Assignment 1

Community Detection - Empirical study

Social Media Analytics
CS G519
in
Master of Engineering
by

Pratyush Thakur (2020H1030121P)

Parth Pandya (2020H1120287P)

Under the supervision of
Dr. Poonam Goyal

Department of Computer Science and Information Systems

Birla Institute of Technology and Science

Pilani, Pilani Campus, Rajasthan (India)

November, 2021

- **Statistics**

	Nodes	Edges	Average Path Length	Average Clustering Coefficient
Karate	34	78	2.4081	0.5706
Dolphins	62	159	3.3569	0.2589
Jazz	198	2740	2.2351	0.6171

- **Implementation**

❖ **Link to the Code: https://github.com/parthapandya410/SMA_A1**
❖ Betweenness-based clustering using the Girvan-Newman algorithm :

1] Karate Club

```
GIRVAN-NEWMAN ALGORITHM

karate club network :
([1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22], [3, 9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34])
([1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22], [3, 9, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34],
[10])
([1, 2, 4, 8, 12, 13, 14, 18, 20, 22], [3, 9, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34], [5, 6, 7, 11, 1
7], [10])
([1, 2, 4, 8, 12, 13, 14, 18, 20, 22], [3, 25, 26, 28, 29, 32], [5, 6, 7, 11, 17], [9, 15, 16, 19, 21, 23, 24, 27, 30, 31, 33,
34], [10])
([1, 2, 4, 8, 13, 14, 18, 20, 22], [3, 25, 26, 28, 29, 32], [5, 6, 7, 11, 17], [9, 15, 16, 19, 21, 23, 24, 27, 30, 31, 33, 34],
[10], [12])
([1, 2, 4, 8, 13, 14, 18, 20, 22], [3, 25, 26, 28, 29, 32], [5, 6, 7, 11, 17], [9, 15, 16, 19, 21, 23, 24, 30, 31, 33, 34], [1
0], [12], [27])
([1, 2, 4, 8, 14, 18, 20, 22], [3, 25, 26, 28, 29, 32], [5, 6, 7, 11, 17], [9, 15, 16, 19, 21, 23, 24, 30, 31, 33, 34], [10],
[12], [13], [27])
([1, 2, 4, 8, 14, 18, 20, 22], [3, 25, 26, 28, 29, 32], [5, 6, 7, 11, 17], [9, 16, 19, 21, 23, 24, 30, 31, 33, 34], [10], [12],
[13], [15], [27])
([1, 2, 4, 8, 14, 18, 20, 22], [3, 25, 26, 28, 29, 32], [5, 6, 7, 11, 17], [9, 19, 21, 23, 24, 30, 31, 33, 34], [10], [12], [1
3], [15], [16], [27])
([1, 2, 4, 8, 14, 18, 20, 22], [3, 28, 29], [5, 6, 7, 11, 17], [9, 19, 21, 23, 24, 30, 31, 33, 34], [10], [12], [13], [15], [1
6], [25, 26, 32], [27])

Number of clusters 11
Modularity of karate club:
0.3159105851413545
Time taken :  0.22612738609313965
```

We have got **0.315910** Modularity score for karate club dataset using Girvan - Newman algorithm with 11 clusters. Runtime for this algorithm is 0.2261.

2] Jazz club

```
(['1', '10', '100', '101', '102', '103', '104', '105', '106', '107', '108', '109', '11', '110', '111', '112', '113', '114', '11
6', '117', '118', '119', '12', '121', '122', '123', '124', '125', '126', '127', '128', '129', '13', '130', '131', '132', '134',
'135', '136', '137', '138', '139', '14', '140', '141', '142', '143', '144', '146', '147', '149', '15', '150', '151', '153', '15
4', '155', '156', '158', '159', '16', '161', '162', '163', '164', '166', '167', '168', '169', '170', '171', '172', '173', '17
4', '175', '176', '177', '178', '179', '18', '181', '182', '183', '184', '185', '186', '187', '188', '189', '19', '190', '191',
'192', '193', '194', '195', '196', '197', '198', '2', '20', '22', '23', '24', '26', '27', '28', '29', '3', '31', '32', '33', '3
4', '35', '36', '38', '39', '4', '40', '41', '42', '43', '44', '45', '46', '48', '49', '5', '50', '51', '52', '53', '54', '55',
'56', '57', '58', '59', '60', '61', '62', '63', '64', '65', '66', '67', '68', '69', '7', '70', '71', '72', '73', '74', '75', '7
6', '77', '78', '79', '8', '80', '81', '82', '83', '84', '85', '86', '87', '88', '89', '9', '90', '91', '93', '94', '95', '96',
'97', '98', '99'], ['115', '152', '157', '6'], ['21'], ['148', '25', '92'], ['17'], ['30'], ['133', '180', '37', '47'], ['12
0'], ['145'], ['160'], ['165'])

Modularity of jazz club:
0.00933966913527636
Number of clusters 11
Time taken :  37.18859601020813
```

We have got **0.009332** Modularity score for the jazz club dataset using Girvan - Newman algorithm with 11 clusters. Runtime for this algorithm is 37.1885.

3] Dolphin club

```
(['Beak', 'Bumper', 'Fish', 'Oscar', 'PL', 'SN96', 'TR77'], ['Beescratch', 'DN63', 'Knit', 'Mus', 'Notch', 'Number1'], ['CCL',
'Double', 'Zap'], ['Cross', 'Five', 'Haecksel', 'Jonah', 'MN105', 'MN60', 'MN83', 'Patchback', 'SMN5', 'Topless', 'Trigger', 'V
au'], ['DN16', 'DN21', 'Feather', 'Gallatin', 'SN89', 'SN90', 'TR82', 'Upbang', 'Wave', 'Web'], ['Fork'], ['Grin', 'Hook', 'Kri
ngel', 'SN100', 'SN4', 'SN63', 'SN9', 'Scabs', 'Shmuddel', 'Stripes', 'TR120', 'TR88', 'TR99', 'TSN103', 'Thumper'], ['Jet', 'M
N23', 'Quasi'], ['Ripplefluke', 'Zig'], ['TSN83', 'Zipfel'], ['Whitetip'])

Modularity of dolphins club:
0.4852853921917645
Number of clusters:  11
Time taken :  0.8036494255065918
```

We have got **0.48528** Modularity score for the dolphins club dataset using Girvan - Newman algorithm with 11 clusters. Runtime for this algorithm is 0.8036.

- ❖ Modularity-based clustering

1] Karate Club

```
MODULARITY BASED CLUSTERING

Karate Club :
[9, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]
[2, 3, 4, 8, 10, 13, 14, 18, 22]
[1, 5, 6, 7, 11, 12, 17, 20]
# clusters formed:3
Modularity of karate club:
0.3806706114398422
Time taken:0.0043599605560302734
```

We have got a 0.3806 modularity score for the karate club dataset using modularity maximization algorithm. Total number of clusters formed using this algorithm is 3 and the run time of the algorithm is 0.04.

2] Jazz Club

```
Jazz Club :
['1', '100', '101', '103', '104', '105', '106', '107', '108', '109', '110', '116', '119', '120', '122', '123', '131', '132', '1
34', '135', '139', '15', '154', '159', '16', '162', '164', '166', '168', '17', '170', '171', '174', '178', '179', '182', '187',
'188', '194', '197', '22', '23', '24', '32', '33', '35', '38', '40', '42', '43', '44', '46', '48', '50', '58', '60', '62', '6
3', '64', '65', '66', '68', '74', '78', '8', '80', '81', '83', '9', '91', '95', '98', '99']
['102', '115', '124', '126', '128', '13', '133', '136', '138', '140', '144', '145', '147', '148', '149', '152', '153', '155',
'156', '157', '160', '163', '167', '169', '172', '173', '176', '18', '180', '181', '184', '189', '191', '198', '21', '25', '2
6', '27', '28', '29', '3', '37', '39', '4', '41', '45', '47', '5', '51', '6', '73', '75', '76', '77', '79', '85', '86', '88',
'90', '92', '96', '97']
['10', '11', '111', '112', '113', '114', '117', '118', '12', '121', '125', '127', '129', '130', '137', '14', '141', '142', '14
6', '150', '151', '158', '161', '165', '175', '177', '183', '185', '186', '19', '190', '192', '193', '195', '196', '2', '20',
'30', '31', '34', '36', '49', '52', '53', '54', '55', '56', '57', '61', '67', '69', '7', '70', '71', '72', '84', '87', '89', '9
3', '94']
['143', '59', '82']
# clusters formed:4
Modularity of jazz club:
0.41724718951462525
Time taken:0.27942562103271484
```

We have got a 0.4172 modularity score for the karate club dataset using modularity maximization algorithm. Total number of clusters formed using this algorithm is 4 and the run time of the algorithm is 0.2794.

3] Dolphins Club

```
Dolphins Club :
['Beak', 'Bumper', 'Fish', 'Fork', 'Grin', 'Hook', 'Kringel', 'SN4', 'SN63', 'SN9', 'SN96', 'Scabs', 'Shmuddel', 'Stripes', 'TR
120', 'TR77', 'TR88', 'TR99', 'TSN103', 'TSN83', 'Thumper', 'Whitetip', 'Zipfel']
['Beescratch', 'DN16', 'DN21', 'DN63', 'Feather', 'Gallatin', 'Jet', 'Knit', 'MN23', 'Mus', 'Notch', 'Number1', 'Oscar', 'PL',
'Quasi', 'Ripplefluke', 'SN90', 'TR82', 'Upbang', 'Wave', 'Web', 'Zig']
['CCL', 'Cross', 'Double', 'Five', 'Haecksel', 'Jonah', 'MN105', 'MN60', 'MN83', 'Patchback', 'SMN5', 'Topless', 'Trigger', 'Va
u', 'Zap']
['SN100', 'SN89']
# clusters formed:4
Modularity of dolphin club:
Time taken:0.017766714096069336
0.4954906847039278
```

We have got a 0.4954 modularity score for the karate club dataset using modularity maximization algorithm. Total number of clusters formed using this algorithm is 4 and the run time of the algorithm is 0.0177.

- ❖ Spectral clustering

1] Karate Club

```
Karate Club :

Clusters :
[[1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22, 29], [5, 6, 7, 11, 17], [24, 25, 26, 28, 32], [9, 10, 15, 16, 19, 21, 23, 27, 30, 31,
33, 34]]
Number of Clusters Found :
4
Modularity Score: 0.4101742274819198
Time taken :  0.1768660545349121
```

We have got a 0.4101 modularity score for the karate club dataset using the graph Laplacian algorithm. Total number of clusters formed using this algorithm is 4 and the run time of the algorithm is 0.1768.

2] Dolphin Club

```
DOLPHIN CLUB
Clusters :
[['CCL', 'Cross', 'Double', 'Five', 'Haecksel', 'Jonah', 'MN105', 'MN60', 'MN83', 'Patchback', 'SMN5', 'SN100', 'Topless', 'TR9
9', 'Trigger', 'Vau', 'Zap'], ['DN16', 'DN21', 'Feather', 'Gallatin', 'Ripplefluke', 'SN89', 'SN90', 'TR82', 'Wave', 'Web', 'Zi
g'], ['Bumper', 'Fork', 'Grin', 'Hook', 'Kringel', 'Scabs', 'Shmuddel', 'SN4', 'SN63', 'SN9', 'Stripes', 'Thumper', 'TR120', 'T
R88', 'TSN103', 'TSN83', 'Whitetip', 'Zipfel'], ['Beak', 'Beescratch', 'DN63', 'Fish', 'Jet', 'Knit', 'MN23', 'Mus', 'Notch',
'Number1', 'Oscar', 'PL', 'Quasi', 'SN96', 'TR77', 'Upbang']]
Number of Clusters Found :
4
Modularity Score : 0.4833076223250663
Runtime :  0.1768660545349121
```

We have got a 0.4833 modularity score for the karate club dataset using the graph Laplacian algorithm. Total number of clusters formed using this algorithm is 4 and the run time of the algorithm is 0.1768.

3] Jazz Club

```
Jazz club
Clusters :
[['7', '8', '9', '10', '30', '36', '37', '49', '56', '57', '58', '62', '63', '64', '65', '66', '67', '68', '69', '70', '71',
'72', '73', '74', '75', '76', '78', '79', '80', '81', '83', '85', '86', '87', '88', '90', '91', '92', '93', '94', '95', '9
6', '97', '98', '99', '100', '101', '102', '103', '104', '105', '106', '107', '111', '117', '118', '120', '121', '134', '13
5', '136', '140', '154', '156', '162', '163', '189', '191', '193', '194', '198'], ['59', '60', '61'], ['1', '2', '3', '4',
'5', '6', '77', '82', '84', '89', '108', '109', '110', '112', '113', '114', '115', '116', '119', '122', '123', '124', '125',
'126', '127', '128', '129', '130', '131', '132', '133', '141', '142', '143', '144', '145', '146', '147', '148', '149', '15
0', '151', '152', '157', '158', '159', '160', '164', '165', '166', '167', '168', '169', '170', '171', '172', '173', '174',
'175', '176', '177', '178', '179', '187', '188', '192', '195'], ['11', '12', '13', '14', '15', '16', '17', '18', '19', '20',
'21', '22', '23', '24', '25', '26', '27', '28', '29', '31', '32', '33', '34', '35', '38', '39', '40', '41', '42', '43', '4
4', '45', '46', '47', '48', '50', '51', '52', '53', '54', '55', '137', '138', '139', '153', '155', '161', '180', '181', '18
2', '183', '184', '185', '186', '190', '196', '197']]
Number of Clusters :
4
Modularity Score : 0.0035469391017102915
Runtime :  0.2202155590057373
```

We have got a 0.4833 modularity score for the karate club dataset using the graph Laplacian algorithm. Total number of clusters formed using this algorithm is 4 and the run time of the algorithm is 0.1768.

- **Comparison**

Algorithm	**Dataset**	**No. of Clusters**	**Modularity Score**	**Run time**
Girvan-Newman	Karate Club	11	0.315910	0.2261
	Dolphin Club	11	0.48528	0.8036
	Jazz Club	11	0.009332	37.1885
Modularity Maximization	Karate Club	3	0.3806	0.0043
	Dolphin Club	4	0.4954	0.01776
	Jazz Club	4	0.4172	0.2794
Graph Laplacian	Karate Club	4	0.4101	0.1768
	Dolphin Club	4	0.4833	0.1768
	Jazz Club	4	0.003	0.2202

- **Observation**

  We have run three different algorithms on three different datasets. From our experiments we have observed that the Dolphin club network has the most dense network of nodes in the same module whereas jazz club has the sparse network of nodes within the same module. Another observation that we did was that betweenness based algorithms takes higher time compared to modularity based and spectral algorithms whereas modularity maximization takes the least time to execute.