PageRank Optimization for Graph Clustering

Group2:
Alp Ege Baştürk
Ecem İlgün
Gürkan Bor
Gündüz Hüseynli

Literature Review

Random Walk With Restarts

Personalized PageRank

Clustering and Link Prediction

Survey on the Topic

- RANDOM WALKS IN SOCIAL NETWORKS AND THEIR APPLICATIONS: A SURVEY

  Purnamrita Sarkar, Andrew W. Moore

Academic Work Related to RWWR

- Improving Random Walk Estimation Accuracy with Uniform Restarts
- Supervised Random Walks: Predicting and Recommending Links in Social Networks
- Local Graph Clustering by Multi-network Random Walk with Restart
- Random Walks with Restarts for Graph-Based Classification: Teleportation Tuning and Sampling Design

RWWR Continued

- Fast Random Walk with Restart and Its Applications
- Fast and Accurate Random Walk with Restart on Dynamic Graphs with Guarantees
- An Improved Random Walk Based Clustering Algorithm for Community Detection in Complex Networks

PPR and Related Work

- Mean Field Analysis of Personalized PageRank with Implications for Local Graph Clustering
- Personalized PageRank Clustering: A graph clustering algorithm based on random walks
- Second-order random walk-based proximity measures in graph analysis: formulations and algorithms

Clustering and Link Prediction

- From biological to social networks: Link prediction based on multi-way spectral clustering
- The Link Prediction Problem for Social Networks
- Link prediction based on local random walk
- Local Graph Clustering and Applications to Graph Partitioning
- Local Clustering Algorithm for Massive Graphs and its Application to Nearly-Linear Time Graph Partitioning
- Parallel Local Graph Clustering
- Memetic Graph Clustering

PageRank Formulation

Adding tuning factor to re-normalize the probabilities of P

$$P_{t+1} = \alpha M P_t + (1 - \alpha) P_r$$

$$p_j = \frac{p_j^\lambda}{\sum p_k^\lambda}$$

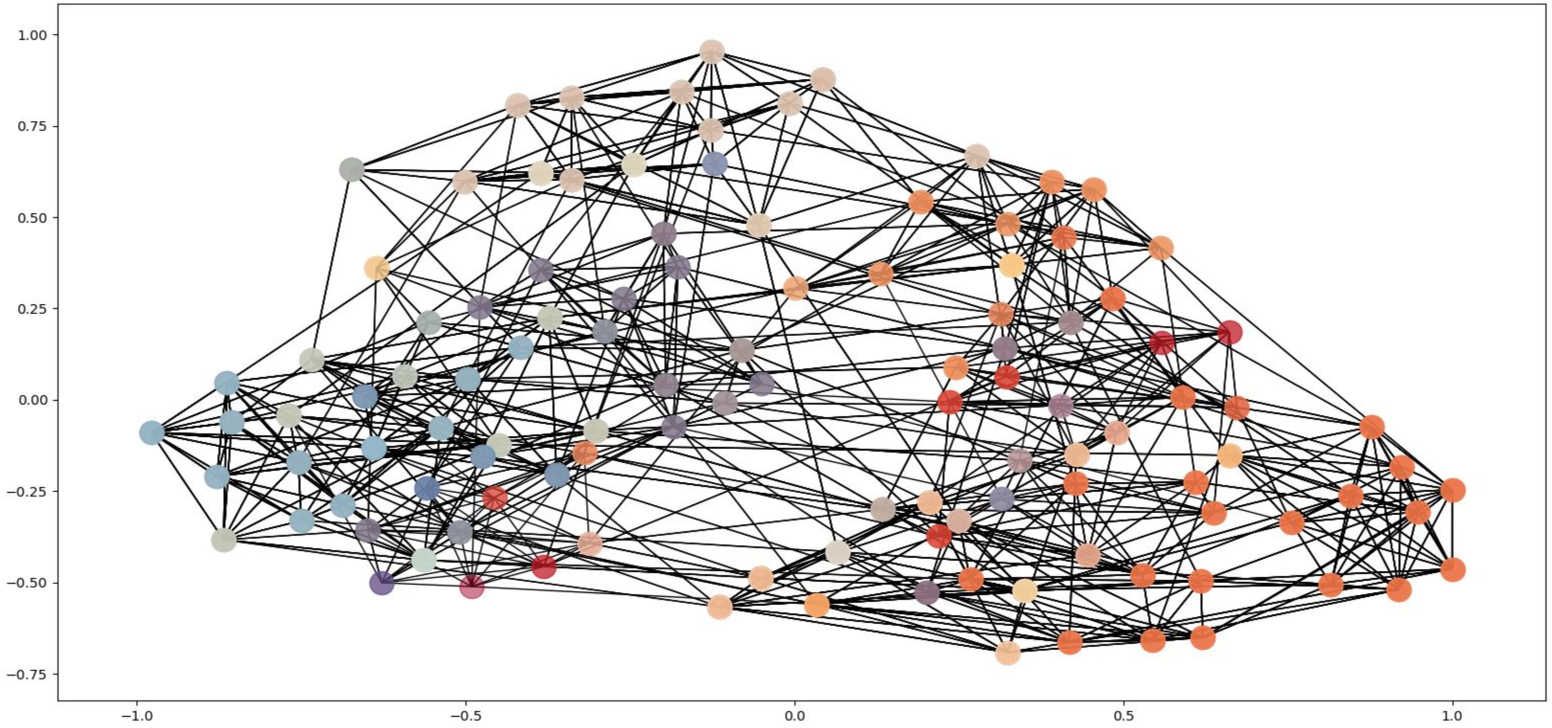
Clustering Tests

Test1	
lambda	1.2
sim_threshold	0.01
cluster_merge_threshold	0.4
NumClusters_PRM	11
NumClusters_PR	11
Conductance_PRM	20.72456
Conductance_Normal	21.16332

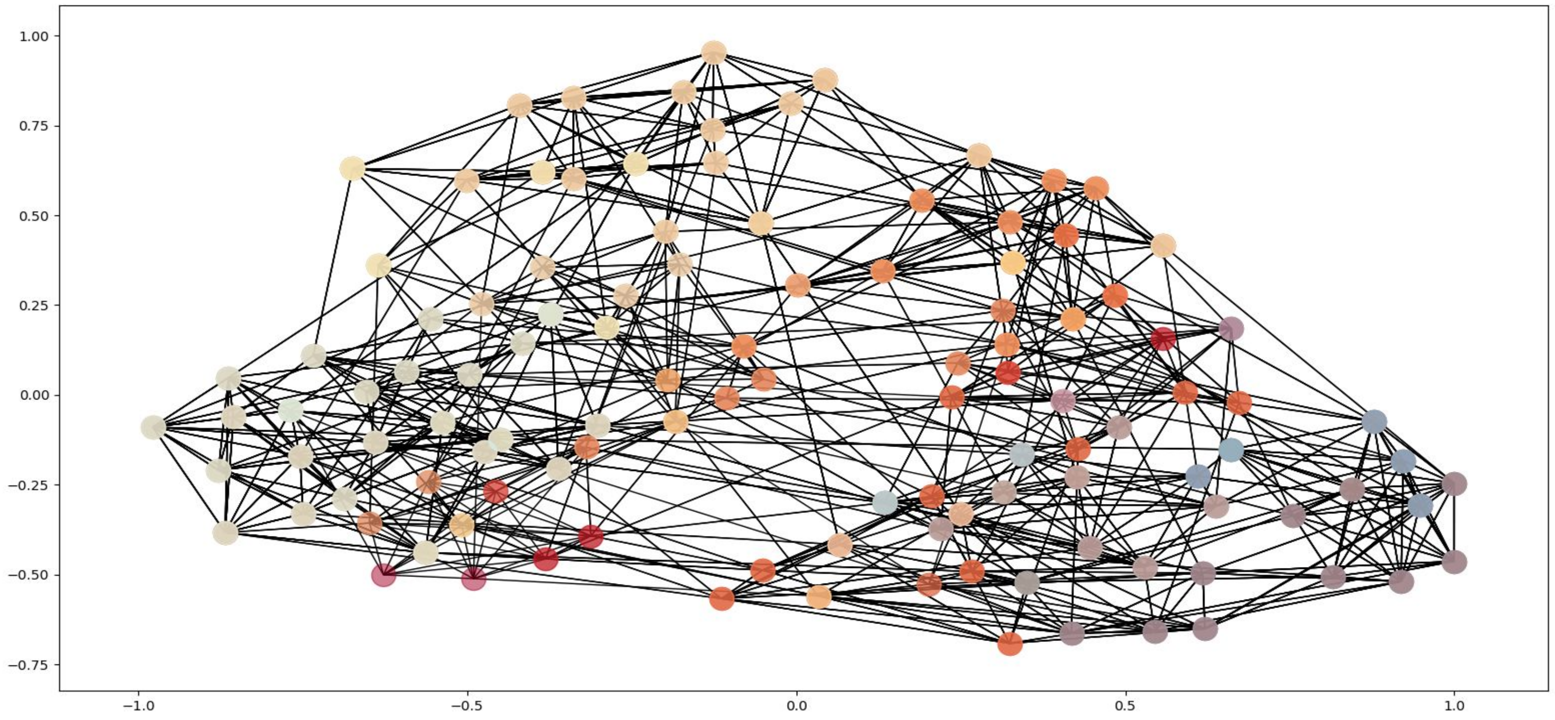
Test1.2	
lambda	2
sim_threshold	0.01
cluster_merge_threshold	0.4
NumClusters_PRM	12
NumClusters_PR	11
Conductance_PRM	26.8223
Conductance_Normal	21.16332

Algorithms were applied on FootballTSE graph [1], [2].

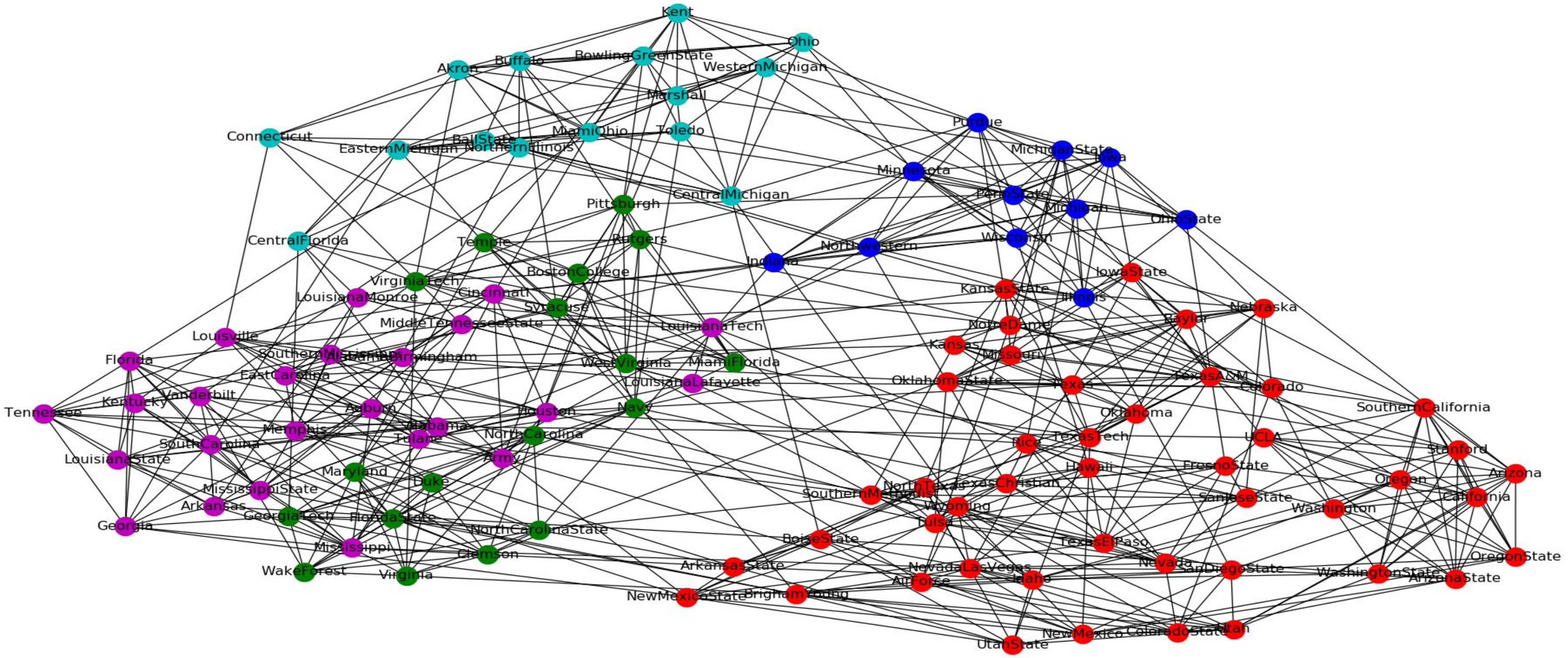
RandomWalk



RandomWalk Modified



Girvan-Newman

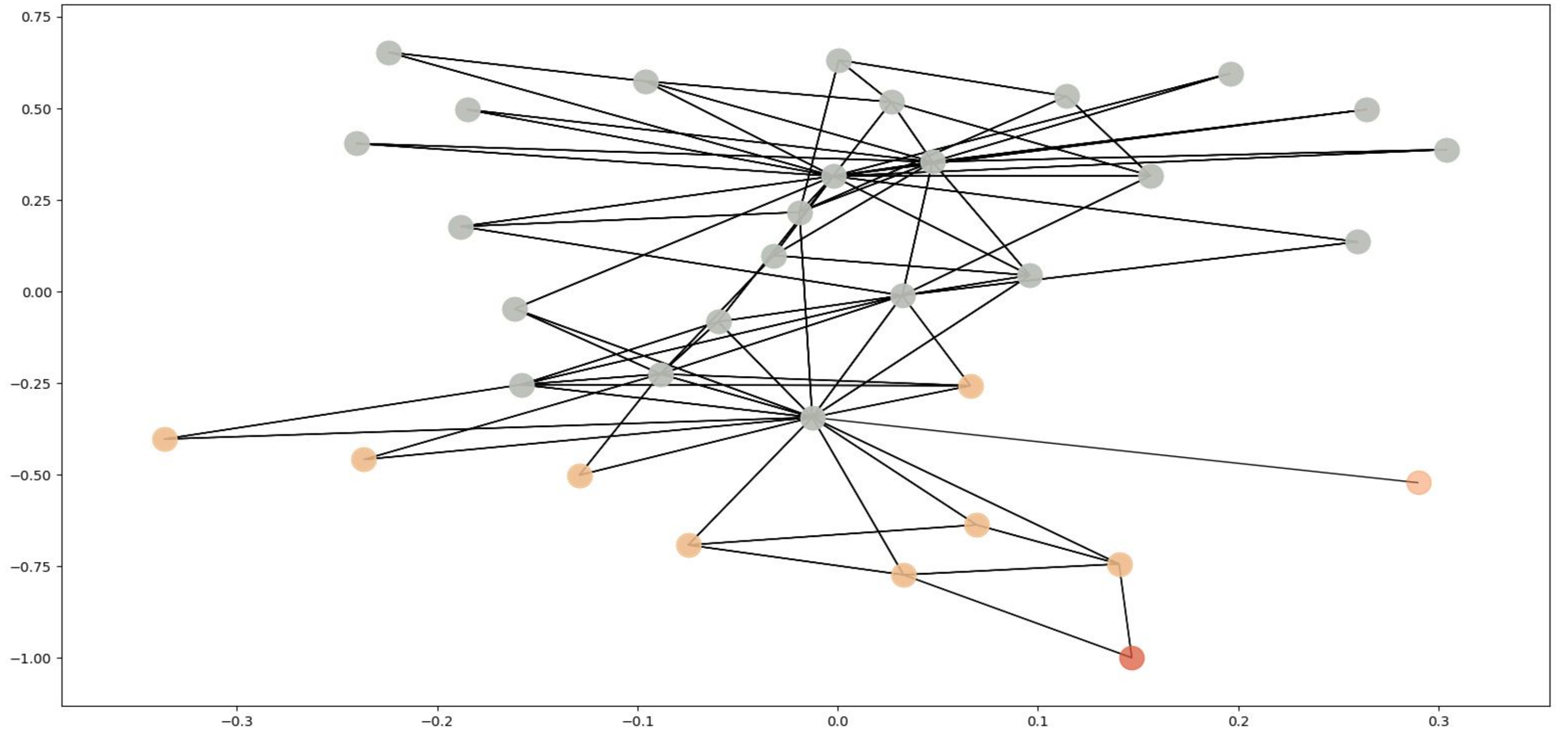


Clustering Tests

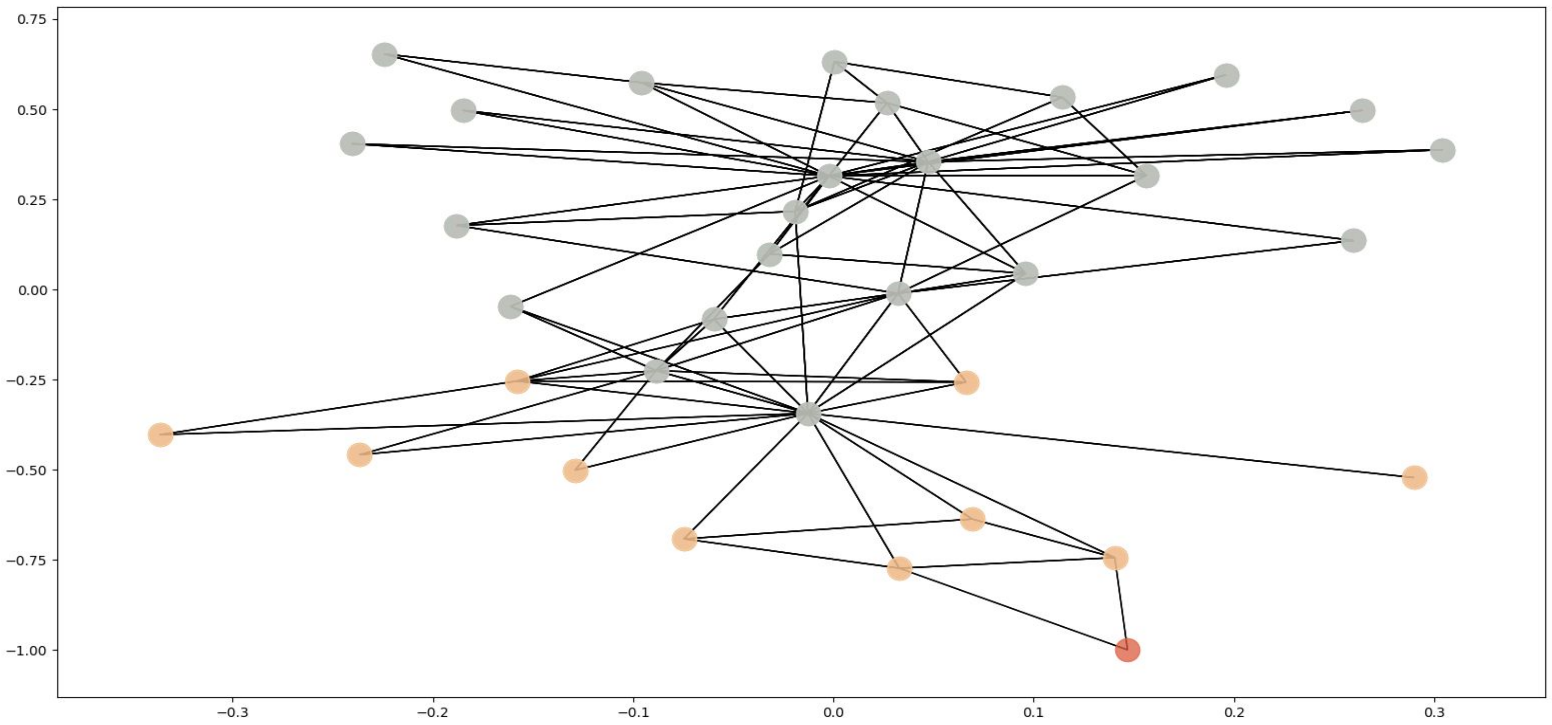
Test2	
lambda	1.2
sim_threshold	0.01
cluster_merge_threshold	0.4
NumClusters_PRM	4
NumClusters_PR	4
Conductance_PRM	4.6422
Conductance_Normal	3.1752

Algortihms were applied on Zachary’s Karate club graph [3].

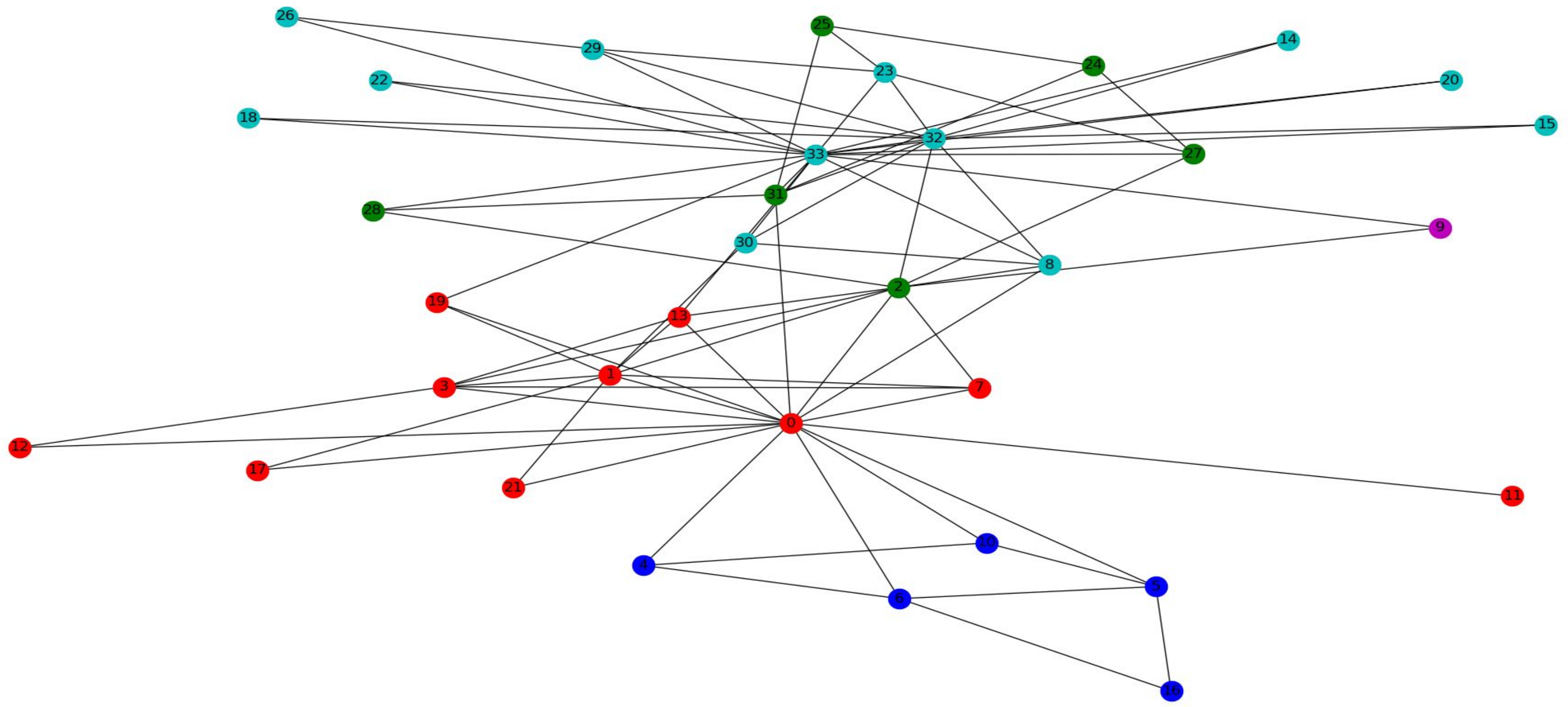
RandomWalk



RandomWalk Modified



Girvan-Newman



Fast Random Walk with Restart [5]

Two traditional methods of calculating PR:

- PreCompute (2)
- OnTheFly (1)
- Blk []

$$\vec{r}_i = c\tilde{\mathbf{W}}\vec{r}_i + (1-c)\vec{e}_i \tag{1}$$

$$\begin{aligned} \vec{r}_i &= (1-c)(I - c\tilde{\mathbf{W}})^{-1}\vec{e}_i \\ &= (1-c)\mathbf{Q}^{-1}\vec{e}_i \end{aligned} \tag{2}$$

Fast Random Walk with Restart

The Proposed Algorithm:

Me:

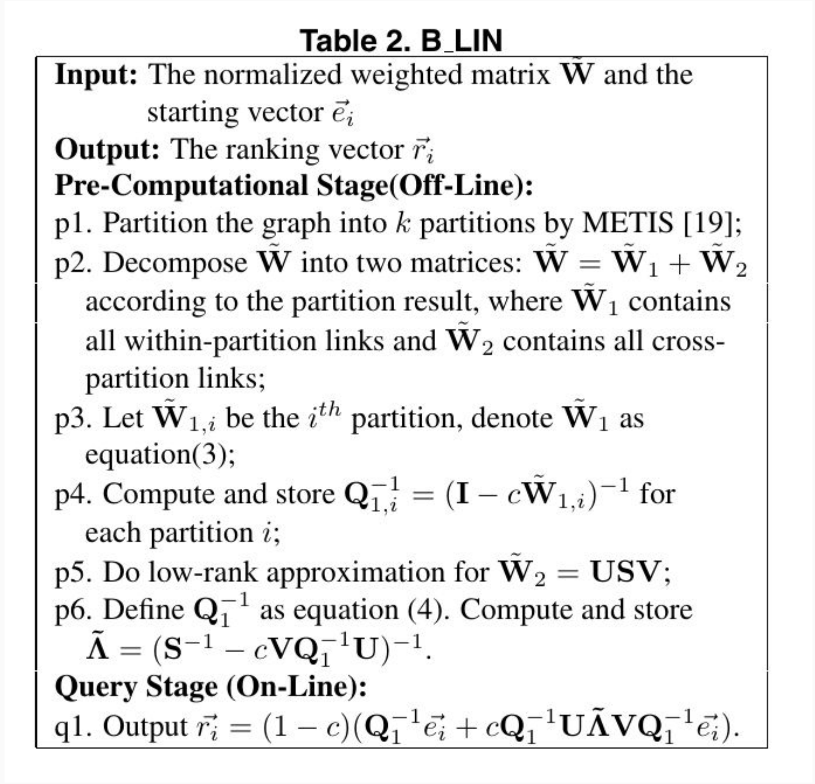


Table 2. B_LIN

```
Input: The normalized weighted matrix $\tilde{\mathbf{W}}$ and the
       starting vector $\vec{e}_i$
Output: The ranking vector $\vec{r}_i$
Pre-Computational Stage(Off-Line):
p1. Partition the graph into $k$ partitions by METIS [19];
p2. Decompose $\tilde{\mathbf{W}}$ into two matrices: $\tilde{\mathbf{W}} = \tilde{\mathbf{W}}_1 + \tilde{\mathbf{W}}_2$
    according to the partition result, where $\tilde{\mathbf{W}}_1$ contains
    all within-partition links and $\tilde{\mathbf{W}}_2$ contains all cross-
    partition links;
p3. Let $\tilde{\mathbf{W}}_{1,i}$ be the $i^{th}$ partition, denote $\tilde{\mathbf{W}}_1$ as
    equation(3);
p4. Compute and store $\mathbf{Q}_{1,i}^{-1} = (\mathbf{I} - c\tilde{\mathbf{W}}_{1,i})^{-1}$ for
    each partition $i$;
p5. Do low-rank approximation for $\tilde{\mathbf{W}}_2 = \mathbf{USV}$;
p6. Define $\mathbf{Q}_1^{-1}$ as equation (4). Compute and store
    $\tilde{\Lambda} = (\mathbf{S}^{-1} - c\mathbf{V}\mathbf{Q}_1^{-1}\mathbf{U})^{-1}$.
Query Stage (On-Line):
q1. Output $\vec{r}_i = (1-c)(\mathbf{Q}_1^{-1}\vec{e}_i + c\mathbf{Q}_1^{-1}\mathbf{U}\tilde{\Lambda}\mathbf{V}\mathbf{Q}_1^{-1}\vec{e}_i)$.
```

Fast Random Walk with Restart

What I understood:

[This page is intentionally left blank]

1. Divide the graph into smaller portions
2. And divide the matrix into two matrices, one containing all links inside the partitions, and one that contains cross sectional links
3. Compute $\mathbf{Q}_{1,i}^{-1} = (\mathbf{I} - c\tilde{\mathbf{W}}_{1,i})^{-1}$ for each partition
4. Define W and Q as below equations:

$$\tilde{\mathbf{W}}_1 = \begin{pmatrix} \tilde{\mathbf{W}}_{1,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{W}}_{1,2} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \mathbf{0} & \tilde{\mathbf{W}}_{1,k} \end{pmatrix} \tag{3}$$

$$\mathbf{Q}_1^{-1} = \begin{pmatrix} \mathbf{Q}_{1,1}^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{1,2}^{-1} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{Q}_{1,k}^{-1} \end{pmatrix} \tag{4}$$

Fast Random Walk with Restart

In the end Fast Random Walk page rank is calculated by this formula:

$$\vec{r_i} = (1-c)(\mathbf{Q}_1^{-1}\vec{e_i} + c\mathbf{Q}_1^{-1}\mathbf{U}\tilde{\mathbf{\Lambda}}\mathbf{V}\mathbf{Q}_1^{-1}\vec{e_i}).$$

Benefits: Online computational cost, pre computational cost and storage decreased.

Fast Random Walk with Restart

The reasons we diverged from this paper

- Too old (2006)
- Complex optimizations
- Hard to implement, and improve

Approximate Personalized Page Rank

Also uses the block structure of large networks. Runtime is proportional to the cluster sizes inside network, rather than the whole network size.

APPR [6]

Algorithm:

- Pick a single seed node and run PPR with for that seed node
- Sort the seed nodes in decreasing page rang score from step 1
- Perform sweeping over the nodes to find clusters

Mean Field Analysis of PPR

There is concentration to the mean field model of PPR when the size of subgraph and number of seeds scale linearly with the graph size [4].

$$\bar{\pi} = \left[\bar{\pi}_0 \mathbf{1}_k^T \ \bar{\pi}_1 \mathbf{1}_{m-k}^T \ \bar{\pi}_2 \mathbf{1}_{n-m}^T\right],$$

$$\frac{k(n)p(n)}{\log(n)} \to \infty$$

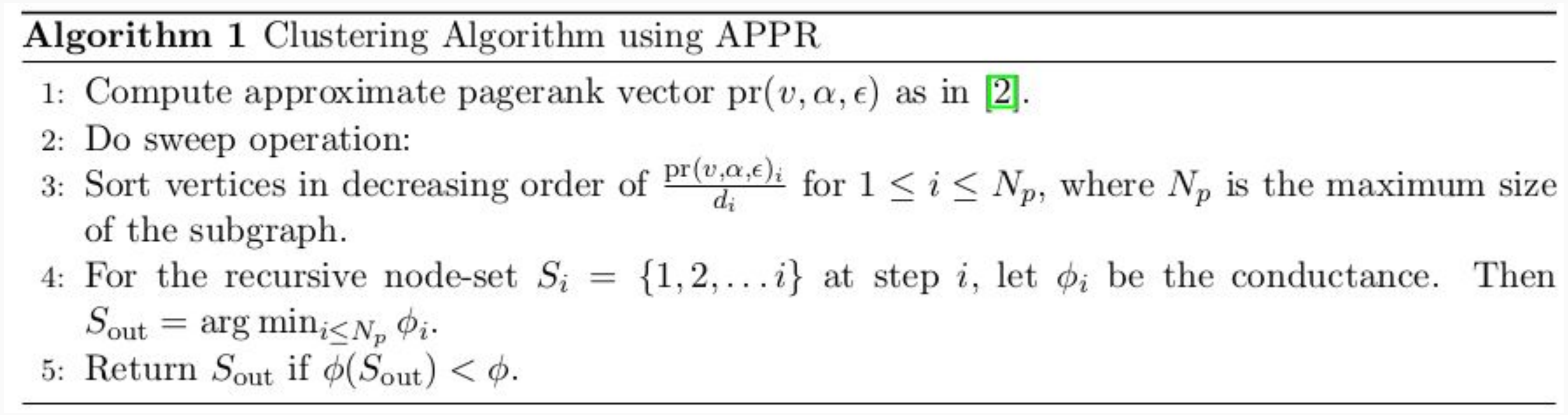
$$\bar{\pi}_2 = \frac{(1-\alpha)\alpha p}{(mq + (n-m)p)\left(1 - \frac{\alpha(n-m)}{n}\right) - \alpha m\left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}.$$

Using (6) and (7), one easily retrieves $\bar{\pi}_1$ and $\bar{\pi}_2$. Namely, we have

$$\bar{\pi}_1 = \frac{(1-\alpha)\alpha\left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}{(mq + (n-m)p)\left(1 - \frac{\alpha(n-m)}{n}\right) - \alpha m\left(q - \frac{\alpha(n-m)}{n}(q-p)\right)},$$

and

$$\bar{\pi}_0 = \frac{1-\alpha}{k} + \frac{(1-\alpha)\alpha\left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}{(mq + (n-m)p)\left(1 - \frac{\alpha(n-m)}{n}\right) - \alpha m\left(q - \frac{\alpha(n-m)}{n}(q-p)\right)}.$$



Algorithm 1 Clustering Algorithm using APPR

```
1: Compute approximate pagerank vector $\mathrm{pr}(v, \alpha, \epsilon)$ as in [2].
2: Do sweep operation:
3: Sort vertices in decreasing order of $\frac{\mathrm{pr}(v,\alpha,\epsilon)_i}{d_i}$ for $1 \leq i \leq N_p$, where $N_p$ is the maximum size
   of the subgraph.
4: For the recursive node-set $S_i = \{1, 2, \ldots i\}$ at step $i$, let $\phi_i$ be the conductance. Then
   $S_{\text{out}} = \arg\min_{i \leq N_p} \phi_i$.
5: Return $S_{\text{out}}$ if $\phi(S_{\text{out}}) < \phi$.
```

[4]

Conclusion

We have focused on quality and improved it under some conditions. But results depend on the input graph.

We have tried to improve the time of the pagerank but couldn’t improve much. Modified version is faster if added factor causes it to converge faster.

References

[1] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

[2] T.S. Evans, "Clique Graphs and Overlapping Communities", J.Stat.Mech. (2010) P12037 [arXiv:1009.0638]

[3] Zachary, Wayne. (1976). “An Information Flow Model for Conflict and Fission in Small Groups1”. Journal of anthropological research. 33. 10.1086/jar.33.4.3629752.

References

[4] Konstantin Avrachenkov et. al. “Mean Field Analysis of Personalized PageRank with Implications for Local Graph Clustering”. June 20, 2018.

[5] Hanghang Tong et. al. “Fast Random Walk with Restart and Its Applications”. September 2006

[6] R. Andersen, K. Lang and F. Chung, "Local Graph Partitioning using PageRank Vectors," 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)(FOCS), Berkeley, California, 2006, pp. 475-486.