GENERATING LEARNING PATHS FROM JOB POSTINGS VIA BAYESIAN NETWORKS

AUTHORS:

Florin STOICA
Dana SIMIAN
Laura Florentina STOICA
Elena-Cristina RĂULEA

Lucian Blaga University of Sibiu, Romania

INTRODUCTION

Objectives

To develop an automated method for generating learning paths based on skills extracted from job postings, aimed at guiding learners (especially freelancers) towards acquiring market-demanded competencies.

Background

The ENTEEF project addresses the gap between education and labor market needs by analyzing freelancing jobs, identifying key skills, and offering targeted courses and tools to support lifelong learning.

Labor Market Relevance

The proposed method directly responds to employer demands by providing a scalable, data-driven model to support career guidance and the acquisition of relevant skills for digital freelancing and entrepreneurship.

GOALS

Vision	Bridge the gap between education and real-world job market needs in freelancing careers.
Plan	Analyze ~30,000 job offers to automatically extract key skills required by employers.
Support	Use the Competency Assessment Tool (CAT) to identify individual skill gaps and recommend MOOCs.
Ideas	Model skill relationships using Bayesian Networks to guide personalized upskilling paths.
Strategy	Apply Raw Mutual Information (RMI) and Normalized Mutual Information (NMI) to weight skill dependencies.
Team	Collaboration between researchers, educators and project stakeholders within ENTEEF.
Motivation	Help learners decide what to learn next, especially in dynamic and fragmented freelancing markets.
Succes	Deploy a data-driven methodology that generates validated, personalized learning paths in the ENTEEF platform.

RESEARCH FOCUS

Problem and Goal

Within the ENTEEF project, the challenge is to automatically identify the skills demanded by the job market by extracting them from around 30,000 job postings, in order to generate structured learning paths that prepare students and freelancers with competencies aligned to real-world job requirements.

Methodology

A Bayesian Network is used where each skill is represented as a node and edges model dependencies or co-occurrences between skills. The edges are weighted using raw mutual information and normalized mutual information scores to rank skill relationships and guide the generation of optimal learning sequences.

Results and Relevance

This data-driven probabilistic approach produces learning paths that realistically model skill relationships from labor market data, enabling learners to upskill effectively according to market demands, particularly supporting entrepreneurship and freelancing careers.

METHODOLOGY

1st Step — Data Collection and Skill Extraction

- Dataset: ~30,000 job offers from Upwork, collected via custom Python web scraper by ENTEEF team.
- Ethical considerations: Only publicly accessible data, client info anonymized.
- *Natural Language Processing* (NLP) pipeline for skill keyword extraction (cf. Rahayu et al., 2023).
- Normalization of skill names and ranking of top 60 in-demand skills.
- Binary jobs × skills matrix $M_{30,000\times60}$ built: M[i,j]=1 if skill *j* in job *i*, else 0.

2nd Step — Bayesian Network Structure Learning

- Skills modeled as binary variables.
- *Hill-Climbing Search* (HCS) algorithm for structure learning using pgmpy's HillClimbSearch class with BIC scoring (Ankan and Textor, 2024).
- Result: Directed Acyclic Graph (DAG) G where edges represent dependencies (e.g., "HTML" → "CSS").
- *Learned Bayesian Networks* (BN) encodes co-occurrence and potential learning order of skills.

METHODOLOGY

3rd Step Arc Weighting Using Mutual Information

- Two weighting schemes: *Raw Mutual Information* (RMI) and *Normalized Mutual Information* (NMI).
- RMI measures dependency strength between skill pairs based on joint and marginal probabilities.
- NMI normalizes RMI to [0, 1], accounting for entropy to compare connection strengths.
- Pruning thresholds applied: RMI<0.009, NMI<0.03 to filter weak connections
- Result: Reduced network with stronger, more meaningful edges.

Statistic	RMI – Full Network	RMI – Pruned Network (threshold: 0.009)	NMI – Full Network	NMI – Pruned Network (threshold: 0.03)
Count	210	85	210	96
Mean	0.016	0.035	0.084	0.169
Median	0.006	0.027	0.024	0.136
Std Dev	0.022	0.023	0.112	0.117
Min	0	0.009	0	0.036
Max	0.102	0.102	0.492	0.492

METHODOLOGY AND CONTRIBUTION

4th Step Learning Path Generation

- Identify target skills for a given job role.
- Extract relevant BN subgraph including target and connected skills.
- Perform weighted graph search from learner's current skills to targets, prioritizing edges with higher weights (e.g., using Dijkstra's or greedy approach).
- Example: Path HTML → CSS → JavaScript → React as a recommended learning sequence.
- Integration with ENTEEF platform: skill sequence linked to MOOCs, skipping mastered skills via Competency Assessment Tool.

Contribution

- Developed an empirical methodology leveraging real-world job market data for skill dependency modeling.
- Applied Bayesian network structure learning with mutual information weighting to uncover skill relationships.
- Proposed a data-driven approach to generate personalized learning paths aligned with actual job requirements.
- Incorporated ethical data collection principles ensuring privacy and compliance.
- Validated pruning and weighting schemes to optimize network clarity and relevance.
- Enabled adaptive, competency-aware recommendation of MOOC sequences to optimize learner progression.

MODEL DEVELOPMENT

Initial Bayesian Network Structure

- The initial Bayesian Network (BN) structure obtained using HillClimbSearch is shown in Figure 1.
- To improve clarity, very weak connections (edges with NMI below a threshold) were filtered out.
- The resulting pruned network is shown in Figure 2, using an NMI threshold of 0.03.

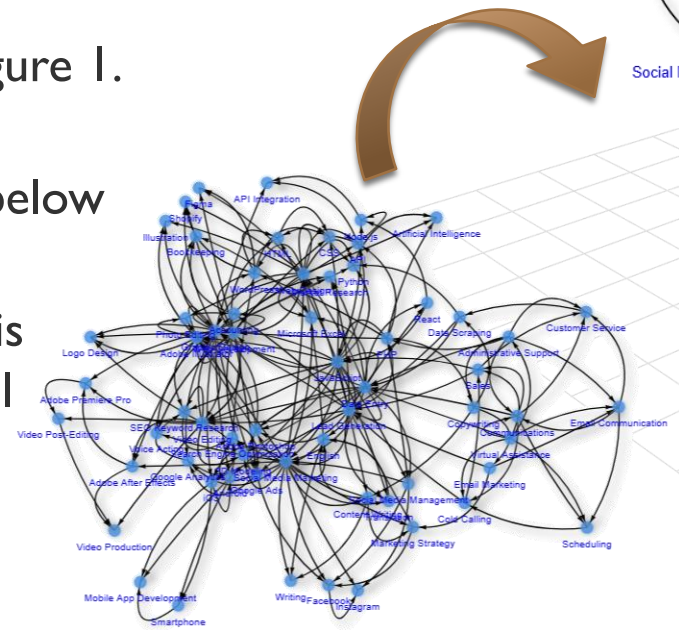


Fig 1. The initial learned Bayesian Network structure

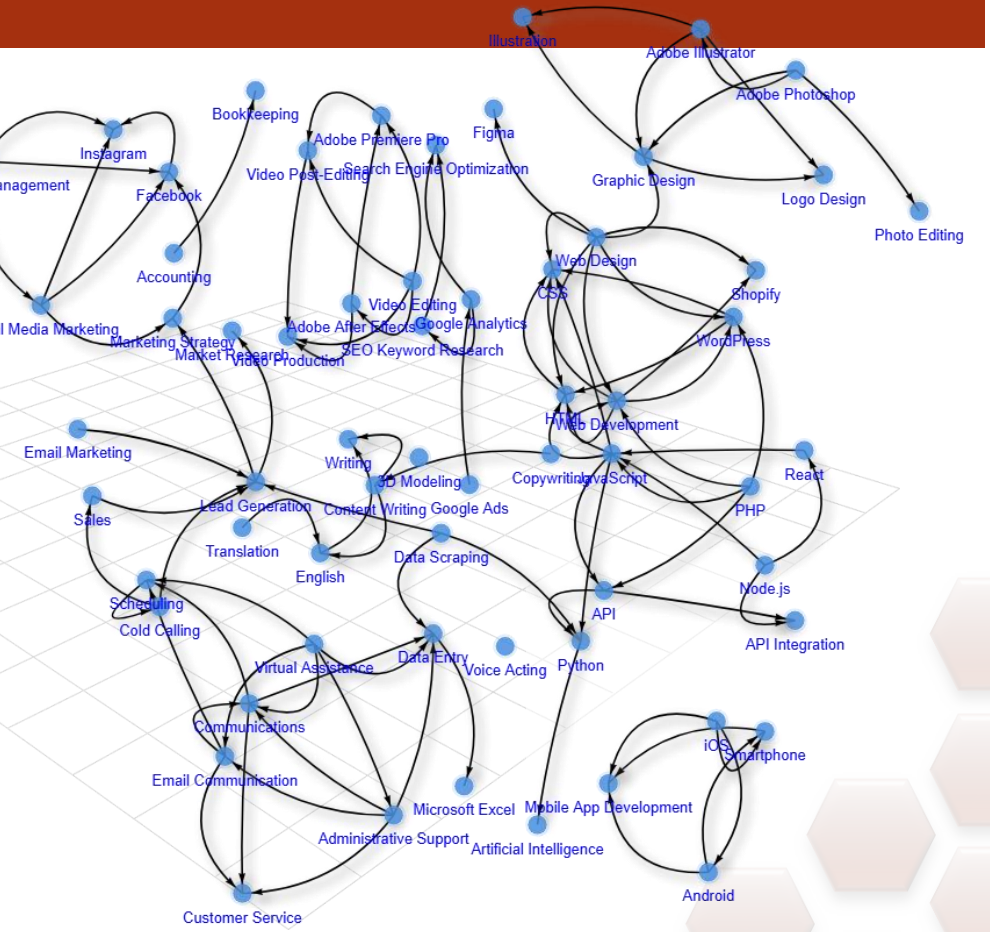
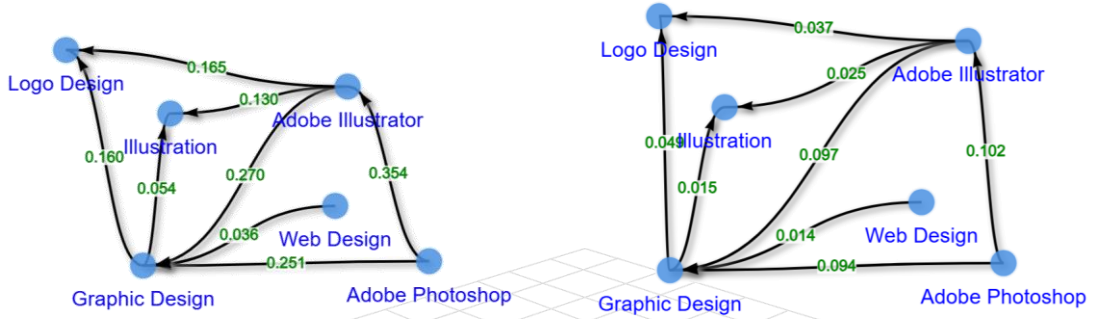


Fig 2. The pruned network, using NMI with threshold 0.03

MODEL DEVELOPMENT AND EVALUATION

Comparison of RMI and NMI weighting approaches

- Extracted subgraphs centered on each skill from both RMI and NMI weighted models.
- For most requested skills, RMI and NMI produce similar BN structures.
- Figure 3 shows identical subgraphs for Graphic Design skill (RMI and NMI).
- Figure 4 and Figure 5 show slight differences for Web Development skill.



a) Arc weighting with NMI method b) Arc weighting with RMI method

Fig 3. The identical extracted subgraphs for Graphic Design skill

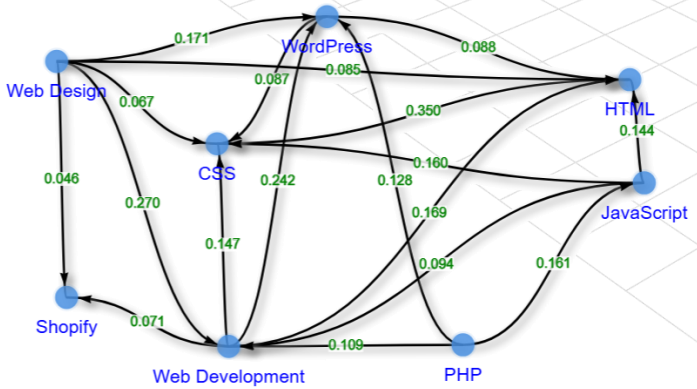


Fig 4. Extracted subgraph of the BN (NMI weights) for Web Development skill

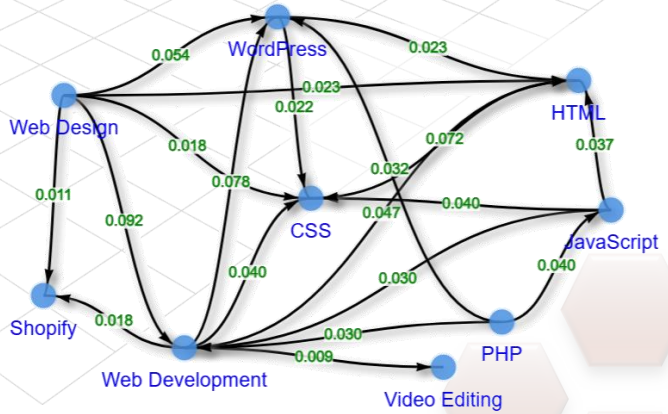


Fig 5. Extracted subgraph of the BN (RMI weights) for Web Development skill

MODEL DEVELOPMENT AND EVALUATION

Automated Comparative Evaluation: Jaccard Similarity

- Similarity of all subgraphs measured via Jaccard similarity on node sets.
- Given a graph G=(V,E) and nodes u,v, Jaccard similarity of their neighborhoods N(u),N(v) is:

$$J(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

- Only skills with node similarity not equal to 1 are shown in Figure 6.

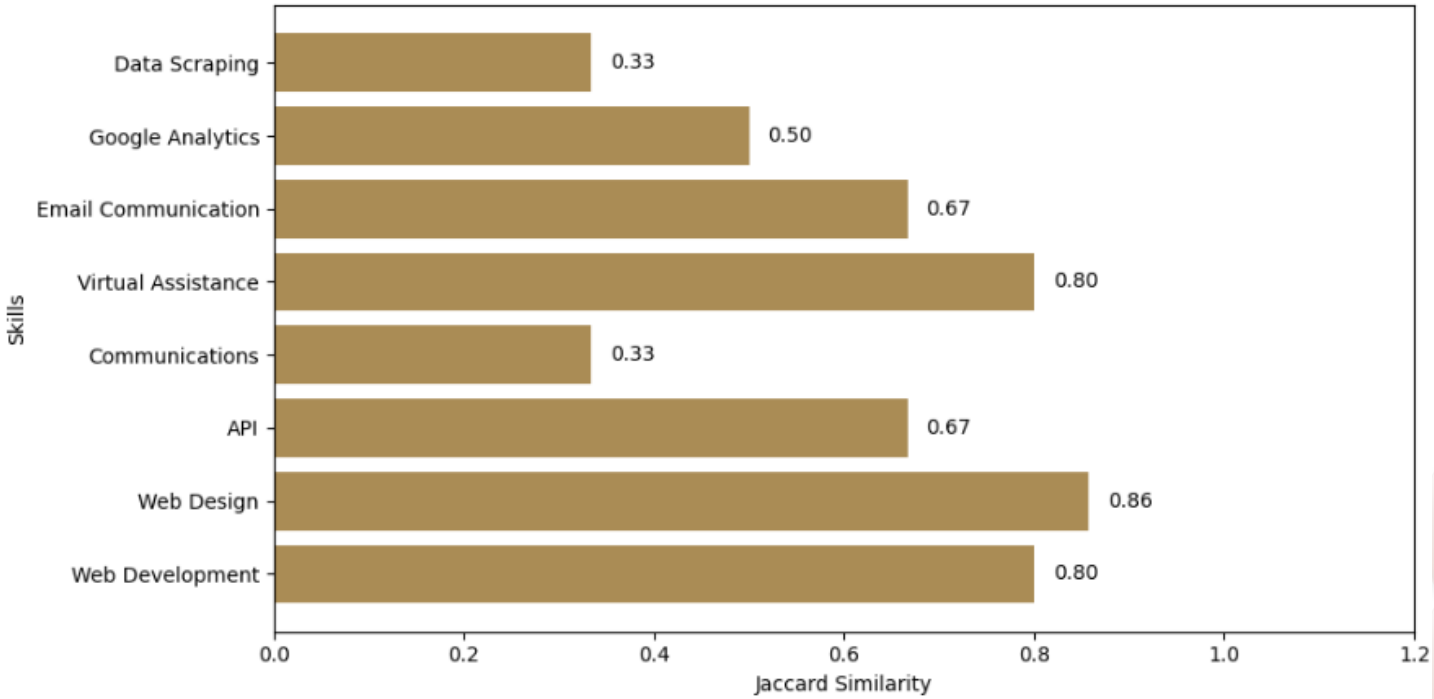


Fig 6. Skills with different Jaccard similarity

DISCUSSION

Hypothesis

- Bayesian Networks learned from large-scale job postings, weighted by mutual information metrics, can generate learning paths reflecting realistic upskilling trajectories aligned with job market demands.

Empirical Support from Real Job Data

- Built from ~30,000 job postings covering diverse freelancing roles; BN structure captures real skill dependencies.

Dual Weighting and Comparison

- Used Raw Mutual Information (RMI) and Normalized Mutual Information (NMI) for arc weights; compared models with Jaccard similarity and subgraph visualization.

Alignment with Learning Sequences

- Generated paths are coherent with BN structure and logical pedagogical progression.

Contribution

- Data-driven, scalable method bridging labor market analytics and educational technology, unlike prior methods based on ontologies or student learning data.

DISCUSSION

Limitations

- Learning paths inferred from job data, not yet validated with real learner outcomes.
- BN edges show statistical dependencies, not guaranteed pedagogical prerequisites.
- System not tested in real learning environments (UX and motivation unknown).
- Empirical pruning threshold may affect path consistency.

Advantages and Implications

- Enables alignment of education with dynamic job market needs, especially in tech and freelancing.
- Supports targeted, demand-driven lifelong learning paths.
- BN can be regularly updated to reflect emerging skills and trends.
- Methodology adaptable to other domains with sufficient job data (e.g., healthcare, cybersecurity).

Summary

- BN-based approach shows potential for personalized, labor market-informed education, but requires future validation via learner outcomes and expert review.

CONCLUSIONS AND FUTURE WORK

Conclusions

- Introduced a novel method combining skill extraction from job postings and Bayesian Networks to generate learning paths.
- The approach automates alignment between labor market needs and educational content in the ENTEEF platform.
- Mutual information-based weighting effectively captures skill relatedness and logical prerequisite structures.

Future Work

- Investigate the NOTEARS algorithm to jointly learn skill dependencies and edge weights via continuous optimization.
- Compare NOTEARS-based paths with current Bayesian models to assess realism and guidance quality.
- Validate generated paths with experts and evaluate learner outcomes.
- Integrate the method into the ENTEEF platform for real-world application.

REFERENCES

- Ankan, A. and Textor J. (2024), 'pgmpy: A Python Toolkit for Bayesian Networks'. Journal of Machine Learning Research, 25(265), 1-8.
- Carroll, D. and Schlippe, T. (2023), 'Connecting Learning Material and the Demand of the Job Market Using Arti-ficial Intelligence', Artificial Intelligence in Education Technologies: New Development and Innovative Practic-es, pp. 282-298.
- Culbertson, M. J. (2016). Bayesian Networks in Educational Assessment: The State of the Field. Applied Psycho-logical Measurement, 40(1), 3–21.
- Dubois, D., Prade, H. and Smets, P. (2008), 'A definition of subjective possibility'. International Journal of Ap-proximate Reasoning, 48(2), 352-364. Available: https://doi.org/10.1016/j.ijar.2007.01.005.
- ENTEEF Project Fostering Entrepreneurship through Freelancing (2025), Erasmus+ Programme Project Website. Available: https://enteef.uek.krakow.pl/.
- Rahayu, N.W., Ferdiana, R. and Kusumawardani, S.S. (2023), 'A systematic review of learning path recommender systems', Education and Information Technologies, 28 (6), 7437-7460. Available: https://doi.org/10.1007/s10639-022-11460-3.
- Shen, H., Liu T. and Zhang Y. (2020). 'Discovery of Learning Path Based on Bayesian Network Association Rule Algorithm', International Journal of Distance Education Technologies, 18(1), 117-130.
- Sklearn Metrics Normalized Mutual Info Score - Scikit-learn 1.7 documentation (2025). Available: https://scikit-learn.org/1.7/modules/generated/sklearn.metrics.normalized_mutual_info_score.html
- Zhang, M., Jensen K. N., Sonniks S. D. and Plank, B. (2022), 'SkillSpan: Hard and Soft Skill Extraction from Eng-lish Job Postings', Proceedings of NAACL 2022 (Association for Computational Linguistics), 4962-4984.
- Zheng, X., Aragam, B., Ravikumar, P. and Xing, E. P. (2018), 'DAGs with NO TA¬E¬RS: Continuous optimization for structure learning'. Advances in Neural Information Processing Systems (NeurIPS 2018), 12 pages.

ACKNOWLEDGEMENTS

Co-funded by the European Union (Project no 2024-1-PL01-KA220-HED-000248152). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them. We gratefully acknowledge the contributions of the ENTEEF team members: Bartłomiej Balsamski, Jakub Kanclerz, Mukhammad Andri Setiawan, Ahmad Fathan Hidayatullah and Ratih Dianingtyas Kurnia for their valuable work in the primary data collection and preparation phase of this research. Their efforts in gathering and organizing job posting data provided a critical foundation for the analysis and modelling presented in this study.

ENTEEF Fostering Entrepreneurship through Freelancing

Thank You!

AUTHORS:

Florin STOICA
Dana SIMIAN
Laura Florentina STOICA
Elena-Cristina RĂULEA

Lucian Blaga University of Sibiu, Romania

ENTEEF Fostering Entrepreneurship through Freelancing

Contact author: florin.stoica@ulbsibiu.ro