Uppsala University

Introduction to Computational Social Science

1DL007

Web Scraping and Network Analysis of Science and Technology Programmes at Uppsala University

Authors:
Stina Brunzell
Theodora Moldovan
Ida Nilsson

March 15, 2024

1 Introduction

Using unobtrusive data collection methods, we collected data on the Bachelor's and Master's programmes offered by the Faculty of Science and Technology at Uppsala University. This was followed by a network analysis to map out the connections between programmes based on shared courses. The goal of the analysis was to visualise how programmes are interlinked through common courses, illustrating their similarities. A community detection algorithm was applied to further identify groups of similar programmes.

Network analysis of university courses has previously been conducted to identify courses deemed to have a crucial impact on student progress [1]. The experiment used data from the University of New Mexico (UNM) and revealed that the cruciality of courses is distributed according to a power law. Moreover, the UNM framework was extended to study the complexity of curricula within universities. These results suggest that studying a network of courses can be beneficial when compiling a programme or evaluating courses crucial for graduation rates.

2 Methods

Unobtrusive methods allow researchers to collect data without interfering with the subjects under study. This type of data collection refers to collecting data that has initially not been intended for research and re-purposing it [2]. The data used for our project was collected using one such method, namely web scraping. In scraping, the HTML structure of individual web pages identified by a Uniform Resource Locator (URL) is used to extract metadata and content with the help of regular expressions. The university's website is structured such that the landing page of each programme includes a link to the course outline, which in turn holds a list of all courses within the programme. The data were scraped from parts of the website that are not explicitly disallowed by the university's robots file. The scraping was performed responsibly to not overload the website's servers or collect personal data without permission.

Network analysis was performed with the scraped data, constructing an undirected network where the nodes are programmes, with edges representing shared courses. In weighted graphs such as the one in our study, connections can be more or less strong, indicated by the weight of the edge, a numerical quantity. Here, weight has the meaning of proximity — the edge is the strength of the relationship between two programmes.The degree is the node's most important and basic feature, representing the number of edges connected to it. It tells us how well connected and how structurally important the node is [3]. Nodes not connected to any other nodes in the network are called isolates.

We performed micro-level and meso-level analysis on our network. Actor prominence at the micro level looks at a node's degree centrality — the number of ties connected to a node, betweenness centrality — the number of shortest paths, such that the sum of the weights of the edges is minimised, that pass through the node, and eigenvector centrality — the relative score measuring the influence of a node connected to other important nodes [4]. At the meso level, we take a look at network communities. Communities in complex networks are groups of nodes densely connected to each other and sparsely connected to the rest of the network. They are one of the most common mesoscale organisations of real world networks [3]. There are two main types of community detection algorithms: overlapping and not overlapping. The Louvain method for community detection is a type of non-overlapping algorithm which quantifies the quality of a community structure by measuring the density of connections within communities relative to connections between communities [5]. The method focuses on maximizing modularity, a measure of the density of links inside communities compared to links between communities. For example, an isolated node or a very small group of nodes with no or minimal connections to others might form its own community if including it in a larger community would decrease the overall modularity.

3 Results

87 programmes within the Faculty of Science and Technology were collected using the *rvest* package in R. The name of each programme, as well as its URL, number of credits, course code and level were extracted from the filtered courses and programmes search results [6]. Preparatory programmes and

programmes including preparatory courses were filtered out programmatically. The Sino-Swedish Master in Computer Science programme was removed due to not having courses included in its outline. The Bachelor's Programme in Leadership included a duplicate entry in the search results, which was also removed. The 300-credit version of the Master's Programme in Materials Engineering was renamed Integrated Master's Programme in Materials Engineering to differentiate it from the 120-credit programme with the same name. Similarly, the 60-credit Master's Programme in Wind Power Project Management was changed to Magister Programme in Wind Power Project Management to distinguish it from the 120-credit Wind Power Project Management programme.

The individual programme page URLs collected in the previous step were further scraped to extract the location, teaching form, pace of study, language of instruction and the latest course outline URL for each programme. Each outline was scraped to collect information about the courses available within the programme. 2,605 observations were collected, including the course name, number of credits, course code, main field(s) of study, and programme in which the course is included. The Master of Science Programme in Biology has several specialisations characterised by a starting course and one or several other profile courses. These programmes do not have outlines linked on their individual programme pages, but rather have a common web page detailing programme content [7] on the Biology Education Centre page, from which the courses were collected and merged with the previously scraped data to bring the total number of courses collected to 2,648[1].

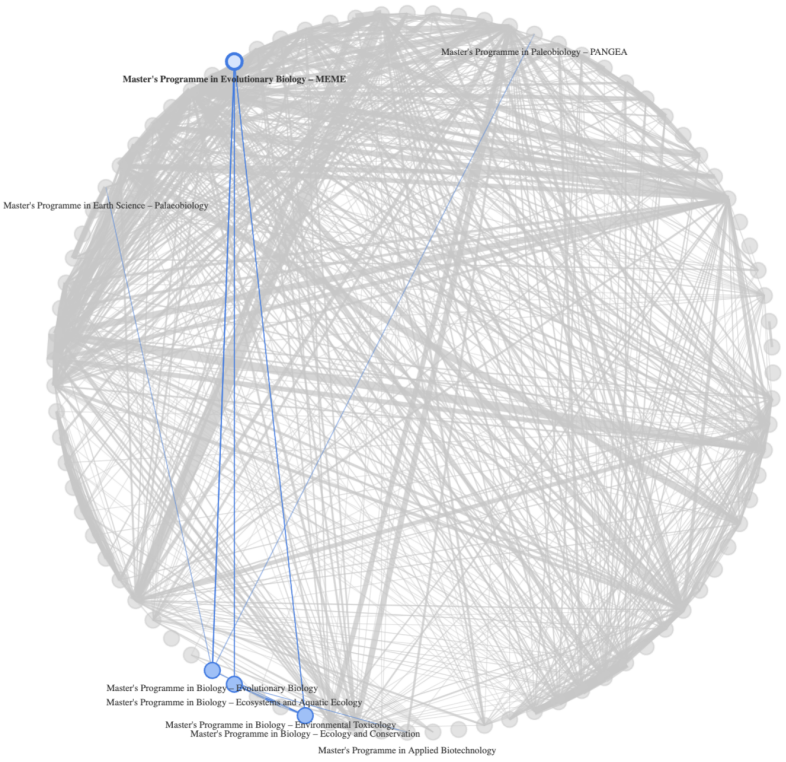

Figure 1: Selection of Master's Programme in Evolutionary Biology – MEME node.

An adjacency matrix and further network analysis identified central programmes, with the Master's Programme in Computer and Information Engineering and Master's Programme in Computer Science sharing the most courses, namely 53 shared courses. The micro level analysis was conducted using the *igraph* package. Table 1 and Table 3 identify the Master's Programme in Engineering Physics as the node with the highest degree and eigenvector centrality respectively. It can be observed that nodes with higher degree and eigenvector centrality are mostly programmes among the field of engineering. The Bachelor's Programme in Biology/Molecular Biology has the highest betweenness centrality (Table 2). Among the programmes with the highest betweenness centrality are cross-disciplinary domains.

[1]Collected data, R scripts and interactive network are available at `https://github.com/todomoldovan/uuprogrammes`.

The Louvain community detection algorithm identified 9 non-overlapping groups in the network. The first community comprises of programmes related to sustainable engineering, as can be seen in Figure 2. Figure 3 shows the second and most well-connected community of programmes related to computer science and mathematics. The third community is made up of Master's level physics and earth science programmes (Figure 4). Chemistry and biology programmes form the fourth community illustrated in Figure 5. Bachelor's programmes in physics compose the community in Figure 6. The add-on Bachelor's programme in Nuclear Engineering, a highly-specialised 60-credit programme can be seen in Figure 7. Figure 8 shows a very small community composed of the Magister and Master Programmes in Wind Power Project Management. The eight community is comprised solely of the Master's Programme in Sustainable Destination Development (Figure 9). Figure 10 showcases another isolate, the Nordic Master in Biodiversity and Systematics.

An edge to-from list was constructed from the adjacency matrix to aid in the creation of a circular graph visualisation using the *visNetwork* R package. Figure 1 demonstrates some of the features of the interactive network: the label representing the Master's Programme in Evolutionary Biology – MEME is in bold and the node is highlighted, while the nodes it is directly connected to are also highlighted. In the interactive plot of the network, it is possible to select either a particular node (programme) to display, as described above, or a community, as shown in Figures 2-10. Moreover, nodes can be dragged and dropped to better visualise connections between nodes of interest to the user.

4 Discussion and conclusion

Web scraping was the only available option to obtain data of interest for the project. Our data source is Uppsala University, which offers open and available access, and no personal data is used. As the data used in the analysis is simply informational, there is little risk for ethical implications.

Our analysis focused on programmes within the Faculty of Science and Technology for a couple of key reasons. Firstly, to maintain the scope of our project at a manageable level, as including all programs from across Uppsala University would have made the project overly broad and less effective in achieving our goal of mapping out a network of shared courses, especially since courses typically do not span multiple faculties. Secondly, the Faculty of Science and Technology offers the largest number of programs, significantly more than any other faculty, providing a rich dataset for constructing an extensive network analysis.

Classical statistical methods assume independence between entities, which is not the case in our collection of programmes. Network analysis, and community detection in particular, allowed us to identify connections between the programmes. The network created provides a comprehensive view of science and technology programmes, offering a valuable resource for prospective students as they decide which program to apply for. By highlighting the overlaps and similarities between programs, this visual tool can aid in the decision-making process by broadening students' horizon. The visual representation of these relationships, especially through interactive graphs, provides an intuitive and engaging means for users to explore the data. Central to our findings was the identification of key programs that serve as hubs within the network, demonstrating high centrality and thus, indicating their pivotal roles in the faculty's academic ecosystem. This insight, coupled with the detection of distinct communities within the network, offers a nuanced understanding of the faculty's structural organisation.

The Master's Programme in Engineering Physics was found to be the top programme by both degree and eigenvector centrality. It is described by Uppsala University as a "broad and genuine engineering education with the opportunity to specialise" [8]. With seven different profiles spanning from hardware for embedded systems to sustainable energy technology to artificial intelligence and to quantum technology, it is not astounding that the Master's programme in Engineering Physics has many connections to programmes in other fields than physics. Overall, the Master's programmes in engineering are highly represented in the top of the centrality measurements. This is likely because they have a similar structure to the Master's Programme in Engineering Physics, having different orientations for the students to choose from after completing the compulsory courses of the programme. In addition, the engineering programmes are all 300 credits, including both Bachelor's and Master's courses and hence, they are expected to have more connections. On the other hand, high betweenness centrality highlights programs that serve as bridges or connectors, fostering interdisciplinary links. The most central node in this regard, the Bachelor's

Programme in Biology/Molecular Biology, represents a connection between biology and chemistry.

The computer science community, in Figure 3, is the most well-connected which could be explained by the fact that analysing data is vital and hence, computer science is incorporated in many programmes. The isolate nature of the Master's Programme in Sustainable Destination Development and Nordic Master in Biodiversity and Systematics (Figure 9 and 10) can be explained by the teaching not taking place in Uppsala, but on the university's Gotland campus and in multiple universities in Scandinavia, respectively [9], [10]. The Bachelor's Programme in Nuclear Engineering is not a full bachelor's programme, but a 60 credit add-on leading to a degree, with a prerequisite of 2 years of study within a Bachelor's or Master's programme in electrical engineering, mechanical engineering or a corresponding set of courses [11]. Therefore the Bachelor's Programme in Nuclear Engineering appears to be an isolate node, while most likely sharing a common prior education to members of the first community (Figure 2), of which the Bachelor's programmes in Electrical and Mechanical Engineering are part of.

Formatting inconsistencies within programme outlines proved to be the biggest challenge, needing extensive data cleaning and some manual data entry for biology programmes, as described earlier. Plans to construct a directed network of courses were abandoned due to the lack of a consistent format in the listing of course prerequisites. It is also important to note that at the time that this project was conducted, the Faculty of Science and Technology was in the midst of restructuring the 7,5 credit curriculum into 5 and 10 credit courses. As a result, many course pages were not populated, and collected course data would have quickly become outdated.

While the data collected through the web scraping is extensive, many possible node attributes such as language of instruction, number of credits, and location are not included in our visual analysis. Future work can expand on building a 2-mode network of programmes and courses, integrating courses' main fields of study into the analysis, and incorporating additional node attributes. Furthermore, the number of shared courses as the sole metric for programme connection is an imperfect measure and points to the need for a more holistic approach that could incorporate qualitative aspects of programme alignment.

References

[1] Ahmad Slim, Jarred Kozlick, Gregory L. Heileman, Jeff Wigdahl, and Chaouki T. Abdallah. Network analysis of university courses. In *Proceedings of the 23rd International Conference on World Wide Web*, page 713–718, 2014. doi: 10.1145/2567948.2579360.

[2] Victoria Yantseva. Unobtrusive data collection methods, 2024. Lecture slides.

[3] Michele Coscia. *The Atlas for the Aspiring Network Scientist*. 2021. doi: 10.48550/arXiv.2101.00863.

[4] Emelie Karlsson. Social network analysis, 2024. Lecture slides.

[5] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

[6] Uppsala University. Courses and programmes. `https://www.uu.se/en/study/search?type=Programme&faculty=Faculty+of+Science+and+Technology`, . Accessed: 2024-02-29.

[7] Uppsala University. Programme content - biology education centre. `https://www.ibg.uu.se/education/master/biology/programme-content`, . Accessed: 2024-02-29.

[8] Uppsala University. Master's programmes in engineering. `https://www.physics.uu.se/education/master-s-programmes-in-engineering`, . Accessed: 2024-03-14.

[9] Uppsala University. Master's programme in sustainable destination development. `https://www.uu.se/en/study/programme/masters-programme-sustainable-destination-development`, . Accessed: 2024-03-14.

[10] Uppsala University. Master's programme in biology – nabis – nordic master in biodiversity and systematics. `https://www.uu.se/en/study/programme/masters-programme-biology-nabis-nordic-master-biodiversity-and-systematics`, . Accessed: 2024-03-14.

[11] Uppsala University. Högskoleingenjörsprogrammet i kärnkraftteknik. `https://www.uu.se/utbildning/utbildningsplan?query=2337`, . Accessed: 2024-03-14.

A Appendix

Table 1: Top 5 programmes by degree centrality.

Programme	Degree centrality
Master's Programme in Engineering Physics	51
Master's Programme in Molecular Biotechnology Engineering	45
Master's Programme in Energy Systems Engineering	44
Master's Programme in Computational Science	39
Master's Programme in Industrial Engineering and Management	39

Table 2: Top 5 programmes by betweenness centrality.

Programme	Betweenness centrality
Bachelor's Programme in Biology/Molecular Biology	566.0150
Master's Programme in Biophysics	372.1019
Master's Programme in Water Engineering	281.6670
Master's Programme in Molecular Biotechnology Engineering	199.2706
Master's Programme in Additive Manufacturing	199.1491

Table 3: Top 5 programmes by eigenvector centrality.

Programme	Eigenvector centrality
Master's Programme in Engineering Physics	1.0000000
Master's Programme in Computer and Information Engineering	0.8488721
Master's Programme in Sociotechnical Systems Engineering	0.7200518
Master's Programme in Computer Science	0.7091363
Master's Programme in Molecular Biotechnology Engineering	0.6235934

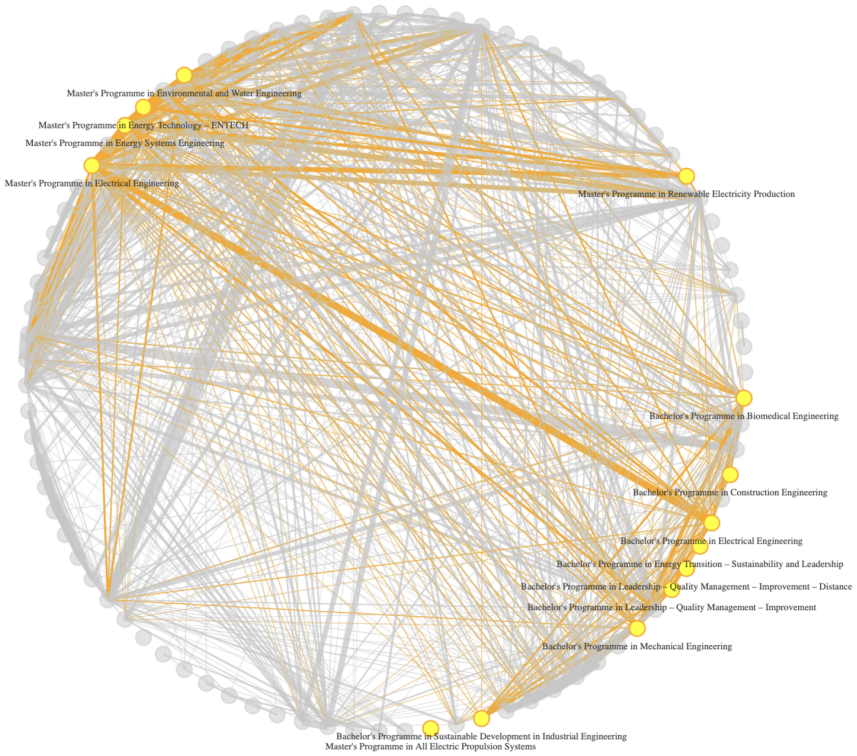

Figure 2: Community 1

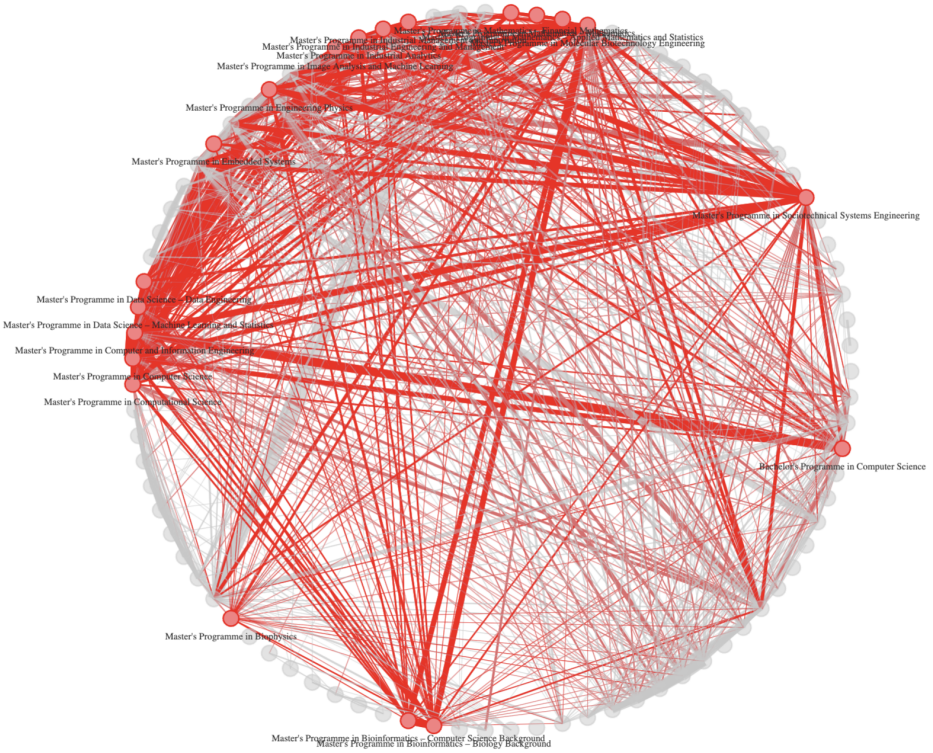

Figure 3: Community 2

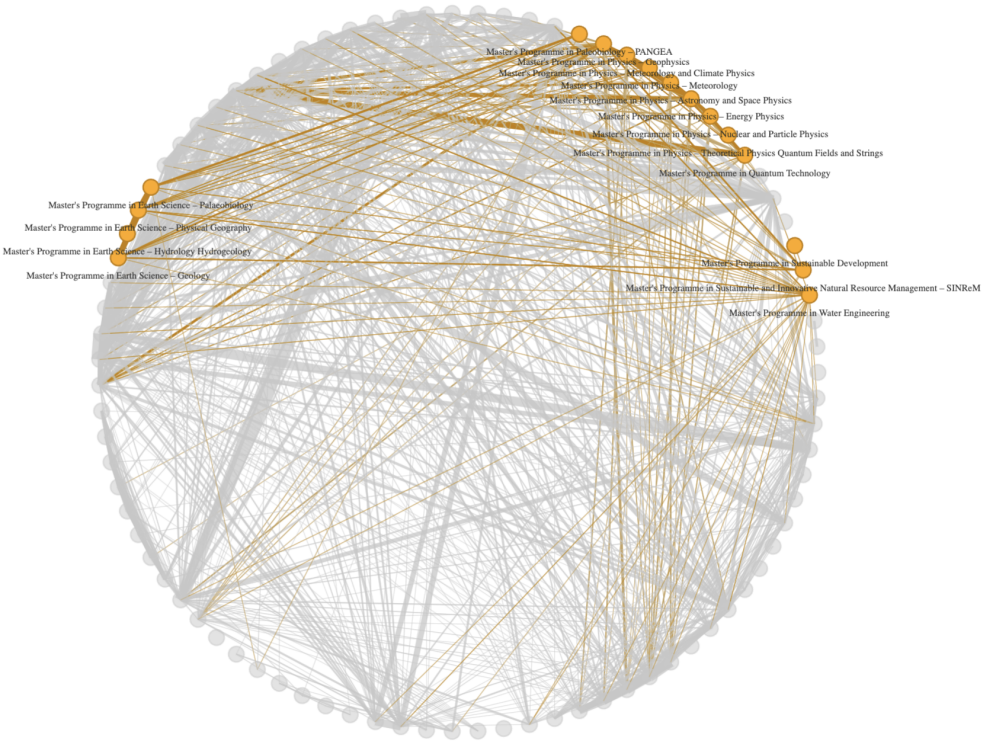

Figure 4: Community 3

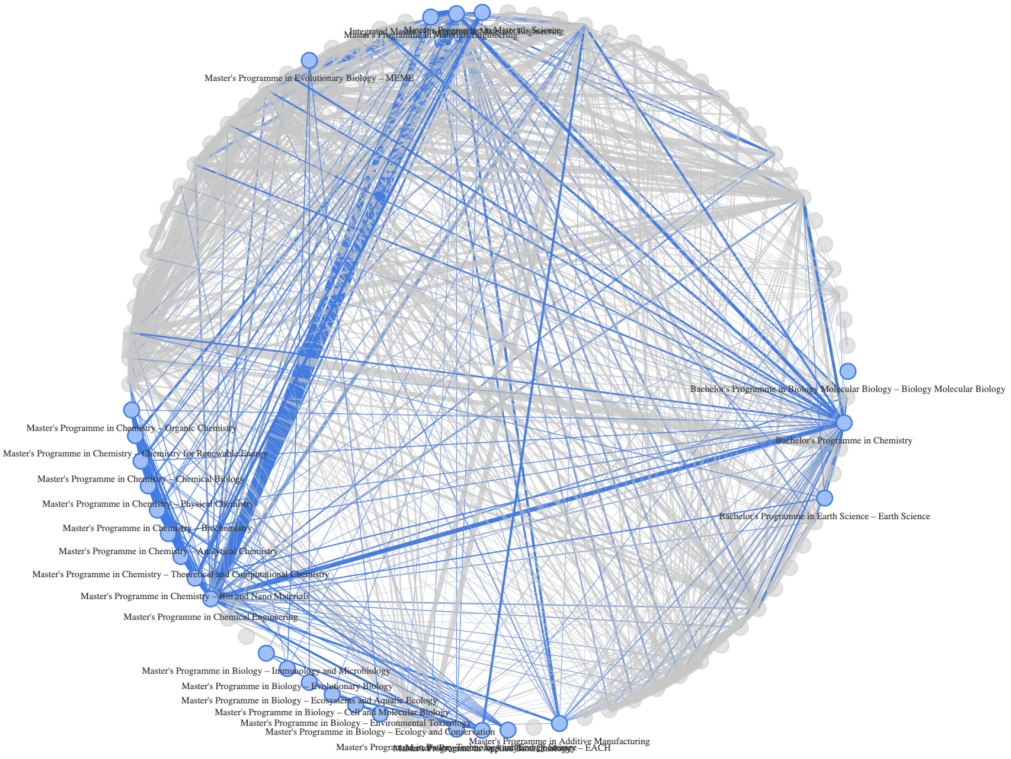

Figure 5: Community 4

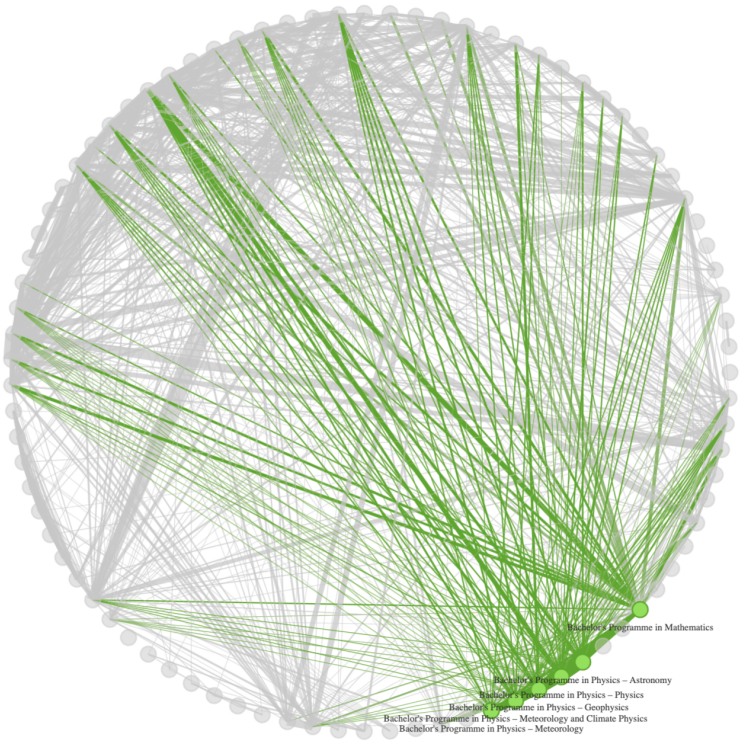

Figure 6: Community 5

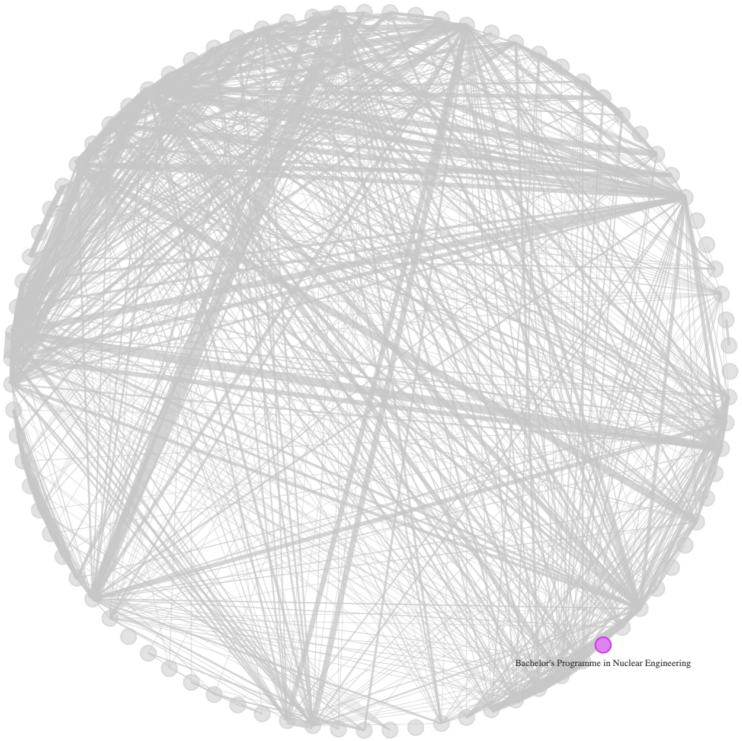

Figure 7: Community 6

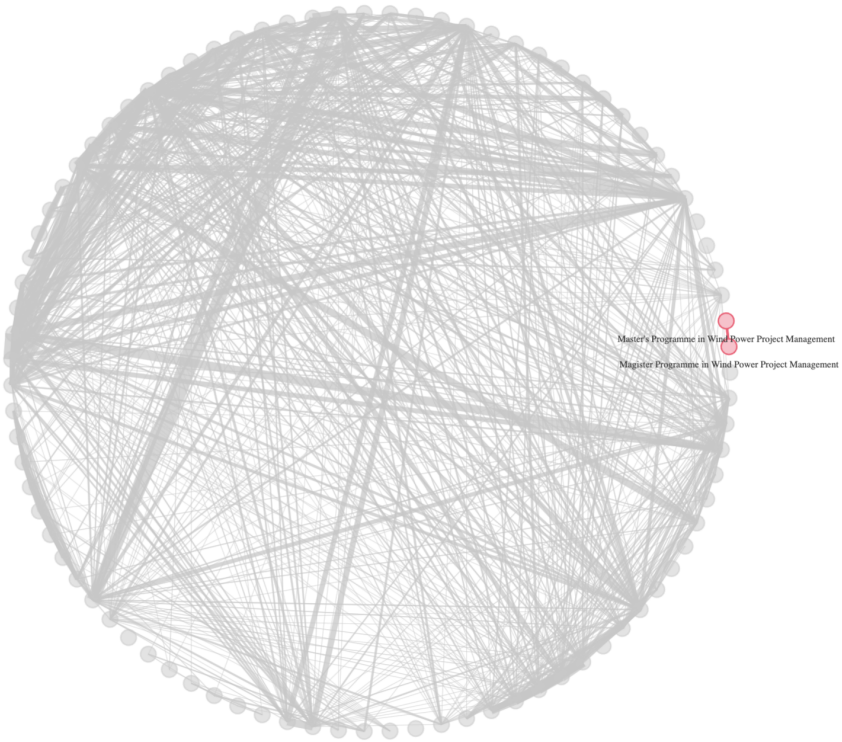

Figure 8: Community 7

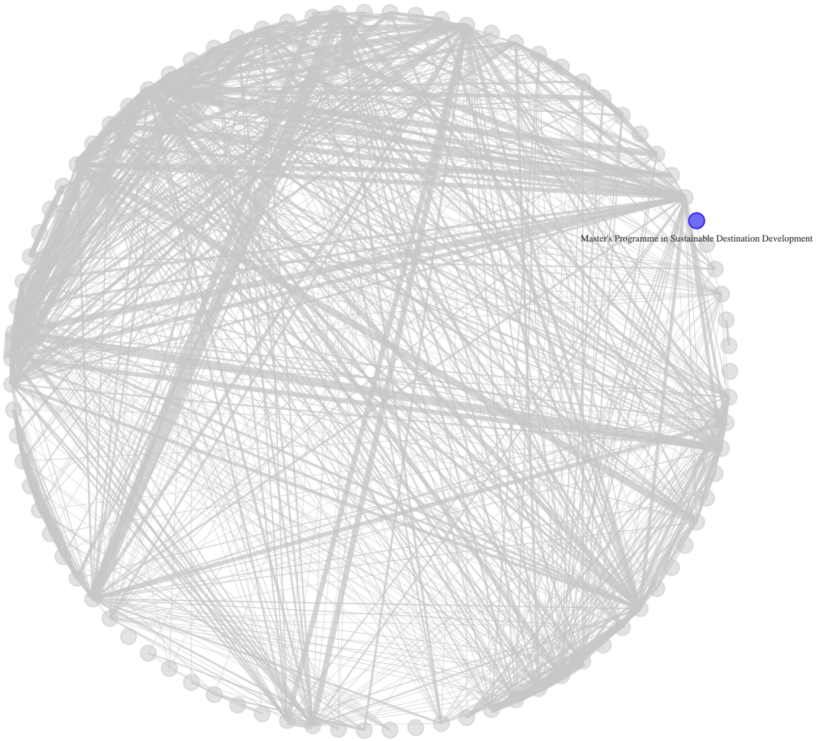

Figure 9: Community 8

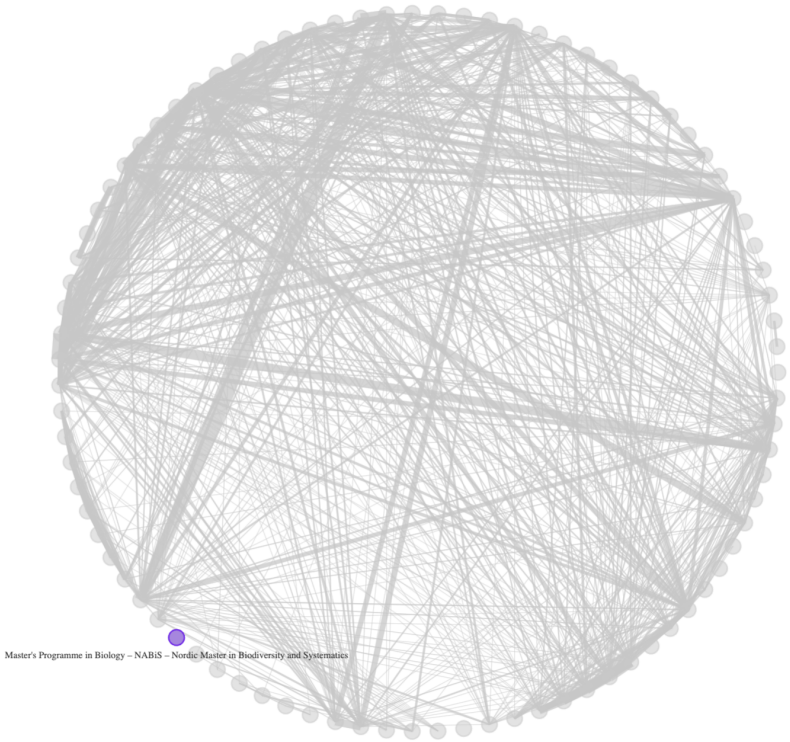

Figure 10: Community 9