Work in Progress: Knowledge Graph Based Analysis of Curriculum-Industry Alignment

Introduction

Traditionally, workforce development in technical fields has followed a cycle: new technologies emerge, industries generate demand, universities adjust curricula accordingly, and graduates with relevant skills enter the labor market. However, as the pace of technological change accelerates, this cycle is facing challenges. Research indicates that nearly half of workers will need retraining for their core skills within five years, as the pace of technological change continues to accelerate [1]. Meanwhile, demand for STEM talent continues to grow. The U.S. Bureau of Labor Statistics projects that STEM occupations will grow 10.4% between 2023 and 2033, compared to 3.6% for non-STEM occupations [2]. When curriculum update cycles cannot keep pace with skill iteration, misalignment between curricula and industry demands can occur, and research suggests this misalignment is a significant factor contributing to graduates' employment difficulties [3]. In this context, systematically analyzing the alignment between curricula and industry demands provides valuable insights for students' career planning and departments' curriculum optimization.

A knowledge graph (KG) is a data structure derived from graph theory, consisting of entities (nodes) and relationships (edges) that represent connections between real-world concepts. KGs are generally heterogeneous, allowing nodes and edges to have different types (for example, a KG of person-movie-director can simultaneously represent relationships such as "a person watched a movie" and "a director made a movie"). KGs have been widely applied in search engines, recommendation systems, and question-answering systems, and have recently been adopted in education for course recommendation and learning path planning [4].

However, existing KG-based research on curriculum-industry alignment has two main limitations. First, most studies analyze only pairwise relationships between courses and skills or between skills and jobs, lacking a unified framework that integrates all three [5]. Second, existing work primarily relies on data from MOOC or job posting websites, focusing on downstream tasks such as course recommendation or job matching, with limited attention to coverage and gap analysis for specific university undergraduate curricula [6].

To address these limitations, this study proposes a KG-based framework for curriculum-industry alignment analysis. The framework models courses, skills, and jobs as a three-layer KG, using real university course syllabus and the O*NET skills database [7] as data sources. Graph structures inherently support path queries, allowing direct answers to practical questions such as: What skills can I learn from this course? What jobs require these skills? What other courses also cover these skills? Building on this, we can quantify skill coverage, identify teaching gaps, assess course-to-job accessibility, and visualize the alignment between curricula and industry demands. This paper presents a case study analyzing a computer science undergraduate program in a public institution, and the proposed framework is generalizable and can be applied to curriculum evaluation in other disciplines and institutions.

Related Work

Knowledge graphs, as a structured knowledge representation method, have been widely applied in education in recent years. [8] proposed KnowEdu, a system that constructs educational knowledge graphs for concept extraction and probabilistic association rule mining for relation identification within the target curriculum. [9] investigated prerequisite relation learning for concepts in MOOCs, proposing a representation learning-based method that better than existing approaches on Coursera datasets. In addition, as preliminary work for current study, our previous work [10] analyzed course prerequisite networks across five public universities using graph-theoretic methods and [11] integrated prerequisite networks with student performance data to identify curricular bottlenecks.

For curriculum-industry alignment analysis, [12] proposed a data analytics framework to bridge the gap between technological education and job market requirements by matching course offerings with labor market skill demands. [13] conducted surveys with software practitioners and compared industry skill expectations against software engineering curriculum, identifying significant gaps in areas such as configuration management and testing. These studies have laid the foundation for this work and motivated the framework proposed in this paper.

Methodology

In terms of Data source, there are two types of data involve in this work: course data and job data. The course data is from public universities' Computer Science (CS) undergrad program. We collected 21 courses' syllabus, including mandatory and elective courses. Each syllabus contains the course contents and learning objectives, which describe the skills students are excepted to acquire. The job data is from the O*NET database [7], the primary source of occupational information developed by the U.S Department of Labor. O*NET provides a standardized job-skill taxonomy, so we selected 152 CS related jobs involving 233 distinct skills.

For the knowledge graph construction, as shown in Figure 1, the KG contains Course, Skill, and Job nodes. Edges are categorized into two types: Course-Skill edges and Skill-Job edges. The Skill-Job edges are directly obtained from O*NET's standardized mappings. For the Course-Skill edges, we used natural language processing (NLP) based text embedding methods to project course content and skill names into a shared semantic space. We evaluated six pre-trained sentence embedding models and selected BAAI/bge-large-en-v1.5, which demonstrated the best performance for technical text. We then computed cosine similarity between each course and all skills. To determine the optimal number of edges per course, we tested different thresholds (top 3, 5, 7, and 10) and found that retaining the top 5 most similar skills for each course yielded the most accurate mappings in our case. The entire graph is stored in Neo4j, with Cypher query language supporting efficient graph traversal for subsequent analysis.

We defined four metrics to evaluate curriculum-industry alignment. **Skill Coverage** measures the proportion of industry skills connected to courses. **Teaching Gaps** identifies skills that have no connection to any course in the knowledge graph, representing missing industry requirements in the curriculum. **Course-to-Job Accessibility** uses graph path analysis to count how many jobs are reachable from each course through skill nodes.

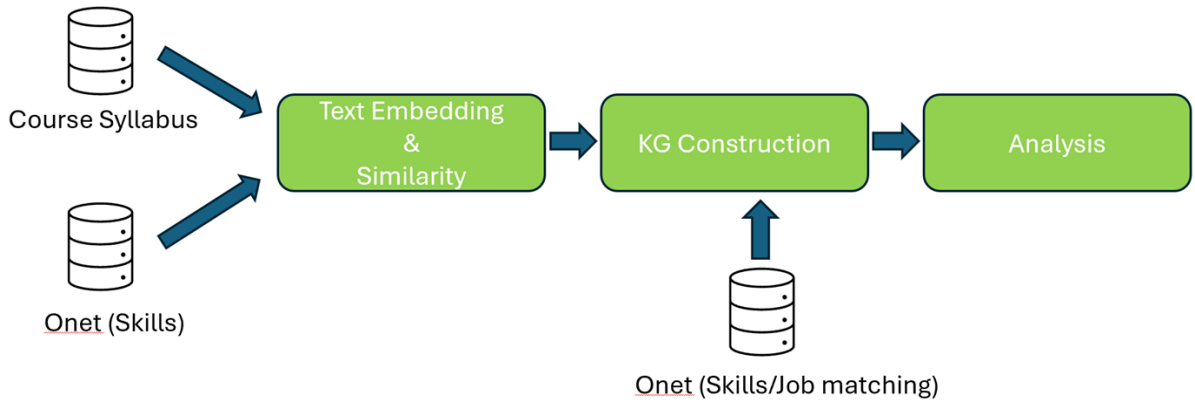

Figure 1: a subsample and the whole network structure

Result

The final KG contains 406 nodes and 2221 edges. Figure 2 illustrates the three-layer structure and a subgraph example. The subgraph shows how machine learning-related courses (as shown in Table 2) CS490ML, CS438, and CS490DL connect to 6 jobs through 5 skills.

To evaluate skill coverage, we filtered skills required by more than 5 jobs, yielding 87 high-frequency skills. Figure 3 visualizes these skills as word clouds, where green and red indicate covered and uncovered skills from the courses, respectively. Font size represents the number of jobs requiring the skill. The curriculum covers 53 of these high-frequency skills (61%), including programming languages, database technologies, development tools, operating systems,Network related skills, and machine learning frameworks. The remaining 34 uncovered high-frequency skills are categorized in Table 1. As shown in Figure 3 (Gap), cloud platforms represented by AWS and Azure, and big data platforms represented by Apache series ecosystem, are the most in-demand but uncovered skills.

Figure 4 presents the course-job skill coverage heatmap, where the x-axis represents courses, the y-axis represents occupations, and color intensity indicates the number of skills a course covers for a given occupation. The heatmap reveals several notable course-job alignments: CS490DL, CS490ML, and CS438 exhibit strong connections to Data Scientists; CS434 links closely to Database Administrators and Database Architects; CS447 corresponds to Network Architects, Network and Computer Systems Administrators, and other network-related positions; CS476 shows a unique association with Bioinformatics Technicians. In contrast, certain occupations such as Blockchain Engineers and Digital Forensics Analysts display lighter colors across most courses, indicating insufficient skill coverage for these positions in the current curriculum.

To further reveal the structural associations between course groups and job groups, we applied spectral biclustering to the job-course alignment (red lines in Figure 4 indicate cluster boundaries, partitioning the matrix into subblocks). The clustering results reveal two clear course-occupation correspondence patterns. The most prominent association appears between the AI/machine learn-

ing course group (CS438, CS490DL, CS490ML) and the data science occupation group (Data Scientists, Computer and Information Research Scientists), where the deep-colored block indicates that these courses precisely cover the core skills required for data science positions. The second clear cluster shows that systems engineering and networking courses (represented by CS286 and CS447, in the bottom-right corner) exhibit high skill coverage for related occupations (Computer Systems Analysts, Computer Network Architects).

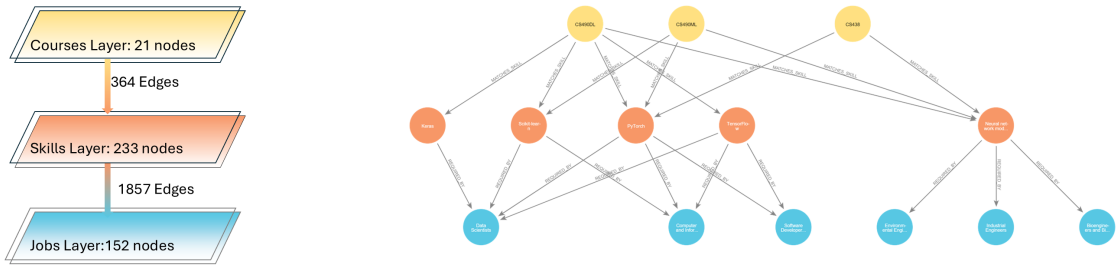

Figure 2: Subsample and Full Network Topology

Figure 3: Covered (Green) VS Gap (red) skills

Table 1: Gap Skills Category

Category	**Skills**
Cloud Platforms	AWS, Amazon EC2, Amazon Redshift, Amazon S3, Amazon DynamoDB, AWS CloudFormation, Microsoft Azure, Red Hat OpenShift
Big Data	Apache Hadoop, Spark, Kafka, Hive, Cassandra
Web Servers	Apache HTTP Server, Tomcat, Maven, Groovy
Backend Frameworks	PHP, Go, Ruby, Bash, Node.js, Swift, Django, .NET Framework, ASP.NET, Angular, React
Database	MongoDB
DevOps	Docker, Kubernetes
Systems & Networking	PowerShell, Windows Server
Frontend & Web	RESTful API

Figure 4: Biclustered Course-Job Skill Coverage Heatmap

Table 2: Course List

Course	**Title**
CS 111	Concepts of Computer Science
CS 140	Introduction to Computing I
CS 150	Introduction to Computing II
CS 234	Database and Web System Development
MATH 224	Discrete Mathematics
CS 240	Introduction to Computing III
CS 286	Introduction to Computer Organization and Architecture
CS 314	Operating Systems
CS 321	Human-Computer Interaction Design
CS 325	Software Engineering
CS 330	Programming Language
CS 340	Algorithms and Data Structures
CS 434	Database Management Systems
CS 438	Artificial Intelligence
CS 447	Networks and Data Communications
CS 476	Bioinformatics Algorithms
CS 490DL	Deep Learning
CS 490ML	Machine Learning
CS 490DIP	Digital Image Processing
CS 490CV	Computer Vision
CS 490CN	Complex Network

Discussion and Future Work

In summary, this paper proposes a knowledge graph-based framework for curriculum-industry alignment analysis, modeling courses, skills, and jobs as a three-layer knowledge graph and using semantic embeddings to automate course-skill matching. Using a computer science undergraduate program as a case study, preliminary results show that the curriculum covers general and practical skills well, while providing a solid foundation for algorithms, theory, and artificial intelligence across both required and elective courses. The 61% coverage rate of high-frequency skills indicates that the curriculum is effective in developing foundational capabilities: students acquire core skills in programming languages, database technologies, network technologies and version control, which are required by a large number of positions. However, the analysis also reveals a clear gap between the curriculum and industry demands. The gaps emerge primarily in tool-based skills such as cloud platforms, big data processing, and DevOps. The framework is generalizable and can be applied to curriculum evaluation in other disciplines and institutions.

Specifically, the 34 uncovered high-frequency skills are concentrated in several areas: cloud platforms, big data processing, DevOps tools, and various backend frameworks and web technologies commonly used in industry. These skills share a common characteristic: they are not foundational computer science concepts, but rather tools and platforms that have evolved in industry to solve practical problems. The curriculum teaches for example: "what is a database and how to write SQL," but industry needs "how to use AWS cloud database services"; the curriculum teaches "what is version control," but industry needs "how to use Docker to package and deploy new code." This gap from concepts to tools reflects a fundamental difference between academic education and industry practice.

The course-to-job accessibility analysis also reveals characteristics of the curriculum. Traditional roles (such as Software Developers and Data Scientists) are reachable from nearly all courses, indicating that the curriculum aligns well with established career paths. More notably, the heatmap shows clear specialized pathways: the networking course (CS447) leads directly to Network Architects, Network Administrators, and other network-related positions; the database course (CS434) connects to Database Administrators and Database Architects; machine learning and deep learning courses (CS438, CS490DL, CS490ML) show strong associations with Computer and Information Research Scientists and Data Scientists; the bioinformatics course (CS476) leads directly to Bioinformatics Technicians. Emerging roles (such as Blockchain Engineers and Digital Forensics Analysts) show lower accessibility, which indicates that the current curriculum does not yet offer courses targeting these areas.

Lastly, the biclustering analysis reveals two clear grouping of course sets to job sets, while the third remaining bicluster appears rather noisy. The clearest association between course groups and job groups that emerges is that connecting the AI and ML courses to the positions of Data Scientist and Computer and Information Research Scientist. The second bicluster is also relatively clean, connecting approximately computing systems oriented courses with associated systems jobs. The usefulness of biclustering is the simultaneous grouping of jobs and courses, while also exhibiting associations between the groups. As such, a high quality biclustering allows departments to create specializations or subgroups of courses based on such industrial associations. However, due to the high level of noise and low quality of the third cluster, one of our future research directions

includes pruning the data to lower the effects of noise.

These findings have implications for both students and curriculum designers. For students, the skills provided by foundational courses are versatile and transferable, while gaps in tool-based skills can be supplemented through self study. For curriculum designers, the issue is not that foundational courses are insufficient, but rather how to help students bridge the ”last mile” from classroom to industry. This could involve incorporating hands-on practice with industry tools into existing courses, or guiding students to understand what additional skills they need to acquire outside the classroom.

Regarding the proposed framework, its novelty lies in the automation and visualization. The framework automates course-skill matching through semantic embeddings, making the analysis scalable to larger curriculum without manual annotation. Meanwhile, the KG representation provides an intuitive visualization of relationships between courses, skills, and jobs, enabling students to directly explore pathways and identify gaps. Yet, only course syllabus are used as input. Incorporating richer course materials such as assignments and textbooks could improve the accuracy of course-skill matching. Additionally, skill matching is constrained to the finite set provided by O*NET, and the database contains some noise, such as outdated technologies that are still retained. We will address these limitations in our future work.

References

[1] World Economic Forum, The Future of Jobs Report 2020,' Oct. 2020.[Online]. https://www.weforum.org/publications/the-future-of-jobs-report-2020/

[2] U.S. Bureau of Labor Statistics, Employment in STEM Occupations, 2024. [Online]. Available: https://www.bls.gov/emp/tables/stem-employment.htm

[3] Aljohani, N., Aslam, A., Khadidos, A. & Hassan, S. Bridging the skill gap between the acquired university curriculum and the requirements of the job market: A data-driven analysis of scientific literature. *Journal Of Innovation & Knowledge*. **7**, 100190 (2022)

[4] Qu, K., Li, K., Wong, B., Wu, M. & Liu, M. A survey of knowledge graph approaches and applications in education. *Electronics*. **13**, 2537 (2024)

[5] Groot, M., Schutte, J. & Graus, D. Job posting-enriched knowledge graph for skills-based matching. *ArXiv Preprint ArXiv:2109.02554*. (2021)

[6] Fettach, Y., Bahaj, A. & Ghogho, M. JobEdKG: An uncertain knowledge graph-based approach for recommending online courses and predicting in-demand skills based on career choices. *Engineering Applications Of Artificial Intelligence*. **131** pp. 107779 (2024)

[7] National Center for O*NET Development, O*NET® 30.1 Database Content License, O*NET Resource Center, 2024. [Online]. Available: https://www.onetcenter.org/license_db.html.

[8] Chen, P., Lu, Y., Zheng, V., Chen, X. & Yang, B. Knowedu: A system to construct knowledge graph for education. *Ieee Access*. **6** pp. 31553-31563 (2018)

[9] Pan, L., Li, C., Li, J. & Tang, J. Prerequisite relation learning for concepts in moocs. *Proceedings Of The 55th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers)*. pp. 1447-1456 (2017)

[10] Yang, B., Gharebhaygloo, M., Rondi, H., Hortis, E., Lostalo, E., Huang, X. & Ercal, G. Comparative analysis of course prerequisite networks for five Midwestern public institutions. *Applied Network Science*. **9**, 25 (2024)

[11] Yang, B., Gharebhaygloo, M., Rondi, H., Banu, S., Huang, X. & Ercal, G. Analysis of student progression through curricular networks: a case study in an Illinois public institution. *Electronics*. **14**, 3016 (2025)

[12] Karakolis, E., Kapsalis, P., Skalidakis, S., Kontzinos, C., Kokkinakos, P., Markaki, O. & Askounis, D. Bridging the gap between technological education and job market requirements through data analytics and decision support services. *Applied Sciences*. **12**, 7139 (2022)

[13] Garousi, V., Giray, G., Tuzun, E., Catal, C. & Felderer, M. Closing the gap between software engineering education and industrial needs. *IEEE Software*. **37**, 68-77 (2019)