Data Mining: Stage 4 Report

Yimeng Lu • 11.14.2017

Graph Mining

Steps

- Preprocessing and graph construction
- Popularity detection
- Community detection
- Link Prediction

Preprocessing and graph construction

Full graphs

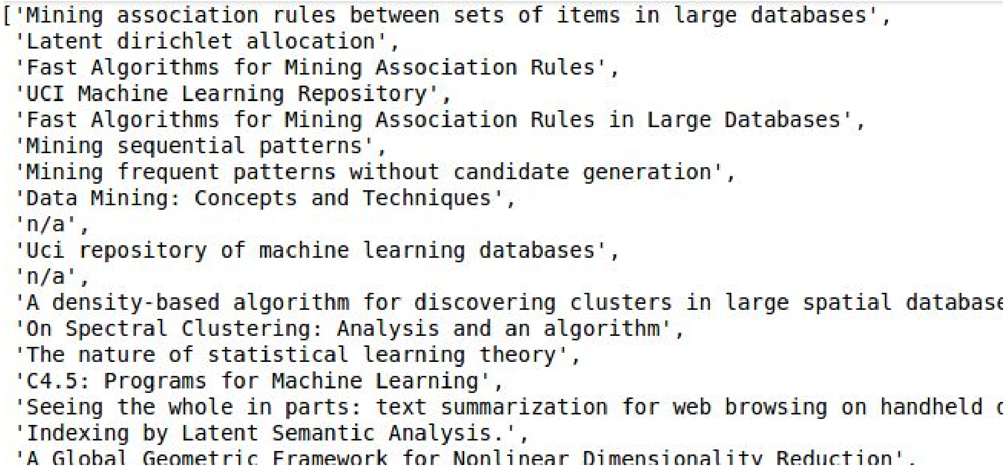
- Undirected: Co-author graph
- Directed: Reference graph

Subgraph

- sub-Co-author graph
  - 5716 nodes
  - 4068 edges

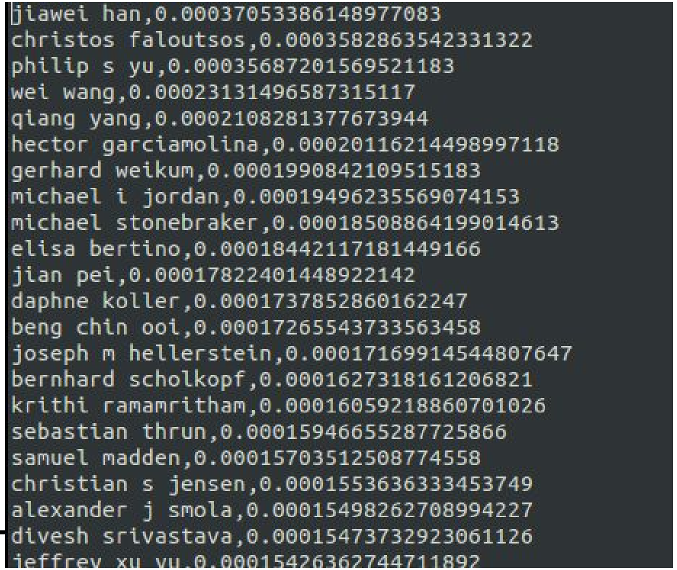
Popularity detection - PageRank

Highly ranked papers



```
['Mining association rules between sets of items in large databases',
 'Latent dirichlet allocation',
 'Fast Algorithms for Mining Association Rules',
 'UCI Machine Learning Repository',
 'Fast Algorithms for Mining Association Rules in Large Databases',
 'Mining sequential patterns',
 'Mining frequent patterns without candidate generation',
 'Data Mining: Concepts and Techniques',
 'n/a',
 'Uci repository of machine learning databases',
 'n/a',
 'A density-based algorithm for discovering clusters in large spatial database
 'On Spectral Clustering: Analysis and an algorithm',
 'The nature of statistical learning theory',
 'C4.5: Programs for Machine Learning',
 'Seeing the whole in parts: text summarization for web browsing on handheld c
 'Indexing by Latent Semantic Analysis.',
 'A Global Geometric Framework for Nonlinear Dimensionality Reduction',
```

Highly ranked authors

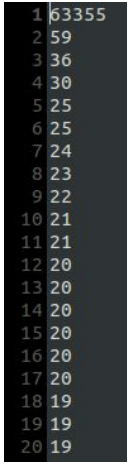


```
jiawei han,0.00037053386148977083
christos faloutsos,0.00035828635423313222
philip s yu,0.00035687201569521183
wei wang,0.00023131496587315117
qiang yang,0.00021082813776739444
hector garciamolina,0.00020116214498997118
gerhard weikum,0.00019908421095151833
michael i jordan,0.00019496235569074153
michael stonebraker,0.00018508864199014613
elisa bertino,0.00018442117181449166
jian pei,0.0001782240144892214 2
daphne koller,0.00017378528601622 47
beng chin ooi,0.00017265543733563458
joseph m hellerstein,0.00017169914544807647
bernhard scholkopf,0.0001627318161206821
krithi ramamritham,0.00016059218860701026
sebastian thrun,0.00015946655287725866
samuel madden,0.00015703512508774558
christian s jensen,0.00015536363334537 49
alexander j smola,0.00015498262708994227
divesh srivastava,0.00015473732923061126
jeffrey xu yu,0.00015426362744711892
```

Community detection

Find max connected components

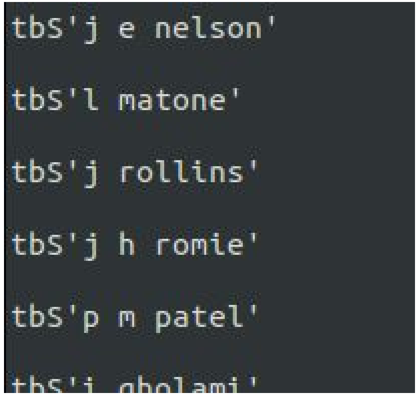
- Component sizes



```
63355
59
36
30
25
25
24
23
22
21
21
20
20
20
20
20
20
19
19
19
```

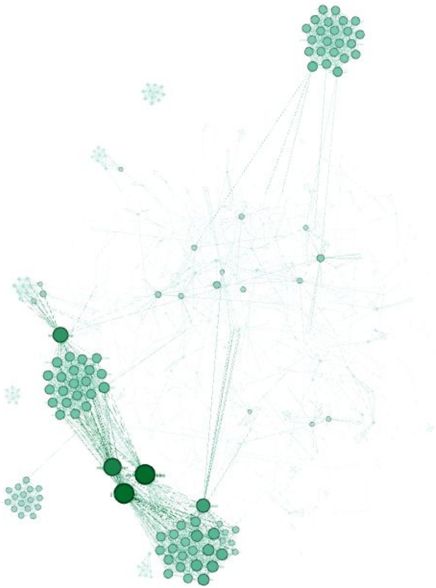
k-clique

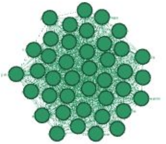
- On subgraph for visualization
- 4 cliques

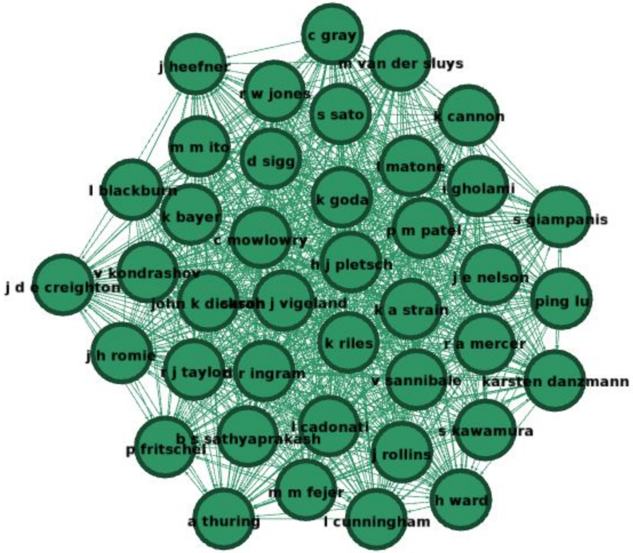


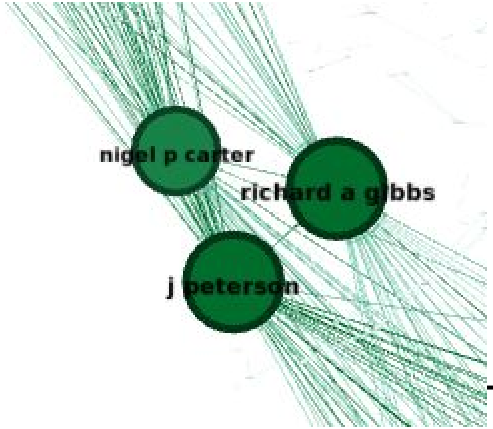
```
tbS'j e nelson'

tbS'l matone'

tbS'j rollins'

tbS'j h romie'

tbS'p m patel'

tbS'i gholami'
```

Visualization



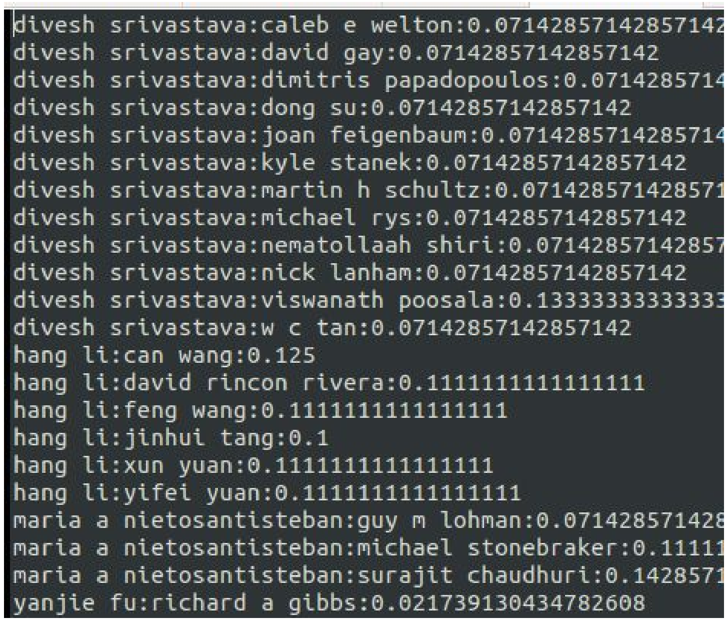






Link prediction

- Jaccard
- Divide into train/test
- Compute score
- 5716 nodes and 4068 edges
- -957.627



```
divesh srivastava:caleb e welton:0.071428571428571428
divesh srivastava:david gay:0.07142857142857142
divesh srivastava:dimitris papadopoulos:0.07142857142
divesh srivastava:dong su:0.07142857142857142
divesh srivastava:joan feigenbaum:0.07142857142857142
divesh srivastava:kyle stanek:0.07142857142857142
divesh srivastava:martin h schultz:0.0714285714285714
divesh srivastava:michael rys:0.07142857142857142
divesh srivastava:nematollaah shiri:0.071428571428571
divesh srivastava:nick lanham:0.07142857142857142
divesh srivastava:viswanath poosala:0.133333333333333
divesh srivastava:w c tan:0.07142857142857142
hang li:can wang:0.125
hang li:david rincon rivera:0.1111111111111111
hang li:feng wang:0.1111111111111111
hang li:jinhui tang:0.1
hang li:xun yuan:0.1111111111111111
hang li:yifei yuan:0.1111111111111111
maria a nietosantisteban:guy m lohman:0.07142857142857
maria a nietosantisteban:michael stonebraker:0.11111
maria a nietosantisteban:surajit chaudhuri:0.1428571
yanjie fu:richard a gibbs:0.021739130434782608
```

Thank you