A Network Tour Of Data Science Final Project

Course Suggester

Cherqui Alexandre

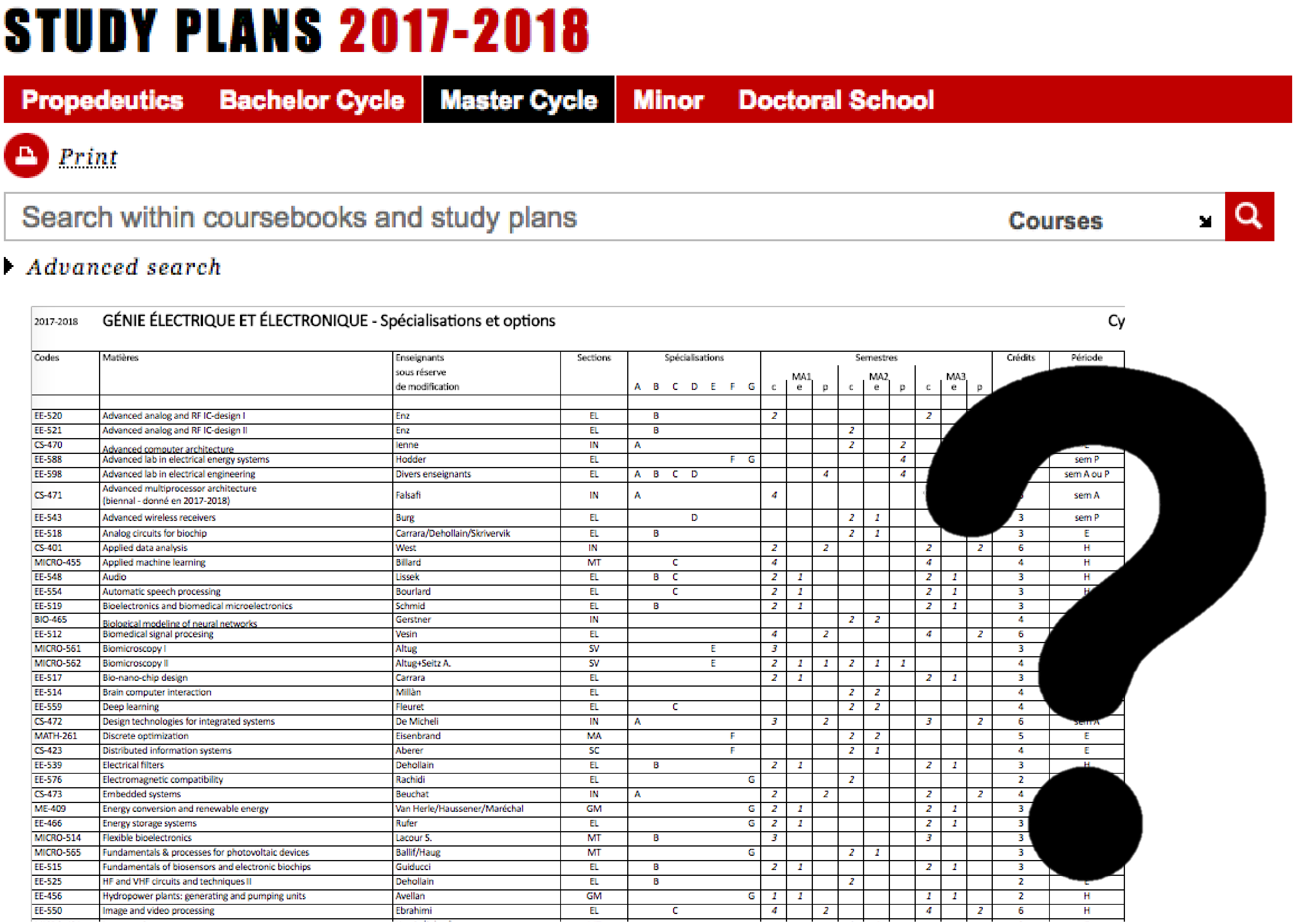
Gusset Frédérick

El Hamamsy Laila

Triquet Thomas

25.01.2018

Course Suggester



STUDY PLANS 2017-2018

Propedeutics Bachelor Cycle Master Cycle Minor Doctoral School

Print

Search within coursebooks and study plans Courses

Advanced search

2017-2018 GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE - Spécialisations et options Cy

Codes	Matières	Enseignants sous réserve de modification	Sections	Spécialisations							Semestres									Crédits	Période
											MA1			MA2			MA3				
				A	B	C	D	E	F	G	c	e	p	c	e	p	c	e	p		
EE-520	Advanced analog and RF IC-design I	Enz	EL		B						2						2				
EE-521	Advanced analog and RF IC-design II	Enz	EL		B									2							
CS-470	Advanced computer architecture	Ienne	IN	A										2		2					E
EE-588	Advanced lab in electrical energy systems	Hodder	EL						F	G						4					sem P
EE-598	Advanced lab in electrical engineering	Divers enseignants	EL	A	B	C	D						4			4					sem A ou P
CS-471	Advanced multiprocessor architecture (biennal - donné en 2017-2018)	Falsafi	IN	A							4										sem A
EE-543	Advanced wireless receivers	Burg	EL				D							2	1					3	sem P
EE-518	Analog circuits for biochip	Carrara/Dehollain/Skrivervik	EL		B									2	1					3	E
CS-401	Applied data analysis	West	IN								2		2				2		2	6	H
MICRO-455	Applied machine learning	Billard	MT			C					4						4			4	H
EE-548	Audio	Lissek	EL		B	C					2	1					2	1		3	H
EE-554	Automatic speech processing	Bourlard	EL			C					2	1					2	1		3	H
EE-519	Bioelectronics and biomedical microelectronics	Schmid	EL		B						2	1					2	1		3	
BIO-465	Biological modeling of neural networks	Gerstner	IN											2	2					4	
EE-512	Biomedical signal procesing	Vesin	EL								4		2				4		2	6	
MICRO-561	Biomicroscopy I	Altug	SV					E			3									3	
MICRO-562	Biomicroscopy II	Altug+Seitz A.	SV					E			2	1	1	2	1	1				4	
EE-517	Bio-nano-chip design	Carrara	EL								2	1					2	1		3	
EE-514	Brain computer interaction	Millàn	EL											2	2					4	
EE-559	Deep learning	Fleuret	EL			C								2	2					4	
CS-472	Design technologies for integrated systems	De Micheli	IN	A							3		2				3		2	6	sem A
MATH-261	Discrete optimization	Eisenbrand	MA						F					2	2					5	E
CS-423	Distributed information systems	Aberer	SC						F					2	1					4	E
EE-539	Electrical filters	Dehollain	EL		B						2	1					2	1		3	H
EE-576	Electromagnetic compatibility	Rachidi	EL							G				2						2	
CS-473	Embedded systems	Beuchat	IN	A							2		2				2		2	4	
ME-409	Energy conversion and renewable energy	Van Herle/Haussener/Maréchal	GM							G	2	1					2	1		3	
EE-466	Energy storage systems	Rufer	EL							G	2	1					2	1		3	
MICRO-514	Flexible bioelectronics	Lacour S.	MT		B						3						3			3	
MICRO-565	Fundamentals & processes for photovoltaic devices	Ballif/Haug	MT							G				2	1					3	
EE-515	Fundamentals of biosensors and electronic biochips	Guiducci	EL		B						2	1					2	1		3	
EE-525	HF and VHF circuits and techniques II	Dehollain	EL		B									2						2	
EE-456	Hydropower plants: generating and pumping units	Avellan	GM							G	1	1					1	1		2	H
EE-550	Image and video processing	Ebrahimi	EL			C					4		2				4		2	6	H

1. Data Acquisition

2. Data Exploration

3. Data Exploitation

4. Conclusion

What is needed ?

Enrolments*

- List of all courses taken by all students between 2007 and 2016
- Standardized data, minimal cleaning required

Course Descriptions*

- Enrich the data with specific course attributes
- Manually input, fastidious to standardize.

*Courtesy of Francesco Pinto, EPFL Data Scientist

1. Data Acquisition

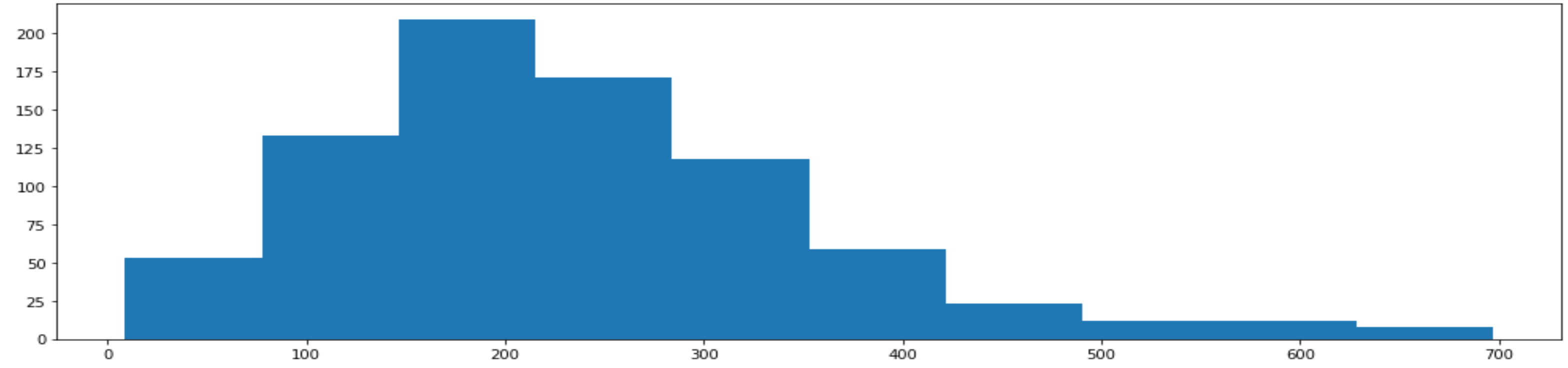
2. Data Exploration

3. Data Exploitation

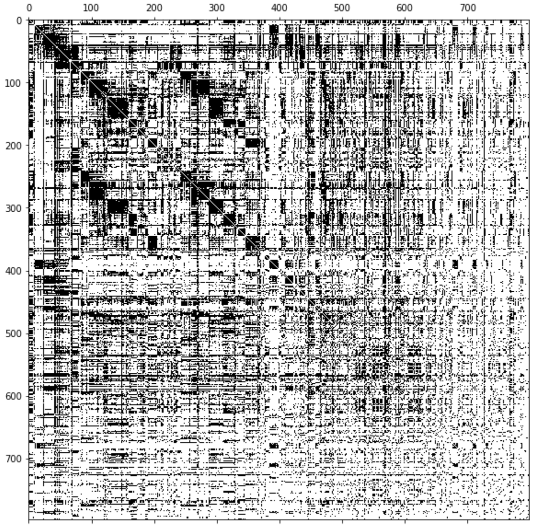
4. Conclusion

Baseline : Enrolments

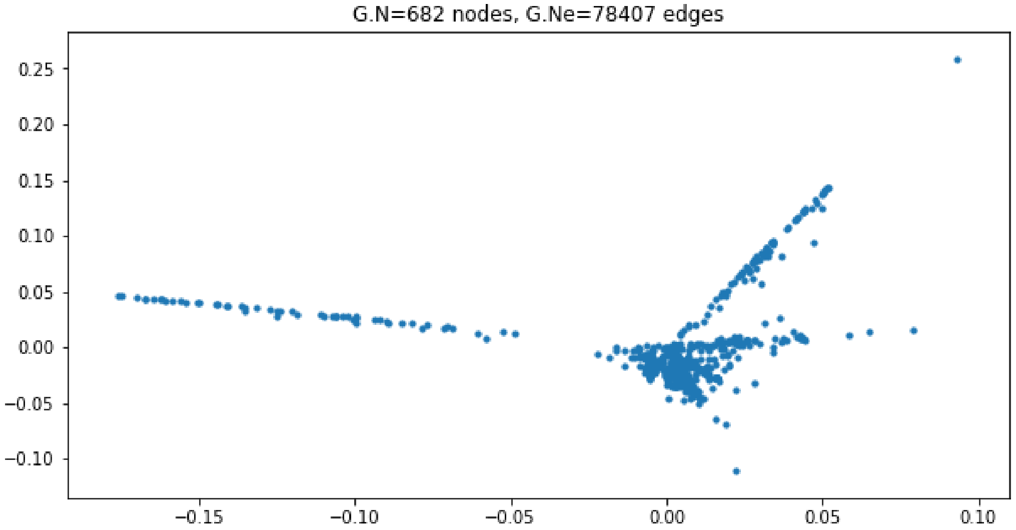
- Weight between two courses proportional to number of students who took both



Unweighted Degree Distribution



Weight Matrix


G.N=682 nodes, G.Ne=78407 edges

Spectral Clustering

1. Data Acquisition

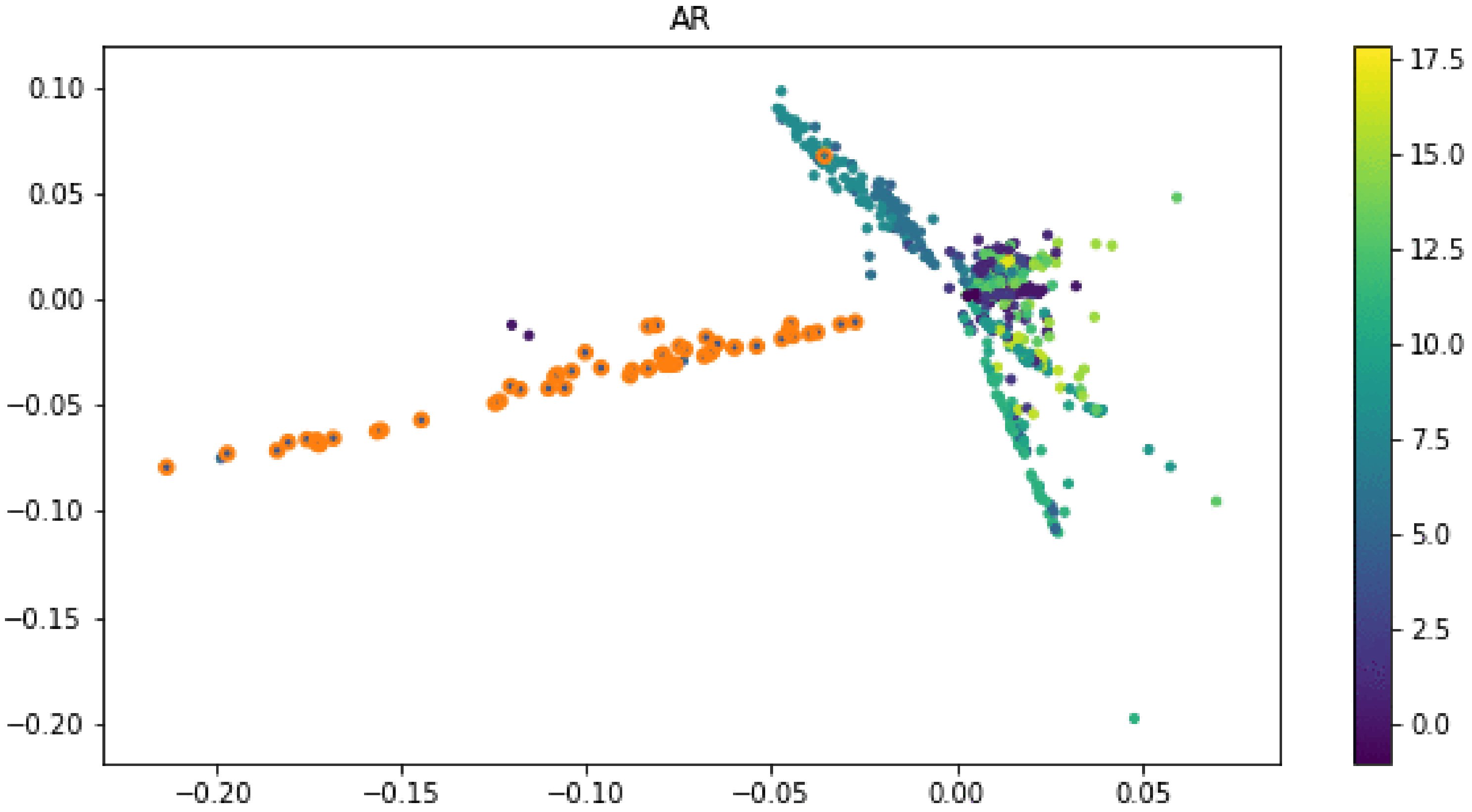
2. Data Exploration

3. Data Exploitation

4. Conclusion

Baseline : Enrolments

- Weight between two courses proportional to number of students who took both



Normalized Laplacian Eigenvectors

Architects are not engineers

1. Data Acquisition

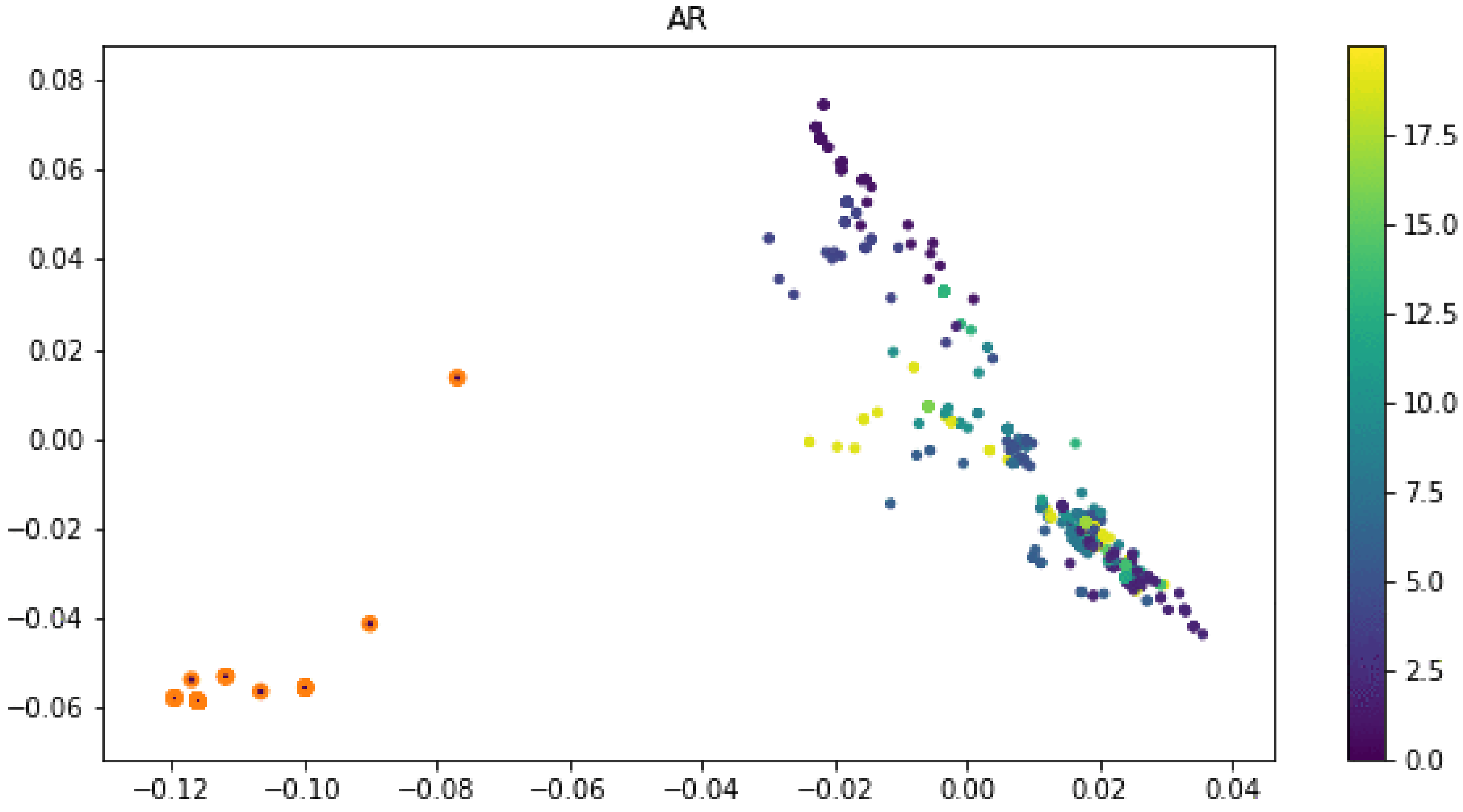
2. Data Exploration

3. Data Exploitation

4. Conclusion

Enriching : Study Plans

- Redundant with the enrolments as students tend to stick to their study plans



Normalized Laplacian Eigenvectors – All Sections
Projection on the 2nd and 3rd eigenvectors

1. Data Acquisition

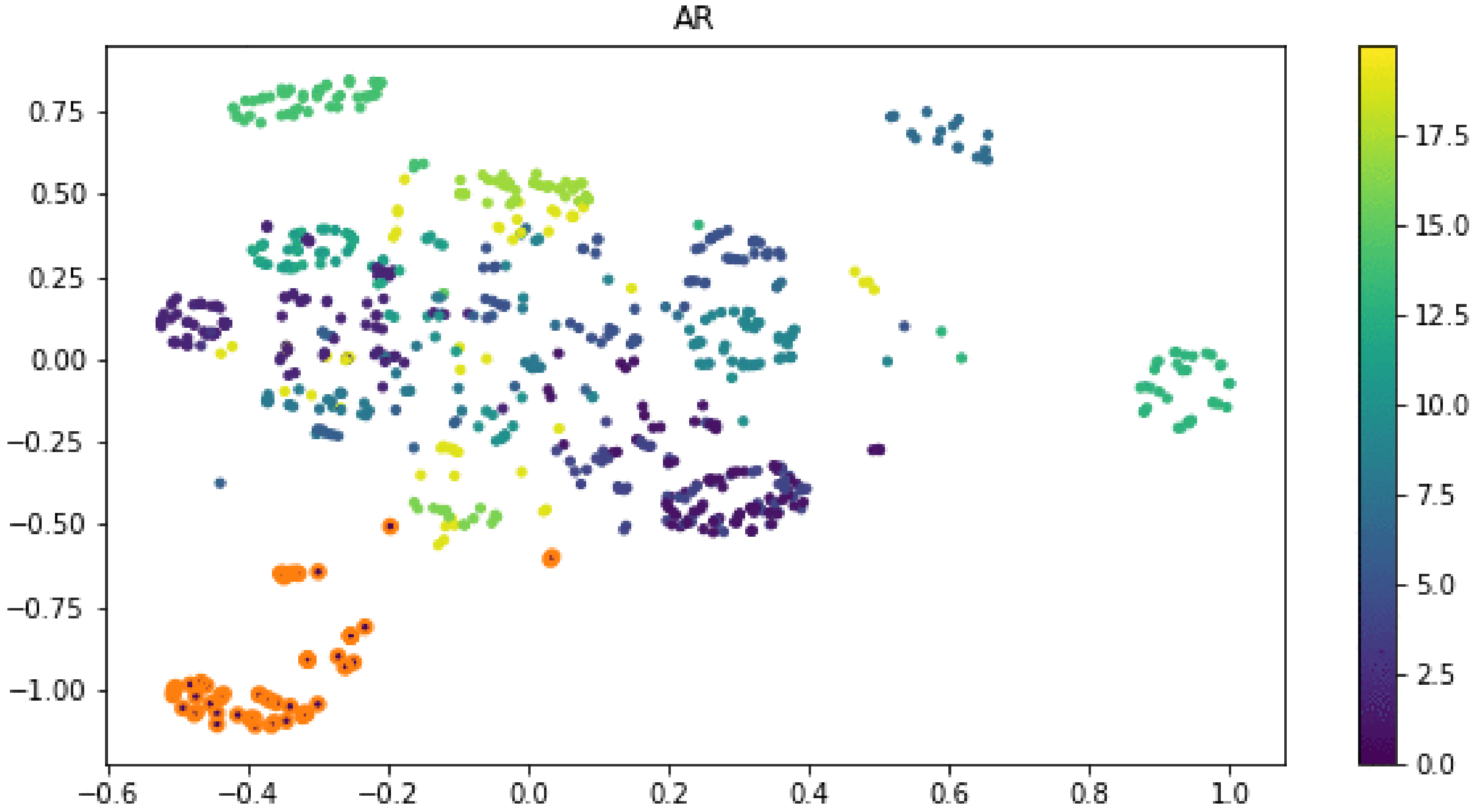
2. Data Exploration

3. Data Exploitation

4. Conclusion

Enriching : Study Plans

- Redundant with the enrolments as students tend to stick to their study plans



Normalized Laplacian Eigenvectors – All Sections
Spring View

1. Data Acquisition

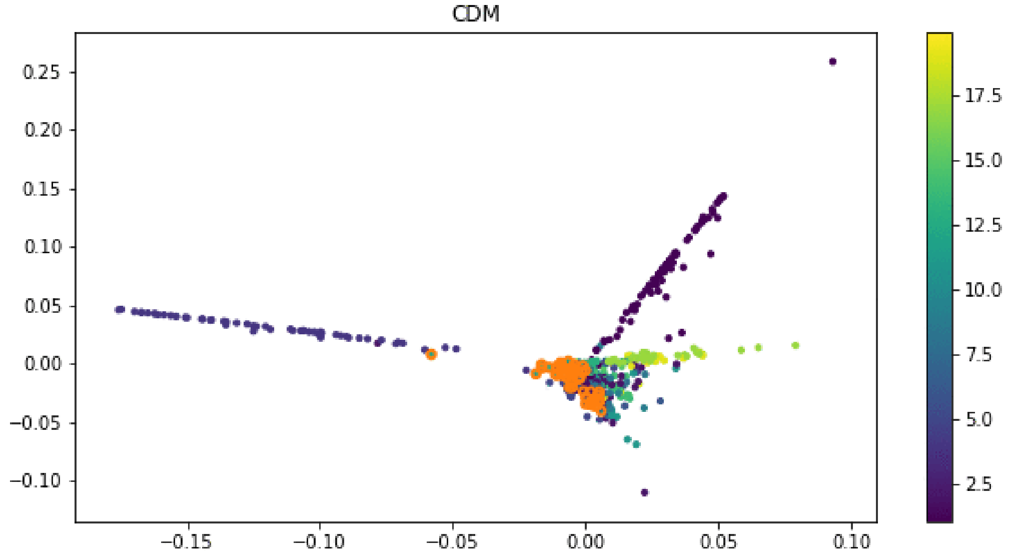
2. Data Exploration

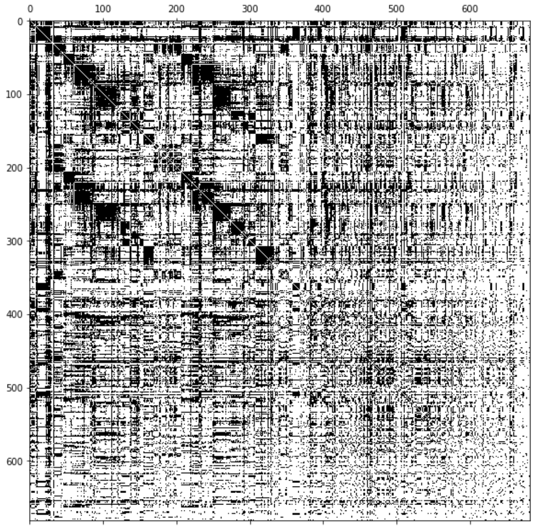
3. Data Exploitation

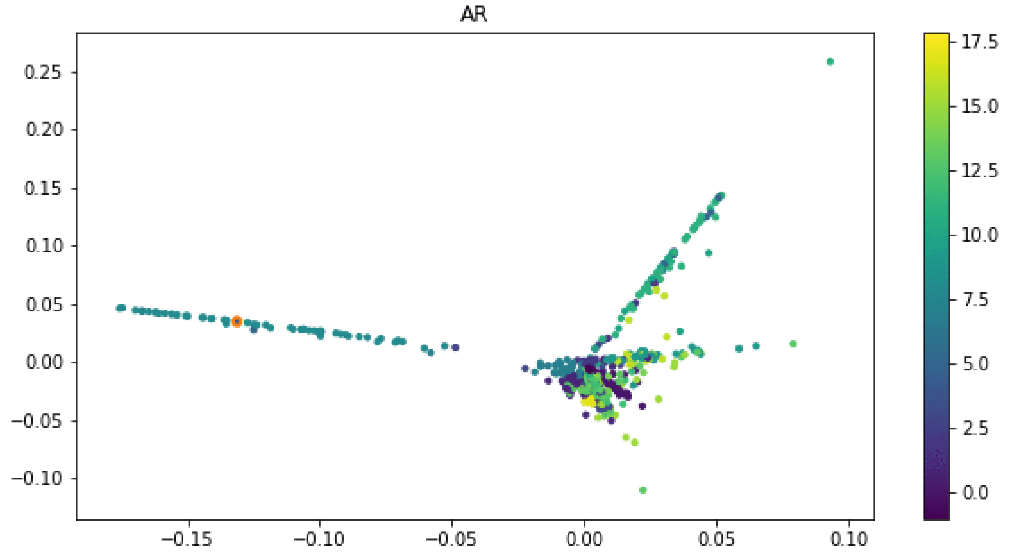
4. Conclusion

Baseline : Enrolments

- Removing Architects
- Spectral clustering on first eigenvectors of the normalized Laplacian
- Distinction between faculties
- Distinction between sections







1. Data Acquisition

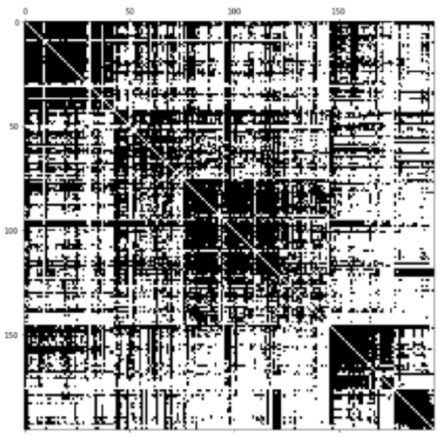
2. Data Exploration

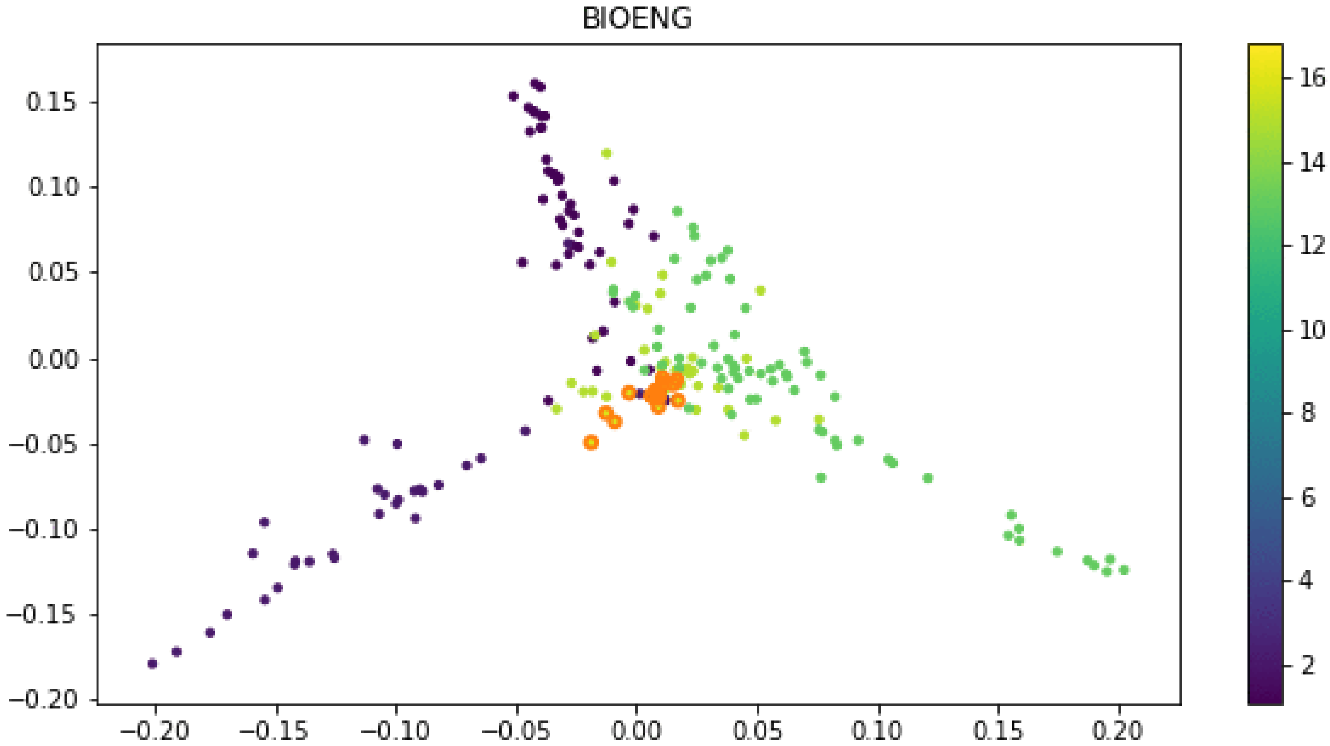
3. Data Exploitation

4. Conclusion

Baseline : Enrolments

- Decision to focus on the STI faculty for Master courses
- Objective : propose a diverse ensemble of courses by enriching this graph





Normalized Laplacian Eigenvectors – STI Faculty Isolating Sections

1. Data Acquisition

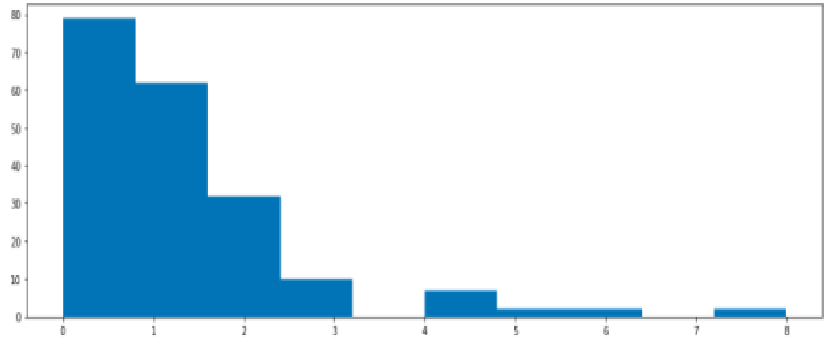
2. Data Exploration

3. Data Exploitation

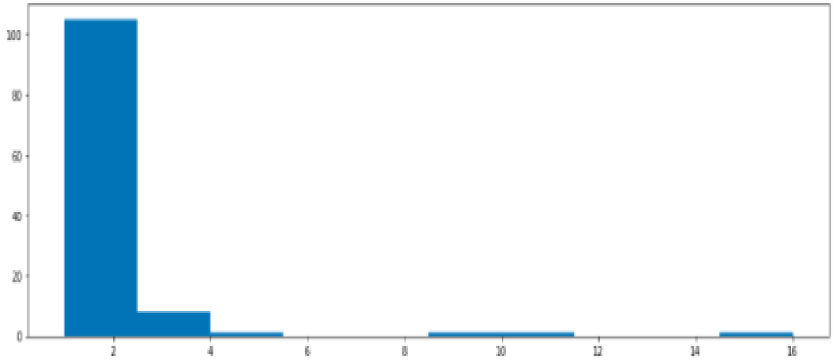
4. Conclusion

Enriching : Professors

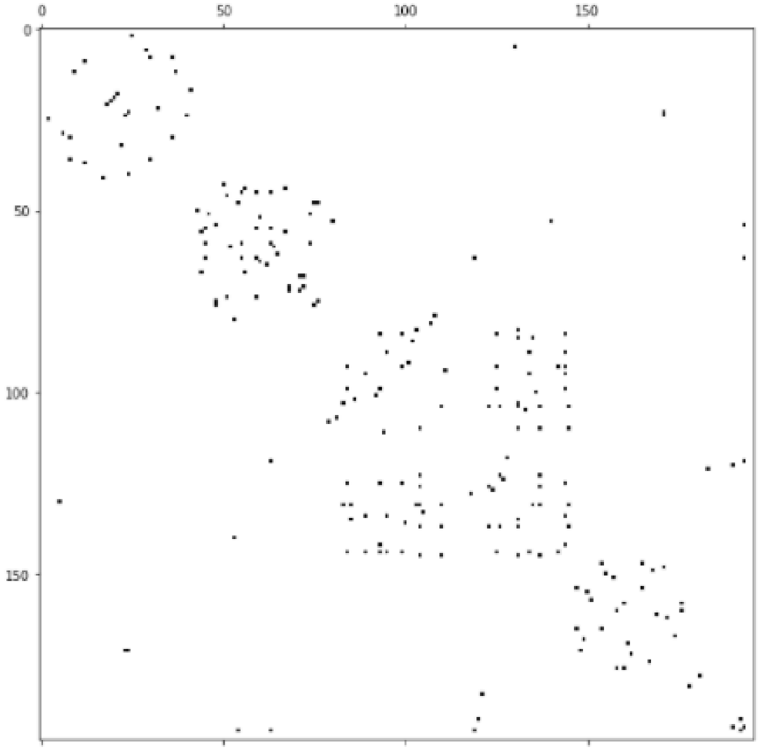
- Sparse



Unweighted degree Distribution



Distribution of Component Sizes



Weight Matrix

- Largest giant component 16 due to courses with multiple professors
- Highest node degree 8

1. Data Acquisition

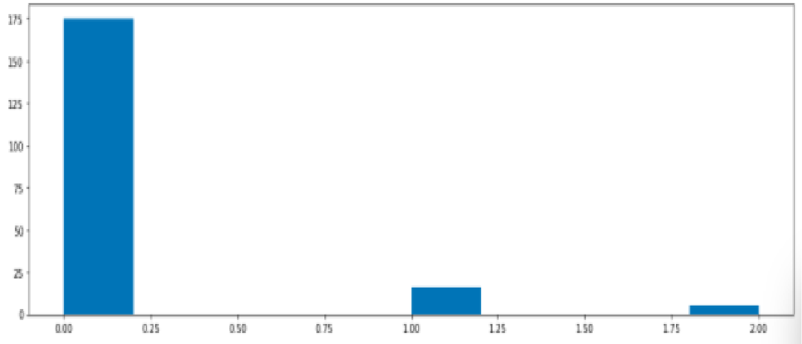
2. Data Exploration

3. Data Exploitation

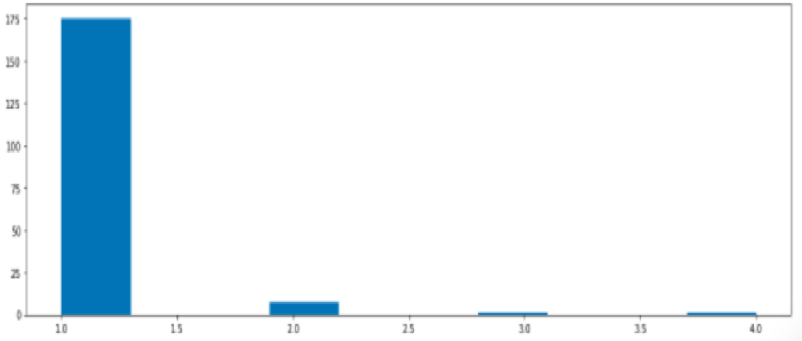
4. Conclusion

Enriching : Assistants

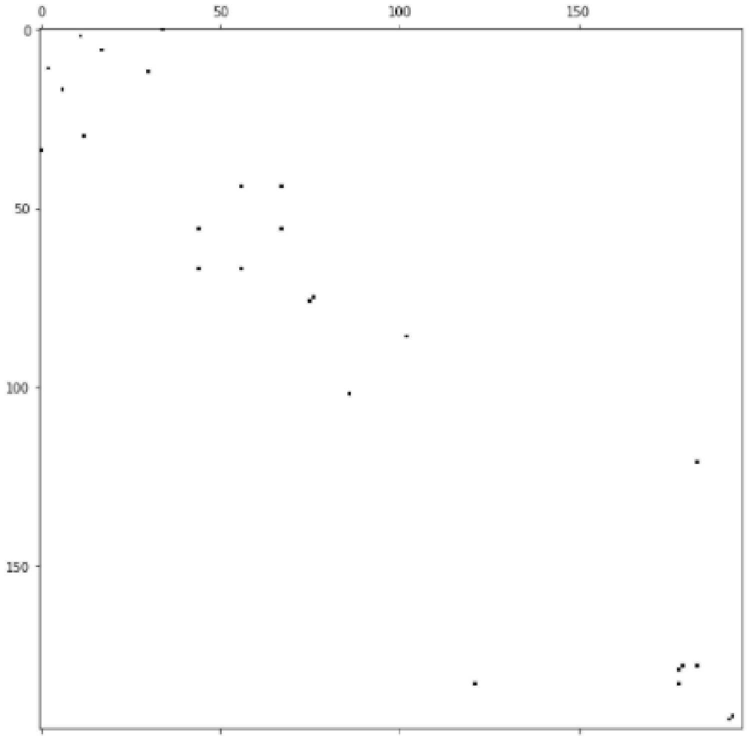
- Even more sparse
- Assistants are not necessarily specified on the course descriptions



Unweighted degree Distribution



Distribution of Component Sizes



- Largest giant component (4) : due to courses with multiple assistants
- Second largest component : 3

1. Data Acquisition

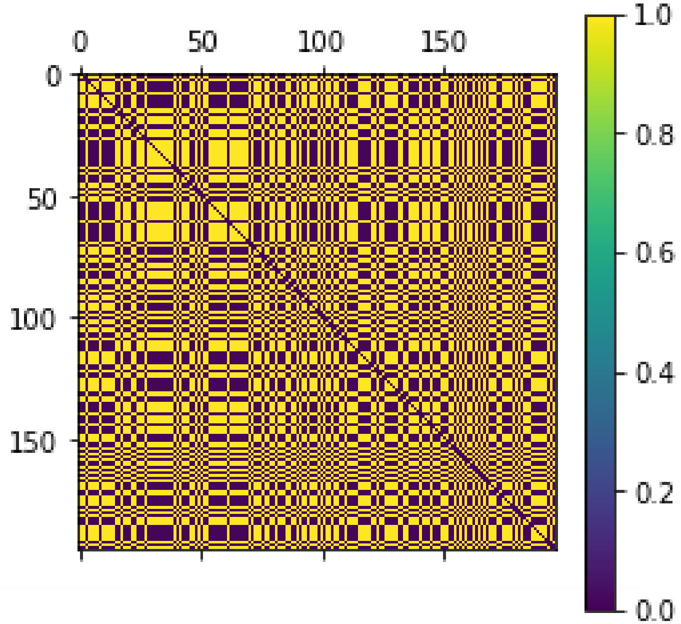
2. Data Exploration

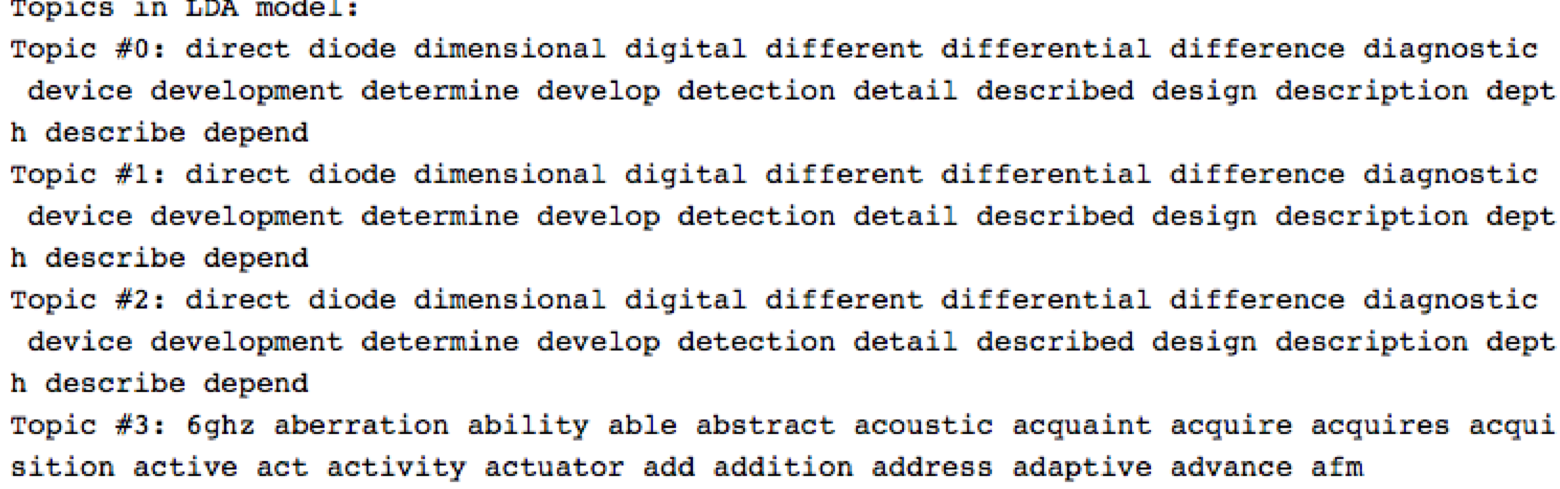
3. Data Exploitation

4. Conclusion

Enriching : Topics

- Keywords, contents, summaries, etc…
- Topic detection to create links between courses
- Matrix factorization to get a Course x Topic matrix
- Graph constructed based on distance between courses





```
Topics in LDA model:
Topic #0: direct diode dimensional digital different differential difference diagnostic
 device development determine develop detection detail described design description dept
h describe depend
Topic #1: direct diode dimensional digital different differential difference diagnostic
 device development determine develop detection detail described design description dept
h describe depend
Topic #2: direct diode dimensional digital different differential difference diagnostic
 device development determine develop detection detail described design description dept
h describe depend
Topic #3: 6ghz aberration ability able abstract acoustic acquaint acquire acquires acqui
sition active act activity actuator add addition address adaptive advance afm
```

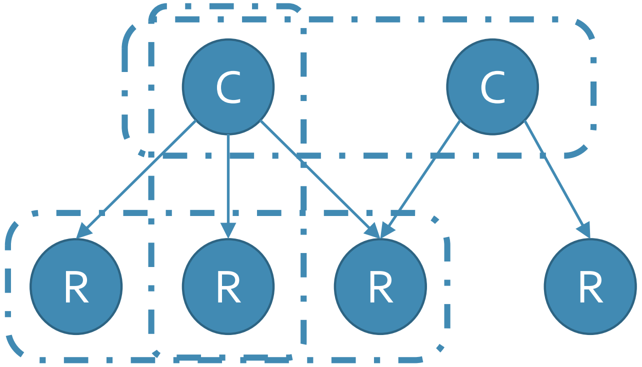
1. Data Acquisition

2. Data Exploration

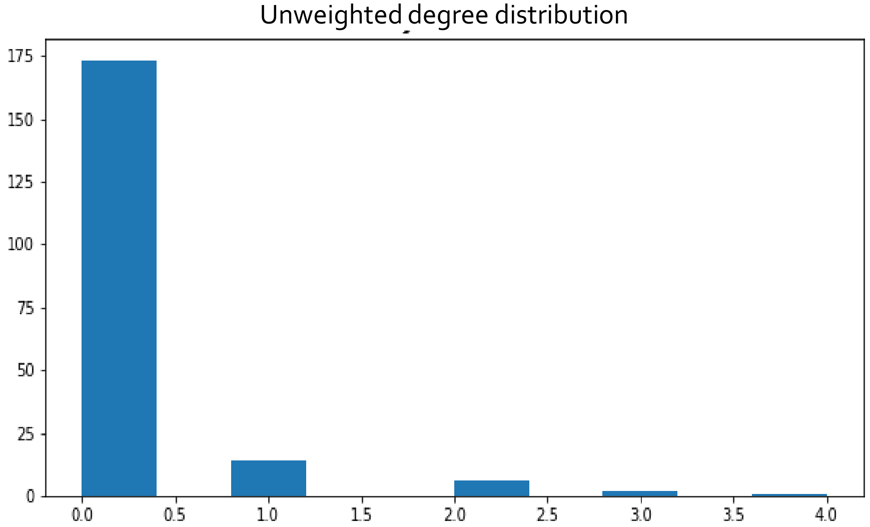
3. Data Exploitation

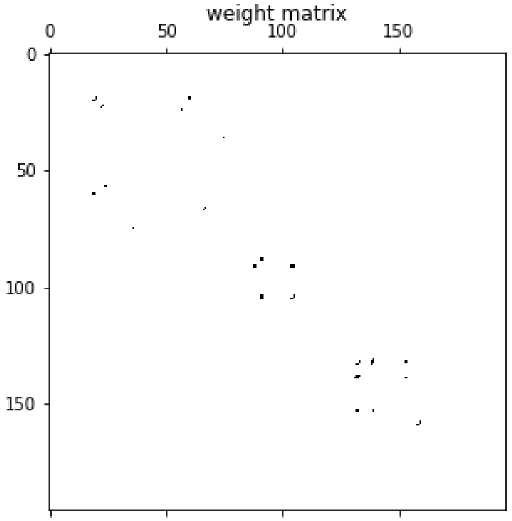
4. Conclusion

Enriching : Requirements



- 3 graphs:
  - Link a course to its requirement.
  - Link the requirements of the same course.
  - Link the courses that have the same requirements.





- Courses with same requirements : not enough connections, courses recommended are irrelevant although in the same section.
- Requirements of same course : immediate neighbours are pertinent
- Course to its requirements : similar results as the requirements of the same course but with less information.

1. Data Acquisition

2. Data Exploration

3. Data Exploitation

4. Conclusion

Enriching : Final Graph

Data	Status	Reason
Students	Kept	Baseline
Study Plans	Discarded	Same information as the enrolments
Requirements	Kept	Links based on common topics of the courses
Keywords	Discarded	Interesting but needed more information
Professors	Discarded	Too sparse
Assistants	Discarded	Too sparse

- Weighted sum of the requirements and the baseline to obtain the final graph
- Empirical weights : not possible to do supervised learning
    - Using the probability of taking two courses would overfit the baseline graph
    - Did not have the data required to do a good metric

1. Data Acquisition

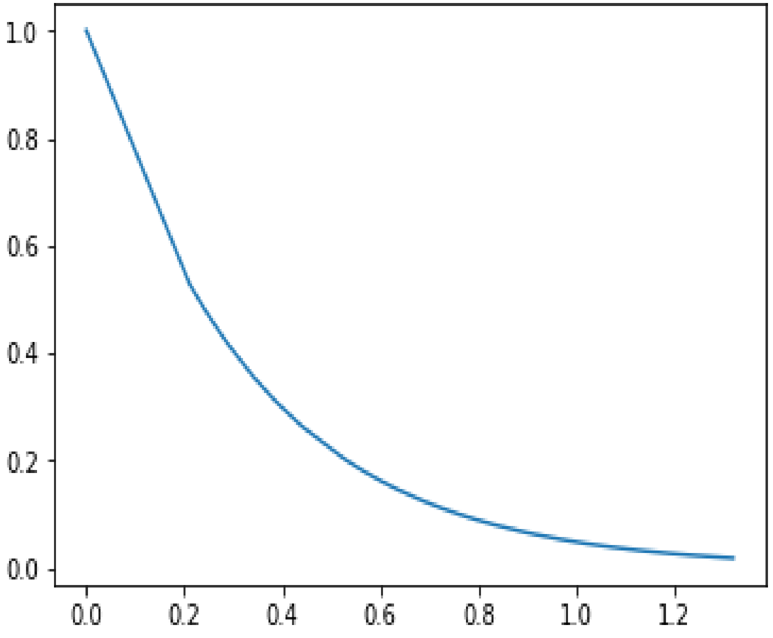
2. Data Exploration

3. Data Exploitation

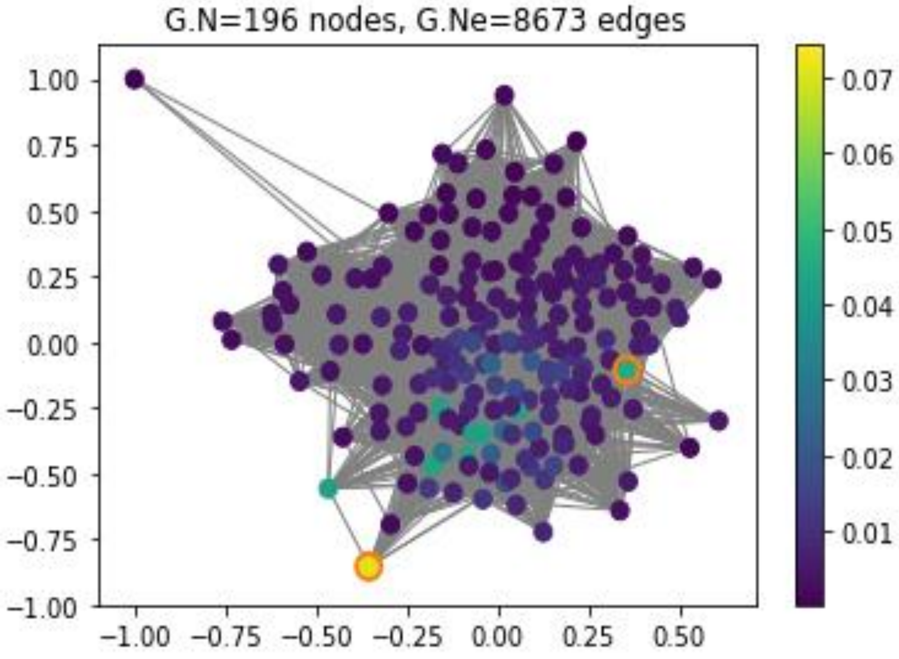
4. Conclusion

Graph Diffusion

- Low pass filter.
- Diffuse from multiple courses.
- Recommendation: sort the courses in descending order



Heat



Diffusion with 2 courses

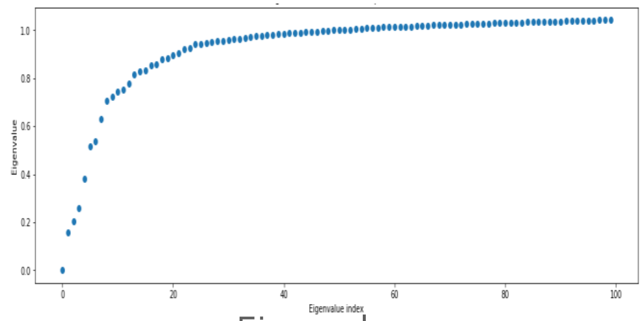
1. Data Acquisition

2. Data Exploration

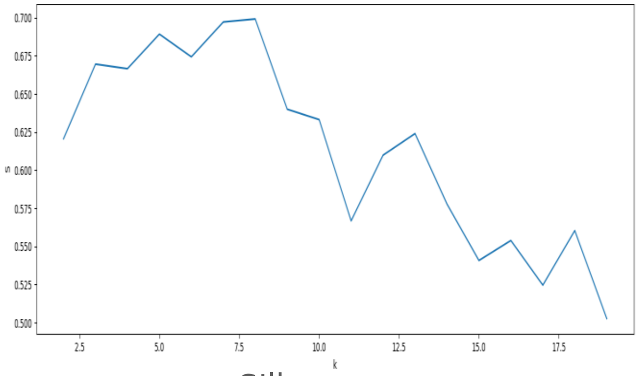
3. Data Exploitation

4. Conclusion

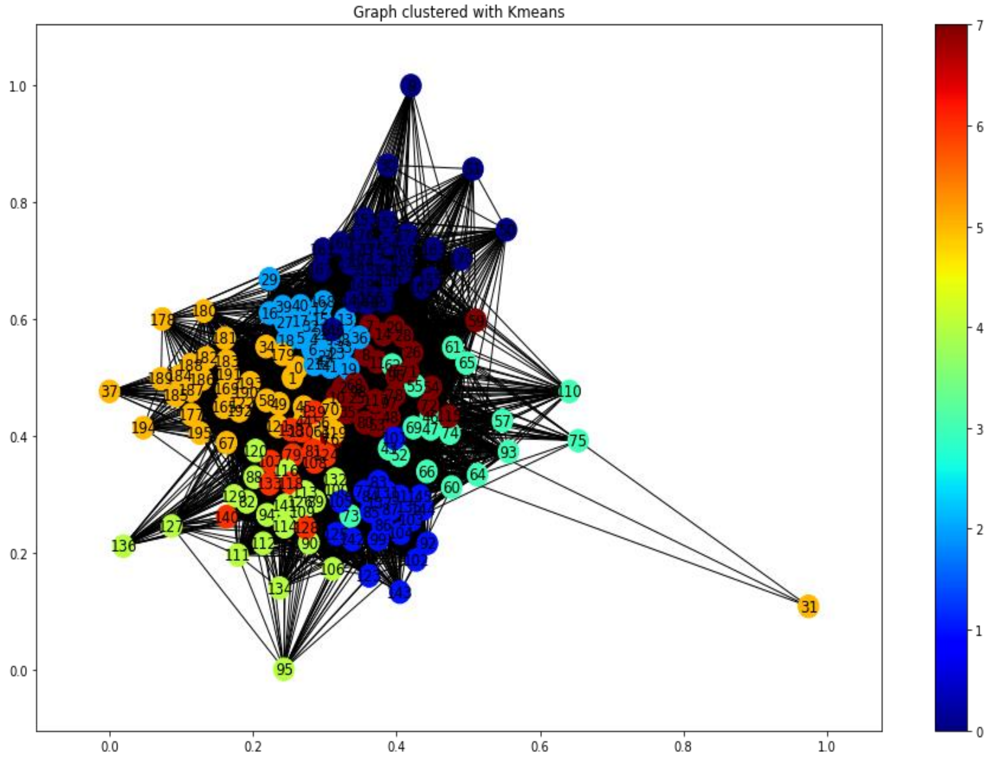
Unsupervised Spectral Clustering With K-Means



Eigenvalues



Silhouette



K-Means Clusters

- Selection of k based on the eigenvalues of the normalized Laplacian
- Selection of k based on the silhouette analysis to assess the distance between clusters

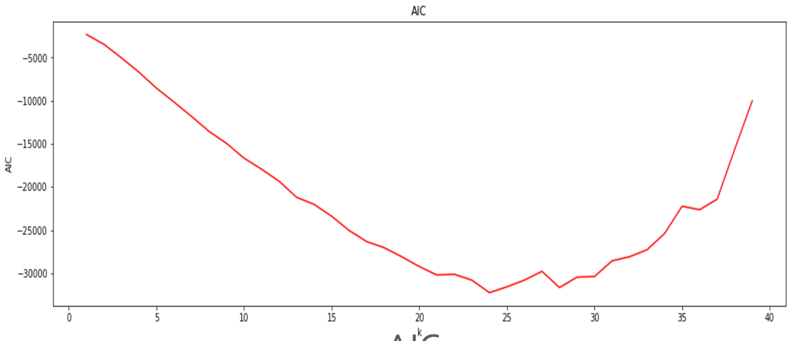
1. Data Acquisition

2. Data Exploration

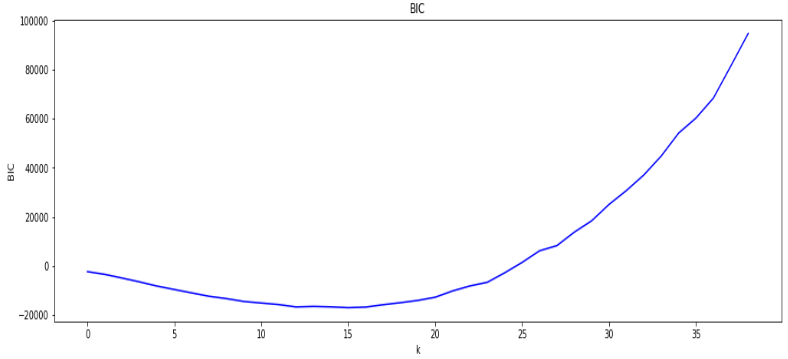
3. Data Exploitation

4. Conclusion

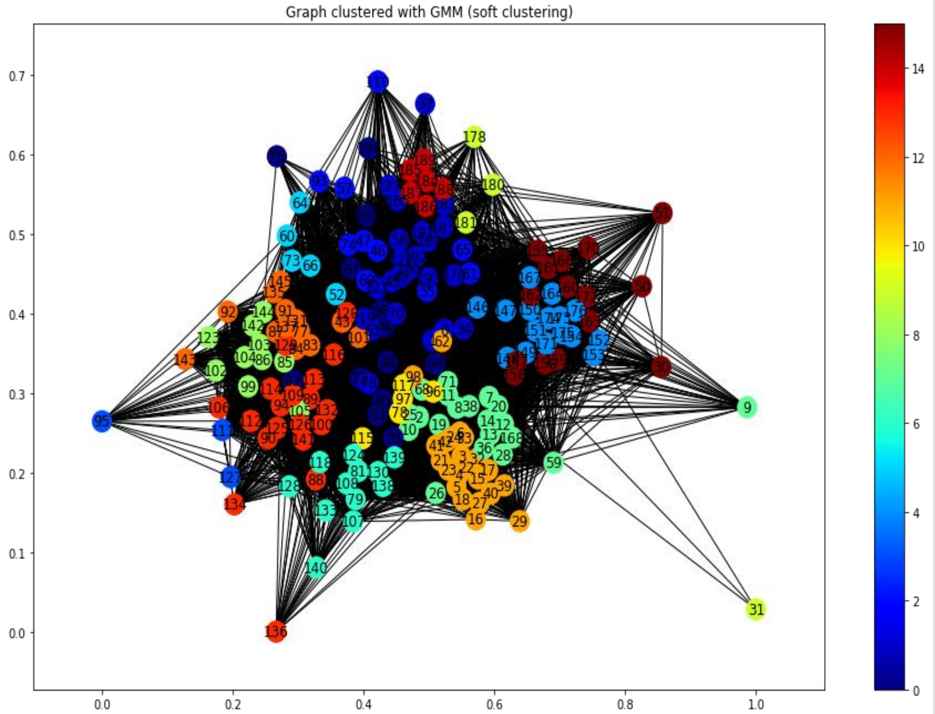
Unsupervised Spectral Clustering With GMMs



AIC



BIC



GMM Hard Clusters

- Selection of k based on AIC and BIC citerions
- 15 clusters with BIC

1. Data Acquisition

2. Data Exploration

3. Data Exploitation

4. Conclusion

Results

- Recommendation From **Automatic Speech Processing**

	Results
K-means	Biomedical signal processing Microwaves Systems and architectures for signal processing Wireless receivers Image analysis and pattern recognition microwaves Speech processing…
GMM soft clustering	Biomedical signal processing Image analysis and pattern recognition Systems and architectures for signal processing Microwaves Speech processing…
Diffusion	Advanced machine learning Applied machine learning Biomicroscopy I Biomicroscopy II Image optics Optical communications…

1. Data Acquisition

2. Data Exploration

3. Data Exploitation

4. Conclusion

Assessment

- Promising outputs
- Risk of biasing the system towards popular courses

Future Improvements

- Course rankings and or ratings for supervised learning
- Other methods of incorporating topics in the final graph
- Reinforcement learning for online adaptation