George Mason University 2015

Unsupervised Academic Curricular Evaluation Through Topic Modeling

Jean Michel Rouly [1], Huzefa Rangwala [1], Aditya Johri [2]

[1] Computer Science, [2] Information Sciences & Technology

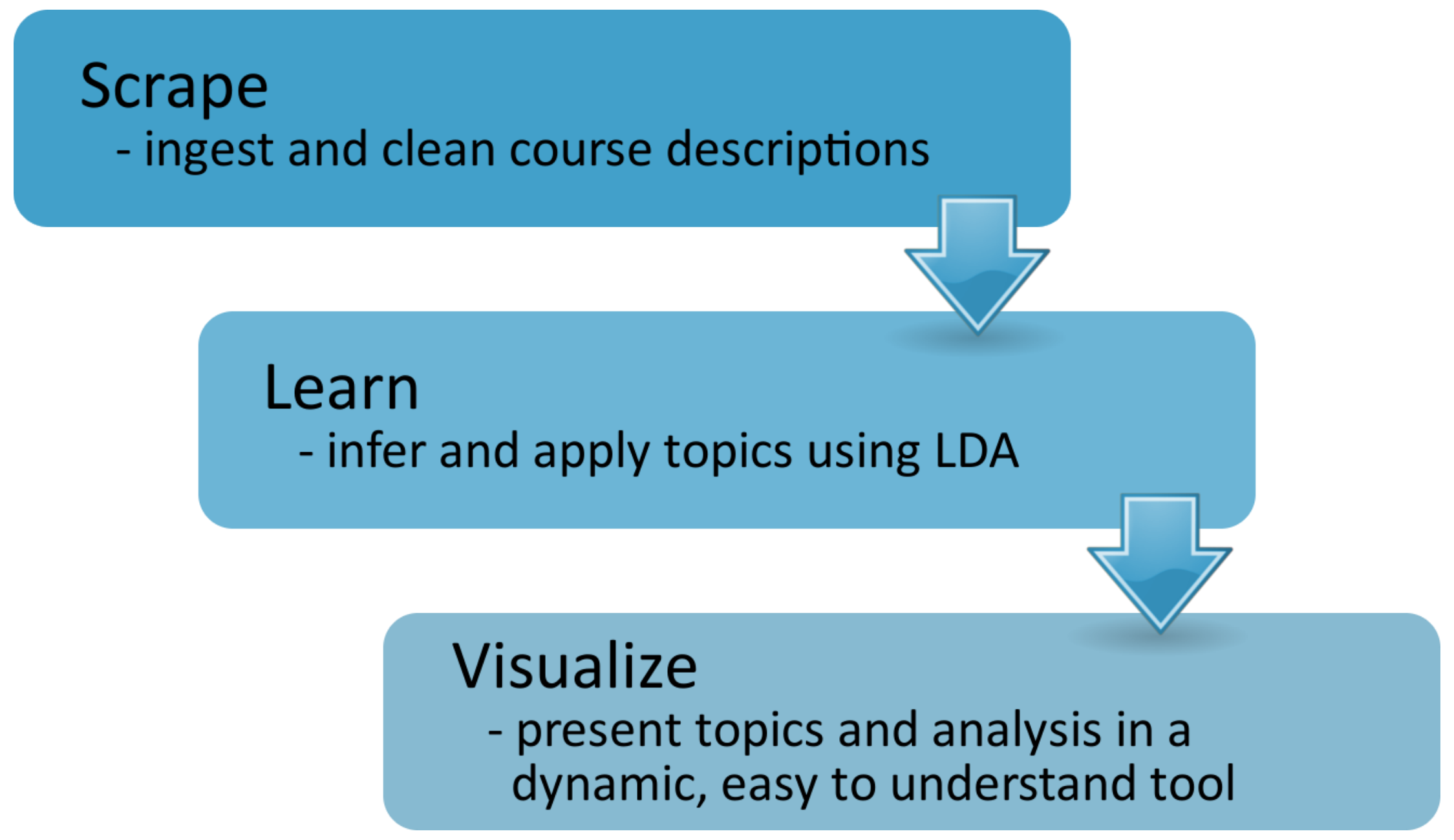
Problem Statement

Computer Science education is an increasingly important field of growth at many universities. As departments grow and change, it becomes necessary to automate the evaluation and comparison process. However, much of the published information about departments is non-standard, natural language text, not easily processed automatically.

Objective

Develop an automated system to infer topics from university course descriptions and provide analysis at the course and departmental levels.

At a Glance



Scrape
- ingest and clean course descriptions

Learn
- infer and apply topics using LDA

Visualize
- present topics and analysis in a dynamic, easy to understand tool

Latent Dirichlet Allocation

Latent Dirichlet Allocation or LDA is a highly effective topic modeling algorithm. Assumes a generative approach to topics, allowing multiple topics to mix within a single document at learned proportions. Requires only the number of topics as input.

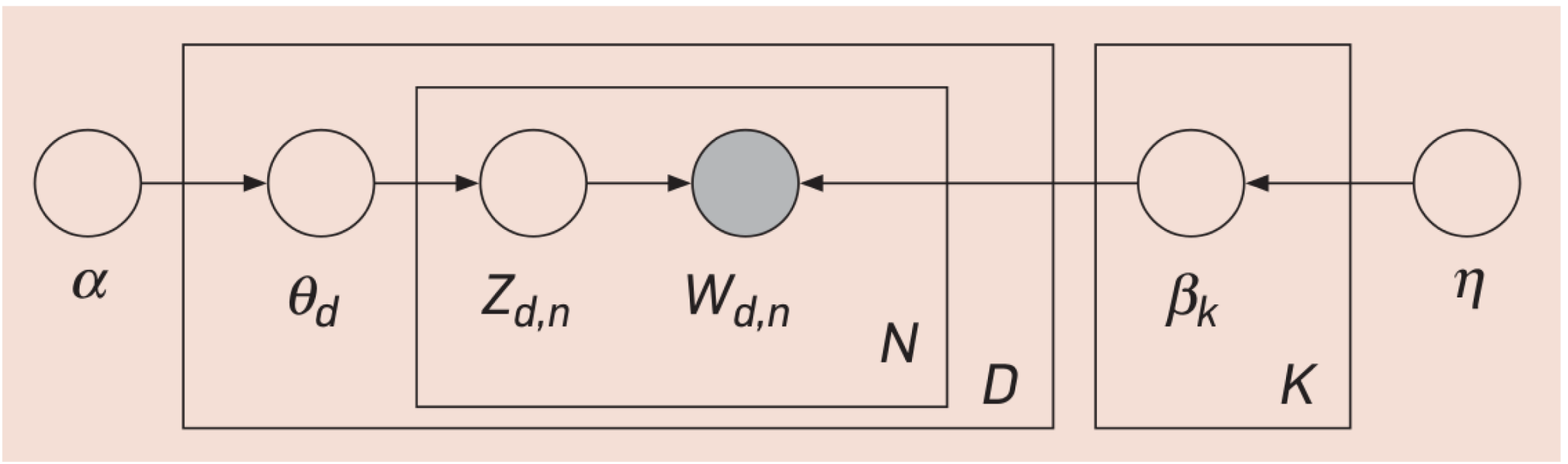


Plate diagram of LDA. Graphic reproduced from Blei 2012 [1].

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}}{w_{1:D}}$$

Posterior probability of topics. Equation from Blei 2003 [2].

University Dataset

University	Course Count
American University	32
George Mason University	145
Kansas State University	83
Louisiana State University	59
Portland State University	190
Rensselaer Polytechnic Institute	61
University of South Carolina	64
Stanford University	69
University of Utah	142
University of Tennessee, Knoxville	29
ACM Exemplar Courses (EC)	68

Computer Science Departments from 10 Universities

Sample Inferred Topics

Descriptions are cleaned, English stop words removed, and aggressively stemmed. Once LDA is run, the inferred topics look something like this.

system, softwar, design, embed, time, real, architectur, interfac, ...

web, develop, design, transform, process, technolog, applic, ...

parallel, program, comput, algorithm, model, share, perform, ...

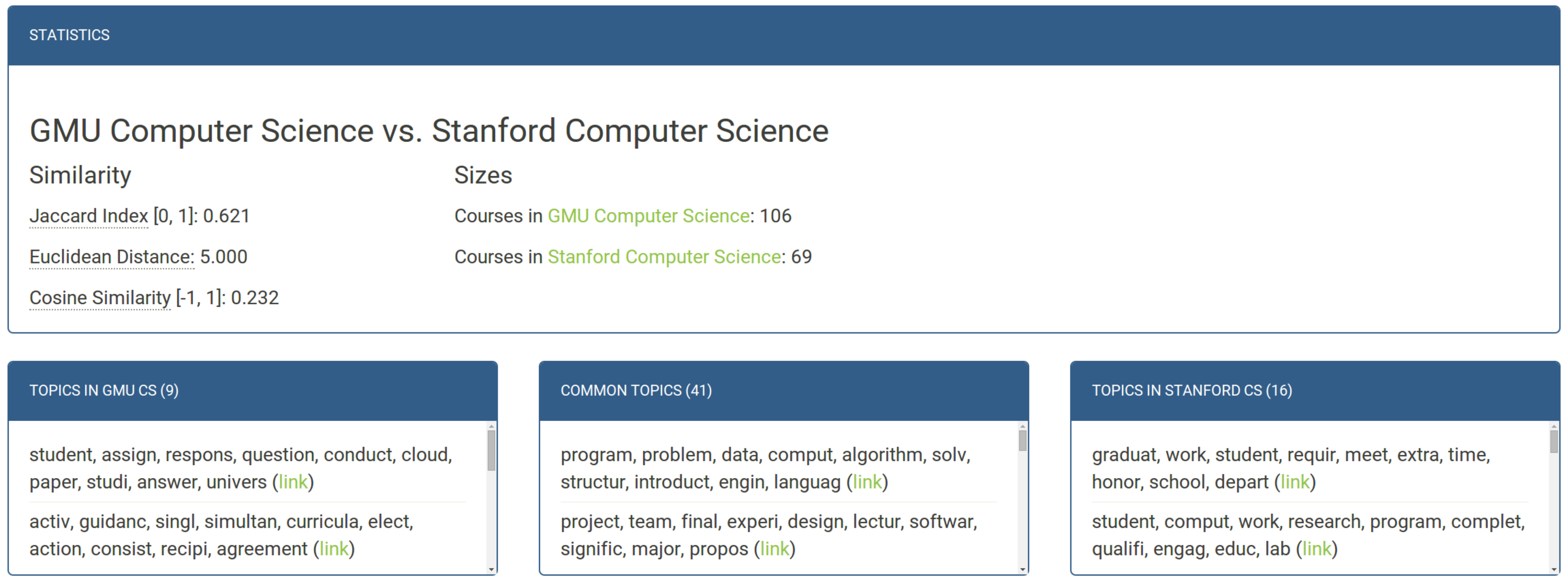
model, system, orient, object, concept, implement, method, ...

method, equat, problem, numer, differenti, linear, includ, system, ...

function, program, higher, order, recurs, write, basic, languag, ...

Visualization Tool: **Trajectory**

Trajectory is a web-based visualization and analytics platform developed as the final module of this project. Users can explore the inferred topics and automatically perform comparisons and evaluations of university departments. The tool is available at **trajectory.rouly.net** and **github.com/jrouly/trajectory**



STATISTICS

GMU Computer Science vs. Stanford Computer Science

Similarity

Jaccard Index [0, 1]: 0.621

Euclidean Distance: 5.000

Cosine Similarity [-1, 1]: 0.232

Sizes

Courses in GMU Computer Science: 106

Courses in Stanford Computer Science: 69

TOPICS IN GMU CS (9)

student, assign, respons, question, conduct, cloud, paper, studi, answer, univers (link)

activ, guidanc, singl, simultan, curricula, elect, action, consist, recipi, agreement (link)

COMMON TOPICS (41)

program, problem, data, comput, algorithm, solv, structur, introduct, engin, languag (link)

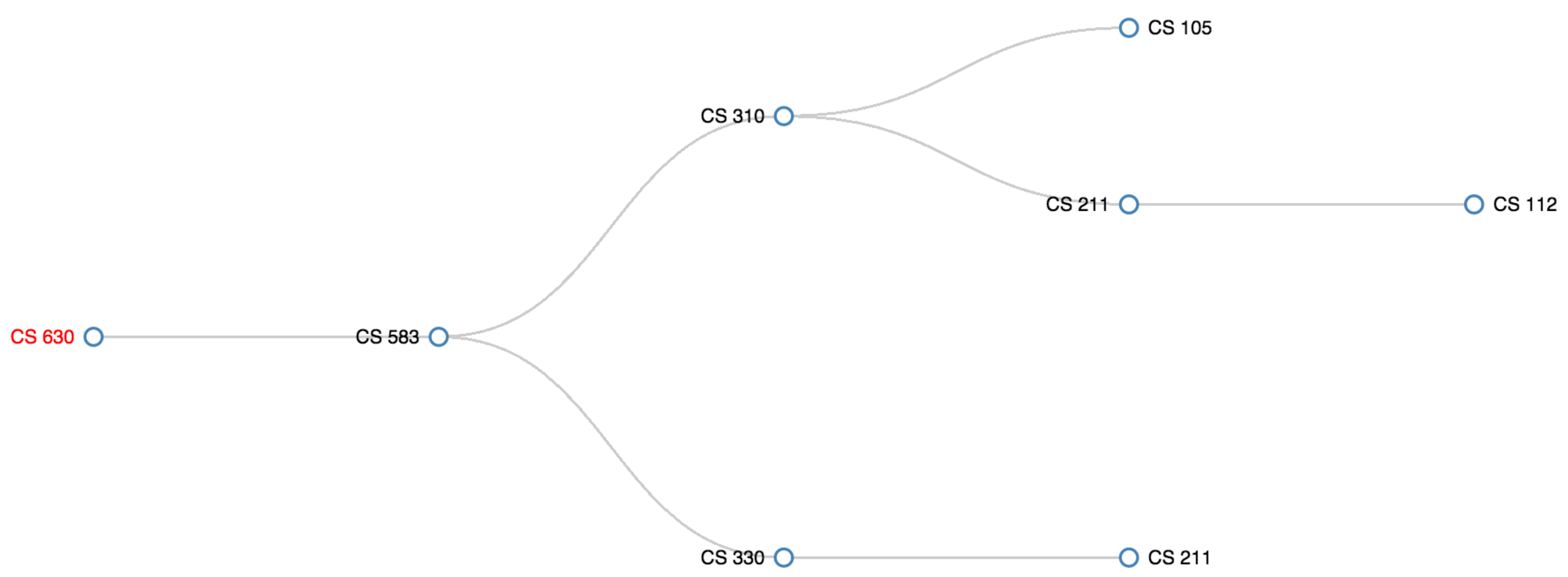
project, team, final, experi, design, lectur, softwar, signific, major, propos (link)

TOPICS IN STANFORD CS (16)

graduat, work, student, requir, meet, extra, time, honor, school, depart (link)

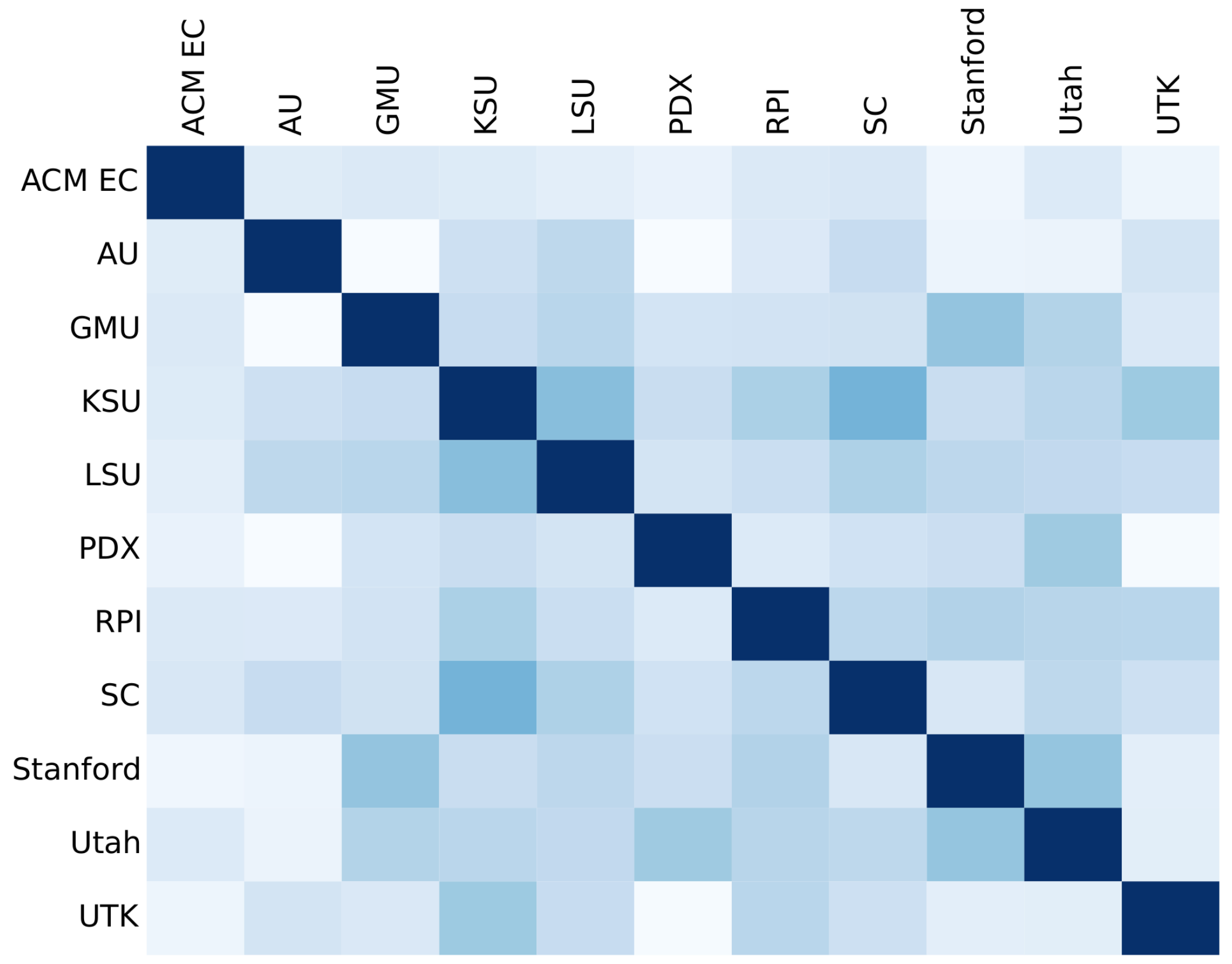
student, comput, work, research, program, complet, qualifi, engag, educ, lab (link)

Departmental comparison screenshot from **Trajectory**. Comparing George Mason University and Stanford University.



Prerequisite detail screenshot from **Trajectory**. Viewing George Mason CS 630.

Departmental Analysis



Pairwise similarity of 10 university CS departments. Similarity value is the Jaccard index of the set of topics taught in courses at each department. Darker shades are more similar.

	Prereq_μ	Prereq_σ
GMU	0.324	0.211
AU	0.278	0.134
KSU	0.273	0.213
Utah	0.257	0.249
UTK	0.201	0.256

Amount of conceptual overlap between courses and their prerequisites. Prereq_μ is average amount of overlap, Prereq_σ is standard deviation. Conceptual overlap calculated by average distance between weighted topic-vectors of courses and their prerequisites (see paper for details).

Future Work

One of the limitations of LDA is its inability to summarize topics with brief, natural language phrases or labels. We hope to address this shortcoming, perhaps by integrating our inferred topics within an existing framework of learning outcomes, specifically Bloom's Taxonomy. [3]

Additional work expanding the study to a larger dataset of universities needs to be completed. Including different types of data, such as job descriptions, might allow a continuous analysis of concepts introduced at the university level and carried into industry as expected skills.

References

[1] D. M. Blei. Probabilistic topic models. *Commun. ACM,* 55(4):77-84, Apr. 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.,* 3:993-1022, Mar. 2003.

[3] D. R. Krathwohl. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41 (4):212-218, 2002.