On Competition for Undergraduate Co-op Placements: A Graph Mining Approach

Yuheng Helen Jiang, Lukasz Golab
University of Waterloo, Canada
EDM 2016
Agenda

• Summary
• Graph Definition
• Data Overview
• Methodology & Results
• Conclusion
Summary

Objective:
• Characterize competition among programs

How:
job interviews (student, job)
Methodology – Graph Definition

- $J_i$: a list of distinct jobs that interviewed students from program $i$
Methodology – Graph Definition

\[ e_{ij} = \frac{|J_i \cap J_j|}{|J_i|} \]

Fraction of jobs in \( J_i \) that also appear in \( J_j \).

Computer Engineering

Electrical Engineering

380

213

278

0.4338

0.3592 = \frac{213}{593}
Data Overview

• One term of interview data in Summer 2014
  – 16,855 student-job interview pairs
  – 2,890 jobs
  – 4,194 students from 93 academic programs

• Job
  – Job title, advertised programs, advertised seniority

• Student
  – Academic program, academic year
Results – Two graphs

• Senior program graph:
  – 88 vertices
  – 1315 pairs of directed edges

• Full program graph vs. Senior program graph
  – Less edges
  – Thicker edges
Motivation -> Methodology

• Unclear differences among academic programs
  – Similar programs: near-cliques, clusters
• Increasing need for multi-disciplinary and well-rounded education
  – Multi-disciplinary programs: outliers
• Areas for job promotion
  – Competing programs: fan-out metric
Results – Near Cliques

- density >= 0.8
- top 5 percent edge weights
- near clique size >= 3
Results – Community Detection
Results – Similar programs

Resolution=1.7
(2 Clusters)

Resolution=1.1
(4 Clusters)

Resolution=0.6
(7 Clusters)

Resolution=0.1
(23 Clusters)
Methodology – Multi-disciplinary programs

• **Multi-disciplinary programs**: Check whether a program has strong connections to other programs from multiple clusters

• **Entropy** of the distribution of edge weights among different communities

\[ \sum_i -p_i \log_2 p_i \]
Methodology – Multi-disciplinary programs

\( p_i \): fraction of total weights of the outgoing edges from a given program to the programs only in cluster \( i \)

\[
\begin{align*}
    P_{red} &= \left( \sum_{i=\text{competitors in red}} e_{\text{Medicinal Chemistry},i} \right) / 3.25 \\
    &= (0.75 + 0.75 + 0.25) / 3.25 \\
    &= 0.54
\end{align*}
\]
Methodology – Multi-disciplinary programs

- $p_{\text{blue}} = 0.54$, $p_{\text{blue}} = 0.23$, $p_{\text{green}} = 0.15$, $p_{\text{purple}} = 0.08$
- Entropy_{medicinal chemistry} = 1.67
Results – Multi-disciplinary programs

• Question: well-rounded students or sets of specialized students?
Results – Multi-disciplinary programs

- Question: well-rounded students or sets of specialized students?
  - Well-rounded > specialized

**Top 5**
- Science & Business/Biochemistry
- English Literary Studies
- Science & Business/Environmental Biology
- Science & Business

**Bottom 5**
- Geological Engineering
- Software Engineering
- French
- Mechatronics Engineering
- Civil Engineering
Methodology – Competing programs

Set fan-out: fraction of jobs that interviewed students from the given program which also interviewed at least one student from another program

\[
\text{Set Fan Out}_i = \frac{\left| \bigcup_{j \neq i} (J_i \cap J_j) \right|}{|J_i|}
\]

Set Fan Out_{CompEng} = \frac{213}{593} = 0.3592
Results – Competing programs

16 programs do not have any jobs that interviewed only their students

8 programs have more than 30 percent of the jobs only interviewed their students
Conclusion

• Similar programs
  – Academic programs did not always aligned well with the groups of closely connected programs
  – Clusters can be used to create job categories and academic specializations

• Multi-disciplinary programs
  – Identify programs that may have broad job qualifications

• Competing programs
  – Attract more employers that offer jobs to programs that face high competitions
• Thank you!
• Questions?
Backup Slide
Methodology – Similar programs

- Community detection
  - **Louvain Modularity** [14]: unsupervised
  - Maximize a modularity metric \( Q \)
  - Compares the discovered communities with random connectedness [80]
    - Range: 0 to 1
  - Resolution parameter \( (t) \): controls number and size of clusters
    - Default: 1
    - Different values of \( t \) in steps of 0.1
Results – Multi-disciplinary programs

• Question: well-rounded students or sets of specialized students?