2011 11th IEEE International Conference on Data Mining

Scalable Diversified Ranking on Large Graphs

Rong-Hua Li   Jeffrey Xu Yu
The Chinese University of Hong Kong, Hong Kong
{*rhli,yu*}*@se.cuhk.edu.hk*

***Abstract*—Enhancing diversity in ranking on graphs has been identified as an important retrieval and mining task. Nevertheless, many existing diversified ranking algorithms cannot be scalable to large graphs as they have high time or space complexity. In this paper, we propose a scalable algorithm to find the top-K diversified ranking list on graphs. The key idea of our algorithm is that we first compute the Pagerank of the nodes of the graph, and then perform a carefully designed vertex selection algorithm to find the top-K diversified ranking list. Specifically, we firstly present a new diversified ranking measure, which can capture both *relevance* and *diversity*. Secondly, we prove the submodularity of the proposed measure. And then we propose an efficient greedy algorithm with linear time and space complexity with respect to the size of the graph to achieve near-optimal diversified ranking. Finally, we evaluate the proposed method through extensive experiments on four real networks. The experimental results indicate that the proposed method outperforms existing diversified ranking algorithms both on improving diversity in ranking and the efficiency of the algorithms.**

I. Introduction

Ranking nodes in graphs is a fundamental task in information retrieval, data mining, and social network analysis. It has a large number of applications such as ranking web-pages in search engine [1], and enhancing personalized service for web search [2]. Most of existing graph-based ranking algorithms are based on the stationary distribution of the random walk on graphs, such as the Pagerank algorithm [1] and its variants [2]. The idea of random walk based ranking algorithms is that the node of a graph should be ranked higher if there are more high ranked nodes link to it. This basic idea has been a crucial criteria for designing ranking algorithm on graphs and also has been successfully applied in many applications.

However, as discussed in some previous work [3][4], this design criteria will lead to many similar nodes in the top-K ranking list because it only considers the centrality of the nodes. Therefore, it will reduce the ranking performance when the applications need to incorporate diversity into the top-K ranking results. Take Flickr (www.flickr.com), which is a well known photo shared website, as an example. Users in Flickr can make friends and join in diverse groups in terms of different interests. Consider a mining task for finding top-K prestige users with diverse groups in Flickr social network. In general, we can use Pagerank algorithm to rank the users, and then find the top-K users based on their Pagerank scores. However, the top-K users yielded by the Pagerank algorithm typically involve many *redundant users* who join in the same interests group, thereby they cannot meet our objective. To this end, we should take the *group coverage* of the top-K ranking list into account for designing ranking algorithm. In other words, the ranking algorithm in this case should yield diversified ranking results so as to cover as many groups as possible.

Recently, improving diversity in top-K ranking results has attracted much attention as it has a variety of applications in information retrieval and data mining areas. There exists a large body of work on search results diversification both in text and graph-type datasets respectively. In this paper, we focus on enhancing diversity in ranking on graph-type datasets. We are interested in finding the top-K ranked nodes that are both *relevant* and dissimilar to each other. Here, the *relevance* of the nodes is measured by their Pagerank score.

In the literature, there exist two frameworks to diversify ranking on graphs. The first one is based on a greedy vertex selection procedure [3][5], and the second framework is based on a so-called vertex reinforced random walk [4]. In particular, the greedy vertex selection procedure chooses a vertex with maximum random walk based ranking score at one time, and then removes the selected vertex from the graph. To get the top-K ranking list, this process will repeat K times. To the best of our knowledge, there are two algorithms based on this framework: the Grasshopper algorithm [3] and the manifold rank with stop points algorithm [5]. Both of this two algorithms have empirically shown that they can improve diversity in ranking on graph-type data. However, the major drawback of this type algorithms is that they have cubic time complexity, thus they cannot be scalable to large graphs. Another drawback of this type algorithms is that they lack of theoretical explanation for the algorithm why it can improve diversity in the ranking results. Some improvements of this point have been achieved in the second framework [4]. In [4], the authors proposed a diversified ranking algorithm, namely DivRank, based on a vertex reinforced random walk. They also presented an optimization explanation for DivRank that it can achieve diversity in ranking. However, their optimization explanation is only suitable for undirected graphs, it cannot be used in directed graphs. In addition, the convergence property of DivRank is not clear, because it resorts to some approximation strategies to the original vertex reinforced random walk. Another drawback of DivRank is that it cannot be

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.126

IEEE computer society

scalable to large graphs. The reason is twofold. On the one hand, the DivRank dynamically update the transition matrix at each iteration. This procedure will result in a full transition matrix, thus it cannot be stored in main memory when the graph is very large. On the other hand, the full transition matrix will increase the computational cost for the matrix-vector multiplication.

In this paper, we propose a novel diversified ranking framework on graphs. The basic idea of our framework is that we first calculate the Pagerank of the vertices of the graph, and then perform a carefully designed vertex selection algorithm to find the top-K diversified ranking list based on the predefined diversified ranking measure. The key challenges in our framework are (1) how to define a reasonable diversified ranking measure that can capture both relevance and diversity, and (2) how to develop an efficient vertex selection algorithm that can optimize the diversified ranking measure. To this end, firstly, we propose a modified definition of *expansion* on graph to capture the diversity of the nodes. The key intuition is that if the nodes have large *expansion*, then the nodes will be dissimilar to each other, thus leading to diversity. Secondly, based on this definition, we propose a novel diversified ranking measure by combining relevance and diversity. Thirdly, we show that the proposed measure is a nondecreasing submodular set function. Based on the submodularity of the proposed measure, we design an efficient greedy algorithm with linear time and space complexity w.r.t. the size of the graph to find the top-K diversified ranking list. Finally, we compare our proposed method with four existing algorithms on four real networks. The experimental results show that the proposed approach outperforms the existing methods in terms of enhancing diversity in ranking and scalability of the algorithms.

Further related work: Diversity has been recognized as an important criteria in information retrieval. There are a large body of work on query or search results diversification [6][7]. In document retrieval, one of a well-known method is maximal marginal relevance (MMR) proposed by Carbonell and Goldstein [6], which achieves diversity by maximizing a linear combination function that captures both dissimilarity among the results and relevance w.r.t. the query. After Carbonell and Goldstein's work, many approaches to results diversification have been proposed in recent years. Agrawal, etc. [8] formulate the query results diversification as a submodular function maximization problem. Gollapudi, etc. [9] present several axioms for query results diversification. All the above mentioned methods primarily address to documents data. For an excellent survey on query results diversification, we point out to [7].

II. Proposed method

There exist many ranking algorithms [3][4][5] that aiming at improving diversity. However, as our analysis in the introduction, none of them can be scalable to large-scale graphs. To this end, in this section, we propose a scalable diversified ranking algorithm on graphs which captures both relevance and diversity. Here the relevance denotes that there are a certain number of percentage of nodes in the top-K ranking list produced by the diversified ranking algorithm should appear in the top-K ranking list yielded by the Pagerank algorithm. We first give some important notations and definitions, and then describe our proposed algorithms.

A. Notations and definitions

Consider a graph $G = (V, E)$, with a set of nodes V and a set of edges E, where the size of nodes is $n = |V|$. We now give some useful definitions as follows.

Definition 2.1: Let S be a set of nodes, then the *expanded set* of S is $N(S)$ such that $N(S) = S \cup \{v \in (V-S) | \exists u \in S, (u, v) \in E\}$, where the symbol "$-$" in $V-S$ denotes the set minus operator.

Definition 2.2: The *expansion* of set S is defined as $|N(S)|$, where $N(S)$ is defined in Def. 2.1 and $|N(S)|$ denotes the cardinality of $N(S)$. And, the *expansion ratio* is defined as $\sigma = |N(S)|/n$.

It is worth mentioning that our definition of expansion is based on the topological structure of the graph and we do not care about the weights of edges in graph. Also, our definition is suitable for any graphs with or without direction. Intuitively, the larger expansion ratio of a set of nodes implies the nodes more scatter in the graph, and thus resulting in better diversity. Based on this key intuition, we propose a diversified ranking measure and an efficient greedy algorithm in the following subsections.

B. Problem formulation

The most commonly used criteria of combining relevance and diversity is the so-called maximum marginal relevance (MMR) [6], which is a linear combination of relevance and diversity and is wildly used in many document retrieval systems. In MMR, a document has high marginal relevance means that it is both relevant to the query and also dissimilarity to the previously selected documents. Similarly, in graph, nodes with high diversified rank should (1) have high Pagerank scores, and (2) be dissimilar to the other selected nodes. Our definition of expansion ratio can be deemed as a diversity measure. The rationale is that the node that results in high expansion ratio will be far away from the selected nodes to some extent, thus it is dissimilar to the selected nodes. Therefore, we aim to find a subset S with K nodes such that the nodes in S has high Pagerank scores as well as the expansion ratio $|N(S)|/n$ is maximum. Formally, our goal is to maximize the following diversified ranking measures.

$$F(S) = \sum_{u \in S} w_u + \lambda \frac{|N(S)|}{n} \tag{1}$$

where w_u denotes the Pagerank score of node u, and $\lambda \in [0, 1]$ is a parameter that is used to tradeoff relevance and diversity. The first term in Eq. (1) is the sum of the Pagerank scores over the ranking results, which reflects the relevance of the ranking results. However, the second term is the expansion ratio of the ranking results. As discussed before, better expansion ratio implies better diversity. Hence, Eq. (1) captures both relevance and diversity.

Note that Eq. (1) does not take into account the ordering of the top-K ranking list. This is because our definition is based on a mild assumption that the users generally focus on all K results. And this assumption is typically reasonable in many practical applications [3][4][5]. However, in Section II-E, we will show that our proposed algorithm still yields an ordering results based on both relevance and diversity score of the node.

To summarize, our problem of finding top-K diversified ranking on graph can be formalized as follows

$$\begin{array}{ll} \arg\max\limits_{S\subseteq V} & F(S) \\ s.t. & |S| = K \end{array} \tag{2}$$

C. Submodularity

Many existing diversified ranking measures exhibit submoularity [8], resulting in an efficient greedy algorithm with $1 - 1/e$ approximation ratio to maximize it. Here we prove that our proposed measure ($F(S)$) is a nondecreasing submodular set function. We give the definition of the nondecreasing submodular set function [10] as follows.

Definition 2.3: Let V be a finite set, a real valued function $f(S)$ on the set of subsets of V is called a nondecreasing submodular set function if the following condition holds.

Condition: For any subsets S and T such that $S \subseteq T \subseteq V$ and node $j \in V$, we have $\rho_j(S) \geq \rho_j(T) \geq 0$. where $\rho_j(S) = f(S \cup \{j\}) - f(S)$.

The following theorem show that Eq. (1) is a nondecreasing submodular function with $F(\phi) = 0$. Due to space limits, all the proofs in this paper are omitted.

Theorem 2.1: *The set function $F(S)$ defined in Eq. (1) is a nondecreasing submodular function with $F(\phi) = 0$.*

D. Generalizations of diversified ranking measure

The proposed diversified ranking measure ($F(S)$) only considers the neighborhood information of set S. Naturally, we can generalize $F(S)$ by taking the k-step nearest neighborhood into account, thus resulting in our generalized diversified ranking measure, which is denoted by $F_k(S)$. For convenience, we first give the definitions of k-step expanded set and k-step expansion as follows.

Definition 2.4: $N_k(S)$ is a *k-step expanded set* of S such that $N_k(S) = S \cup \{v \in (V - S) |\ \exists u \in S, d(u, v) \leq k\}$, where $d(u, v)$ denotes the length of the shortest path (the minimum number of edges) from u to v.

Definition 2.5: The *k-step expansion* of S is defined as $|N_k(S)|$. And the *k-step expansion ratio* is defined as $\sigma_k = |N_k(S)|/n$.

Given the definition of k-step expansion, the generalized diversified ranking measure $F_k(S)$ is defined as follows.

$$F_k(S) = \sum_{u \in S} w_u + \lambda \frac{|N_k(S)|}{n} \tag{3}$$

Obviously, $F(S)$ is a special case of $F_k(S)$ when $k = 1$. Like $F(S)$, $F_k(S)$ is also a nondecreasing submodular function. Similarly, we have the following theorem.

Theorem 2.2: *The set function $F_k(S)$ defined in Eq. (3) is a nondecreasing submodular function with $F_k(\phi) = 0$.*

E. The greedy algorithm

As proved before, the top-K diversified ranking problem aims to maximize a nondecreasing submodular function with a cardinality constraint. Unfortunately, maximizing a nondecreasing submodular function with cardinality constraint has shown to be NP-hard [10]. Consequently, there is no hope to optimally solve the top-K diversified ranking problem in polynomial time unless P=NP. Nevertheless, there exists a greedy algorithm that can approximately solve the submodular function maximization problem in polynomial time [10]. Since $F(S)$ and $F_k(S)$ have shown to be submodular, the top-K diversified ranking problem can be approximately solved by an efficient greedy algorithm. In the following, we mainly focus on $F(S)$ and similar techniques can be easily generalized to $F_k(S)$. We present our greedy algorithm for the diversified ranking problem in Alg. 1.

Algorithm 1: Greedy algorithm for top-K diversified ranking problem

```
Input: Graph $G = (V, E)$, $K$, damping factor $\alpha$, and
       adjacency matrix $A$
Output: A set of $K$ nodes
Calculate the Pagerank vector $w$
$S = \phi$
while $|S| \leq K$ do
    Find $u_{\max} = \arg\max_{u \in (V-S)} w_u + \frac{\lambda}{n}|N(\{a\}) - N(S)|$
    $S = S \cup \{u_{\max}\}$
end while
return $S$
```

In Alg. 1, we first compute the Pagerank vector as the initial ranking, which measures the relevance of the nodes. Then the algorithm chooses one node with maximum $\rho_u(S) = w_u + \frac{\lambda}{n}|N(\{a\}) - N(S)|$ at one step, and this procedure will repeat K times. As mentioned before, our proposed algorithm will naturally produce an ordering ranking list according to $\rho_u(S)$. Since $\rho_u(S)$ satisfies the nondecreasing properties, Alg. 1 is indeed a reasonable ranking procedure that the node with high ranking score will appear in the top ranking list.

Theoretically, we have the following theorem.

Theorem 2.3: *The greedy algorithm (Alg. 1) is a $1 - 1/e$-approximation algorithm for the top-K diversified ranking problem (Eq. (2)).*

It is worth mentioning that the $1 - 1/e$-approximation factor is tight [11]. In other words, there are no other polynomial-time algorithms that can get a more tight approximation factor unless P=NP.

Complexity analysis of the greedy algorithm: The time complexity of Alg. 1 is $O(|E| + K|V|)$. In the line 1 of Alg. 1, it will take $O(|E|)$ time to compute the Pagerank vector. The time consuming from line 3 to line 5 is $O(|E| + K|V|)$. To this objective, we can create a list with size $|V|$ that dynamically maintains $\rho_u(S)$ for each node u. The time requirement for creating the list is $O(|E|)$ and the dynamic update procedure takes $O(|V|)$ time. For the space complexity, our algorithm takes $O(|V| + |E|)$. In terms of this analysis, the proposed algorithm has linear time and space complexity w.r.t. the size of the graph. Therefore, it can be scalable to large scale graphs. It is important to note that the time complexity of maximizing $F_k(S)$ is not linear when $k > 1$. A naive implementation of maximizing $F_k(S)$ is that we can first construct a new graph such that any two nodes u and v of the new graph have a edge (u, v) if u can reach v in k ($k > 1$) steps in the original graph, and then perform Alg. 1 on it. The construction of the new graph can be implemented by Floyd algorithm [12], resulting in $O(|V|^3)$ time complexity. And performing Alg. 1 on the new graph will take $O(|V| + |E|')$ time complexity, here $|E|'$ denotes the number of edges of the new graph. Hence, the time complexity of this naive algorithm is $O(|V|^3)$. However, in practice, many social networks are extremely sparse, thus the computational cost of this naive algorithm is acceptable. Our experiments show that the algorithm of maximizing $F_k(S)$ with 2-step expansion is very efficient.

III. Experiments

In this section, we evaluate the effectiveness and efficiency of the proposed approach.

Datasets: We conduct our experiments on four diverse real networks. (1) Collaboration networks. We select three collaboration networks from Stanford network collections [13]: GrQc, HepTh, and CondMat. GrQc, HepTh, and CondMat are collaboration networks originally collected from the e-print arXiv archive. (2) The social network of Flickr: in our experiments, we employ the Flickr dataset from ASU social computing data repository [14]. The detailed statistical information of our datasets are presented in Table I.

Table I
Summary of the datasets

name	nodes	edges
GrQc	5242	28,980
HepTh	9,877	51,971
CondMat	23,133	186,936
Flickr	80,513	5,899,882

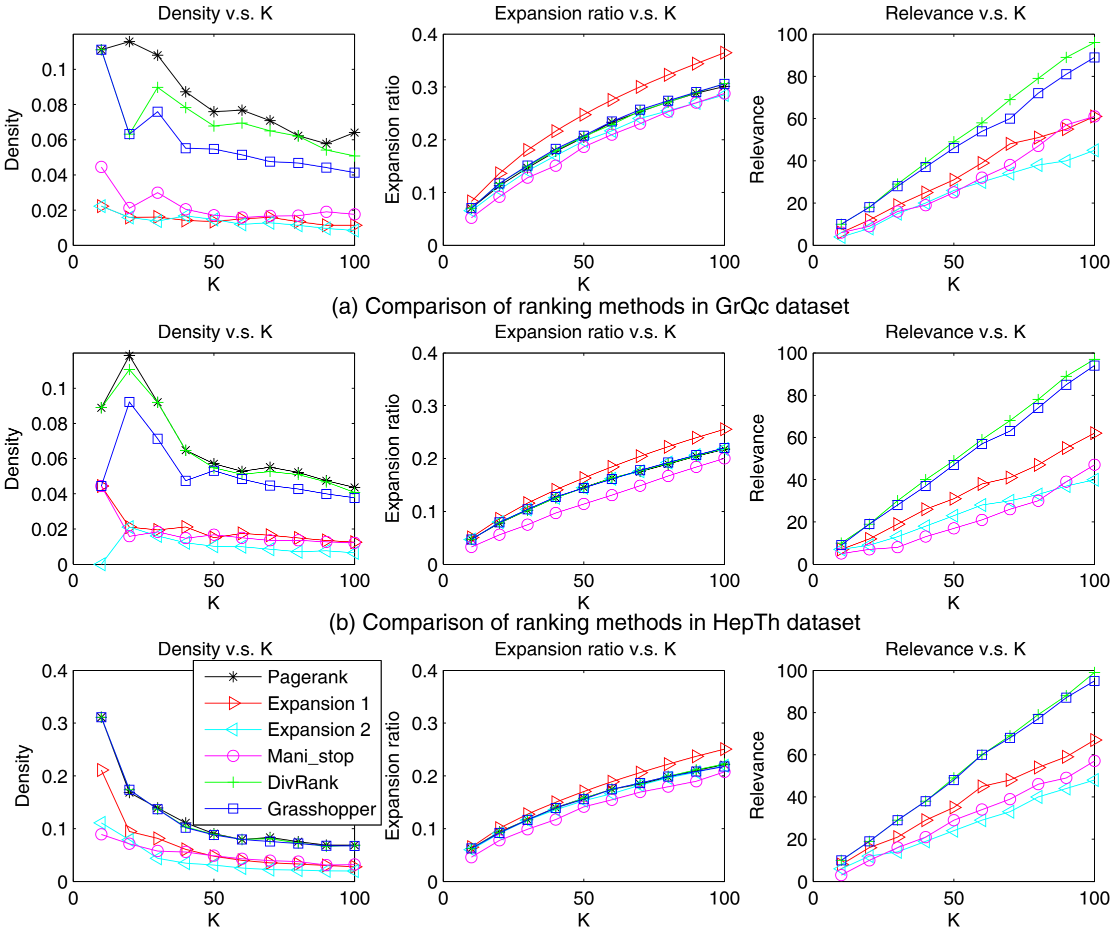
Evaluation metrics: In the literature, there is no wildly accepted measures for diversity [15]. In our experiments, we employ two metrics to measure the diversity. One is proposed in [4], which makes use of the density of the induced subgraph by the top-K ranked nodes. The density of a graph is a ratio that is equal to the number of edges appearing in the graph divided by the maximum possible number of edges in the graph. Intuitively, the density inversely measure the diversity of the top-K ranked nodes. The second metric is the expansion ratio which defined in Def. 2.2. The reason is because the larger expansion ratio of the top-K ranked nodes indicates the nodes more scatter in the graph, thus resulting in better diversity. For the relevance metric, we use the cardinality of the intersection between the top-K ranking list yielded by the diversified ranking algorithms and the top-K ranking list produced by the Pagerank algorithm. Specifically, the relevance Rel can be calculated as $Rel = |R_{div} \cap R_{pagerank}|$, where R_{div} denotes the top-K ranking list generated by the diversified ranking algorithm and $R_{pagerank}$ denotes the top-K ranking list yielded by the Pagerank algorithm.

Baselines: We compare our proposed algorithm with four baselines under diversity and relevance metrics defined above. The four baselines involves: (1) Pagerank: Pagerank is a natural competitor of our algorithm, which can be served as a baseline for evaluating relevance. (2) Grasshopper: Grasshopper is a diversified ranking algorithm that leverages an absorbing random walk to achieve diversity [3]. (3) Manifold ranking with stop points (Mani_stop): The Mani_stop algorithm was proposed in [5], which is very similar to the Grasshopper algorithm. (4) DivRank: The DivRank [4] makes use of the stationary distribution of a vertex reinforced random walk to rank nodes. It is worth mentioning that there are two various implementation of DivRank, namely pointwise DivRank and cumulative DivRank respectively. But as reported in the original paper [4], this two algorithms achieve similar ranking performance. Hence, we choose to implement the pointwise DivRank in our experiments.

Parameter settings and experimental environment: In our proposed algorithm, there are two parameters: the damping factor α for computing Pagerank, and the parameter λ to tradeoff relevance and diversity. We set $\alpha = 0.85$ as it is wildly used in web search. For parameter λ, we set it to 1. This is because the parameter λ is not very sensitive in our experiments. For all parameters of the baseline methods, we set them as the same as their original papers respectively. All the experiments are conducted on a Window Server 2007 with 4xDual-Core Intel Xeon 2.66 GHz CPU, and 4G memory. All algorithms are implemented by MATLAB 2009 and Visual C++ 6.0.

A. *Testing in collaboration networks*

In this experiment, we implement the proposed algorithms based on 1-step and 2-step expansion, namely Expansion 1 and Expansion 2 respectively, and compare them with four baselines described above over three collaboration networks. Notice that Expansion 1 aims to maximize Eq. (1), and Expansion 2 maximizes Eq. (3), where $k = 2$. The results are shown in Fig. 1.



(a) Comparison of ranking methods in GrQc dataset

(b) Comparison of ranking methods in HepTh dataset

(c) Comparison of ranking methods in CondMat dataset

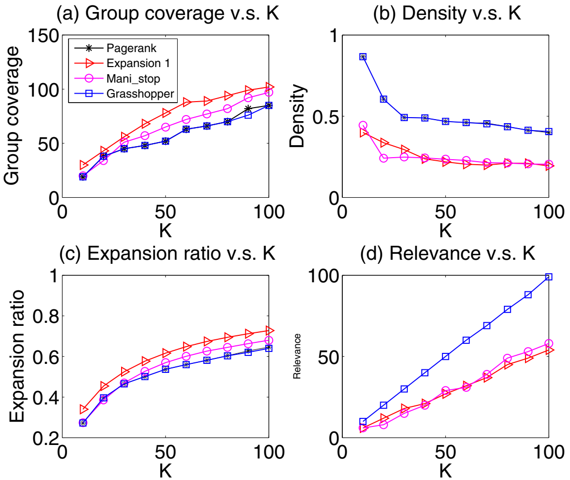
Figure 1. Comparison of various diversified ranking algorithms in collaboration networks (GrQc, HepTh, and CondMat). Here Expansion 1 denotes the 1-step expansion based algorithm, and Expansion 2 denotes the 2-step expansion based algorithm.

From Fig. 1, we can clearly see that the Expansion 2 achieves the best diversity under the density metric, followed by the Expansion 1, Mani_stop algorithm, and then Grasshopper, DivRank and Pagerank algorithm. Interestingly, there exists a significant gap between the Expansion 1 and Grasshopper algorithm, implying that the Grasshopper and DivRank algorithm exhibit poor performance to enhance diversity in collaboration networks. Instead, under the expansion ratio measure, Expansion 1 outperforms the competitors. And Expansion 2, Grasshopper, DivRank, and Pagerank achieve nearly equivalent diversity. Also it is worth noting that the Mani_stop algorithm even performs worse than the Pagerank algorithm. Hence, under the expansion measure, all algorithms except Expansion 1 are not very effective for boosting diversity. For the relevance, we can observe that the best algorithm is DivRank, followed by Grasshopper, Expansion 1, Mani_stop, and Expansion 2. In terms of this observation, improving diversity will reduce relevance if we measure the relevance by Pagerank. Hence, in practice, we should seek a reasonable tradeoff between relevance and diversity. In this experiment, we find that Expansion 1 can achieve this end, as it considerably improves diversity but do not significantly sacrifice the performance of relevance.

B. Testing in Flickr social network

Now we test our proposed algorithm in Flickr dataset. Our goal is to find the top-K users that not only have higher Pagerank scores, but also cover as many interest groups as possible. Hence, in addition to the diversity measures described above, we also introduce the “group coverage” as a new diversity measure in this experiment. Intuitively, the more groups are covered by the ranking list the more diversity it has. In this experiment, the DivRank algorithm cannot be conducted because its quadratic space complexity. In addition, we omit the results of the Expansion 2 as its performance is not better than Expansion 1. Our results are shown in Fig. 2.

From Fig. 2, we can observe that the Expansion 1 achieves the best diversity under the group coverage and expansion ratio measures, followed by the Mani_stop algorithm, and then Grasshopper and Pagerank. However, under the density measure, the proposed algorithm obtain the best performance when $K \geq 40$. When $K < 40$, the Mani_stop algorithm achieves the best diversity. It is wroth noting that the Grasshopper algorithm gets the same diversity as the Pagerank algorithm, indicating that the Grasshopper algorithm are not effective for improving diversity in ranking on Flickr social network. For the relevance measure, we find that the Grasshopper algorithm achieves the best relevance. The proposed algorithm and Mani_stop algorithm achieve comparable relevance. This results consist with the observations of the former experiments that enhancing diversity could diminish relevance.


(a) Group coverage v.s. K

(b) Density v.s. K

(c) Expansion ratio v.s. K

(d) Relevance v.s. K

Figure 2. Comparison of ranking algorithms in Flickr dataset.

To summarize, the experimental results suggest that there is no free lunch to diversify ranking on graphs. A reasonable tradeoff between diversity and relevance is more necessary in practice. The Expansion 1 achieves a good balance between diversity and relevance than the existing methods. Furthermore, the Expansion 1 is easy to implement and its time and space complexity is linear w.r.t. the size of the graph. Consequently, it can be used for diversified ranking on large graphs.

C. Time comparison

We compare the running time of various diversified ranking methods over four different network datasets. We take the average on the running time of the ranking algorithm over different K. Table II shows our results. From Table II, we can observe that Pagerank is the best efficient algorithm. Expansion 1 achieves competitive efficiency with the Pagerank algorithm. This is because Expansion 1 has linear time complexity w.r.t. the size of the graph. Besides, Expansion 2 is also very efficient on the medium-sized graphs, as this graphs are extremely sparse. Rather, the efficiency of DivRank, Mani_stop, and Grasshopper are quite poor. This results confirm our time and space complexity analysis in the previous sections.

Table II
COMPARISON ON AVERAGE RUNNING TIME (IN SECOND) OF VARIOUS RANKING ALGORITHMS.

	GrQc	HepTh	CondMat	Flickr
Pagerank	0.02	0.056	0.17	2.71
Expansion 1	0.06	0.12	0.39	10.30
Expansion 2	0.10	0.28	1.34	–
DivRank	52.66	228.49	1426.8	–
Mani_stop	215.57	933.55	4675.30	51824
Grasshopper	227.51	1027.30	3418.7	52777

ACKNOWLEDGMENTS

The work was supported by grants of the Research Grants Council of the Hong Kong SAR, China No. 419008 and 419109.

REFERENCES

[1] S. Brin and L. Page, "Pagerank: Bringing order to the web," Stanford Digital Library Project, Tech. Rep., 1997.

[2] T. H. Haveliwala, "Topic-sensitive pagerank," in *WWW '02*.

[3] X. Zhu, A. B. Goldberg, J. V. Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *HLT-NAACL'07*.

[4] Q. Mei, J. Guo, and D. R. Radev, "Divrank: the interplay of prestige and diversity in information networks," in *KDD '10*.

[5] X. Zhu, J. Guo, X. Cheng, P. Du, and H. Shen, "A unified framework for recommending diverse and relevant queries," in *WWW '11*.

[6] J. G. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *SIGIR '98*.

[7] M. Drosou and E. Pitoura, "Search result diversification," *SIGMOD Rec.*, vol. 39, pp. 41–47, 2010.

[8] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *WSDM '09*.

[9] S. Gollapudi and A. Sharma, "An axiomatic approach for result diversification," in *WWW '09*.

[10] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions-i," *Mathematical Programming*, vol. 14, pp. 265–294, 1978.

[11] U. Feige, "A threshold of ln n for approximating set cover," *J. ACM*, vol. 45, pp. 634–652, 1998.

[12] T. H. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*. MIT Press, 2011.

[13] J. Leskovec, "Standford network analysis project," 2010. [Online]. Available: http://snap.standford.edu

[14] R. Zafarani and H. Liu, "Social computing data repository at ASU," 2009. [Online]. Available: http://socialcomputing.asu.edu

[15] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," *SIGIR Forum*, vol. 43, 2009.