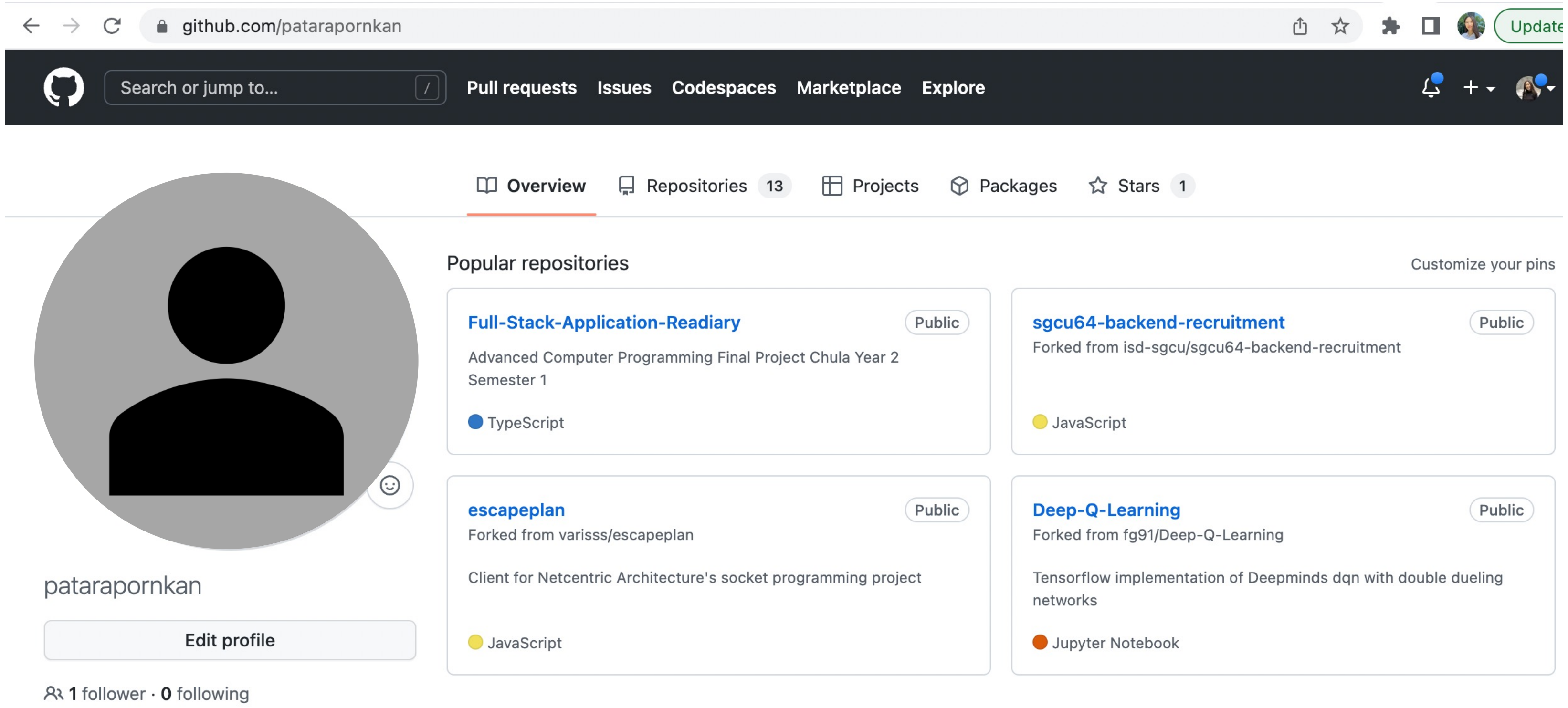
Analysis of GitHub Social Network

Pataraporkan Anantarangsi
QSS41: Analysis of Social Networks
Dartmouth College
Fall 2022

Outline

- Background
- Dataset
- Analysis
- Conclusion

GitHub



pataraporkan

1 follower · 0 following

Popular repositories

Full-Stack-Application-Readiary Public
Advanced Computer Programming Final Project Chula Year 2 Semester 1
● TypeScript

sgcu64-backend-recruitment Public
Forked from isd-sgcu/sgcu64-backend-recruitment
● JavaScript

escapeplan Public
Forked from varisss/escapeplan
Client for Netcentric Architecture's socket programming project
● JavaScript

Deep-Q-Learning Public
Forked from fg91/Deep-Q-Learning
Tensorflow implementation of Deepminds dqn with double dueling networks
● Jupyter Notebook

What is GitHub

- link to Git (version control)
- host repositories

Connections

- follower, following
- receive notifications about their activity and discover projects in their communities
- better reputation

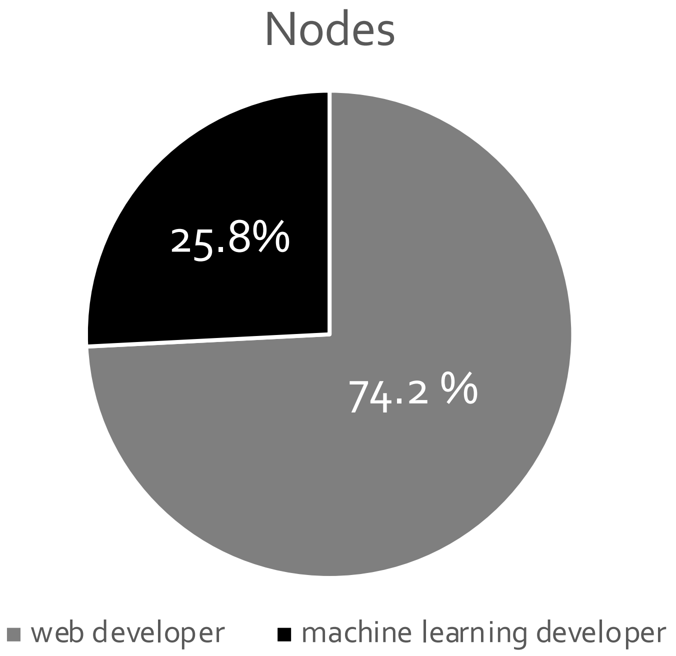
Dataset

- **GitHub Social Network**
- **Source**: Stanford Network Analysis Project (SNAP)
- **Node:** GitHub accounts that have starred at least 10 repositories

  attributes:
  - id
  - name
  - ml_target: 0= web developer, 1 = machine learning developer
  - features
- **Edge:** reciprocal connection (following and follower of each other)

Network Basic Stats

- **37,000 nodes**
- **289,003 edges**
- density: 0.001
- transitivity: 0.013
- 1 component
- 4,005 features

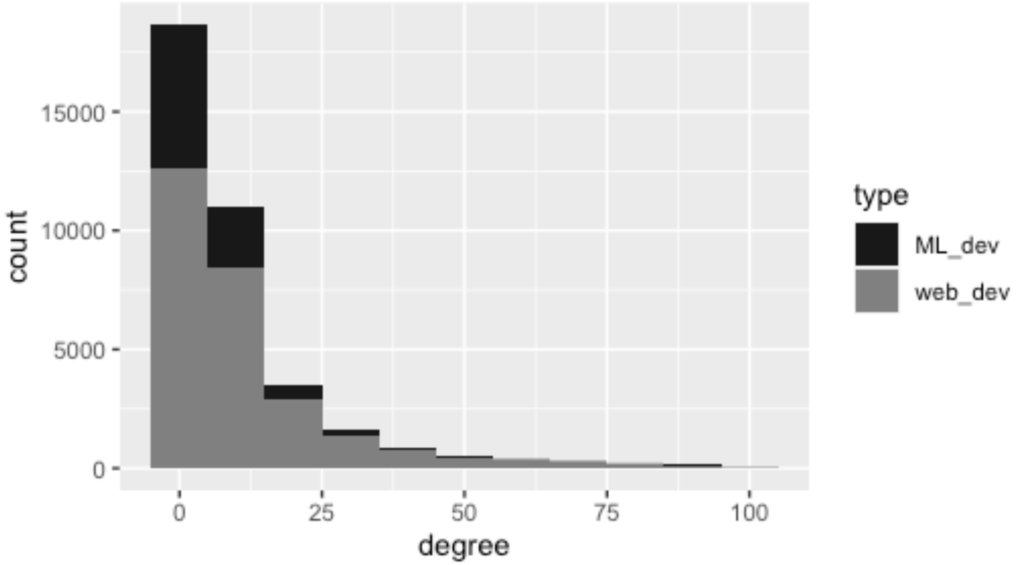


Questions

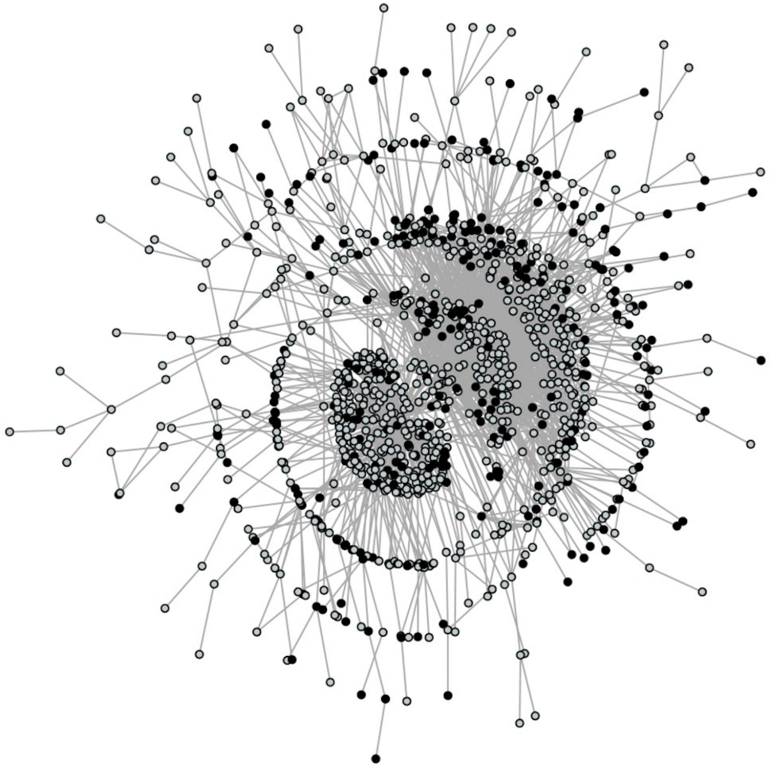
- Are network characteristics of web and machine learning developers different?
- Is there a higher tendency for a user to connect to the same type of user?
- Is there a higher tendency for similar users, in terms of number of shared features, to connect to each other?

Degree Distribution

Type	min	1st quartile	median	3rd quartile	max	mean	Standard deviation
Web developer	1	3	6	15	9548	17.66	92.77
Machine learning developer	1	2	4	9	967	8.632	22.25



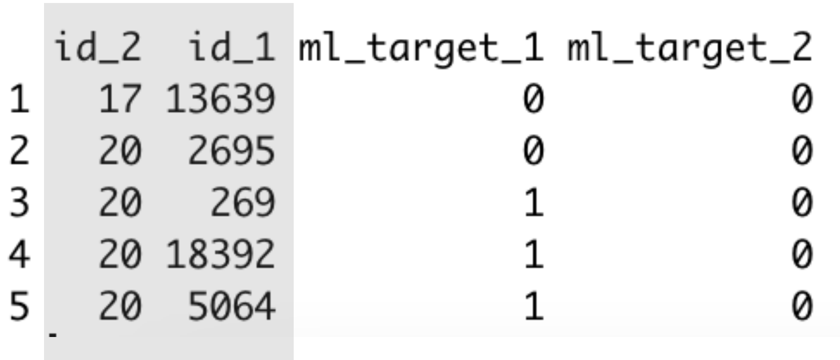
Degree Distribution



Findings

- free-scale network
- very positively skewed
- preferential attachment
- mean degree of web developers is higher

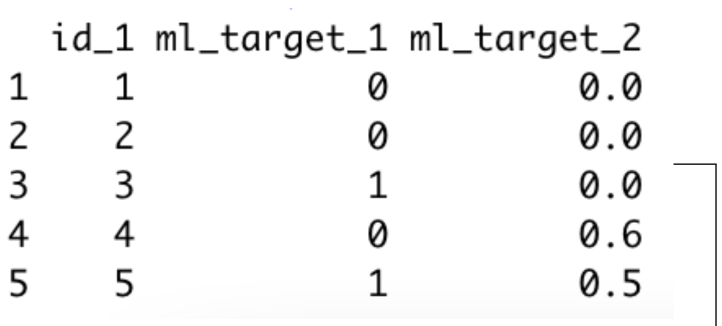
Tendency of GitHub user to connect to users of the same type



```
   id_2  id_1 ml_target_1 ml_target_2
1    17 13639           0           0
2    20  2695           0           0
3    20   269           1           0
4    20 18392           1           0
5    20  5064           1           0
```

edge

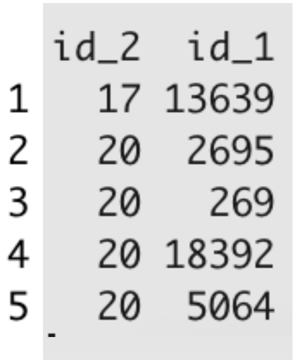
join node’s attribute



```
  id_1 ml_target_1 ml_target_2
1    1           0         0.0
2    2           0         0.0
3    3           1         0.0
4    4           0         0.6
5    5           1         0.5
```

mean of alter’s ml_target

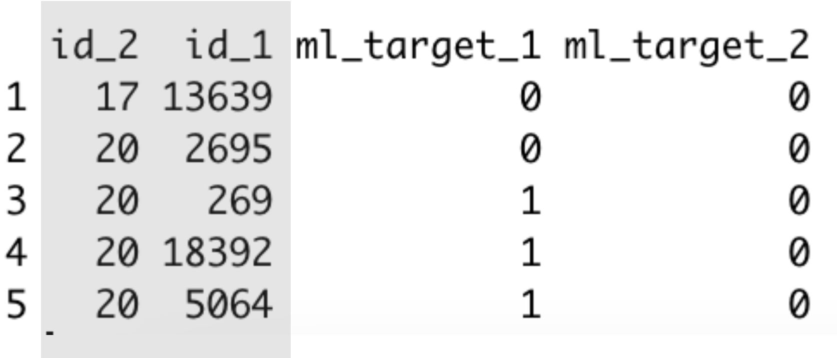
Tendency of GitHub user to connect to users of the same type



```
  id_2  id_1
1   17 13639
2   20  2695
3   20   269
4   20 18392
5   20  5064
```

edge

Tendency of GitHub user to connect to users of the same type

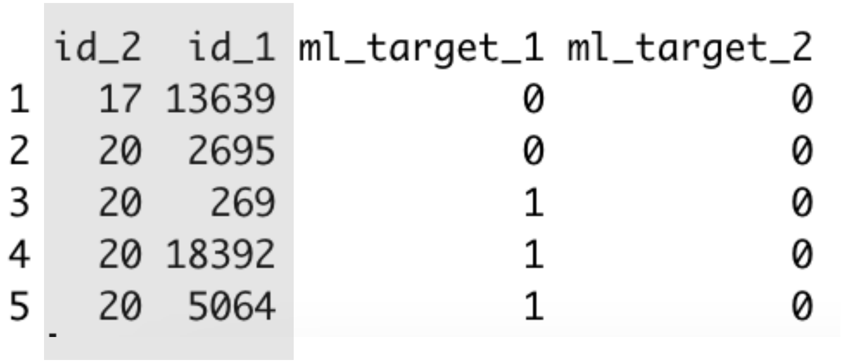


```
  id_2  id_1 ml_target_1 ml_target_2
1   17 13639           0           0
2   20  2695           0           0
3   20   269           1           0
4   20 18392           1           0
5   20  5064           1           0
```

edge

join node’s attribute

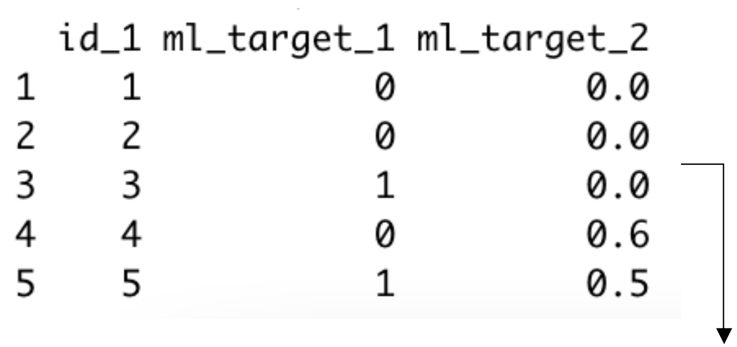
Tendency of GitHub user to connect to users of the same type



```
  id_2  id_1 ml_target_1 ml_target_2
1   17 13639           0           0
2   20  2695           0           0
3   20   269           1           0
4   20 18392           1           0
5   20  5064           1           0
```

edge

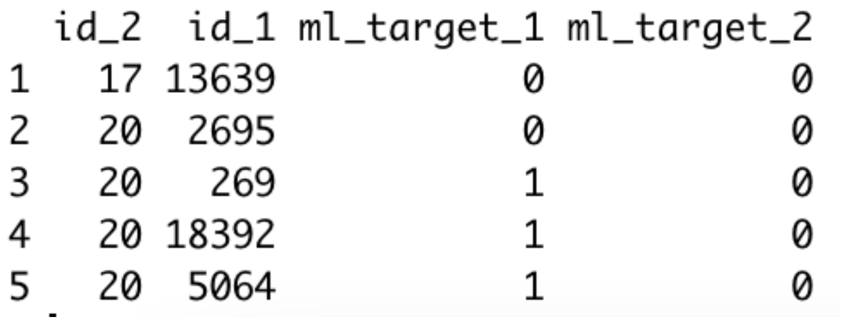
join node’s attribute



```
  id_1 ml_target_1 ml_target_2
1    1           0         0.0
2    2           0         0.0
3    3           1         0.0
4    4           0         0.6
5    5           1         0.5
```

mean of alter’s ml_target

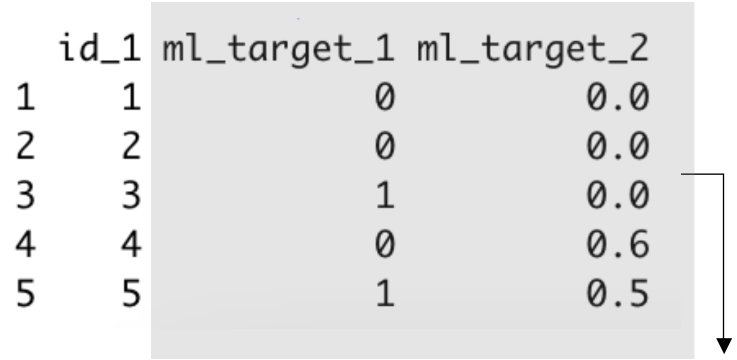
Tendency of GitHub user to connect to users of the same type



```
  id_2  id_1 ml_target_1 ml_target_2
1   17 13639           0           0
2   20  2695           0           0
3   20   269           1           0
4   20 18392           1           0
5   20  5064           1           0
```

edge

join node’s attribute

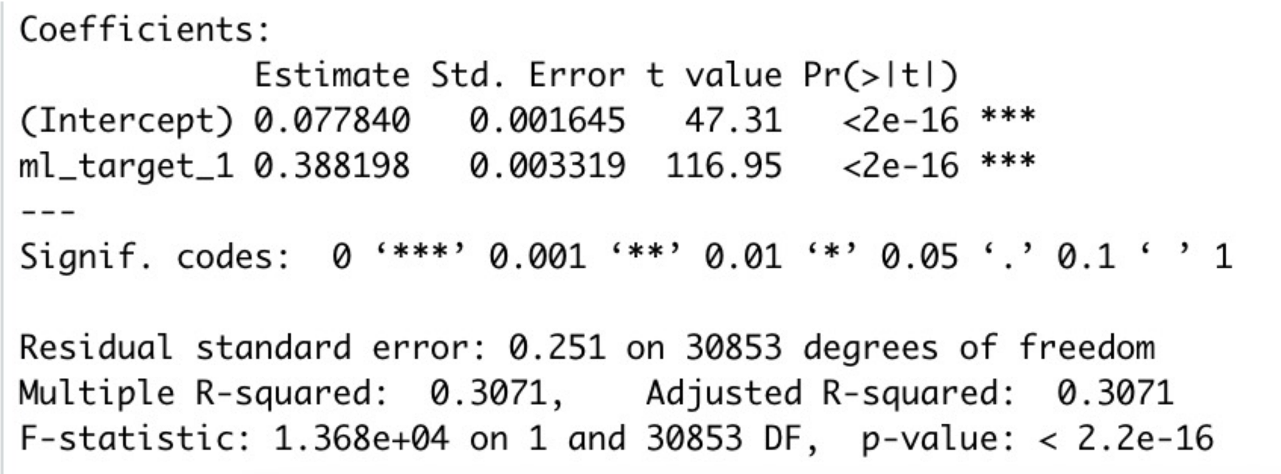


```
  id_1 ml_target_1 ml_target_2
1    1           0         0.0
2    2           0         0.0
3    3           1         0.0
4    4           0         0.6
5    5           1         0.5
```

mean of alter’s ml_target

Tendency of GitHub user to connect to users of the same type

Linear regression



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.077840   0.001645   47.31   <2e-16 ***
ml_target_1 0.388198   0.003319  116.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.251 on 30853 degrees of freedom
Multiple R-squared:  0.3071,	Adjusted R-squared:  0.3071
F-statistic: 1.368e+04 on 1 and 30853 DF,  p-value: < 2.2e-16
```

Pearson correlation

= 0.55

- the correlation is modelately high

Findings

- A machine learning developer is 38.8% more likely to connect to a machine learning engineer
- homophily
- work related -> more interested -> connection

Prediction of connection based on feature similarity

Investigation

- Are nodes with more similar features more likely to connect to each other

Approach

- Jaccard similarity
  - features are very sparse
- Pairwise similarity
  - 37,000 so there are 1.369×10^9 pairs of nodes -> computationally expensive
  - Prune network
    - consider only nodes that have degree higher than 50
  - 1,844 nodes
    - 1.699×10^6 possible pairs of node
    - 48, 428 existing edges

Prediction of connection based on feature similarity

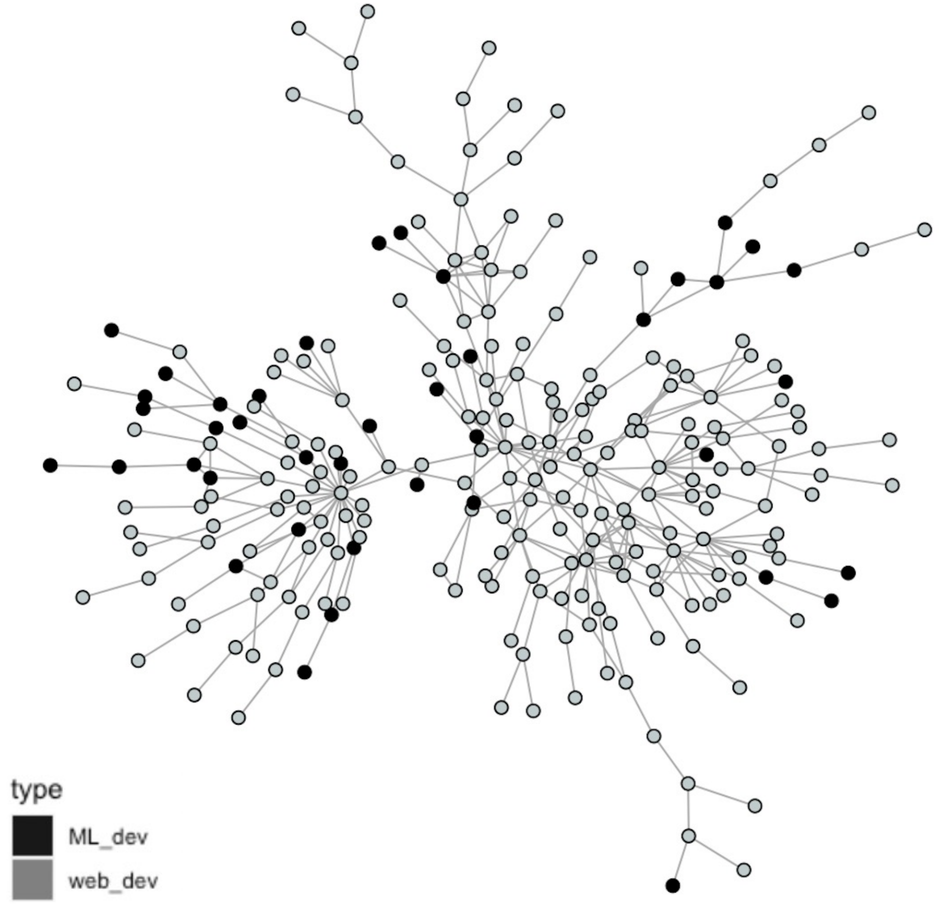
Jaccard similarity based on features

	Average	Standard deviation
Edges	0.2022	0.0747
Possible connections that do not exist	0.1928	0.0718

Findings

- Jaccard similarity of existing connections is **not significantly higher** than that of non-existing pairs
- Shared features cannot be predictors of ties in GitHub networks

Conclusion



- Preferential attachment
- GitHub users are more likely to connect to users of the same type
- Highly similar users do not have more tendency to connect to one another
- Bias
  - only consider those who have starred at least 10 repositories
  - undirected edge

References

B. Rozemberczki, C. Allen and R. Sarkar. Multi-scale Attributed Node Embedding. 2019.