Collaboration Network Analysis

R04921049 柯劭珩

B01901121 李律慈

Outline

- Data Source
- Purpose
- Process
- Visualization
- Result & Observation
- Conclusion

Data Source

- From Algorithms Class, Fall 2015
- Need to specify collaboration (with who)
  in all HW problems (25 in total)
- All the collaboration together forms a social network
  - 152 Students as nodes
  - 866 Collaboration Relationship (Directed)
  - Edge weight = The problem's weight in grades
  - Aggregated to weighted simple graph

Purpose

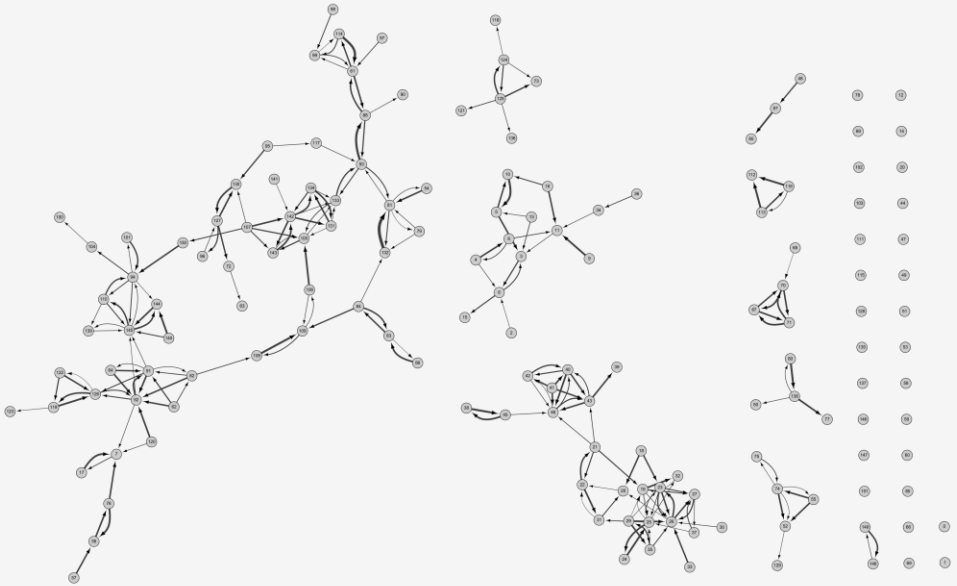
- Analyze the Collaboration Network
- Obtain network parameters
  - in degree & out degree
  - eigen centrality (who is more influential)
  - page rank
  - hubs and authorities in HITS
- Relation between parameters and HW score!
- **Can we predict HW score base on role in the network?**

Process

- Use *Python & networkX* to get following:
  - in-degree & out-degree (aggregated / unaggregated)
  - aggregated edge weight
  - eigen centrality
  - page rank
  - hub/authorities by HITS
- Use *Cytoscape* & above parameters to visualize
- Use *Weka* to mine the data

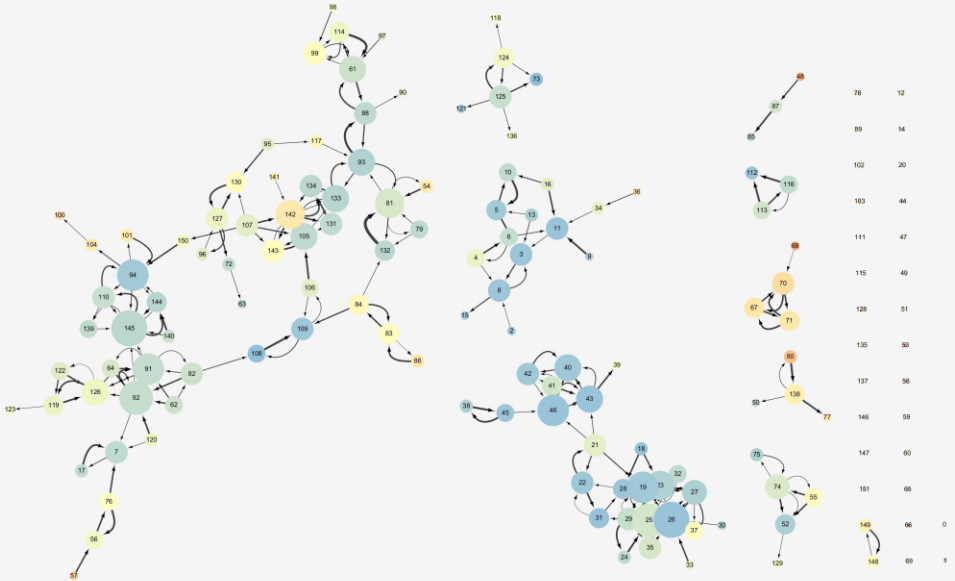
Edge thickness =
Edge weight

Visualization



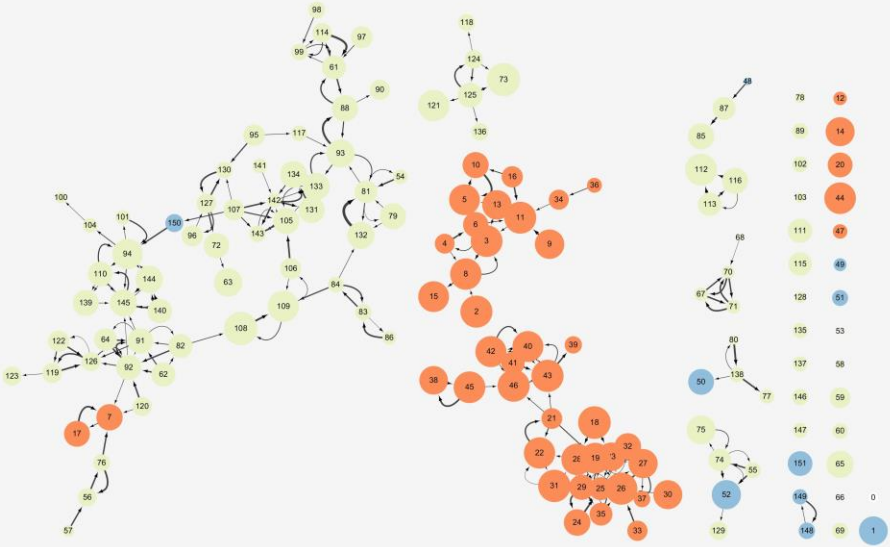
Node color = Grade
Node size = Degree

Visualization



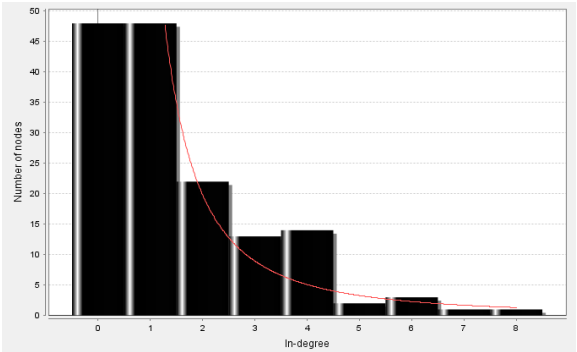
Node color = Identity
Node size = Grade

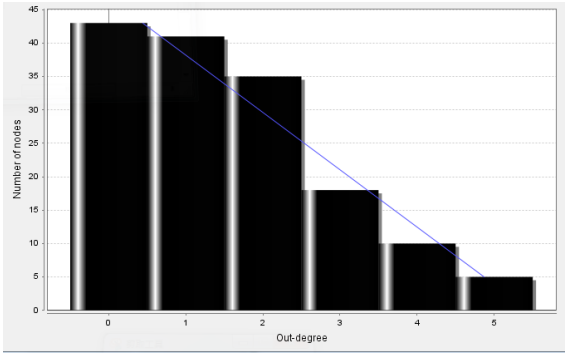
Visualization



Network Analyzing

- 152 nodes (28 isolated), 242 weighted edges
- Clustering Coefficient = 0.174 (Very high)
- 10 connected components (omitting singletons)
- In-degree like power law, out-degree like straight line



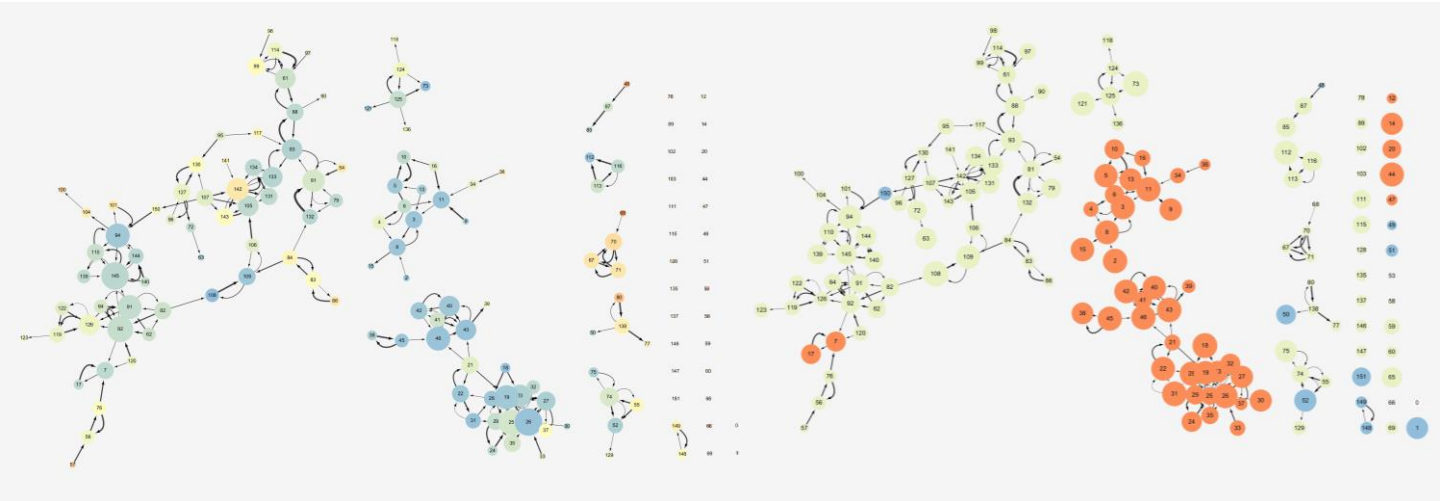


Data mining result

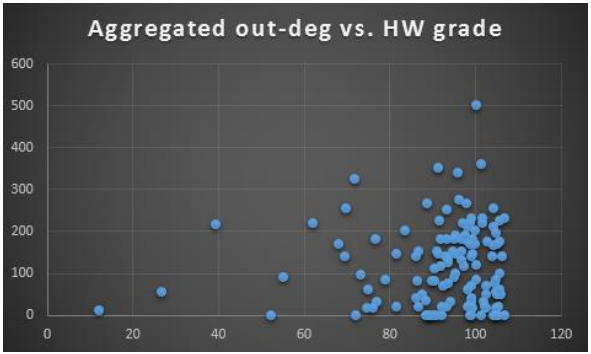
- Discretize grades into {High, Medium, Low} and do classification
  - *J48 decision tree* classifies 80% using identity, 83% not using
  - Does not do well in cross-validation (around 50%)
  - Separating attribute:
    - Using identity: Identity(!), and then In-Degree
    - Not using: Page-rank
- Discretize all parameters and run *Apriori*
  - Find a bunch of good rules between network parameters (expected)
  - No good rules about HW grades

Observation

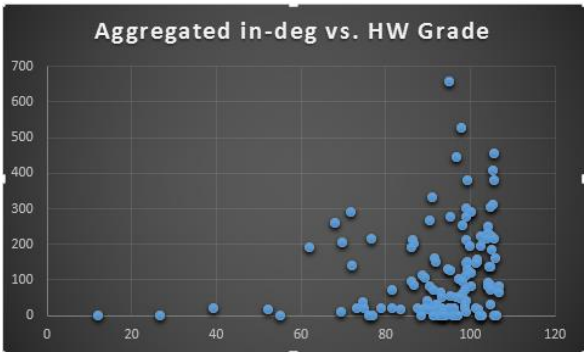
- **Identity is more related to grade than role in network!**

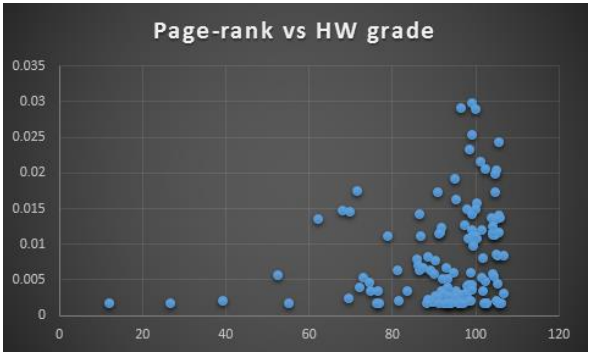


Observation



- ~~High Grades → High Collaboration~~ (Not necessarily)
- High Collaboration → High Grades, especially as a source



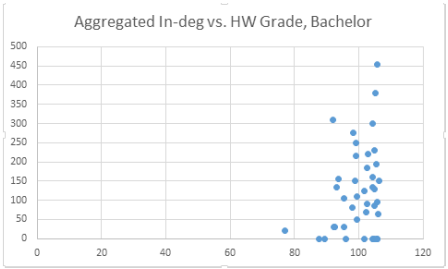


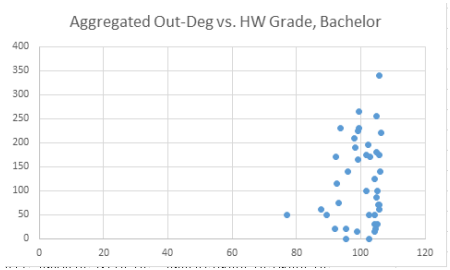
Split the data

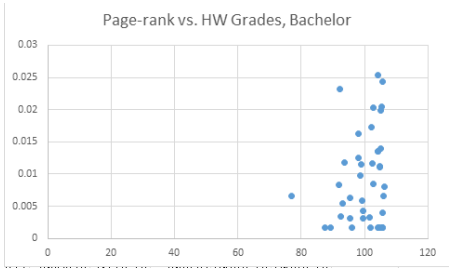
- Identity plays an important rule
- How about separating the data by identity?
- 46 Undergraduated, 95 Graduated, 11 Others(omitted)
- Results similar to mixed data
  - Around 80% class-rate, does not do well in cross-validation
  - Separating attribute: In-Degree, Out-Degree, Hub

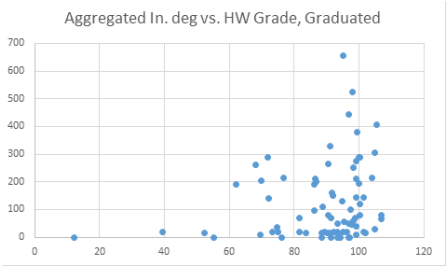
Observation

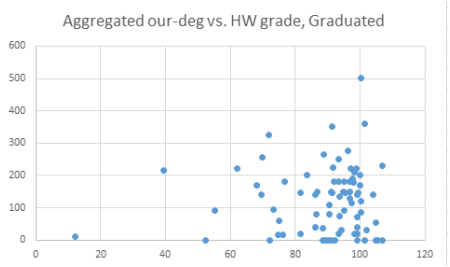
- ~~High Grades → High Collaboration~~ (Not necessarily)
- High Collaboration → High Grades, especially as a source

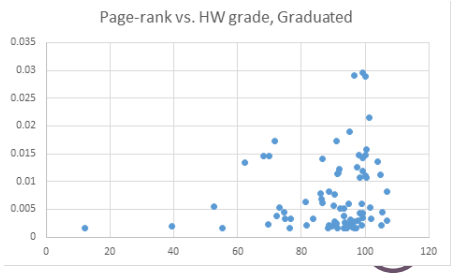












Conclusion

- In this course, identity (graduated or under) plays a big role
- Collaboration helps grade, but not vice versa

- Since the above relation is one-sided,
  data mining algorithms generate poor results

- Drawbacks: Lack of volume, sparse and small network
  -> not enough instances for training