Facebook Pages Network Analysis

Professor Feng Mai - BIA 658 Social Network Analysis

Naman Jogani
Malvika Kandi
Rana Putta
Yug Patel

Problem statement and Objective

Problem Statement

Oftentimes, social network analysis pertains to studying users on social media platforms such as Facebook and Twitter but we rarely see network analysis on the facebook pages that exist.

Our team was curious to find out if we could perform a network analysis on facebook pages.

Objective

- How many edges does each page have?
- What is the most connected page?
- Which of the pages represent a tighty connect cluster?
- The few pages that connect separate clusters
- Understanding the patterns of interaction

Facebook Dataset

- The dataset was obtained from GitHub from a user.

- He has used a custom function for which when a seed set of users are given, it recursively samples additional users that are "liked" by the original set.

- The seed user given for the function was "Bill Gates".

- The function built a list of users until it hit a maximum depth of 2.

- The data retrieved from the facebook pages was written into a text file with two attributes.

- The first attribute with the user themself and the second attribute was the friend/liked page.

Data Description and EDA

Dataset was collected from GitHub

Dataset only contained two columns.

Column 1 represented a facebook page (A) that liked another facebook page (B) in column 2.

Edges were formed by directed edges of the facebook pages as nodes.

The dataset had 5062 nodes and 6060 edges.

```
 1  Scholastic→Scholastic Librarians
 2  Scholastic→Scholastic Art & Writing Awards
 3  Scholastic→National Council for Teachers of English
 4  Scholastic→Scholastic Australia
 5  Scholastic→This is Teen
 6  Scholastic→Scholastic Parents
 7  Scholastic→Hunger Games Page
 8  Scholastic→Scholastic Teachers
 9  Scholastic→Reading Is Fundamental (RIF)
10  Scholastic→Scholastic Reading Club
```

Data Description and EDA

Initial Dataset:

Number of nodes: 5062
Number of edges: 6060
Average degree: 2.3943
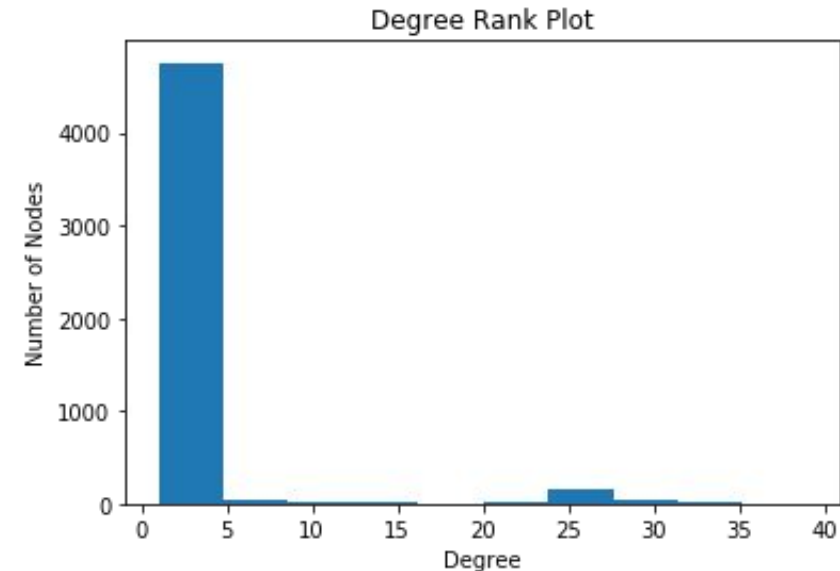Network density: 0.00047309040687414414

Filtered dataset:

Remove all nodes with a degree less than 10
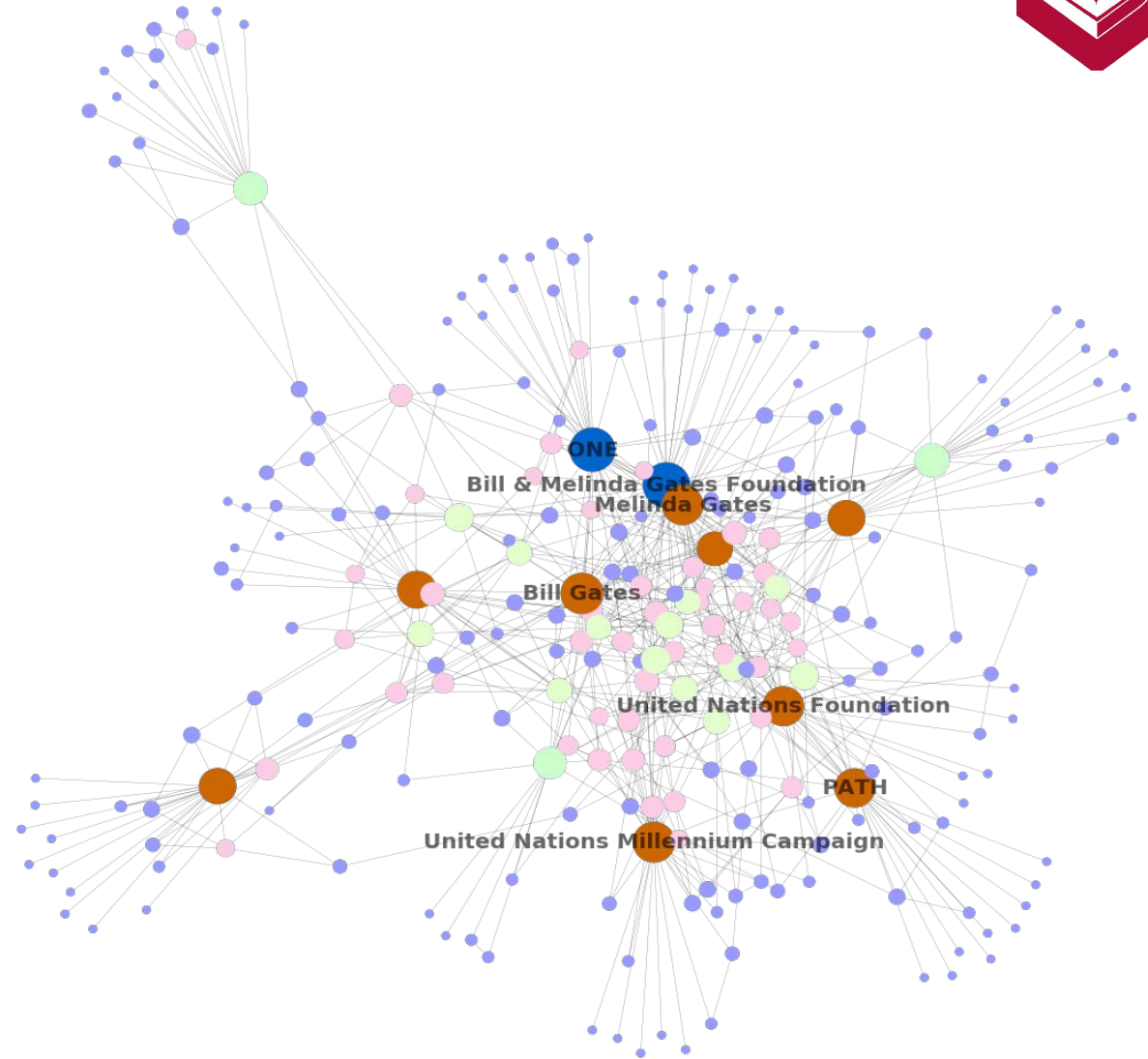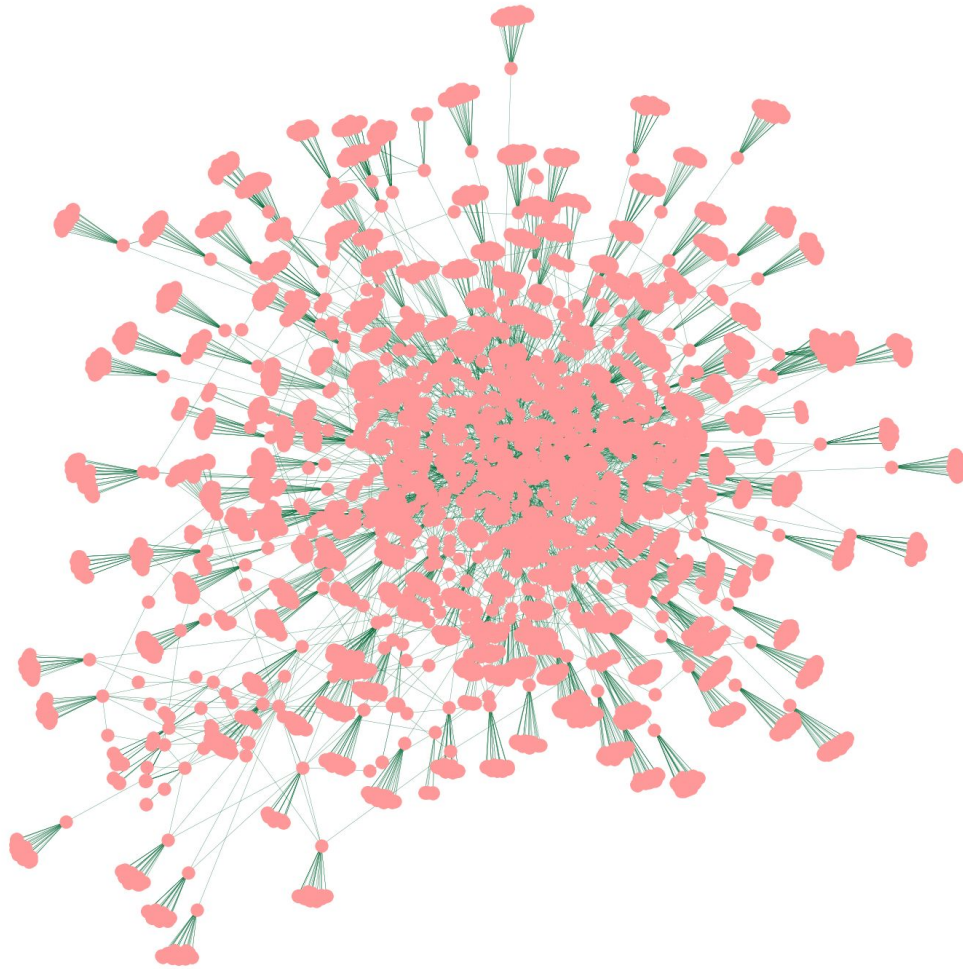Number of nodes: 259
Number of edges: 598
Average degree: 4.6178
Network density: 0.01789829696806441

Facebook Network Graph, Degree > 10

Subgraph Centrality Measures

1. Degree centrality

2. Closeness centrality

3. Betweenness centrality

4. Transitivity (Clustering Coefficient)

5. Pagerank centrality

6. Eigenvector centrality

Degree Centrality

[('**Bill & Melinda Gates Foundation'**, 0.135),

('**ONE**', 0.127),

('**Bill Gates**', 0.112),

('United Nations Millennium Campaign', 0.108),

('Melinda Gates', 0.104),

('United Nations Foundation', 0.104),

('PATH', 0.100),

('TED', 0.093),

('Microsoft', 0.085),

('End Polio Now', 0.085)]
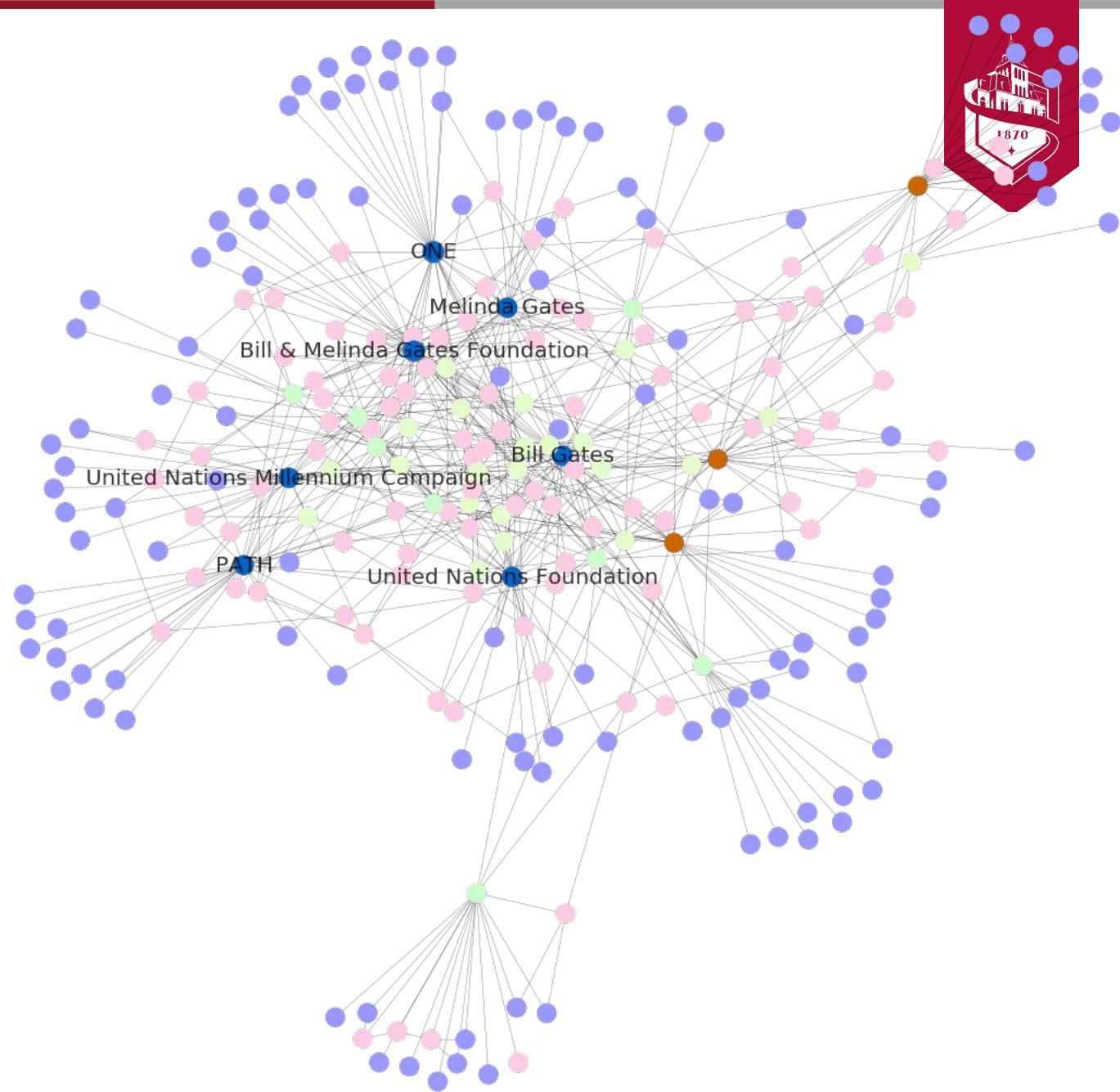
Closeness Centrality

[('**Bill Gates**', 0.528),

('**Melinda Gates**', 0.412),

('**Bill & Melinda Gates Foundation**', 0.406),

('ONE', 0.403),

('United Nations Millennium Campaign', 0.393),

('United Nations Foundation', 0.388),

('RESULTS AU', 0.388),

('Vox', 0.387),

('Malaria No More', 0.387),

('TED', 0.385)]
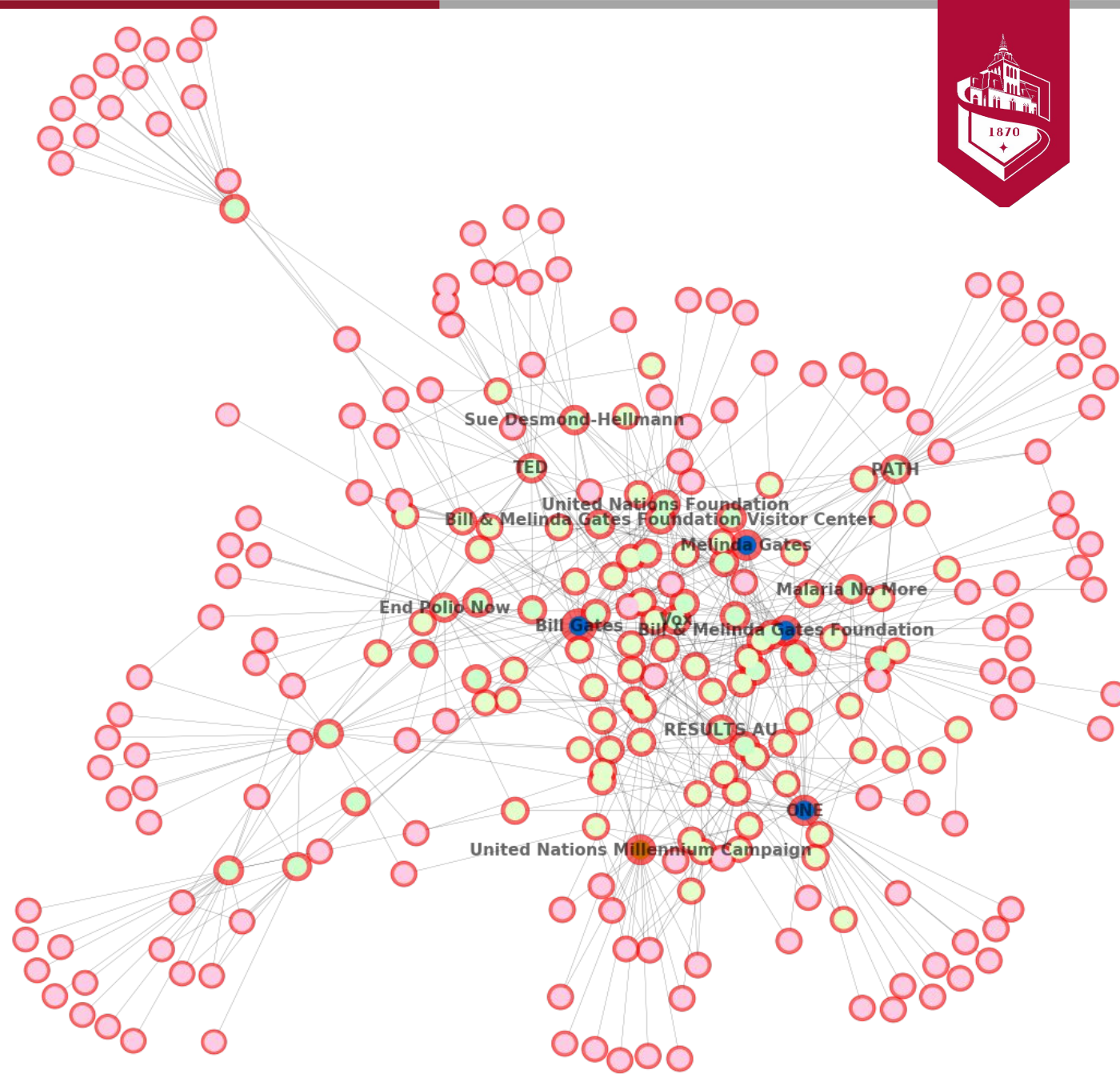
Betweenness Centrality Graph

Edge betweenness

('**Bill Gates**', 0.424),

('**ONE**', 0.134),

('**United Nations Millennium Campaign**', 0.115),

('Bill & Melinda Gates Foundation', 0.114),

('Big History Project', 0.113),

('PATH', 0.107),

('Microsoft', 0.105),

('Melinda Gates', 0.096),

('United Nations Foundation', 0.094),

('TED', 0.088)
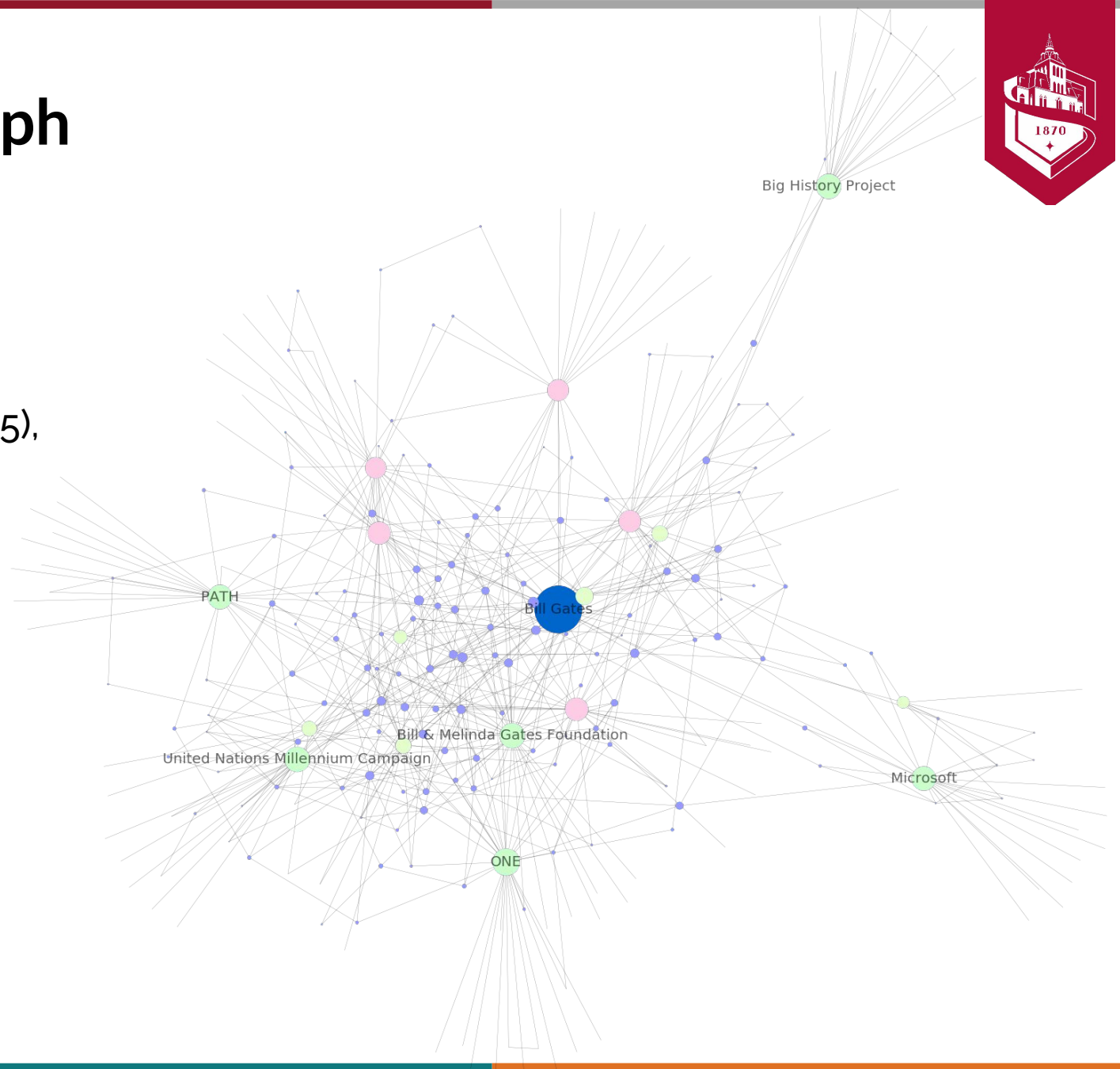
Transitivity

Clustering coefficient

[('公益財団法人日本ユニセフ協会', 1.0),

('**NPR Extra'**, 1.0),

('**Scholastic'**, 1.0),

('Seattle Met Magazine', 1.0),
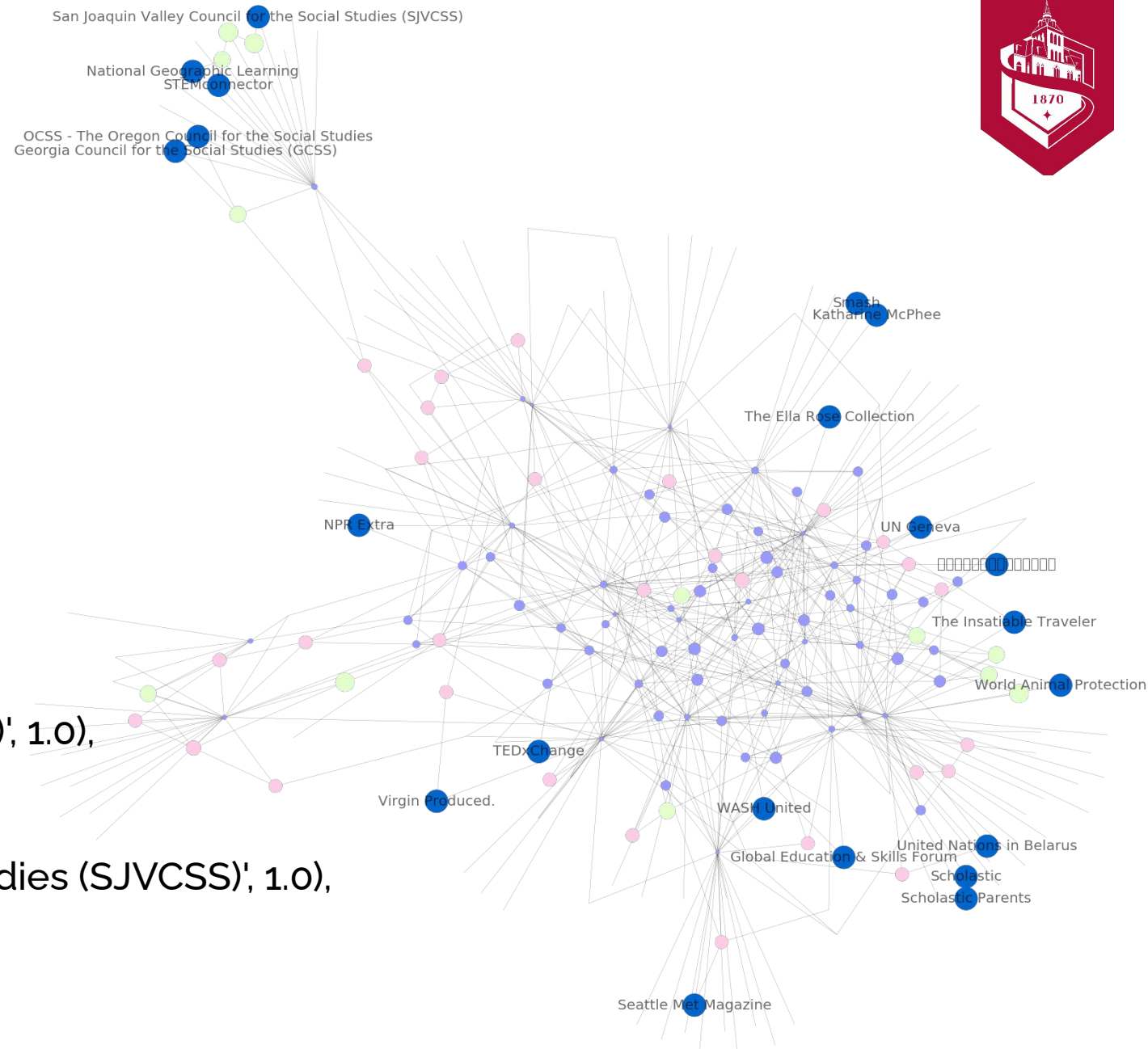
('TEDxChange', 1.0),

('STEMconnector', 1.0),

('Georgia Council for the Social Studies (GCSS)', 1.0),

('The Ella Rose Collection', 1.0),

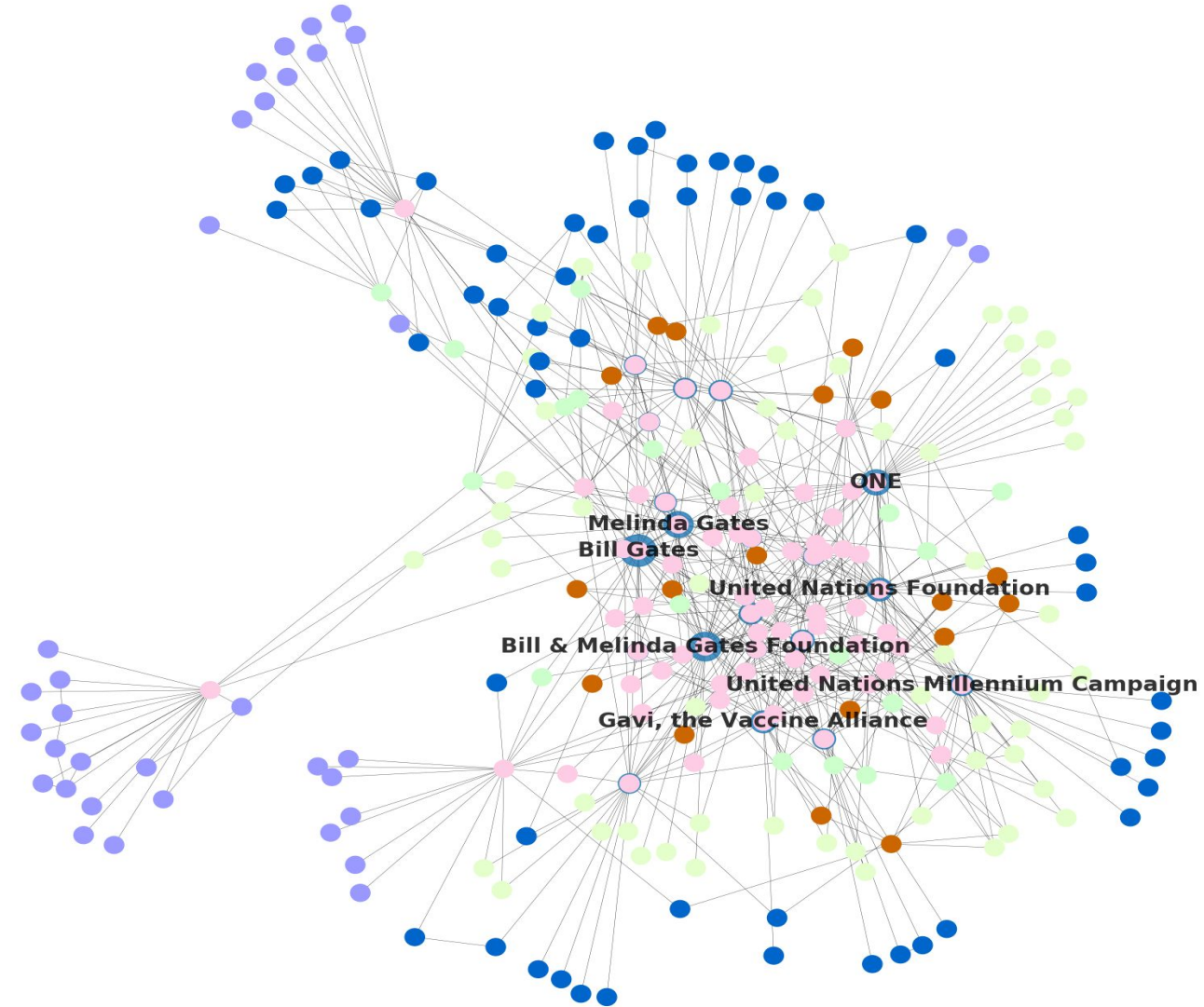('San Joaquin Valley Council for the Social Studies (SJVCSS)', 1.0),

('UN Geneva', 1.0)]

Eigenvector Centrality

[('**Bill Gates**', 0.345),

('**Bill & Melinda Gates Foundation**', 0.283),

('**Melinda Gates**', 0.230),

('ONE', 0.218),

('United Nations Foundation', 0.178),

('Gavi, the Vaccine Alliance', 0.167),

('United Nations Millennium Campaign', 0.151),

('Global Citizen', 0.148),

('RESULTS AU', 0.144),

('PATH', 0.143)]

Community Detection

We used Python-Louvain community detection algorithm which detected 10 communities

Subject Matter of the communities:

Community 1: Bill & Melinda Gates Foundation, UNICEF UK, One Billion Hungry, Giving Tuesday

Community 2: United Nations, Institute for Health Metrics and Evaluation (IHME) - Health
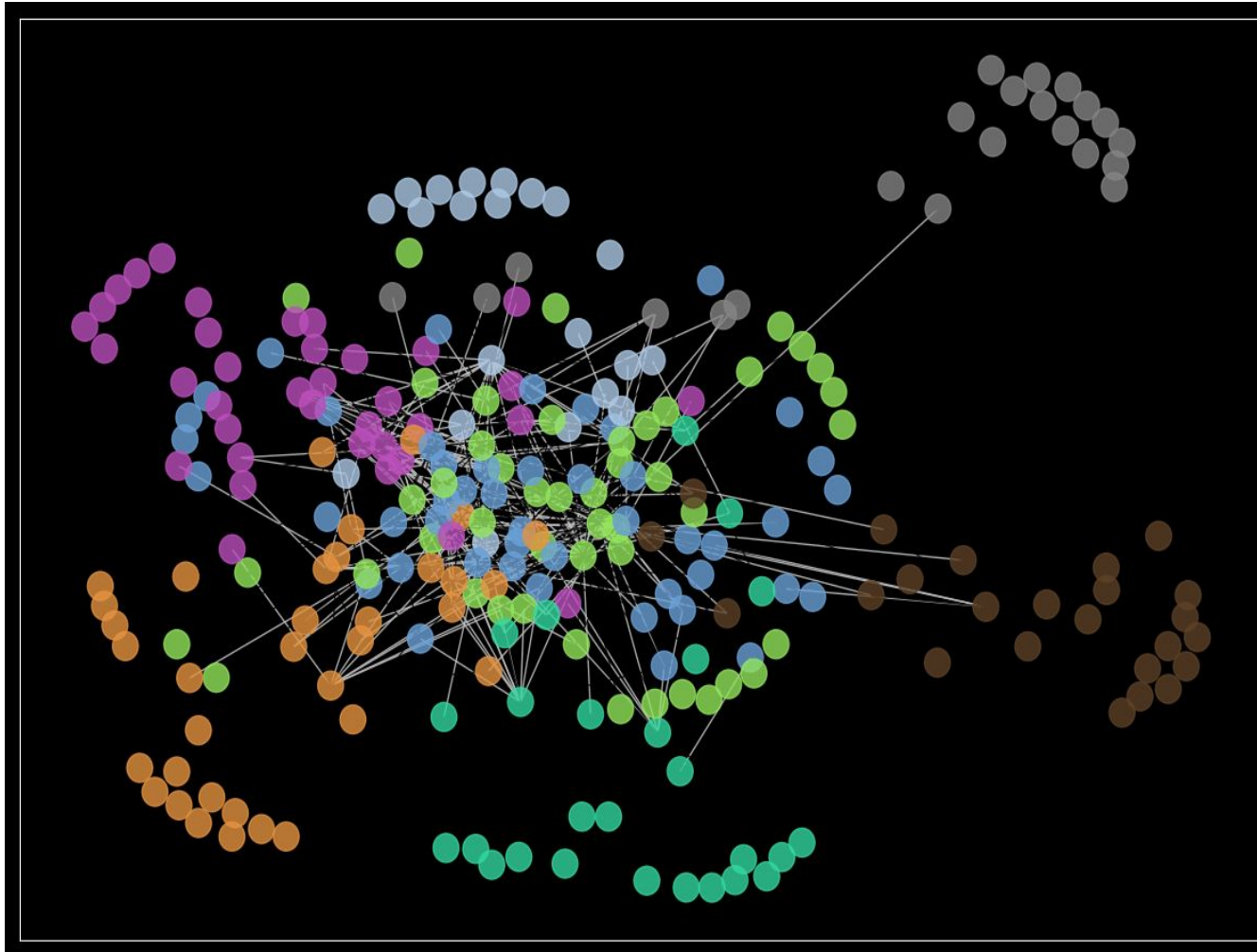
Community 3: Bill Gates, Malaria No More UK, World Health Summit, Global Poverty Project, Speak Up Africa

Community 4: Windows, ASUS, Microsoft, HP, MSN, Acer - Technology

Community 5: Big History Project, National Geographic Education, Hubble Space Telescope, High Tech teachers - Educational
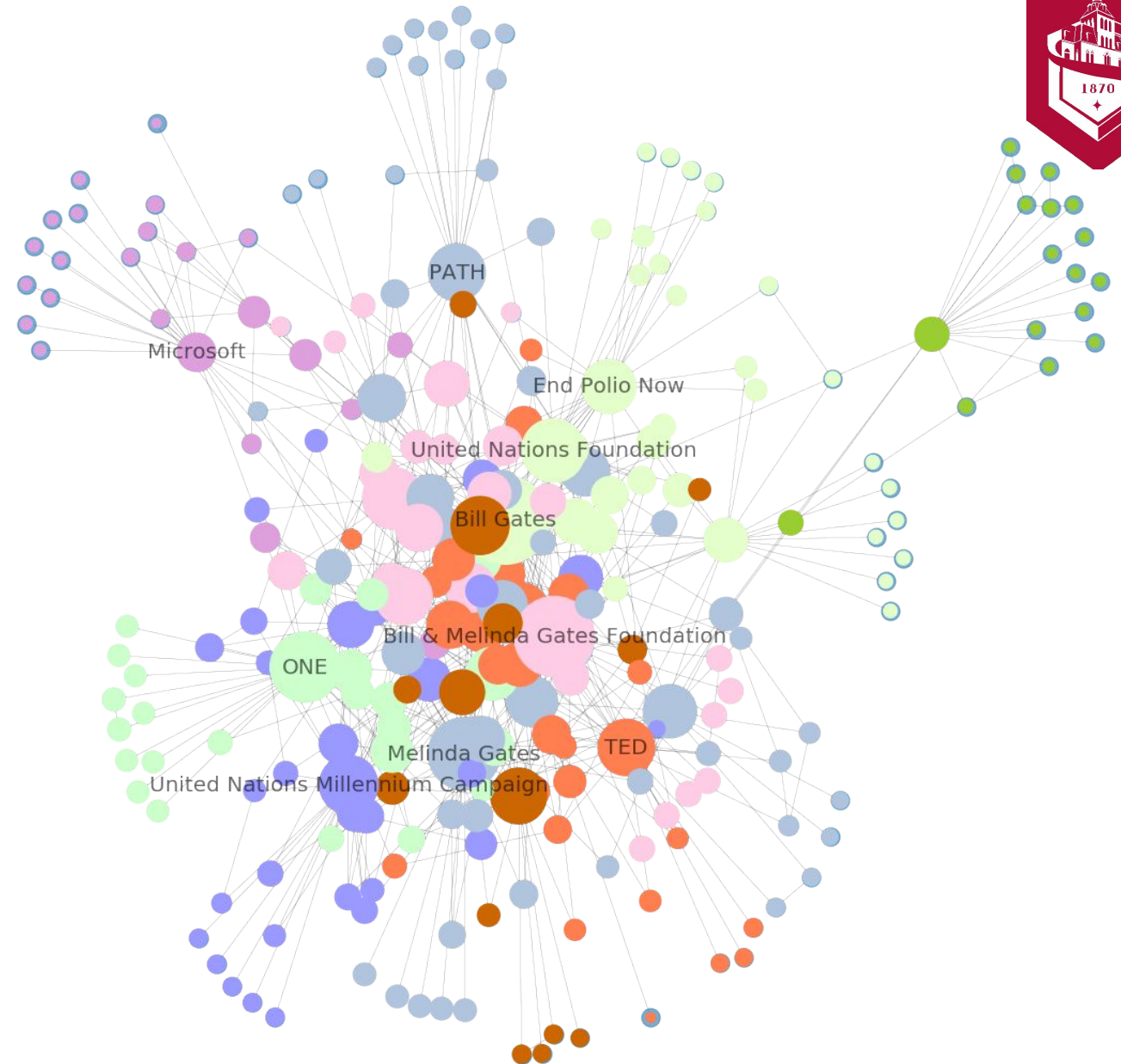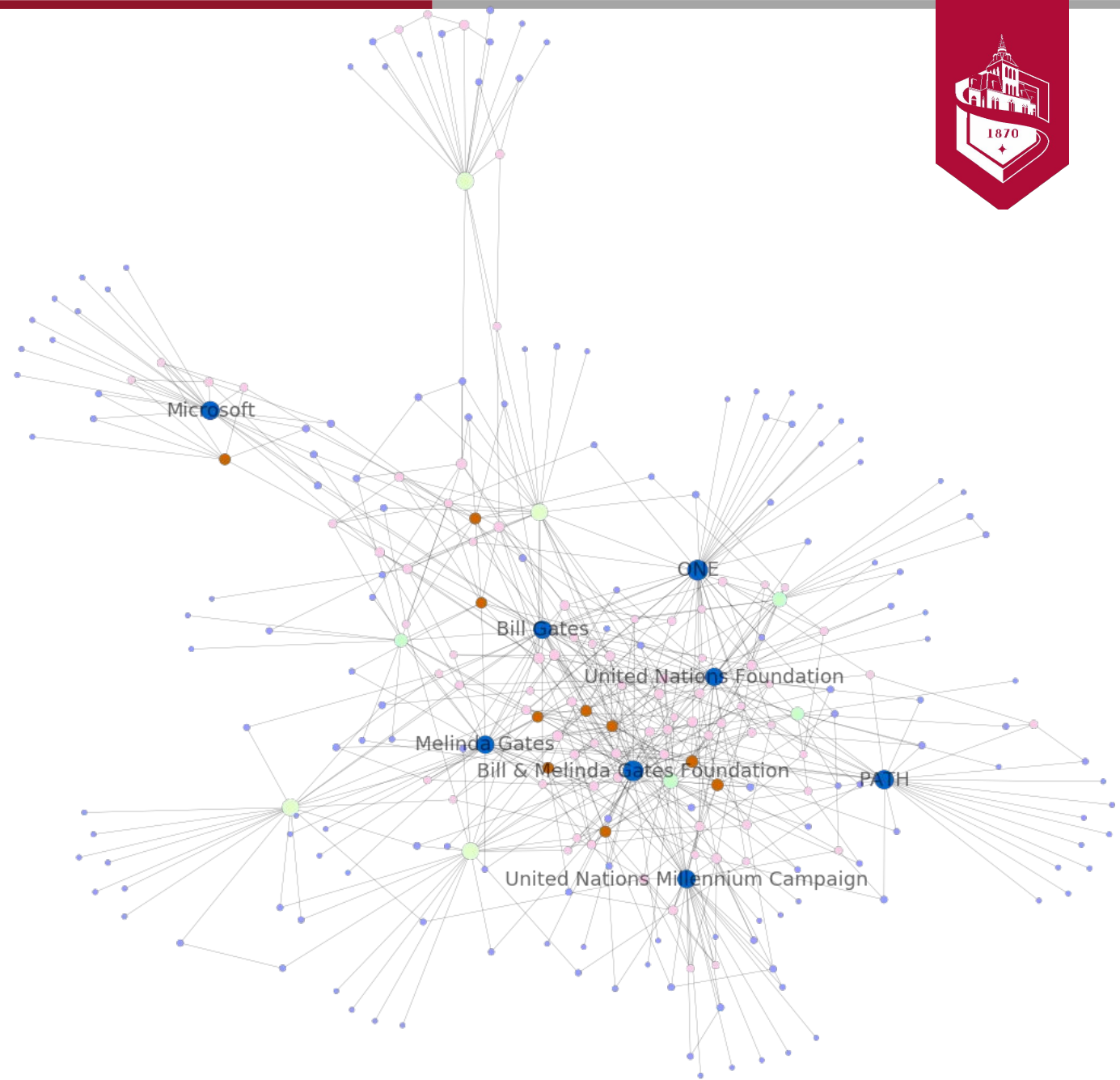
External Edges Analysis

PageRank Centrality

[('**ONE**', 0.026),

('**Bill & Melinda Gates Foundation**', 0.026),

('**PATH'**, 0.022),

('Microsoft', 0.021),

('United Nations Millennium Campaign', 0.020),

('United Nations Foundation', 0.020),

('Bill Gates', 0.020),

('Melinda Gates', 0.020),

('Big History Project', 0.018),

('Rotary International', 0.018)]

Link Prediction

We split the dataset into 30% test (hidden links) and 70% train to run link prediction.

Three types of similarity measures were used to recommend links.

For each node 10 links were recommended and respective Mean Average Precision was calculated.

Results:-

Adamic Adar similarity - 0.0021

Shortest path similarity - 0.0016

Jaccard similarity - 0.0003

An example for node 'Scholastic' :-

```
> recommend_example <- recommend_links(node_name = 'Scholastic', sim_matrix = jaccard_sim, n_recommend = 5)
> recommend_example
[1] "Clifford The Big Red Dog" "I SPY"                    "The Magic School Bus"     "Fubiz"
[5] "Elizabeth Elton"
> # Test
> h[['Scholastic']]
$Scholastic
+ 5/1814 vertices, named, from 57be73a:
[1] This is Teen                 Klutz Books                  Scholastic Canada
[4] National Student Poets Program Clifford The Big Red Dog

> apk(5, actual = unlist(names(h[['Scholastic']][[1]])), predicted =recommend_example)
[1] 0.2
```

Conclusion

Key network findings for Facebook Pages dataset

Five different communities were identified.

For the centrality measures, pages 'Bill Gates', 'Melinda Gates' and their foundation have highest values.

Nodes such as 'Scholastic' and 'NPR Extra' rank high on transitivity measure.

Several pages that have appeared on the top on every centrality measure did not have a transitivity score of 1. In fact, they had a transitivity score of >> 0.1.

Future research

An interesting insight would be to build a recommendation system by using a supervised learning technique to predict the user's next page visit.

stevens.edu