Semantic Network as a Method for Visual Text Representation of Tacit Relationships

PRATIK SHRIVASTAVA and VAISHNAVI PADALA | NETWORK ANALYSIS

"Information is giving out; communication is getting through"

-Sydney J. Harris

We thank Prof. Jana Diesner and Shubham Mishra for helping us communicate.

1. Keywords

Semantic- network- text mining- topic- attribute - relatedness – covert-possibility

2. Abstract

Text is a common medium to convey information but is often not good in representing covert knowledge that requires understanding relationships between multiple items within and outside of the document. An example is that of a course catalog that lists short course descriptions. The intuition is that there is a network structure to the vocabulary used in this text and the relations among the vocabulary can form a network that helps in understanding the domain landscape better. These relations are not trivial for human cognitive processing, and thus we explore semantic network as a tool to aid comprehension. We also intend to identify through the network, any gaps in course offerings of the program with respective to predetermined research themes and vision of the program. As a result, we have been able to find difference in network interpretation vs user interpretation highlighting covert information. The results are discussed further below.

3. Introduction

The project aims to address the problem of comprehension of the landscape of a new domain in education through semantic networks for visual text analysis.

"Comprehension is a higher cognitive process of the brain that searches relations between a given object or attribute and other objects' attributes and relations in the long-term memory, and establishes a representational model for the object or attribute by connecting it to appropriate clusters of memory. It is recognized that although knowledge and information are powerful, before any information can be possessed and processed, it should be comprehended properly."[1]
(Wang,Y and Gafurov,D, 2003)

As per Wikipedia's, "A semantic network, or frame network, is a network that represents semantic relations between concepts. This is used as a form of knowledge representation. It is a directed or undirected graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between concepts."

We propose using a Semantic network to uncover hidden relationships among courses along various attributes to aid knowledge building and explorative analysis This helps comprehend the domain's landscape for better decision making of students and management [2]. Such awareness can provide a valuable competitive advantage in the industry. In addition, it opens up possibilities and connections to advance research and collaboration that one may not have imagined before.

A student's mental model for decision-making process in course selection involves multiple parameters like the class structure, tools taught industry relation, relevant job descriptions, professor's research, underlying concepts and themes, similarity and differentiation between courses. This process of understanding domain itself
Is crucial for career planning. The course catalog adheres to certain guidelines such as short description, minimum specifications and a domain specific or teaching specific description of courses. Thus gathering this initial information is not only time taking but also requires complex information seeking and communication skills, which may act as a barrier for entry and timely access of resources.
A person accrues such tacit knowledge by prolonged participation in the organization or domain. Particularly, international students may not yet be aware of the resources or cultural nuances of a higher education system in a different country. Beginners, career switchers or interdisciplinary students are all prone to the challenge. This calls for a clear, learner centric and timely communication of course information to ensure an equitable educational attainment. The problem also extends to prospective students' ease of comprehending the course offerings and possibilities in the program.

4. Background and Context

"The concept of semantic networks has a long history (Quillian, 1968) and opened up a basis for knowledge modeling and representation" [3]. Comprehension involves being aware of the metastructure of the content which requires that the vocabulary is presented in a way that aids this cognitive process of comprehension. Text structure can be visualized to show similar, contrasting, analogical concepts. We try to depict this structure in each course description to align with student's information seeking behavior.

This involves:

- An intersection of text mining, data modeling (ontologies) and network analysis, i.e Semantic networks and relational content analysis.
- Thematic analysis- Human coding of content.

These will help in identifying the themes that are conflicting, cooperating or augmenting each other among the courses. “Network Text Analysis (NTA)” describes such a concept in a paper by Professor Jana Deisner . “The technique is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them” (Sowa, 1984)” [4]

There has been significant research conducted in the area of citation analysis, media content analysis and digital humanities at the intersection of networks and text mining. [5, 6]. While there is a high incidence of social network and text analysis applications, there is relatively less applications and research about structure of data, particularly ontologies and semantics.

The two questions we ask in this context are:

- Are the course descriptions reflective of the pathways shown on the ischool website?
- In the present text descriptions, are students able to identify clusters of all possible similar or complementary courses

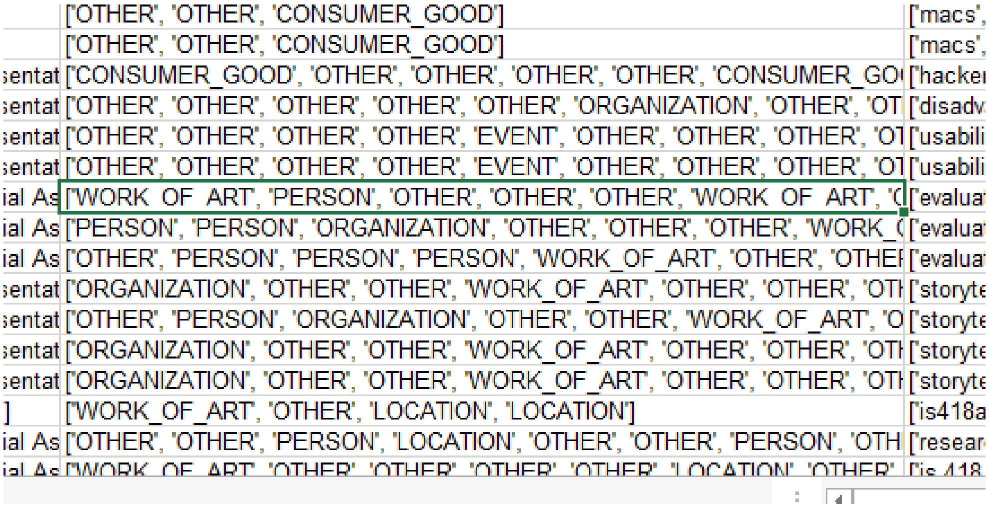
b. It has been identified that there exists research that deals with curriculum development and curricular topic study from a learner centric view based on keywords that align with student’s objectives rather than body of knowledge or teaching perspective [8].

5 .Data

Description:

The initial dataset with course descriptions data has been collected from the Spring 2018 Courses list and can be found here on GitHub. The dataset comprises of 146 rows containing, Course ID, Course Name, Course description and Instructors. The data was collected and cleaned manually.

Next, in our finalized approach, we used Google NLP to extract entities from the course descriptions, along with a column for automatic categories for the entities. However, upon examining the categories it is found that they are not relevant to the context of the network and the use case of course description. For Example, some categories shown are Person, organization, work of art, location, consumer goods and so on. Hence, we first defined 19 categories based on information found in sources like curricular topic study, curriculum development, and Lis and Analytics body of language (7). This combined dataset showing both Google NLP categories and human coded categories can be found here.



	['OTHER', 'OTHER', 'CONSUMER_GOOD']	['macs',
	['OTHER', 'OTHER', 'CONSUMER_GOOD']	['macs',
sentat	['CONSUMER_GOOD', 'OTHER', 'OTHER', 'OTHER', 'OTHER', 'CONSUMER_GO	['hacker
sentat	['OTHER', 'OTHER', 'OTHER', 'OTHER', 'OTHER', 'ORGANIZATION', 'OTHER', 'OT	['disadv
sentat	['OTHER', 'OTHER', 'OTHER', 'OTHER', 'EVENT', 'OTHER', 'OTHER', 'OTHER', 'OT	['usabili
sentat	['OTHER', 'OTHER', 'OTHER', 'OTHER', 'EVENT', 'OTHER', 'OTHER', 'OTHER', 'OT	['usabili
ial As	['WORK OF ART', 'PERSON', 'OTHER', 'OTHER', 'OTHER', 'WORK OF ART', 'O	['evaluat
ial As	['PERSON', 'PERSON', 'ORGANIZATION', 'OTHER', 'OTHER', 'OTHER', 'WORK_O	['evaluat
ial As	['OTHER', 'PERSON', 'PERSON', 'PERSON', 'WORK_OF_ART', 'OTHER', 'OTHER	['evaluat
sentat	['ORGANIZATION', 'OTHER', 'OTHER', 'WORK_OF_ART', 'OTHER', 'OTHER', 'OTH	['storyte
sentat	['OTHER', 'PERSON', 'ORGANIZATION', 'OTHER', 'OTHER', 'WORK_OF_ART', 'O	['storyte
sentat	['ORGANIZATION', 'OTHER', 'OTHER', 'WORK_OF_ART', 'OTHER', 'OTHER', 'OTH	['storyte
sentat	['ORGANIZATION', 'OTHER', 'OTHER', 'WORK_OF_ART', 'OTHER', 'OTHER', 'OTH	['storyte
]	['WORK_OF_ART', 'OTHER', 'LOCATION', 'LOCATION']	['is418a
ial As	['OTHER', 'OTHER', 'PERSON', 'LOCATION', 'OTHER', 'OTHER', 'PERSON', 'OTH	['resear
ial As	['WORK_OF_ART', 'OTHER', 'OTHER', 'OTHER', 'OTHER', 'LOCATION', 'OTHER',	['is 418

Image 1: Categories generated by Google NLP

Then we extracted a bag of common entities among all course descriptions programmatically. Following this, we grouped most of the words from the bag of words into a suitable category manually. This dataset can be found here. Here, each common word is present in one or more courses.

For identifying categories, we followed a learner centric approach:
Existing workflow of a student in decision-making is not an isolated activity and faces following challenges:

- Shuffling multiple web pages- linked data needed, more details needed from website, syllabus archive, research themes page, google
- Redundantly asking other students for tacit course information
- Identifying topics and skills common across multiple courses
- Similarity or difference between courses
- Missing possibilities that are not considered based on user search terms.
- Disambiguation of similar sounding concepts or words for beginners, interdisciplinary or career changers who do not have domain knowledge.
  (Ex: Model in database vs model in statistics)
- Multidisciplinary- vocabulary context mismatch
- Communication gaps and barriers for international students seeking guidance
- Gaps in description
- Relatively new field
- Course description is domain specific or school specific but not student specific.

We have identified a few factors that students wish to know from course descriptions. For this, we conducted informal interviews.

- Beginner vs advanced
- Tools used
- Relevant job descriptions
- Complementary and necessary topics
- Lecture only vs group project/ paper/final exam/case study
- Sequence- Prerequisites
- Credit hours
- Professor’s network
- Number of assignments
- Speakers in class
- Relevant industry functions and Job descriptions
- Relevant Research park companies hiring for the course role
- Alumni and company they are in, and how the course connects to their work

Among these, we have chosen attributes relating to application domain, data lifecycle and type of class due to the limited scope of the project.

DOMAIN	SKILL -DATA LIFECYCLE	CLASS TYPE
1. Media 2. Teaching 3. Arts 4. Medical 5. Business 6. Social Aspects 7. Security 8. GIS 9. LIS 10. Miscellaneous	1. Requirements/research/user /consulting/synthesis 2. Design 3. Information Organization 4. Analytics 5. Presentation 6. Tools 7. Programming	1. Theory Based 2. Practical Training

Table 1: User defined categories

An example of categories and keywords:

Categories	Keywords
Presentation	storytelling, exhibits, audio and visual, visualization, presentations
Programming	Python, R, SQL, C++, C#, Java, html, RDF, turtle, NoSQL, datalog
Tools	Tableau, UML, Protege, OpenRefine, Jupyter notebooks, Gephi, Netlogo, UCInet. Weka, Vegalite, ArcGIS, YesWorkflow,
Social Aspects	Race, Gender, Communities, Youth, Ethnicity, Country name, regions, literacy, culture, sea, land, history, adults, children, genealogy, Population, Government, hacktivism, lives, politics, local, regional, family, heritage, school, church

Table 2: Categories and keywords define by user

5. Methods

The network: Course IDs form the nodes of the network.
For edges, we explored:

First, we wanted to identify ways in which two courses are similar, to convey this relationship through the edges. We chose cosine score to find out the similarity score for any two courses. We decided to assign a cutoff threshold, above which two courses would be considered similar.
However, in this approach we were not able to find specific details of why the two courses are similar. We could not comprehend in what dimension or attribute they were similar, in other words, the semantic meaning of the relationship was missing. Therefore, we decided on a different approach where we drew an edge between the courses if they had common entities. Next, for every common entity, we drew an edge and labelled the edge based on the category that the common entity belonged to. We also calculated the edge weights by simply counting the number of common categories. Since, we were using the CourseID as our nodes, and common entities or edges, there was a redundancy in the data, due to the same course description. Hence, we removed that redundancy / duplicate values by creating the nodes with the course name and common entities. This dataset was generated using pandas DataFrame and can be found here. The dataset was then loaded into Gephi for further analysis. We then found 89 nodes and 2210 edges in our network. The network consisted of 6 components, 5 of which were isolates and 1 connected component of 84 nodes.

For analysis, we came up with two approaches:

1. **Visual exploratory analysis:**

   Ground truth validation from student perspective: Based on the network drawn, we noted observations about the relationships among courses. Our initial intuition was that there are covert relationships that are not obvious to beginners by reading through the text. We identified a few non-intuitive observations and crosschecked with students in the MSIM program. We noted down a comparison between the two outcomes to highlight possible gaps and hence establish evidence to our initial hypothesis that certain relationships may not be visible through simple text descriptions.

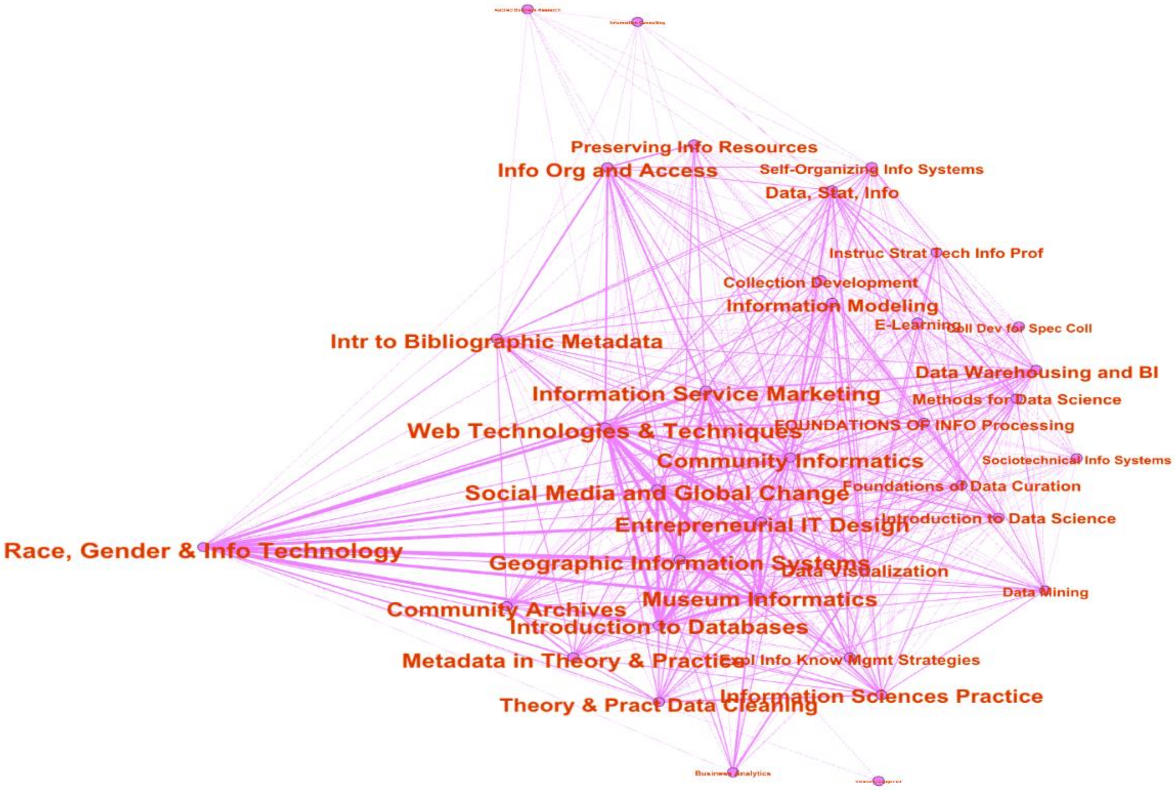

Applied Business Research
Information Consulting
Preserving Info Resources
Info Org and Access
Self-Organizing Info Systems
Data, Stat, Info
Instruc Strat Tech Info Prof
Collection Development
Information Modeling
E-Learning
Coll Dev for Spec Coll
Intr to Bibliographic Metadata
Data Warehousing and BI
Information Service Marketing
Methods for Data Science
Web Technologies & Techniques
FOUNDATIONS OF INFO Processing
Community Informatics
Sociotechnical Info Systems
Social Media and Global Change
Foundations of Data Curation
Entrepreneurial IT Design
Introduction to Data Science
Race, Gender & Info Technology
Geographic Information Systems
Data Visualization
Data Mining
Community Archives
Museum Informatics
Introduction to Databases
Metadata in Theory & Practice
Info Know Mgmt Strategies
Theory & Pract Data Cleaning
Information Sciences Practice
Business Analytics

Image 2: One of the communities formed in the network. Dominant theme: Analytics

2. Quantitative analysis:

From organizational network analysis perspective:
In the second method, we analyzed the nature of the course distribution, themes and subgroups to look into the Ischool's topics topology.

For this section, we recorded the following metrics:

- Connected components
- Degree distribution
- Clustering coefficient
- Average path length
- Communities formed

6. Results

Visual exploratory analysis based on user perspective:

User Experiment:

- A student from MSIM had been requested to mention two topics of interest.
- The student mentioned ‘anything related to Presentation skills and Privacy’.
- Then she was directed to the course catalog to identify one course that helps learn the same. Next, the student was instructed to pick other possible courses that may discuss these topics and she picked a few titles (she picked them without looking into the description). We then looked related courses up by the network.

All observations have been recorded as follow:

<table>
<tr><th>First Course identified by user</th><th>User identified relationships</th><th>User perspective</th><th>Time taken</th></tr>
<tr><td>Storytelling</td><td>Social media analytics, data science storytelling, DV,
storytelling</td><td>Presentation skill needed</td><td>~15 minutes</td></tr>
<tr><td colspan="4"><b>Based on the network</b>: Looked up ‘Presentation’ category edges:
Museum Informatics, Exhibit design, Entrepreneurial IT design , Data science storytelling, DV, storytelling, Introduction to Databases, Global Health Informatics, Geographic Information Systems, Social Media and Global Change, Intro to Technology in LIS, Information Service Marketing, DW and BI, Oral History, Community Informatics, Literacy, Reading, and Readers, Rare Book and Spec Colls, Issues Scholarly Communication, BA, computers and cultures, Race, Gender & Info Technology, web technologies and technique, Pro Comm Lib & other Info Pro</td></tr>
<tr><td>Info Assurance</td><td>Information Ethics</td><td>Anything related to security</td><td>~7 Minutes</td></tr>
</table>

Based on the network : Looked up edges with ethics and security categories: Library book preservation, information ethics, local, regional, global IS, libraries info and society, Race, Gender & Info Technology, Computers and Cultures, Web Technologies & Techniques, Museum Informatics.

Table 3: Observations for visual exploration of the semantic network

Quantitative analysis for organizational network analysis:

The below are the network metrics calculated using Gephi and the details for the same can be found on GitHub as well.

- Connected components-
- Degree distribution
- Clustering coefficient
- Communities formed

Connected components: From the below graph, we can observe that the network consisted of 6 components, 5 out of which were isolates and didn't form any edge.

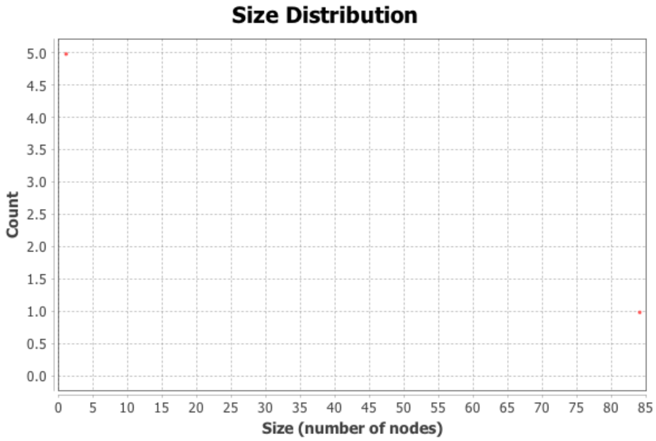


Image 3: Connected components

Degree distribution: We notice that it is a skewed distribution.

The top three are Adult Popular Literature (77), Intro to Technology in LIS(76), Libraries Info and Society(76)

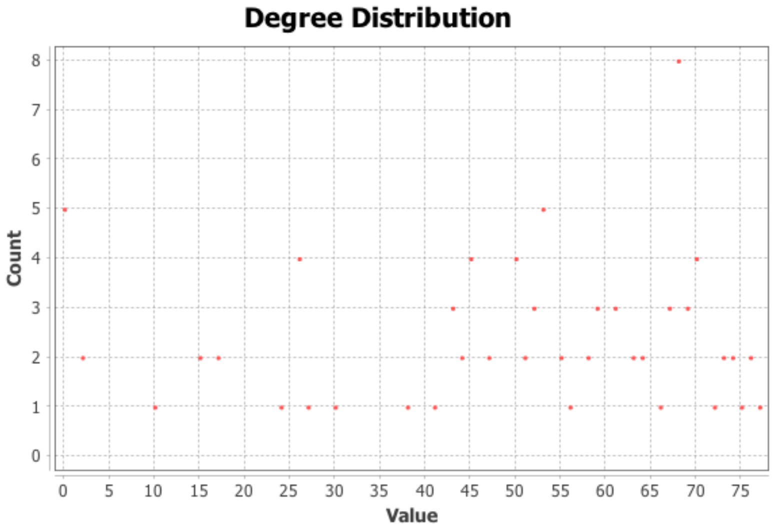


Image 4: Degree distribution

Clustering coefficient: A high value of clustering coefficient is noted (0.84) implying a strongly connected graph with small average path length resembling a small world topology.

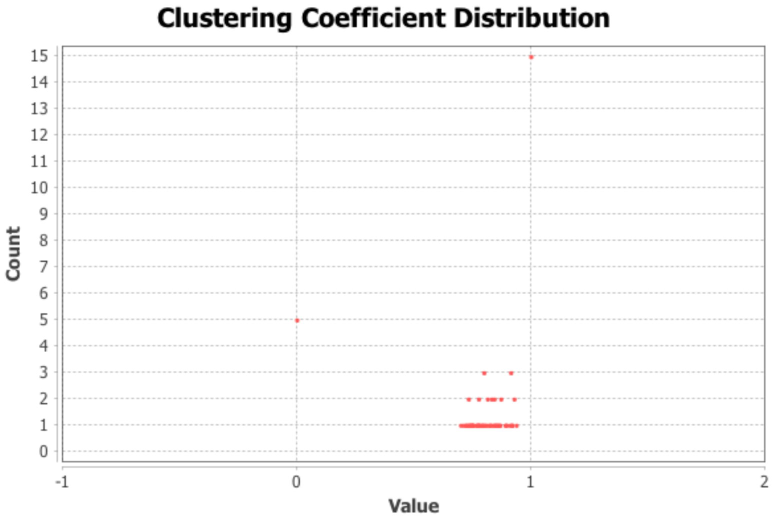


Image5: Clustering coefficient

Communities formed: It is noticed that 8 communities are formed, 5 of which are isolate. The three communities seem to be about analytics, Library science and social aspects as dominant themes.

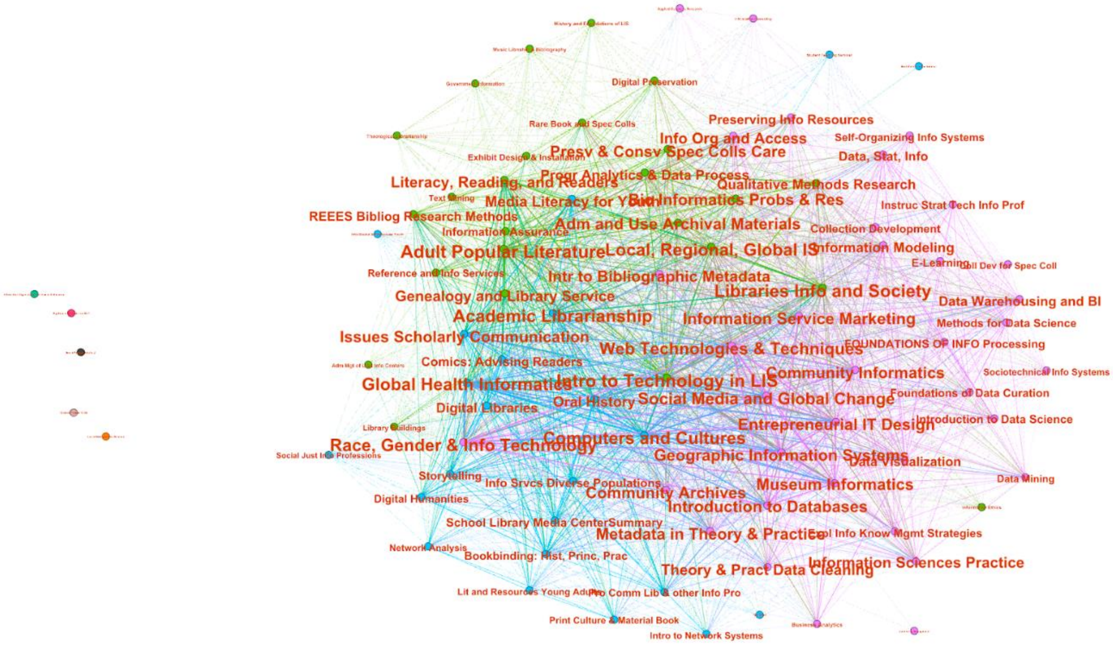

History and Foundations of LIS
Music Librarianship Bibliography
Government Information
Digital Preservation
Theological Librarianship
Rare Book and Spec Colls
Preserving Info Resources
Info Org and Access
Self-Organizing Info Systems
Exhibit Design & Installation
Presv & Consv Spec Colls Care
Data, Stat, Info
Progr Analytics & Data Process
Literacy, Reading, and Readers
Qualitative Methods Research
Text Mining
Media Literacy for Youth
Bio Informatics Probs & Res
Instruc Strat Tech Info Prof
REEES Bibliog Research Methods
Information Assurance
Adm and Use Archival Materials
Collection Development
Adult Popular Literature
Local, Regional, Global IS
Information Modeling
E-Learning
Coll Dev for Spec Coll
Reference and Info Services
Intr to Bibliographic Metadata
Libraries Info and Society
Genealogy and Library Service
Data Warehousing and BI
Academic Librarianship
Information Service Marketing
Methods for Data Science
Issues Scholarly Communication
Web Technologies & Techniques
FOUNDATIONS OF INFO Processing
Adm Mgt of Lib Info Centers
Comics: Advising Readers
Community Informatics
Sociotechnical Info Systems
Global Health Informatics
Intro to Technology in LIS
Foundations of Data Curation
Oral History
Social Media and Global Change
Digital Libraries
Entrepreneurial IT Design
Introduction to Data Science
Library Buildings
Computers and Cultures
Race, Gender & Info Technology
Geographic Information Systems
Social Just Info Professions
Data Visualization
Storytelling
Info Srvcs Diverse Populations
Museum Informatics
Data Mining
Community Archives
Digital Humanities
Introduction to Databases
School Library Media CenterSummary
Metadata in Theory & Practice
Ecol Info Know Mgmt Strategies
Network Analysis
Bookbinding: Hist, Princ, Prac
Theory & Pract Data Cleaning
Information Sciences Practice
Lit and Resources Young Adults
Pro Comm Lib & other Info Pro
Print Culture & Material Book
Business Analytics
Intro to Network Systems

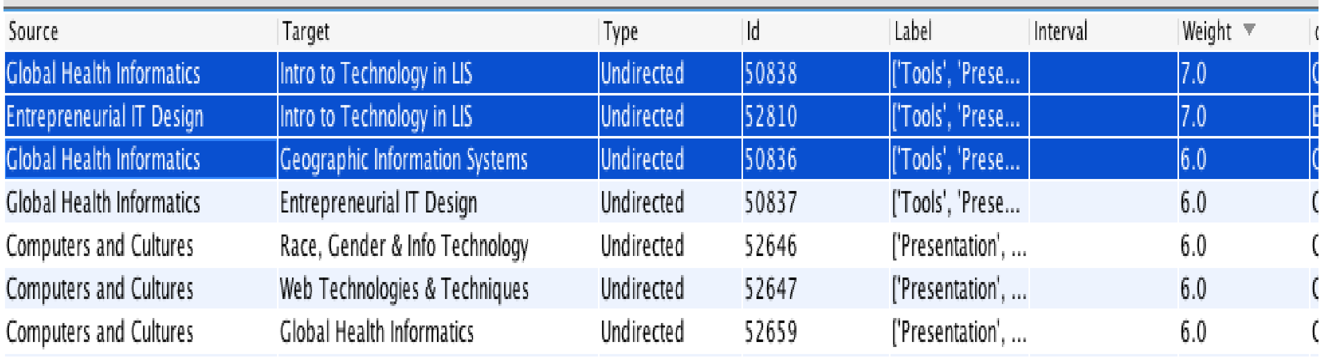
Image6: Communities formed

Betweenness Centrality, Closeness Centrality, modularity and clustering for top 10 nodes with highest degree:

Label	Degree	Closeness centrality	Betweeness centrality	Modularity class	Clustering
Adult Popular Literature	77	0.932584	46.709026	1	0.699932
Intro to Technology in LIS	76	0.922222	33.687553	1	0.721754
Libraries Info and Society	76	0.922222	44.145617	1	0.711228
Academic Librarianship	75	0.912088	40.64571	6	0.722162

Local, Regional, Global IS	74	0.902174	32.057087	1	0.733062
Global Health Informatics	74	0.902174	67.977033	6	0.733062
Computers and Cultures	73	0.892473	25.879009	6	0.753044
Race, Gender & Info Technology	73	0.892473	63.385536	2	0.747336
Web Technologies & Techniques	72	0.882979	31.761579	2	0.745696

Table 4: Betweenness Centrality, Closeness Centrality, modularity and clustering for top 10 nodes with highest degree



Source	Target	Type	Id	Label	Interval	Weight ▼	c
Global Health Informatics	Intro to Technology in LIS	Undirected	50838	['Tools', 'Prese...		7.0	C
Entrepreneurial IT Design	Intro to Technology in LIS	Undirected	52810	['Tools', 'Prese...		7.0	E
Global Health Informatics	Geographic Information Systems	Undirected	50836	['Tools', 'Prese...		6.0	C
Global Health Informatics	Entrepreneurial IT Design	Undirected	50837	['Tools', 'Prese...		6.0	C
Computers and Cultures	Race, Gender & Info Technology	Undirected	52646	['Presentation', ...		6.0	C
Computers and Cultures	Web Technologies & Techniques	Undirected	52647	['Presentation', ...		6.0	C
Computers and Cultures	Global Health Informatics	Undirected	52659	['Presentation', ...		6.0	C

Table5: Edges with highest weights

8. Conclusion

1. From table number 3, it is evident that students may have limited peripheral understanding of the topic of their interest and may be missing possible courses with same theme as their interest but in different applications or context. As noted, the network-identified courses based on course description text are more in number and non-intuitive than student identified.
2. From the network metrics and the table number 4, it is evident that a lot of importance is given to LIS courses and are central to the course distribution.

This is intuitive as MSIM is a new program and new courses are being introduced now.

3. The graph is well connected and has short average path length meaning the focus of the department is well concentrated and interrelated which helps in developing a comprehensive understanding of the domain.
4. We noted 3 major communities. Broadly, the Green represents LIS, the Purple for Analytics and Blue for Social impact. Some surprising findings are that collection dev, community archives, bibliographic Metadata, Race, gender and technology are grouped with Analytics and technology. It is indeed an important aspect of analytics to discuss about aspects like metadata and diversity/ethics but is not intuitively picked up by students in the MSIM track.
5. The fact that social impact appears as a separate group reassures the vision of the school based in Human Centric information sciences.
6. Further, data mining and text mining are not grouped together, which is counter intuitive. This is because Text mining description has little focus on technology aspects.
7. We notice that some of the isolates are R and D infrastructure, History of LIS and Grad Bio informatics. A few isolates are due to incomplete descriptions and some because they are truly unique courses. This gives insight into text quality or gaps for other related courses to be added in support of the existing isolates.
8. Finally, based on the school website, the electives for Analytics track are suggested as 12 courses. But According to the network, 36 courses can be considered for this track.

In conclusion, the semantic network has been able to act as a tool to both inform students in course selection and the program in course management and organizational analysis with respect to courses available.Thus the findings are relevant to prospective and current students as well as the program management.
If combined with student surveys and market trends, it can be a powerful medium to identify and bridge gaps in both student satisfaction and employability trends.

8. Limitations:

1. Our dataset consists of the courses for Spring 2018, which is a big limitation.
2. As we expand our scope for the data collection, manual addition of categories would not be efficient.
3. No authentic field research, qualitative analysis or user experiments have been conducted for problem statement or visual analysis. Problem identification and definition is informal and not representative, particularly of LIS students.
4. Categories have been defined based on amateur online research but not from Industrial/Organizational or educational psychology domain expertise.
5. The current methods are difficult to scale.
6. Word variations (stemming) for entities has not been taken into account resulting in loss of relations between courses.
7. User study has been based on only style of search.

8. Future Work:

1. The network can be extended to include other entities in the program including research topics, professors, relevant job descriptions and other attributes identified in course selection decision-making process. It can also be built into an app for prospective students and current students to explore.
2. We can devise a quality metric for the course catalog based on semantic meaning conveyed
3. The sematic network visualization can be extended to other applications in learning to address communication issues of students with Autism or learning disabilities, and in teaching new languages.

9. References:

1. Wang, Y., & Gafurov, D. (2003, August). The cognitive process of comprehension. In *Cognitive Informatics, 2003. Proceedings. The Second IEEE International Conference on*(pp. 93-97). IEEE.
2. Arnold, S., Burke, D., Dörsch, T., Loeber, B., & Lommatzsch, A. (2014, October). News Visualization based on Semantic Knowledge. In *International Semantic Web Conference (Posters & Demos)* (pp. 5-8).
3. Drieger, P. (2013). Semantic network analysis as a method for visual text analytics. *Procedia-social and behavioral sciences*, *79*, 4-17.
4. Diesner, J., & Carley, K. M. (2004, July). Using network text analysis to detect the organizational structure of covert networks. In *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference* (Vol. 3). NAACSOS.
5. De Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek* (Vol. 27). Cambridge University Press.
6. Rogers, E. M., & Kincaid, D. L. (1981). Communication networks: Toward a new paradigm for research.
7. Australian Library and Information Association. (2003). The Library and Information Sector: Core Knowledge, Skills and Attributes.
8. Emes, C., & Cleveland-Innes, M. (2003). A journey toward learner-centred curriculum. *The Canadian Journal of Higher Education*, *33*(3), 47.