Compare Network Detection Algorithms with Real Datasets

project 1

소셜네트워크 분석

2020 - 1

컴퓨터공학부

2014-17719

장원이

Social Network Analysis

Project 1 Report

Introduction

본 Project는 Social Network에서 Community를 Dectection하는 Algorithm들을 비교하고 분석한다. 비교분석에 사용되는 알고리즘은 CNM(Clauset, Newman, Moore) Algorithm, Louvain Algorithm, Label Propagation Algorithm 총 3가지이다. 사용한 Dataset은 'Zachary's karate club', 'Dolphin social network', 'GEMSEC-Facebook'이다. 비교분석의 척도는 Modularity, Runtime, Edge Coverage, Performance를 사용하였다. Project는 Python 3.0을 이용하여 진행되었다.

Description of method, measures

◎ Algorithm

Girvan-Newman Algorithm, CNM Algorithm, Louvain Algorithm은 강의에서 다뤘던 Algorithm들 이므로 자세한 설명은 생략한다. 각각의 시간복잡도는 O(M^2*N), O(N^2*log(N)), O(N*logN)이다.

Label Propagation Algorithm은 위의 3가지 Algorithm과 달리 Modularity같은 특별한 지표를 이용하지 않는다. Agglomerative한 방법이며, 내가 속한 community는 내 주변사람들의 community일 확률이 높다는 직관이 바탕이 된다. 간단하게 Algorithm을 정리하자면 다음과 같다.

1. 모든 node들은 각자의 label을 가진 상태에서 시작한다.
2. node순서를 random하게한 list를 만든다.
3. list 순서대로 node를 선택한 후, 그 주변 node의 label중 가장 높은 빈도의 label(max freq label)로 변경한다.(같다면 random하게)
4. 모든 node가 유일한 max freq label을 가질 때까지 2,3을 반복한다.

위와 같은 Algorithm을 통해 같은 Label을 가진 node들을 하나의 community로 볼 수 있다. random한 요소가 많아 결과가 변동이 있고, quality도 높지는 않지만 선형시간안에 수행가능하다는 아주 큰 장점이 있다. 이 장점은 엄청나게 큰 network를 분석할 때 더욱 중요해진다.

이 Algorithm의 문제점으로는 Bipartite Graph에 적용을 하면 label이 진동하며 무한히 반복되는데, 이는 Asyncronous하게 알고리즘을 조금 변경하면 해결된다. 이 둘을 절충한 semi-synchronous 한 방법도 있으므로 Graph의 특성을 고려하여 사용하면 된다.

◎ Measure

Modularity

Modularity is one measure of the structure of networks or graphs. Difference between actual number of edges within partition and expected number of edges within partition.

Runtime

Real time the algorithm runs.

Edge Coverage

The coverage of a partition is the ratio of the number of intra-community edges to the total number of edges in the graph.

Performance

The performance of a partition is the ratio of the number of intra-community edges plus inter-community non-edges with the total number of potential edges.

Details of experiment

◎ Execution Environment

Macbook pro (13-inch, 2018, Four Thunderbolt 3 Ports)
OS - macOS Catalina 10.15.4
Processor - 2.3 GHz quad core Intel Core i5
Memory - 8GB 2133 MHz LPDDR3
Graphic - Intel Iris Plus Graphics 655 1536MB

◎ Algorithms

Girvan-newman algorithm, CNM algorithm과 Label Propagation Algorithm은 python의 Library인 ‘networkx’에 내장된 함수들을 이용하였다.

Louvain algorithm은 open source로 공개되어있는 ‘python-louvain’ Library를 이용하였다.

◎ Dataset

Zachary's karate club

#Node = 34, #Edge = 78

social network of friendships between 34 members of a karate club at a US university in the 1970s.

Dolphin social network

#Node = 62, #Edge = 159

an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand.

GEMSEC-Facebook (Politicians)

#Node = 5908, #Edge = 41729

These datasets represent blue verified Facebook page networks of different categories. Nodes represent the pages and edges are mutual likes among them.

◎ Visualization

Matplotlib와 networkx를 이용하여 시각화하였다. 같은 partition으로 분류된 node들에게는 같은 색을 칠해 보다 효과적으로 community를 확인할 수 있게 하였다.

◎ Etc

requirement.txt

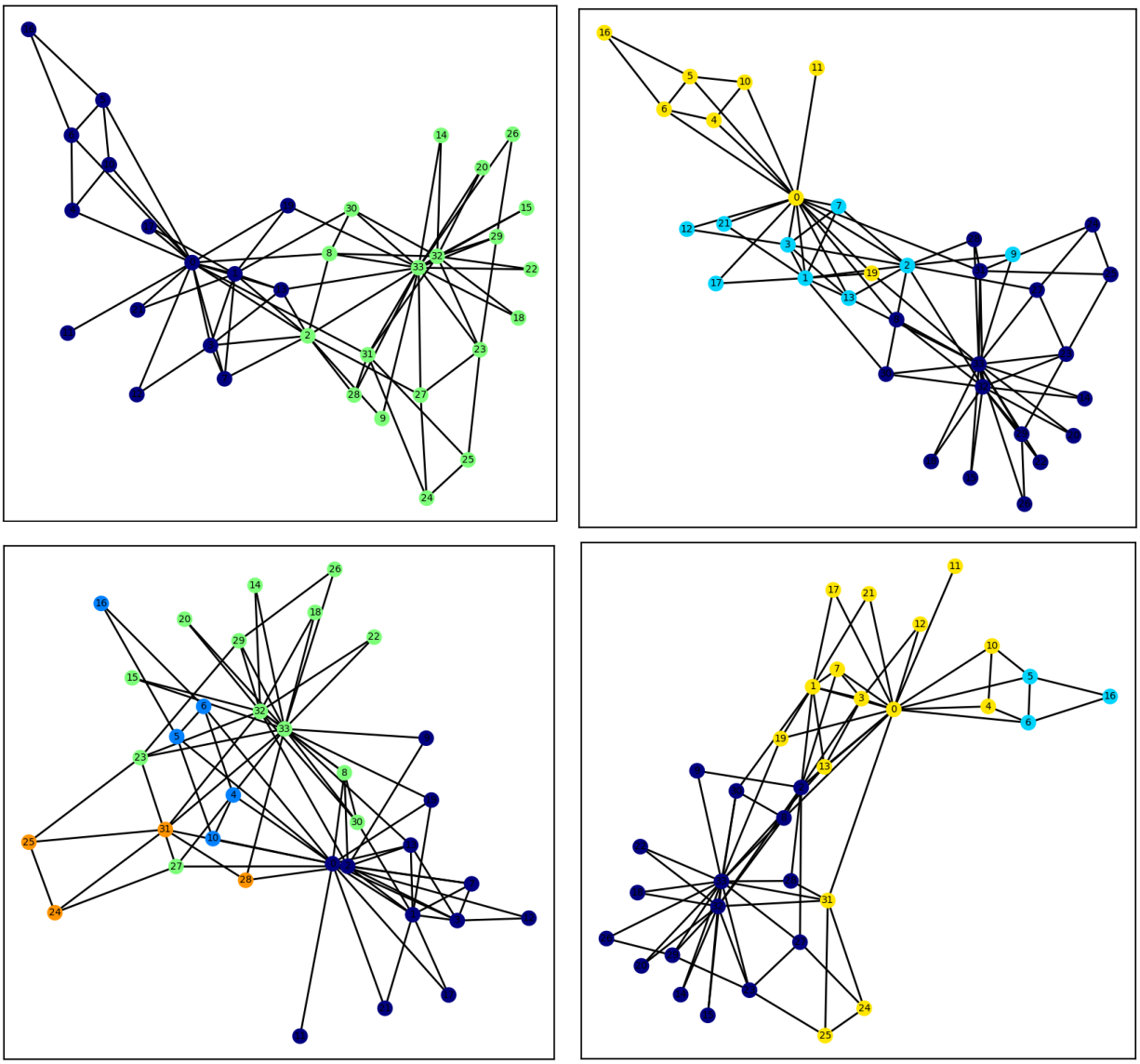
```
cycler==0.10.0
decorator==4.4.2
kiwisolver==1.2.0
matplotlib==3.2.1
networkx==2.4
numpy==1.18.3
pandas==1.0.3
pyparsing==2.4.7
python-dateutil==2.8.1
python-louvain==0.14
pytz==2020.1
six==1.14.0
```

자세한 코드는 https://github.com/JangWony/2020-1_Social-Network-Analysis 에서 확인 가능하다.

Performance analysis

◎ Zachary's karate club (#Node 34)

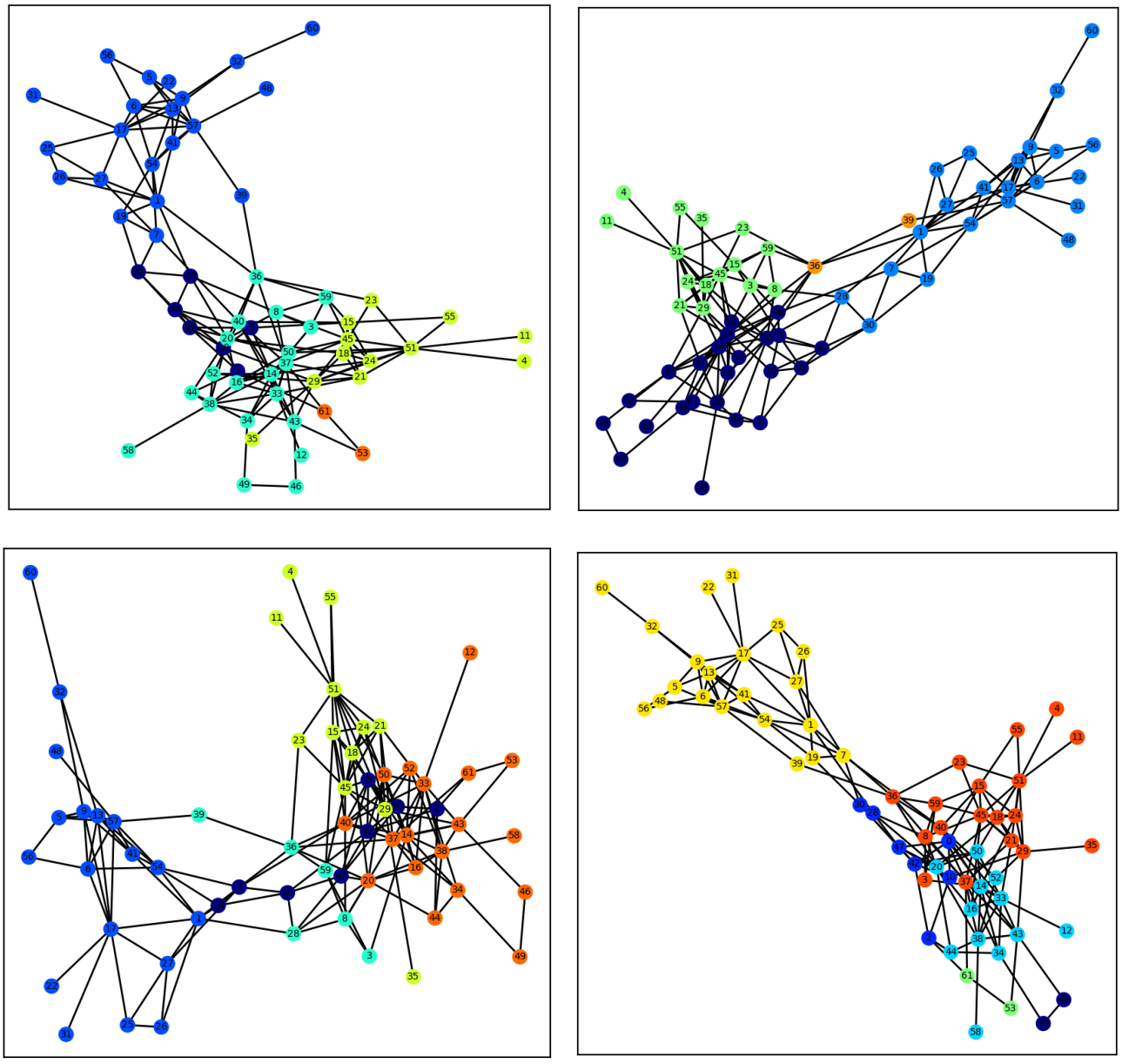
	Runtime(sec)	Modularity	Coverage	Performance
Girvan-Newman	0.06238	0.3600	0.8718	0.6114
CNM	0.00376	0.3807	0.7564	0.7861
Louvain	0.00363	0.4156	0.7564	0.7861
Label Propagation	6.96E-06	0.3251	0.6684	0.7692



좌상단부터 시계 방향으로 **Girvan-Newman, CNM, Label Propagation, Louvain**

◎ Dolphin Social Network (#Node 62)

	Runtime(sec)	Modularity	Coverage	Performance
Girvan-Newman	0.36601	0.5194	0.7987	0.7922
CNM	0.00732	0.4955	0.8239	0.8239
Louvain	0.00415	0.5188	0.7547	0.7547
Label Propagation	8.29E-06	0.4986	0.7610	0.7610



좌상단부터 시계 방향으로 **Girvan-Newman, CNM, Label Propagation, Louvain**

◎ Politician in GEMSEC-Facebook (#Node 5908)

	Runtime(sec)	Modularity	Coverage	Performance
Girvan-Newman	-	-	-	-
CNM	16.8783	0.8093	0.9388	0.8743
Louvain	1.31180	0.8689	0.9368	0.9477
Label Propagation	1.83E-05	0.8126	0.8633	0.9765

Girvan-Newman Algorithm을 사용했을 때 알고리즘의 특성상 시간이 너무 오래 걸리고, 그에 따른 실행환경의 과부하가 심해져 중지하였다.

Conclusion

본 Project에서 수행한 결과를 바탕으로 4가지 Community Detection Algorithm을 비교분석한 결과를 간단히 말하자면 Louvain Algorithm이 Girvan-Newman과 CNM보다 상위호환이었고, Runtime적으로는 Label Propagation이 가장 효과적이었다.

Dataset이 작을 때에는 수행시간이 비슷하기 때문에 quality가 중요한데, Louvain Algorithm은 다른 Algorithm과 비슷하거나 높은 quality를 보여준다. 그리고 Dataset이 커짐에 따라 Louvain Algorithm이 가장 잘 partition하면서도 빠르다는 결과를 얻을 수 있었다.

그리고 Label Propagation의 결과도 눈에 띈다. 다른 Algorithm과 비교해보았을 때 압도적으로 작은 Runtime을 보여준다. Algorithm의 특성상 선형시간안에 partition이 가능한데, 그 결과 비교적 quality가 떨어진다. 엄청 큰 Dataset을 분석할 때 Louvain algorithm으로는 가용한 시간과 computing power가 부족할 때 Label Propagation을 사용하면 quality는 조금 떨어지더라도 빠른 시간 안에 의미있는 결과를 얻을 수 있을 것이다.

최종적으로, 기본적으로는 Louvain Algorithm이 가장 efficient했고, 엄청 큰 규모의 Dataset에서는 Label Propagation Algorithm이 장점을 가진다. 다만 이는 3가지의 Dataset을 통해서만 얻은 결과이므로 Dataset이 가지는 특성에 따라 효과적인 Algorithm이 달라질 수 있음을 알아야한다.

References

Label Propagtion - *Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks." Physical review E 76.3 (2007): 036106.*

Zachary's karate club - *W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977).*

Dolphin social network - *D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology 54, 396-405 (2003).*

GEMSEC-Facebook - *B. Rozemberczki, R. Davies, R. Sarkar and C. Sutton. GEMSEC: Graph Embedding with Self Clustering. 2018.*

Networkx - *networkx.github.io*

etc - D. Easley and J. Kleinberg, "Networks, Crowds and Markets", Cambridge
M. Newman, "Networks: An Introduction", Oxford