CS425 : Algorithms for Web-scale Data Term Project
PageRank Optimization for Graph Clustering

Group 2: Alp Ege Basturk, Gurkan Bor, Ecem Ilgun and Gunduz Huseynli

***Abstract*— In this paper, various random walk based algorithms were analyzed in order to improve performance and quality of clustering.**

CONTENTS

I. PROBLEM DEFINITION

In computer science, graphs help facilitate many problems, by modeling them and creating practical data sets from which researchers can derive answers. Trying to model sample data based on similarity or other constraints become time or space consuming complex operations. Currently, as of the writing of this paper, one of approach is to use Personalized PageRank (PPR) to identify local clusters.

PageRank is derived by Google from the intuitive understanding of the importance of a web page. Simply put; it is the relation between the outgoing and incoming links of a web page. Although first built on the graph models of the world wide web, the idea has found mirrored uses in many other applications of graph algorithms and models. However in order to make sense of data in a given graph, certain traversals and operations need to be conducted on the data set.

The aforementioned approaches more often than not capitalize on a simple idea: random walks. This paper mainly investigates local clusters through the aid of PPR and link predictions. PPR can be easily derived from random walks. Using a seed set of vertices, randomly walking in a graph can identify local clusters.

Original PageRank algorithm suggested by Google, randomly walks into one of the neighbouring vertices with probability $(1-\beta)$, and teleports into a random vertex with the probability of β. Teleportation allows the algorithm to escape from dead ends and cycles. After a long walk, the probability of a random walker to be on a vertex, is same as that vertexs PageRank. PPR, modifies the previous algorithm, by forcing the random walker to teleport to only some chosen seed vertex(s), with the same probability. In its turn, this results in the random walker to teleport to the seed vertex often, and walk close to the seed set. Thus those vertices that are closer to the seed set, have a higher PageRank. In the end, the seed vertices together with the vertices that have PageRank higher than certain threshold can be considered as a single cluster.

Aside from PPR and RWWR, link prediction can also help with the identification of local clusters. Using data contained inside a vertex and its known neighbors, it is possible to predict candidate vertices that could be linked to any given vertex.

II. CONCEPTUAL INFORMATION

A graph consists of vertices and ordered pairs of edges. For each vertex V_i ; $i \in \mathbb{N}$, there might exist an edge E_j ; $j \in \mathbb{N}$, which establishes a connection with another vertex V_k. Graphs in our research models real world relationships between entities.

Graphs are traversed by random walks which give estimates for PageRanks of vertices. PageRank of a given vertex is the relative importance of the vertex when compared to other vertices. Random walks are probabilistic traversals of graphs. PageRank vector r is defined mathematically as;

$$r = \beta M.r + \frac{[1-\beta]}{N}N$$

Where $1-\beta$ denotes the probability of a restart, $\frac{[1-\beta]}{N}N$ is a vector and M is the adjacency matrix interpretation of a given graph.

III. RELATED WORK

To find a topic of interest, as well as to generate a network of useful knowledge about PPR and RWWRs, we conducted a web search in Google Scholar as well as utilizing Bilkent University Librarys Online Journal resources. During our search we came up with a number of resources, which their contents and our rationale behind our choice to utilize or discard them are explained in the following sections.

A. Initial Survey

In our initial inquiries about the topic, the most useful work we had found was a survey paper by Sarkar, P. and Moore, A. W. [19]. This work allowed us to have a precise understanding about the topic and guided us towards a research topic relevant to our interests. The paper provides a detailed overview about random walks as well as PPR, and touches on link predictions as well. Of note: it details the fundamentals behind random walks and expands on the relationship between clustering and random walks. The paper also shows the relationship between conductance, PPR and random walks. Conductance in this context is defined as the measurement of cluster quality generated by a random walk and PPR whence;
For a subset of S of all vertices V , Let $\phi V(S)$ denote conductance of S, and vol(S) = $\sum_{i \epsilon S} d(i)$. Then, conductance is calculated as follows,

$$\Phi V(S) = \frac{E(S, V \backslash S)}{min(vol(S))}$$

For a good quality cluster, conductance needs to be small, where the number of cross edges are small compared to the total number of edges. This is our measurement of cluster qualities as well.

B. Relevant Findings

1) *Improving Random Walk Estimation Accuracy with Uniform Restarts :* In [17] the properties of a hybrid sampling scheme that mixes independent uniform vertex sampling and random walk Random Walk-based crawling is studied. The proposed method is essentially a version of PageRank that preservers time-reversibility. Since we had no constraints on time reversibility in our research, this approach was discarded.

2) *Supervised Random Walks: Predicting and Recommending Links in Social Networks:* [16] deals mainly with how supervised random walks outperform state of the art approaches when trying to predict the occurrence of links. Although a sensible approach, we had sever conceptual problems in the implementation., The supervised random walks were outperforming random walks both in terms of speed and quality, the methods of predictions are not easily applicable to all data. Their idea was to find parameter vector w such that the PageRank scores of vertices in D will be greater than the scores of vertices in L, essentially an optimization problem with strict constraints. Using a regularization parameter (i.e. constraints can be violated) a simplified approach was followed. Furthermore a loss function was introduced that penalizes violated constraints. This function, a viable parameter vector and the regularization parameter are not universal for every data set. As such this approach was discarded.

3) *Parallel Local Graph Clustering:* [15] deals with parallelization of local graph clustering algorithms. Due to time constraints this approach was discarded. Of note, we were also interested in finding a faster algorithm that produced results of higher quality. If we had achieved in aim, we were also interested in parallelizing our algorithm if we had time left over.

4) *Local Graph Clustering by Multi-Network Random Walk with Restart:* In [14] deals with local clustering using random walks with restarts. However the data presented here is changed to multiple connected networks. Stemming from our inexperience in the field we discarded this approach.

5) *Fast and Accurate Random Walk with Restart on Dynamic Graphs with Guarantees:* [13] explores with tracking similarities between vertices in dynamic graphs using RWWR. Since our data models were not dynamic and the methods provided here did not give any significant edge over others we decided to discard this approach.

6) *Mean Field Analysis of PPR with Implications for Local Graph Clustering:* [**?**] deals with a means-field model of PPR on Erds-Rnyi random graphs containing a denser planted Erds-Rnyi sub-graph. This paper's theoretical implications did not prove to aid our experiments, since real world graphs did not contain as much data regarding to the potential communities, and the paper was discarded.

7) *Random Walks with Restarts for Graph-Based Classification: Teleportation Tuning and Sampling Design:* [11] studies methods for semi-supervised classifications over the vertices of a graph. Methods examined here were either out of the scope of this project or were similarly performing to current methods and as such were discarded from our approach.

8) *A Local Clustering Algorithm for Massive Graphs and its Application to Nearly-Linear Time Graph Partitioning:* [10] deals with partitioning large graphs: the local clustering method presented here finds acceptable quality partitions in near linear time. However we found other methods to compare with our work and they were far more efficient. As such this paper was discarded from our scope.

9) *Link Prediction Based on Local Random Walk :* [9] uses the data gathered by local random walks to obtain competitively good link predictions. However our current approaches to local clustering were satisfactory. As such this approach was discarded.

C. Main Findings

On the topic, we have found three papers to be more relevant than aforementioned. A 2013 study done by Symeonidis et al. [18] likens online social networks to protein-protein interaction networks and devises link prediction by performing multi-way spectral clustering. By looking at a select sample of eigenvectors and eigenvalues of a normalized Laplacian matrix, the algorithm finds a multiway partition of the data. This paper and our intention to apply it along with the latter was discarded due to time constraints.

Another work conducted by Tong et al.[7] in 2006 improves on current implementations of random walks with restarts. Random walks with restarts do not scale too well with large graphs. For large sets of data, the algorithm requires $O(n^4)$ space complexity as well as $O(n^3)$ computation complexity. Even with prefetched data, slow response times from queries are an issue. The authors propose a faster solution that exploits linear correlations which utilize low-rank matrix approximations. Furthermore they make use of block-wise community-like structures. The results increase the speed up to 150 times with 90% of quality preserved.

Finally Cai et al.[1] proposes an efficient clustering method that discovers communities in two steps. First is a clustering process that uses random walks to construct initial sets. In this part, it is proposed that adding a tuning factor to the matrix M of PageRank might favor closer vertices and result in closer knit clusters. The proposed factor is $p_j = \frac{p_j^\lambda}{\sum_k p_k^\lambda}$. In the second step clusters that have been identified prior are refined further. The algorithm uses the $O(n^3)$ complexity to perform the computation as mentioned before. $O(m^2)$ complexity is also required to process m clusters in the second step.

IV. WORK DONE

Code for the algorithms was written in Python due to the language's extended support for graph algorithms in terms of libraries. Networkx library was used to process graphs and code skeletons. In addition to Networkx, matplotlib was utilized in the creation of the visualizations of the graphs. Scipy and Numpy libraries were utilized in pagerank implementation and matrix operations respectively. PageRank based algorithm was tested in terms of quality of clustering on well documented graphs, such as Zacharys Karate Club and American College Football [2], [3], [4], [5]. Tests were done using following parameters with 0.85 transition probability. This value for the parameter was chosen due to its effectiveness when compared to other arbitrarily smaller or larger values such as 0.75 or 0.95.

Following graphs were generated using the random walk approach with restart on each vertex. Later, sets of PageRank scores obtained for each vertex were iterated and vertices with scores less than similarity threshold, which was chosen empirically, through trial and error, were removed. This process gave initial clusters around each vertex, which were traversed and merged if their similarity is greater than the cluster merge threshold. Through this merge operations clusters of acceptable quality and quantity were attained. Similarity among the clusters was calculated according to the overlapping coefficient given in Cais paper, which is $\frac{|C_i \bigcap C_j|}{min|C_i|,|C_j|}$ where C is a cluster [1].

In the graphs obtained from random walk, vertices have 0.5 alpha value for transparency so that if a vertex is part of more than one clusters, it will shown as the mix of colors assigned to these vertices to show overlapping.

Following are the parameters and results from the American College Football graph:

	Test 1	Test 2
lambda	1.2	2
sim_threshold	0.01	0.01
cluster_merge_threshold	0.4	0.4
NumClusters_PRM	11	12
NumClusters_PR	11	11
Conductance_PRM	20.72456	26.8223
Conductance_Normal	21.16332	21.16332

TABLE I

TEST PARAMETERS AND RESULTS ON FOOTBALL CLUB GRAPH FOR QUALITY MEASUREMENT

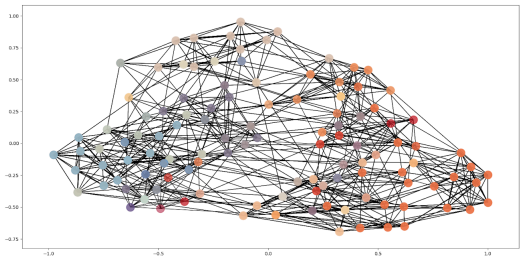

Fig. 1. First graph, generated by using PageRank with power iteration method. No modifications applied

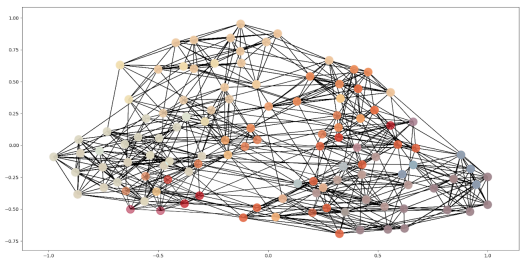

Fig. 2. Second graph, generated with modified PageRank

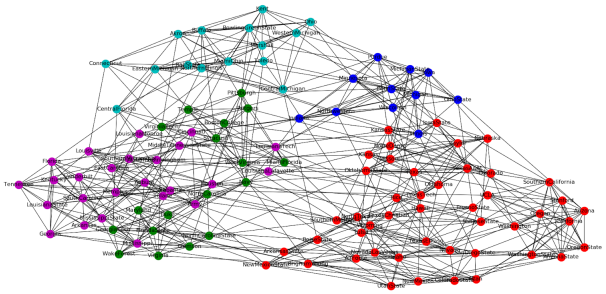

Fig. 3. Third level of Girvan-Newman algorithm [4] applied on the same data as a control of clustering

Following results were obtained from Zachary's Karate Club graph in a similar manner to experimentation on American College Football Graph:

	Test 2
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

TABLE II

TEST PARAMETERS AND RESULTS ON KARATE CLUB GRAPH FOR QUALITY MEASUREMENT

Clusters generated for this graph are shown below in the same order as the previous one:

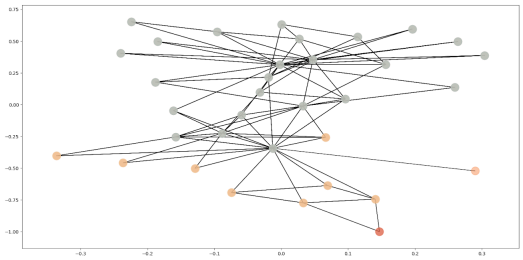

Fig. 4. First graph, generated by using PageRank with power iteration method. No modifications applied

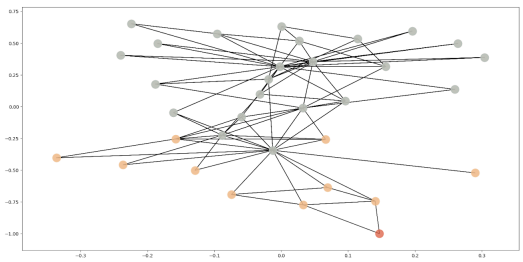

Fig. 5. Second graph, generated with modified PageRank

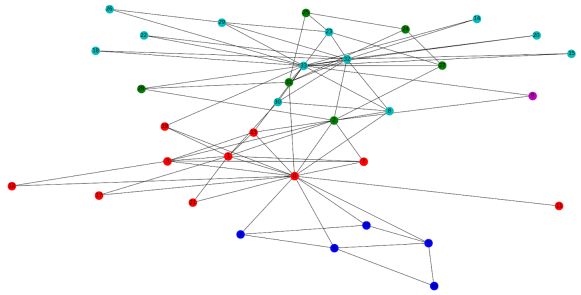

Fig. 6. Third level of Girvan-Newman algorithm [4] applied on the same data as a control of clustering

A. *Successful Findings*

Random walk based algorithms gave detailed information about the clusters. Main clusters are similar to each other and Girvan-Newman [4] control clusters. In addition, random walk based algorithms are able to find overlapping communities. All algorithms gave clusterings close to ground truth for these graphs. This was concluded that figures 1,2 are mainly similar to control clustering on figure 3 obtained by GN algorithm. This is the similar case with figures 4,5 and control graph on figure 5. Though clustering gives 4 clusters which is more than the ground truth value of Zacharys Karate club. However general structure of the clusters is similar, and misclassified cluster is at the edge. Other misclassified cluster is probably on one of the vertices and its color was blended because of the chosen transparency.
Comparing the random walk based graphs, it can be seen that some clusters are more closely knit, which are more similar to GN results. This is the expected outcome from the modification.

B. *Inconclusive Findings*

Measurements about the time were done and modified PageRank was faster in some cases, which might be due to earlier convergence. However, noise might have affected the results since measurements were close to each other.
In addition, conductance results were inconclusive. For large graphs with same number of clusters, both random walk based algorithms gave similar results, though algorithm with modified version had better results. However, unmodified PageRank based algorithm gave better results for small graphs. Thus, relation of conductance could not be concluded.

C. *Unsuccessful Findings*

In order to improve the performance of the implemented algorithm we were looking into papers that proposed solutions that decreased run time significantly. Fast RWWR and Its Applications, written by Tong et al.[7], is one of the papers we focused on early on, since authors proposed a method called BLIN, which increases the run-time efficiency significantly. The authors, building up on previous work done in the field, proposes a new method which recognizes linear correlations and block wise, community-like structure, and exploits these features found in most

real life graphs using low-rank matrix approximation and graph partitioning. The algorithm uses METIS to divide the graph into smaller partitions, and uses both within-partition links as well as cross partition links for calculating the PageRank values for vertices. Thus, their method retains the global PageRank value, opposed to PageRank values calculated by only considering local nodes inside a single partition. Furthermore, the algorithm performs low rank approximation on the matrix that represents cross partition links, to increase the run time efficiency of the algorithm. Although the promising results of the paper, we decided to discard it due to the difficulty in implementing the proposed algorithm, and integrating it to the current one. Furthermore, since the above mentioned paper was published in 2006, we decided to look into more recent papers for improving the speed of our algorithm.
We decided to transform our original graph by partitioning it using METIS, before running our algorithm. However, using slices obtained from METIS algorithm generated no significant difference when compared to our earlier results. Coarsening the partitions obtained from METIS and using local random walks and PPR proved unfruitful.

In [6], it is shown that given a graph and a sub-graph planted inside it, mean field PageRank approximations and the optimal teleportation probability is computable in $O(1)$ time. Their theoretical formulas receive number of seed vertices as variable and some knowledge about the graph and its sub-graph. However, since the paper theoretically planted the sub-graph, there is extra knowledge about it such as density and size, which is need in the calculations. Such knowledge is not available in real world graphs or even the benchmark datasets used in this paper. Since both the optimum teleportation probability and the PageRank scores formulas needed knowledge of the density and size of the sub-graph, they were not integrated into algorithms used in this paper. However, if there would be a way to predict these data about the sub-graph, e.g. cluster, then these formulas could accelerate the calculations.

V. CONCLUSION & FUTURE WORK

Algorithm proposed by Cai et. al.[1] increases the cluster quality and generates closely knit clusters compared to standard PageRank based algorithm. However PageRank based algorithms are slow because of computational complexity. Thus further work on the performance of PageRank calculations such as using approximate PageRank could reduce the time inefficiency of the algorithm.

VI. CONTRIBUTIONS

All group members took part in reading papers initially. Gurkan Bor and Ecem Ilgun have done the survey part. Gunduz Huseynli has focused on the Tong's paper. Alp Ege Basturk has focused on the Cai's paper.

REFERENCES

[1] Bingjing Cai et. al. An Improved Random Walk Based Clustering Algorithm for Community Detection in Complex Networks IEEE International Conference, 2011. https://ieeexplore.ieee.org/document/6083997

[2] Mark Newman. Network data. [Online]. Available: http://www-personal.umich.edu/ mejn/netdata/.

[3] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977).

[4] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

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

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

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

[8] 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.

[9] W. Liu and L. L, Link prediction based on local random walk, EPL (Europhysics Letters), vol. 89, no. 5, p. 58007, Jan. 2010.

[10] D. A. Spielman and S.-H. Teng, A Local Clustering Algorithm for Massive Graphs and Its Application to Nearly Linear Time Graph Partitioning, SIAM Journal on Computing, vol. 42, no. 1, pp. 126, 2013.

[11] D. Berberidis, A. N. Nikolakopoulos, and G. B. Giannakis, Random Walks with Restarts for Graph-Based Classification: Teleportation Tuning and Sampling Design, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[12] K. Avrachenkov, A. Kadavankandy, and N. Litvak, Mean Field Analysis of Personalized PageRank with Implications for Local Graph Clustering, Journal of Statistical Physics, vol. 173, no. 3-4, pp. 895916, May 2018.

[13] M. Yoon, W. Jin, and U. Kang, Fast and Accurate Random Walk with Restart on Dynamic Graphs with Guarantees, Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW 18, 2018.

[14] Y. Yan, D. Luo, J. Ni, H. Fei, W. Fan, X. Yu, J. Yen, and X. Zhang, Local Graph Clustering by Multi-network Random Walk with Restart, Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, pp. 490501, 2018.

[15] J. Shun, F. Roosta-Khorasani, K. Fountoulakis, and M. W. Mahoney, Parallel local graph clustering, Proceedings of the VLDB Endowment, vol. 9, no. 12, pp. 10411052, Jan. 2016.

[16] L. Backstrom and J. Leskovec, Supervised random walks, Proceedings of the fourth ACM international conference on Web search and data mining - WSDM 11, 2011.

[17] K. Avrachenkov, B. Ribeiro, and D. Towsley, Improving Random Walk Estimation Accuracy with Uniform Restarts, Algorithms and Models for the Web-Graph Lecture Notes in Computer Science, pp. 98109, 2010.

[18] P. Symeonidis, N. Iakovidou, N. Mantas, and Y. Manolopoulos, From biological to social networks: Link prediction based on multi-way spectral clustering, Data Knowledge Engineering, vol. 87, pp. 226242, 2013.

[19] P. Sarkar and A. W. Moore, Random Walks in Social Networks and their Applications: A Survey, Social Network Data Analytics, pp. 4377, 2011.