Assignment 2 – CSE 5245

Meraj Ahmed Khan, Anirban Gupta

1. Problem Statement

Find community structures in provided Graphs using Metis, MLRMCL and Clauset-Newman-Moore community detection algorithms using different parameters. Calculate different measures of cluster quality, and perform meta analysis on the computed measures.

2. Proposed Solution

1. Find community structures using MLRMCL with different values of R
2. Find commmunity structure using CNM
3. Find community structure using Metis with the partition parameter specified as number of communities obtained using the previous two algorithms for different parameters.

2. Packages Used

networkx, community

3. Community Detection methods

METIS - Metis community detection works using multilevel recursive-bisection, multilevel *k*-way, and multi-constraint partitioning to generate high quality clusters in reasonable time.

MLRMCL - This is a scalable multi-level variant of R-MCL. The basic idea is to coarsen the graph, run R-MCL on the coarsened graph, and then refine the graph in incremental steps

Clauset, Newman, and Moore algorithm - It's a hierarchical agglomeration algorithm for detecting community structure

5. Experiments

5.1 Dataset

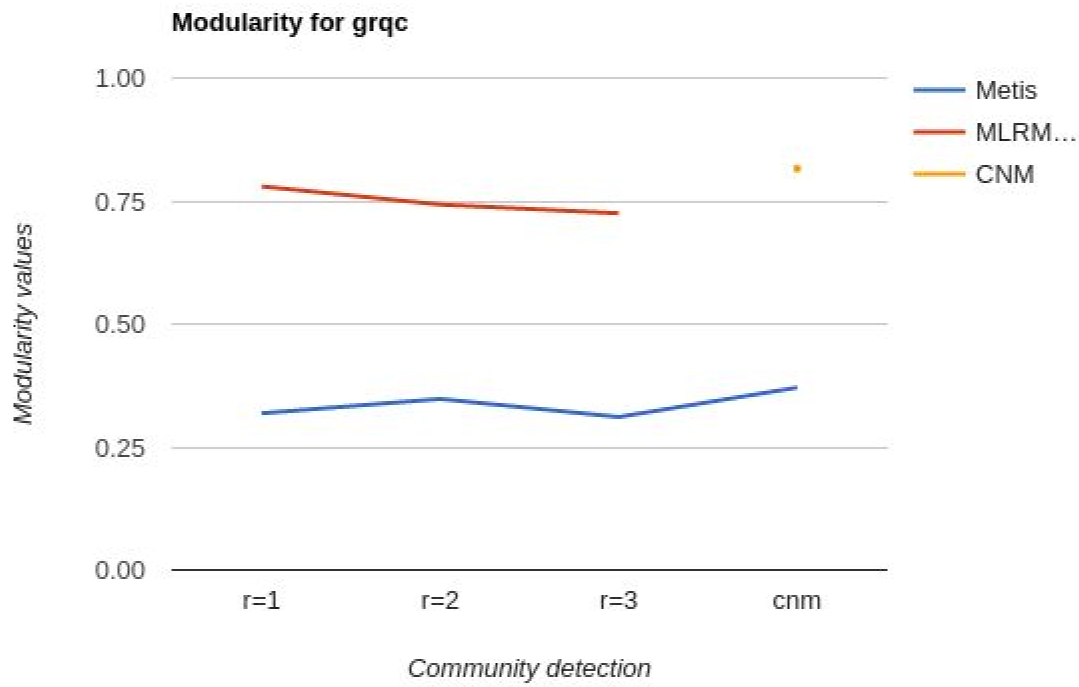
We performed our computations, and analysis on the following four graphs
Directed Graphs - Gnutella-p2p, Wiki-Vote, Arxiv - General Relativity and Quantum Cosmology (ca-GrQc), Youtube

5.2 Assumptions and Adjustments

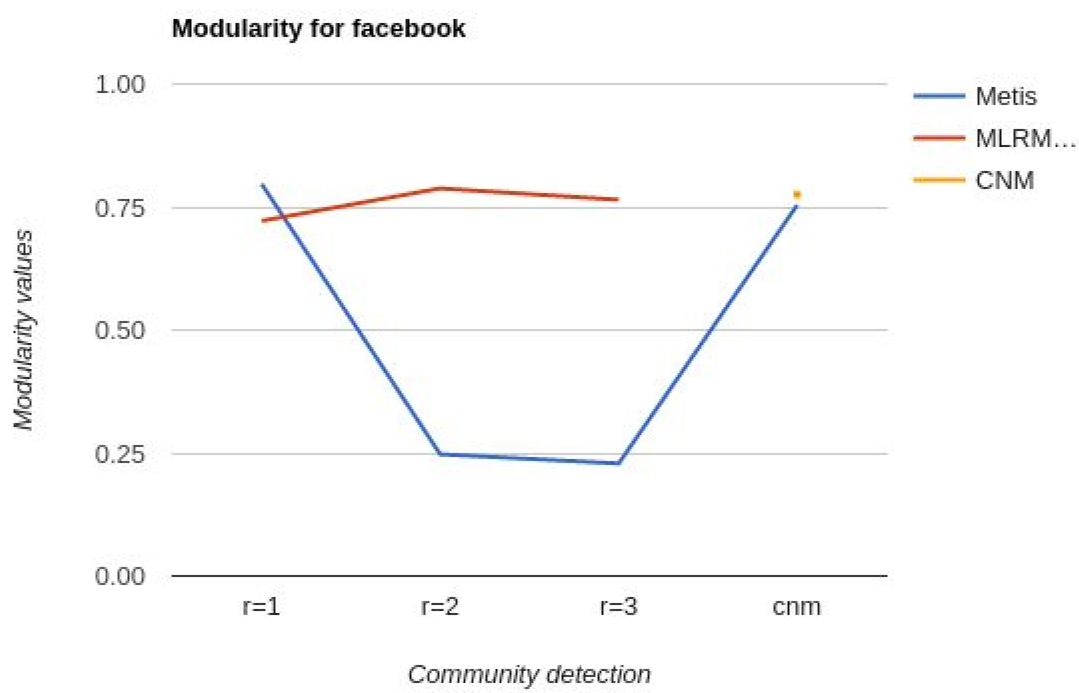
1. 4 of the graphs are disconnected, we print the conductance values across all components for the disconnected graphs in the program, and have included the minimum conductance across all components in the report.
2. We do not consider the components which constitute just one community in conductance calculations.

5.3 Results

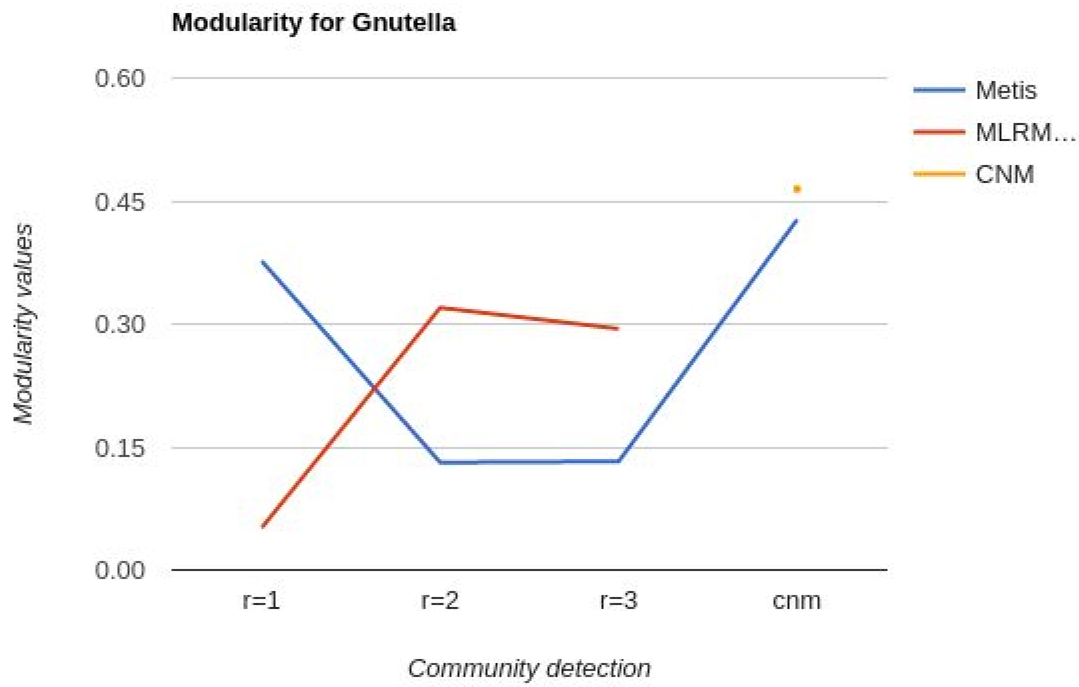
We report the modularity, time taken and average ncut for each graph.

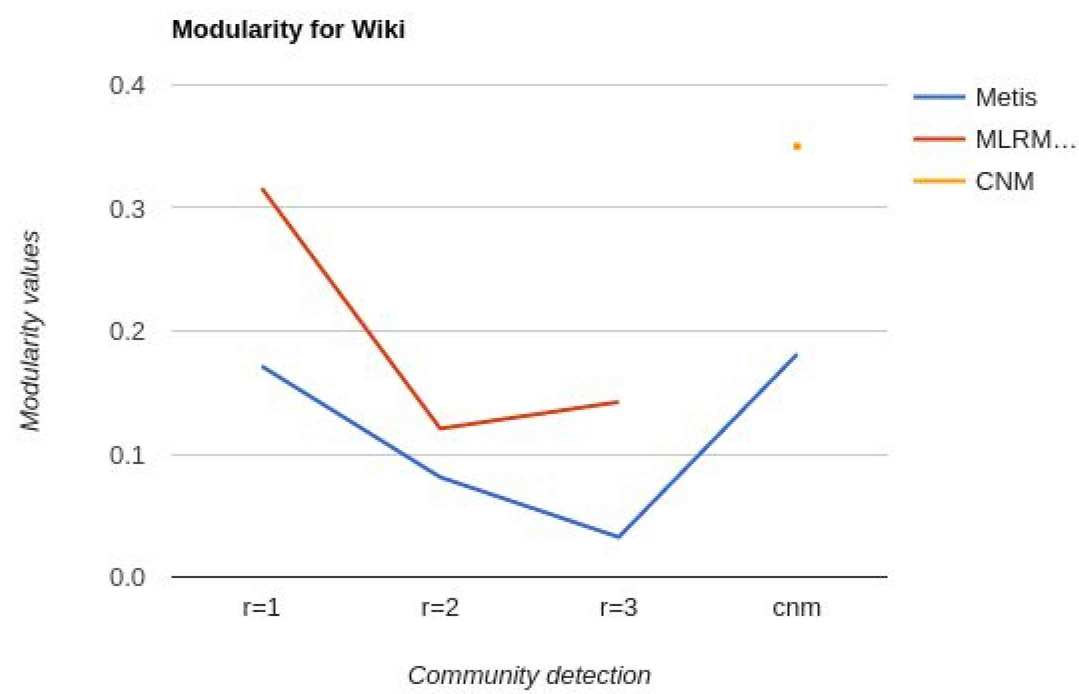


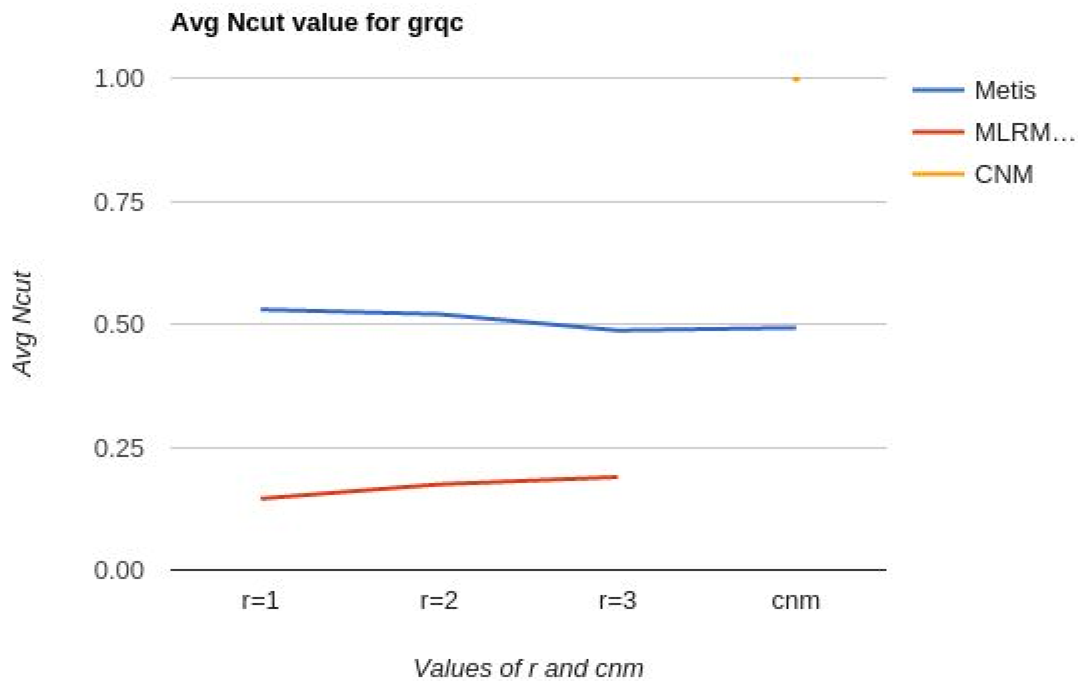
Modularity for grqc
1.00
0.75
0.50
0.25
0.00
Modularity values
r=1
r=2
r=3
cnm
Community detection
Metis
MLRM…
CNM

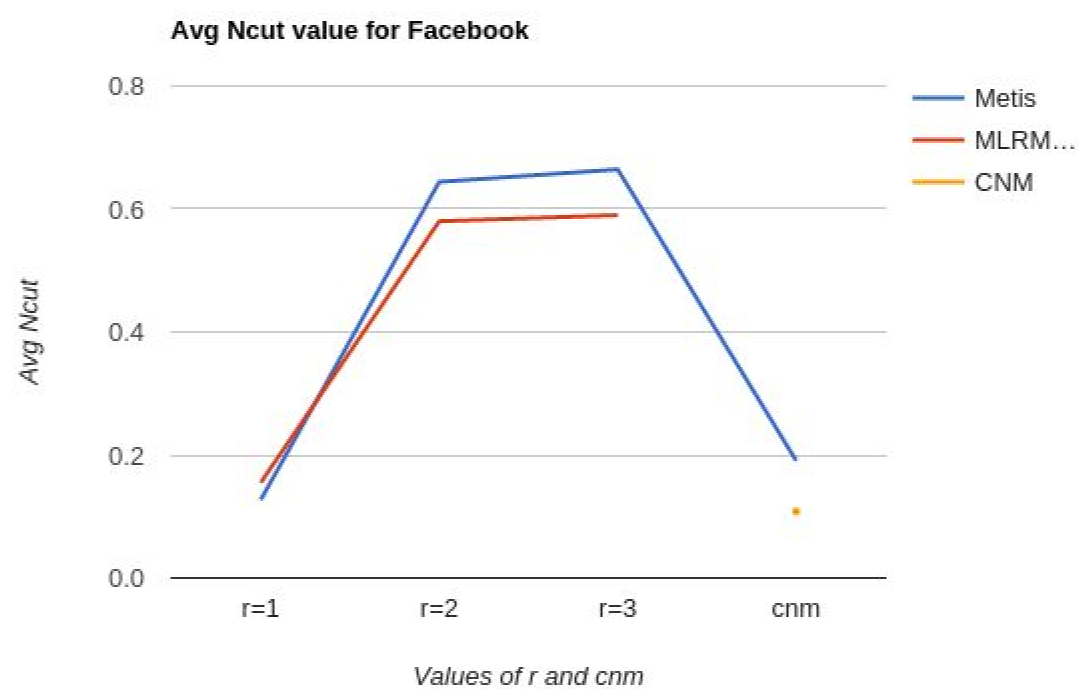


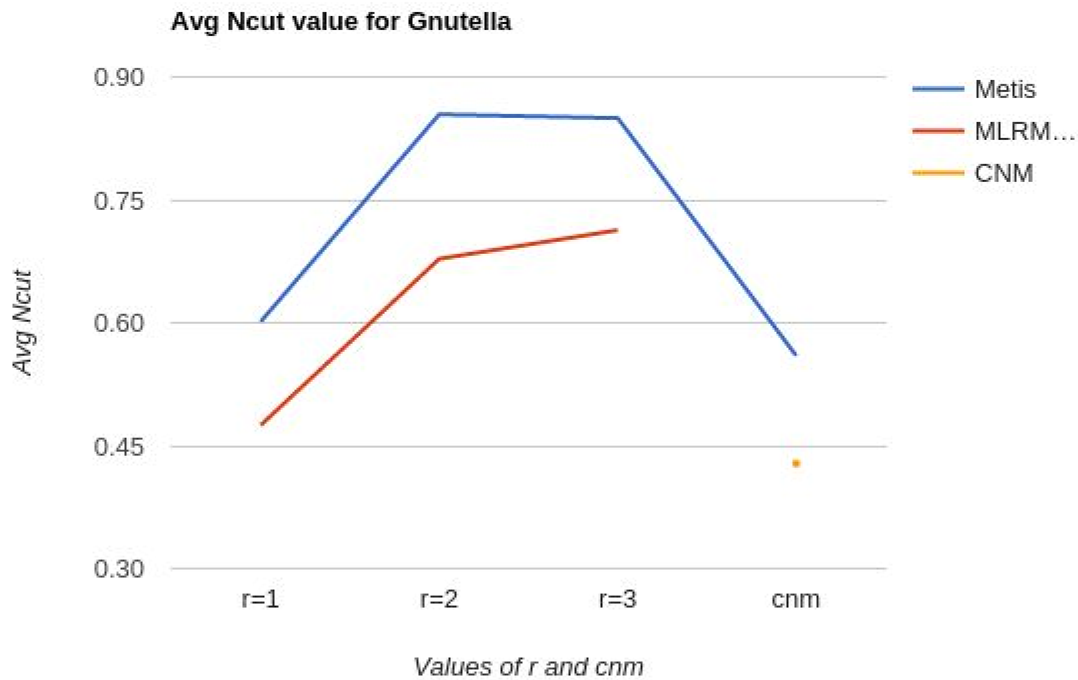
Modularity for facebook
1.00
0.75
0.50
0.25
0.00
Modularity values
r=1
r=2
r=3
cnm
Community detection
Metis
MLRM…
CNM

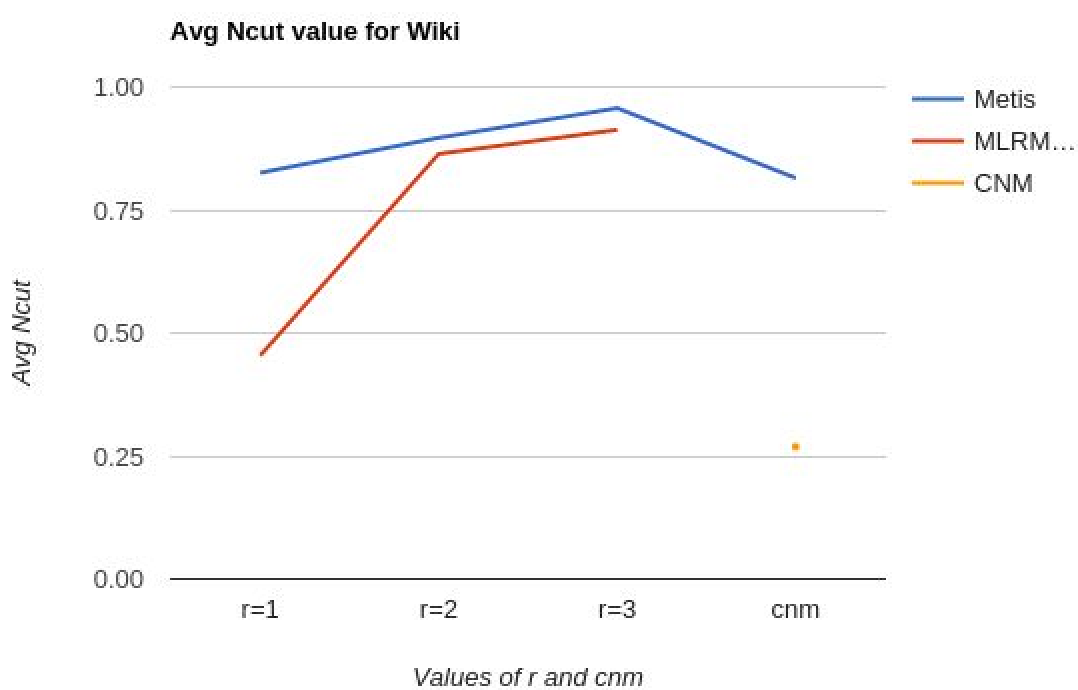

Modularity for Gnutella
0.60
0.45
0.30
0.15
0.00
Modularity values
Metis
MLRM…
CNM
r=1
r=2
r=3
cnm
Community detection

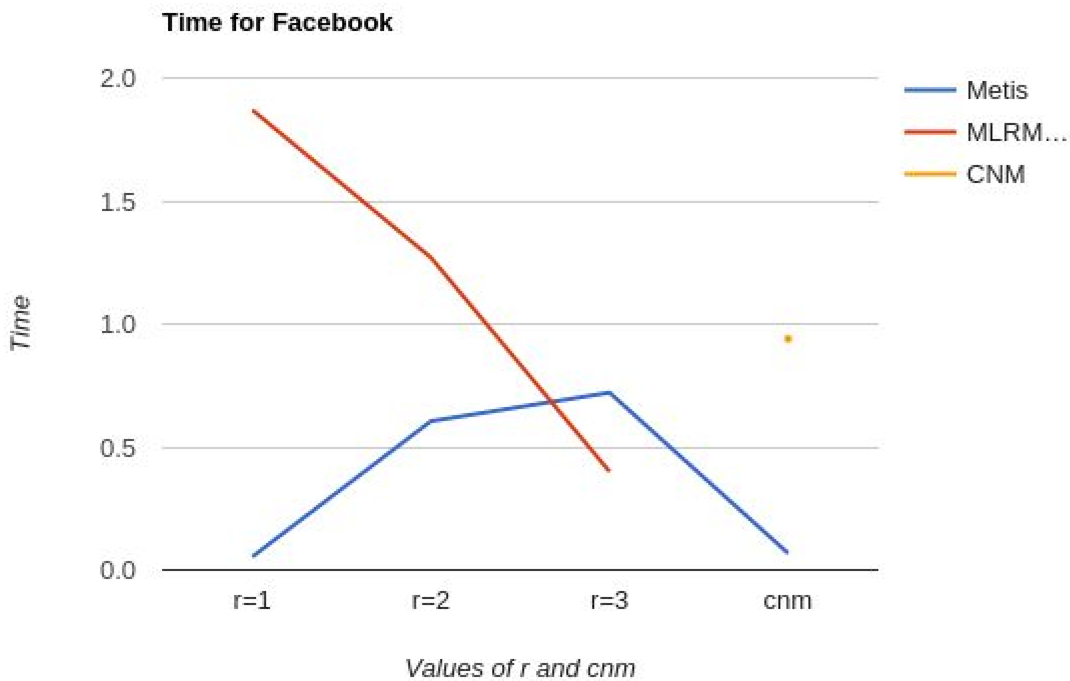

Modularity for Wiki
0.4
0.3
0.2
0.1
0.0
Modularity values
Metis
MLRM…
CNM
r=1
r=2
r=3
cnm
Community detection

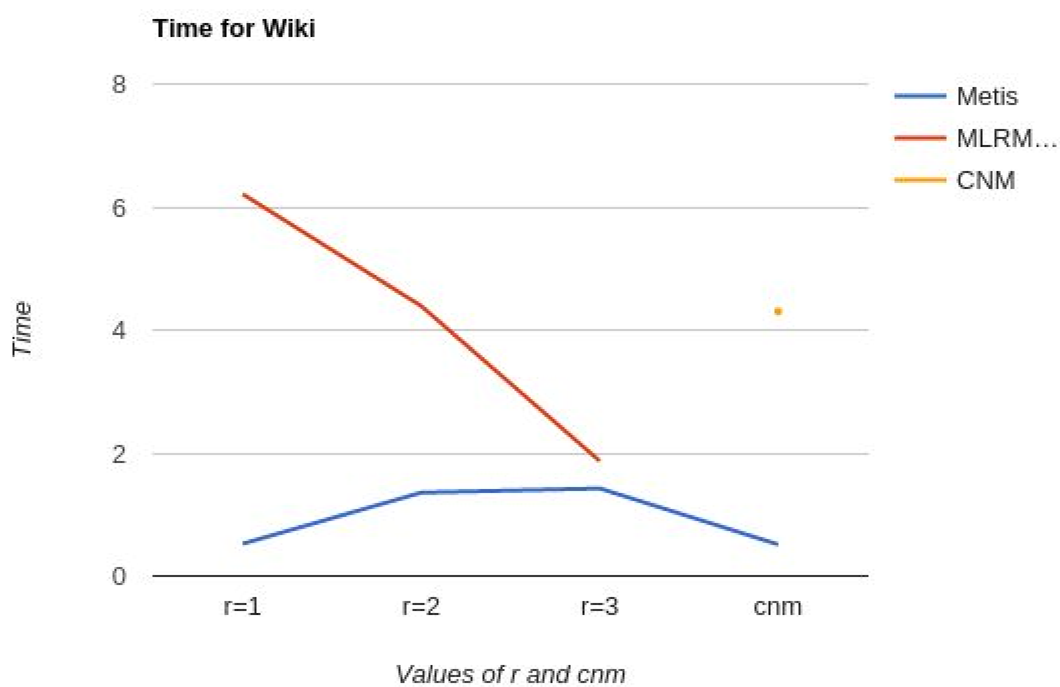

Avg Ncut value for grqc
1.00
0.75
0.50
0.25
0.00
Avg Ncut
Metis
MLRM…
CNM
r=1
r=2
r=3
cnm
Values of r and cnm

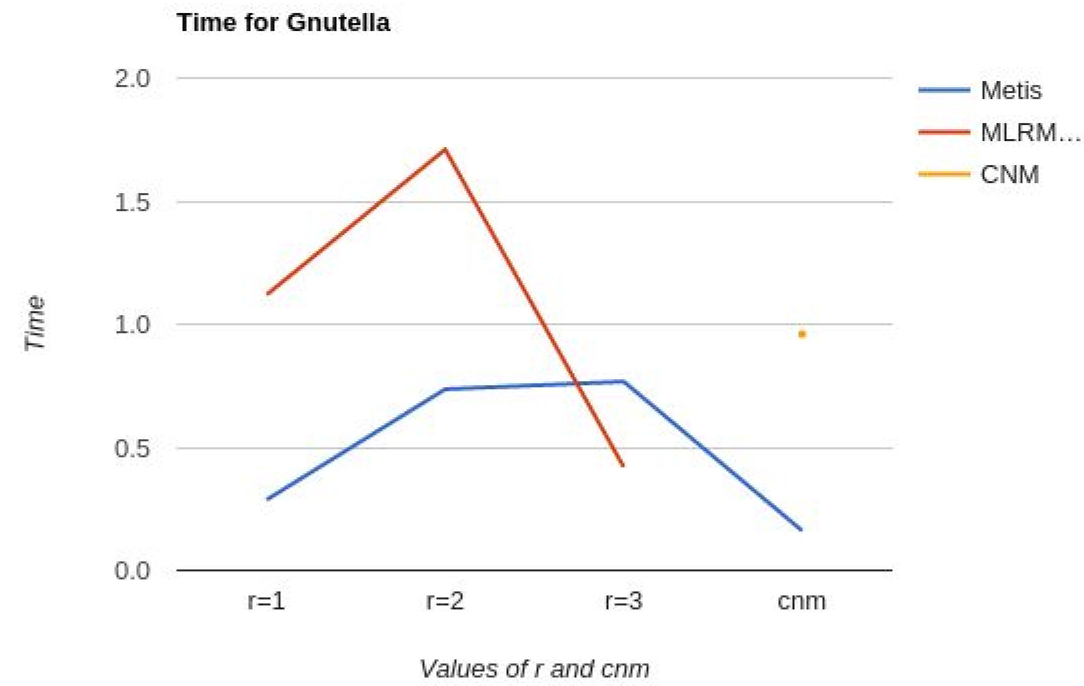

Avg Ncut value for Facebook
0.8
0.6
0.4
0.2
0.0
Avg Ncut
Metis
MLRM…
CNM
r=1
r=2
r=3
cnm
Values of r and cnm

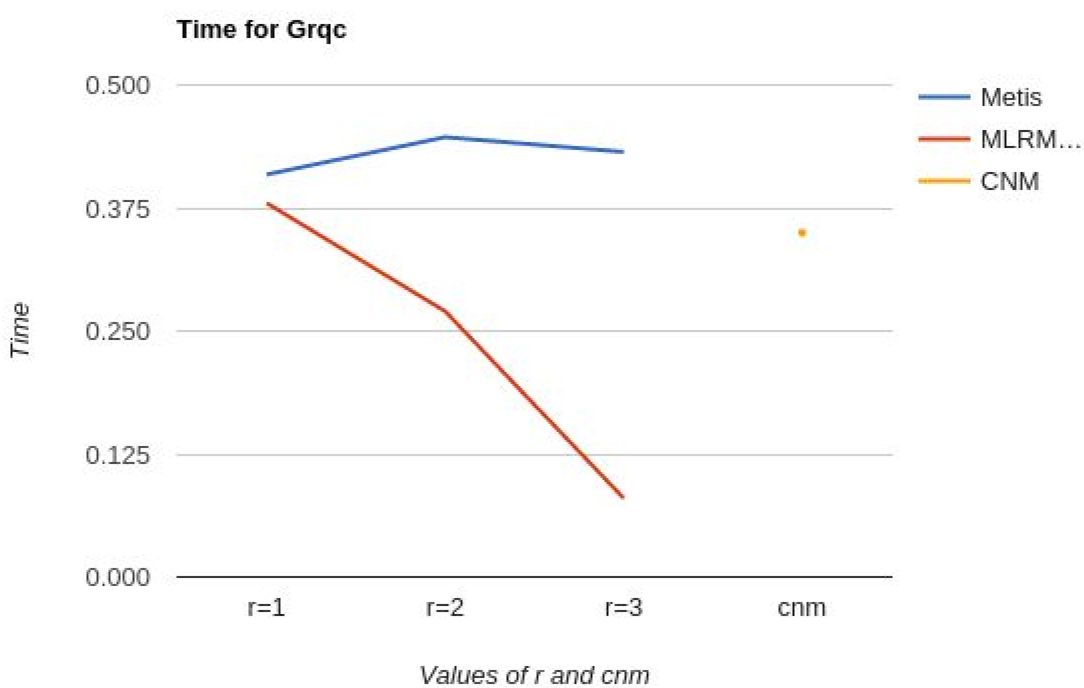

Avg Ncut value for Gnutella
0.90
0.75
0.60
0.45
0.30
Avg Ncut
r=1
r=2
r=3
cnm
Values of r and cnm
Metis
MLRM…
CNM

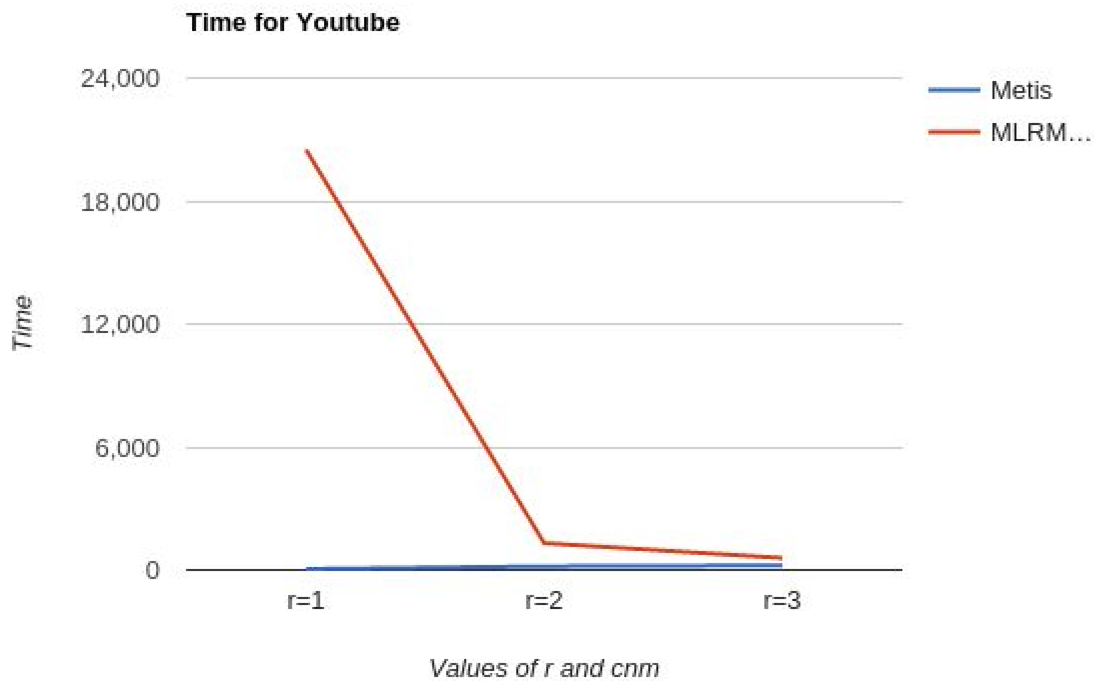

Avg Ncut value for Wiki
1.00
0.75
0.50
0.25
0.00
Avg Ncut
r=1
r=2
r=3
cnm
Values of r and cnm
Metis
MLRM…
CNM


Time for Facebook
2.0
1.5
1.0
0.5
0.0
Time
r=1
r=2
r=3
cnm
Values of r and cnm
Metis
MLRM…
CNM


Time for Wiki
8
6
4
2
0
Time
r=1
r=2
r=3
cnm
Values of r and cnm
Metis
MLRM…
CNM


Time for Gnutella
2.0
1.5
1.0
0.5
0.0
Time
Metis
MLRM…
CNM
r=1
r=2
r=3
cnm
Values of r and cnm


Time for Grqc
0.500
0.375
0.250
0.125
0.000
Time
Metis
MLRM…
CNM
r=1
r=2
r=3
cnm
Values of r and cnm



6. Observations and Insights

1) Modularity

Considering the graphs for modularity,

All the cases for Metis seems to show a dip in graph, as we try to mimic the no of partitions from r=1 to r=3 from MLRMCL. This is pretty interesting, as this shows that the strength of the division of a network decreases according to metis, as one goes through different parttion values based on r. However, the value of modularity in Metis between r=1 and CNM (no of partitions based) turns out to be pretty comparable.

Observing MLRMCL on the other hand, there seems to be an increase in Gnutella, and Facebook, and a decrease in Grqc and Wiki. This is surprising, as we thought Facebook and Grqc would show similar patterns.

We observe that CNM reports the highest score for modularity, among all three community measures. Finally, we also observed, that modularity tends to suffer from not being able to observe small communities, which could be the case for Grqc and Wiki.

2) Avg Ncut value

Now, let's consider the graphs for the average Ncut value.

Besides Grqc, the average Ncut values tend to be pretty close for values of r=1 and r=2, between Metis and MLRMCL.

Also, there seems to be a rise in the value of Ncut, as we increase from r=1 to r=2, and it saturates at r=3. Again, all graphs tend to show a severe dip for the Metis method, with the no of partitions taken from the output of the CNM method. What's more weird is that, we observe the avg Ncut values to be low for the CNM method, in all cases, except for gr-qc.

We feel that the gr-qc could consist of many small modules, thereby showing unique results,a s compared to the rest of the graphs. Also, it seems Metis and MLRMCL are fairly consistent in Grqc, thereby indicating presence of a smaller model, with smaller communities.

3) TIme

Looking at the time graphs, one can observe, that as the no of nodes, and the sizes of the communities increases, Metis starts performing better than MLMCL. Also, as r increases, MLRMCL improves at an amazing rate, pretty much catching up with Metis, by r=3. Again gr-qc seems to show a weird pattern, as compared to the rest of the graphs, with Metis performing very poorly.

This shows that time and time again, one has to be very careful while choosing the intiial no of partitions, as this plays an important role.

CNM timed out for Youtube, so we gave up on that.

4) Conductance and entropy values:

These were tough to analyze, as we faced a lot of problems
a) While calculating conductance, as there were a lot of disconnected communitites, the minimum conductance value would be reported as zero. This wouldn't provide much of an insight about the network, so we decided to return the no of components, the minimum conductance value and the conductance value for each disconnected component of the graph. We filter out values which are equivalent to zero, as they won't help, other than telling us that the graph is indeed disconnected. Conductance values don't seem to show much of a pattern. However, for an increase in r, there is an increase in conductance observed in the graphs.

b)

Values of r	MLRMCL (r=1)	MLRMCL(r=2)	MLRMCL(r=3)	Metis(r=1)	Metis(r=2)
Entropy	0.0042571518	0.00412	0.0037344015	0.003363116	0.002875748

As the youtube dataset is big, entropy takes time to calculate. Also, we had to consider what to do regarding calculating entropy of nodes, which have ground truth shown as being part of multiple communities. In such a case, we decided to do a weighted bias between the communities it belongs to, when we calculate entropy for the graph.

One can observe that the values of entropy are petty consistent for r=1 and r=2, but it takes a serious dip, which shows that there is less disorder, for r=3, thereby, showing healthier predictions for youtube communities, in the case of MLRMCL. Also, the initial no of partitions play an important role in Metis, as there is a significant decrease in entropy, thereby showing more stability.

Work Split up

Anirban Gupta - Conductance, Entropy, Metis, CNM, Avg Ncut

Meraj Ahmed Khan - Conductance, Entropy, MLRMCL, CNM, Modularity