user hovers areas of the graph could be improved with web links to full subject descriptions

Such a visualisation and its corresponding data warrant further exploration and evaluation, especially testing in it's usefulness in prolonged university scenarios amongst academia.

References

Barnes, J., Hut, P., (1986) A hierarchical O(N log N) force-calculation algorithm. Nature, 324(4), December 1986.

Bateman, S., Mandryk, R.L., Gutwin, C., Genest, A.M., McDine, D., Brooks, C. (2010). Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts. In ACM Conference on Human Factors in Computing Systems (CHI 2010), Atlanta, GA, USA. 2573-2582.

Eades, Huang (2000). Navigating Clustered Graphs using Force-Directed Methods, Journal of Graph Algorithms and Applications vol. 4, no. 3, pp. 157-181

Emden R. Gansner, Yifan Hu, Stephen G. Kobourov. GMap: Visualizing graphs and clusters as maps. (2010). In Proceedings of PacificVis'2010. pp.201~208

Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), Usability Inspection Methods, John Wiley & Sons, New York, NY.

Plaisant, C. (2004). The challenge of information visualization evaluation. In Proceedings of the working conference on Advanced visual interfaces (AVI '04). ACM, New York, NY, USA, 109-116.

Newman, M. E. J. (2006). "Modularity and community structure in networks". *Proceedings – National Academy of Sciences USA* **103** (23): 8577-8696

Crossing minimization in extended level drawings of graphs

Christian Bachmaiera,, Hedi Buchnerb, Michael Forster Seok-Hee Hong

Appendix

Below are several images listed in the report.

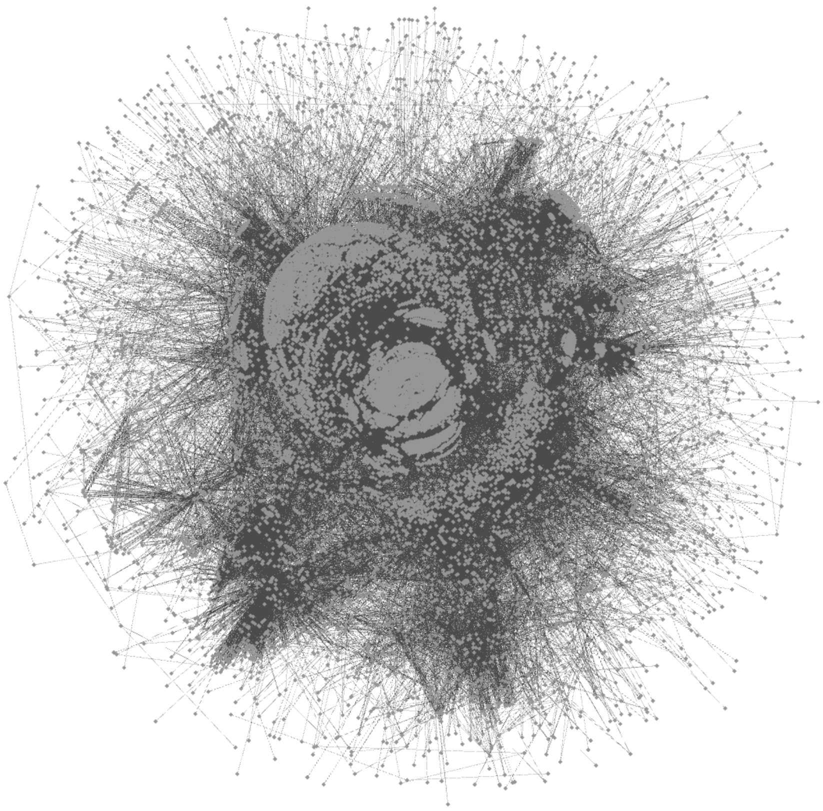


Figure 1. Force-directed graph showing all subjects and courses in the data set as nodes, with edges connecting subjects contained in courses

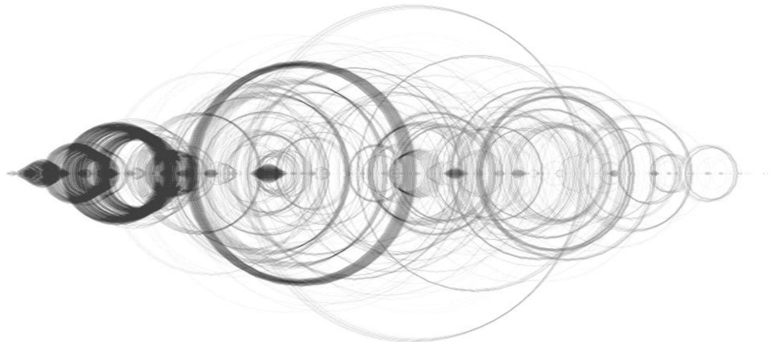


Figure 2. Single-level arc diagram of an example course clustering, with the thickness of arcs corresponding to the number of shared subjects between course clusters

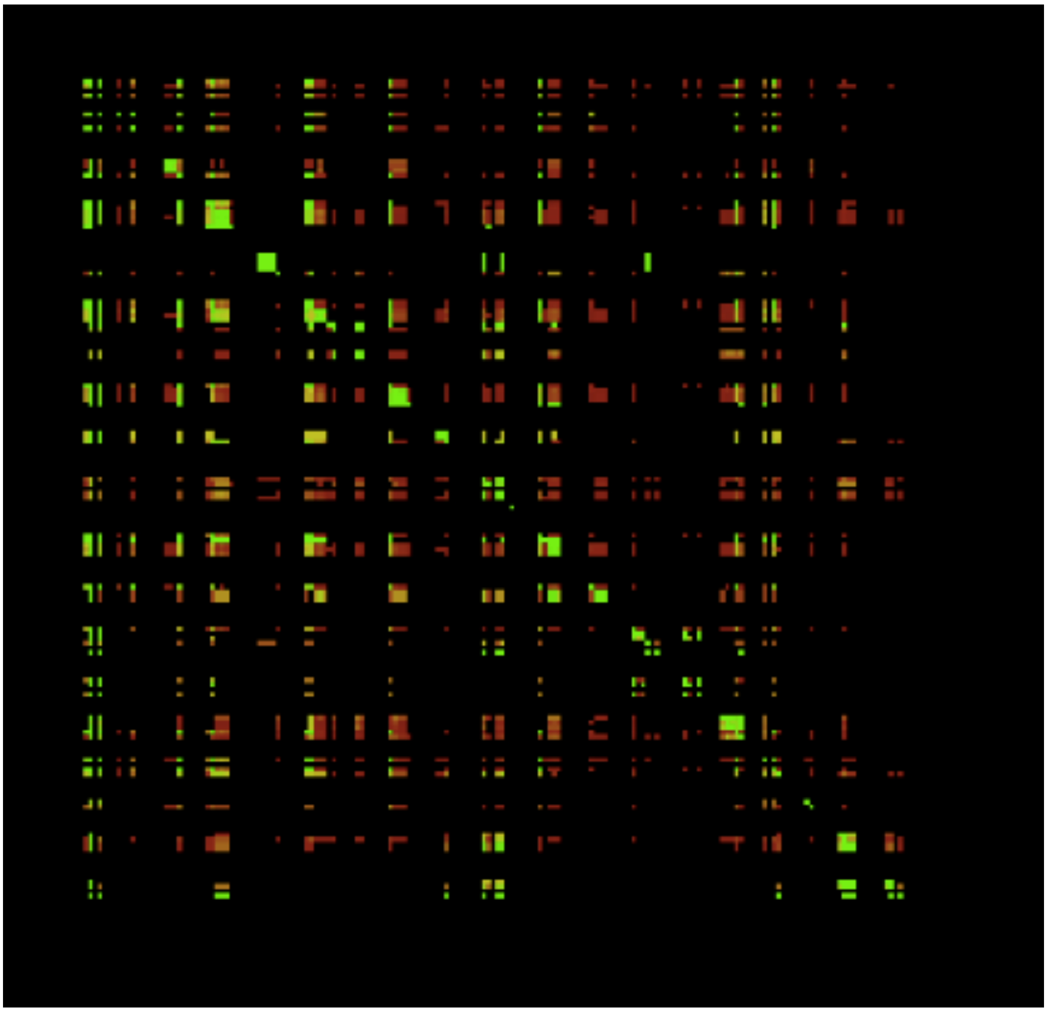


Figure 3 – Matrix representation of number of shared subjects between courses, with courses grouped by department prefix. Green areas indicated high numbers of shared subjects, darker red areas show weaker links.

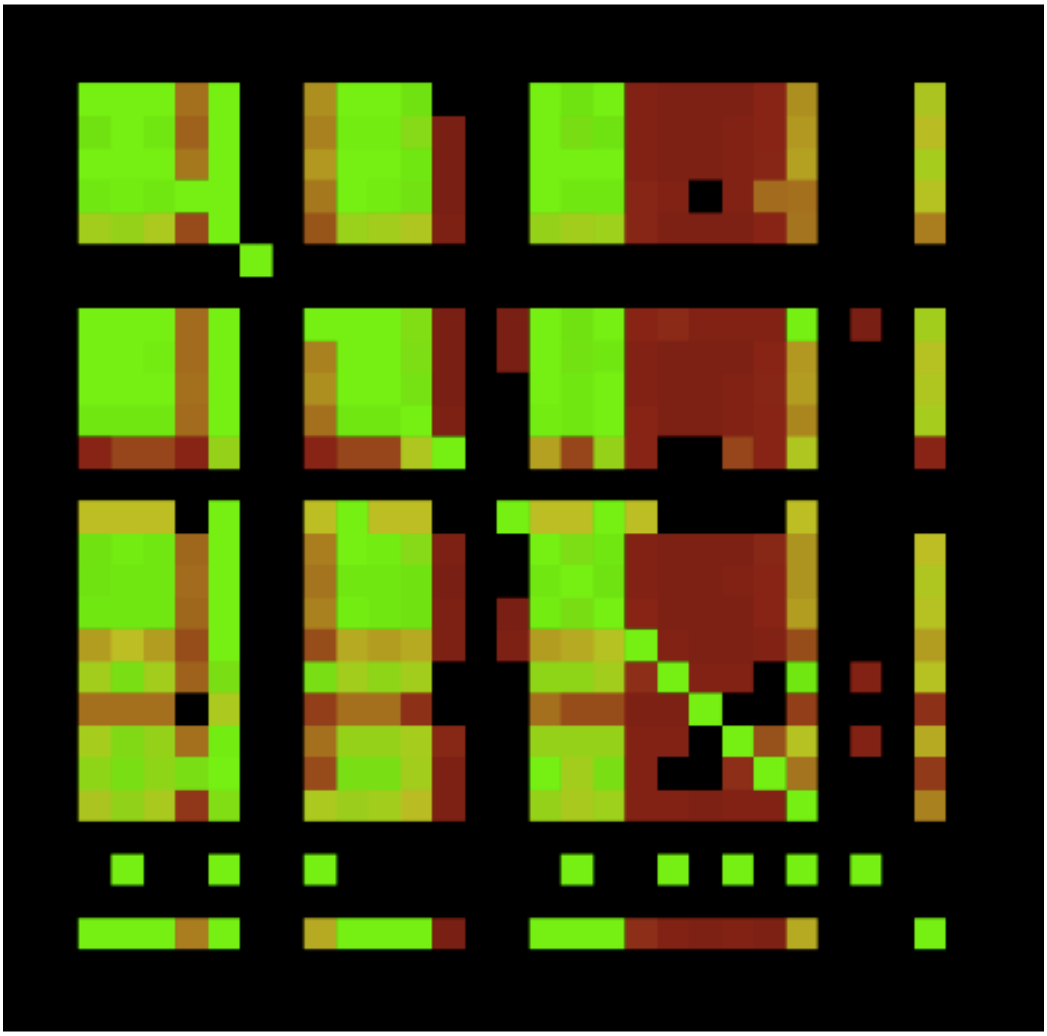


Figure 4 – Matrix representation of number of shared subjects between courses, with courses grouped by faculty. Green areas indicated high numbers of shared subjects, darker red areas show weaker links.

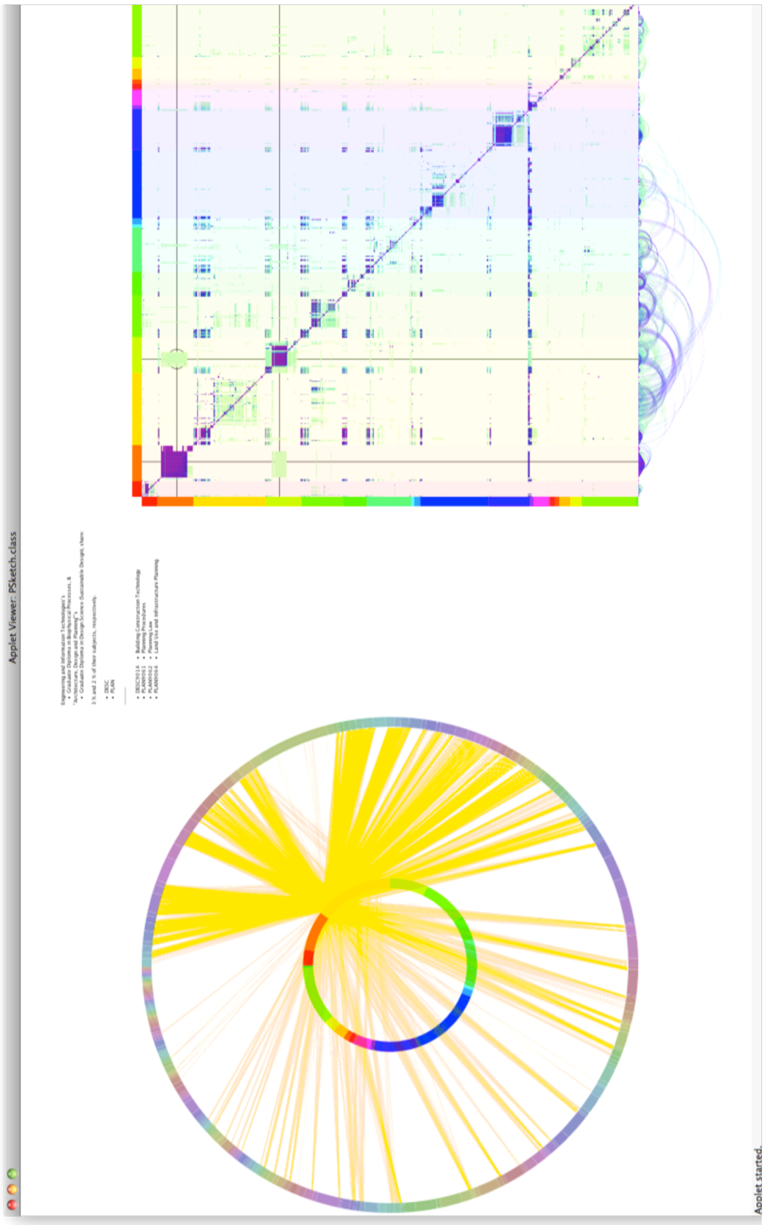


Figure 5 – Screenshot of final visualisation depicting interactive radial drawing and interactive matrix.