The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Ranking Users in Social Networks with Higher-Order Structures

Huan Zhao,[*1] **Xiaogang Xu,**[†1] **Yangqiu Song,**[*] **Dik Lun Lee,**[*] **Zhao Chen,**[‡] **Han Gao**[‡]
[*]Department of Computer Science & Engineering, Hong Kong University of Science and Technology, Hong Kong
[†]College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China
[‡]Tencent Technology (SZ) Co., Ltd., China
[*]{hzhaoaf,yqsong,dlee}@cse.ust.hk [†]xiaogangxu@zju.edu.cn [‡]{gilbertchen,alangao}@tencent.com

Abstract

PageRank has been widely used to measure the authority or the influence of a user in social networks. However, conventional PageRank only makes use of edge-based relations, ignoring higher-order structures captured by motifs, subgraphs consisting of a small number of nodes in complex networks. In this paper, we propose a novel framework, motif-based PageRank (MPR), to incorporate higher-order structures into conventional PageRank computation. We conduct extensive experiments in three real-world networks, i.e., DBLP, Epinions, and Ciao, to show that MPR can significantly improve the effectiveness of PageRank for ranking users in social networks. In addition to numerical results, we also provide detailed analysis for MPR to show how and why incorporating higher-order information works better than PageRank in ranking users in social networks.[1]

Introduction

Online social networks have become the everyday communication and interaction platform for most people. User ranking in social networks is a general problem for opinion leader mining (Song et al. 2007), influence analysis (Tang et al. 2009; Xiang et al. 2013) and social trustworthiness (Wang et al. 2015). Besides customized features such as content and topic (Tang et al. 2009), PageRank (Page et al. 1999) can be considered as a general algorithm for ranking users in social networks. PageRank measures the authority of nodes in the network, which can be used as a measure for opionion leader mining, and influence and trustworthiness analysis (Song et al. 2007; Tang et al. 2009; Xiang et al. 2013; Wang et al. 2015). However, PageRank has several limitations in social network analysis. In a network, there could be some important *structures* that can affect the influence or trustworthiness of the nodes. For example, when mining opinion leaders in a social network, it is obvious that an authority node, which many other users connect to, should be given high influence score. However, there could be a node that connects to multiple such authority nodes but itself is not an authority node. For example, in a social network a journalist could have connections to many VIPs but he/she may not be as famous as the VIPs. Then when the journalist expresses an opionion, it could be read by the VIPs, who would propagate the opinion much farther in the future. Moreover, when analyzing a citation network, even if a paper $\mathcal{P}$ is highly cited by many other papers, we are not sure how $\mathcal{P}$'s influence might be affected if the citing papers have mutual citations among themselves.

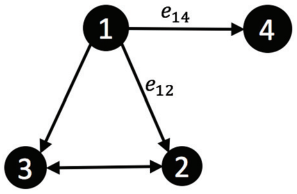


Figure 1: A graph with four nodes.

For trustworthiness, we use a concrete example to show why higher-order structures matter. As shown in Figure 1, an edge e_{ij} means user i follows user j. In traditional PageRank, the initial weights of e_{12} and e_{14} will be set to the same value. However, it can be observed that user 1 trusts user 2 more than user 4 because user 1 also follows user 3, who is a friend of user 2 (user 3 and user 2 follow each other). This indicates that user 1 has a closer relation to user 2 than user 4. Thus, it is unreasonable to set the weights of e_{12} and e_{14} to the same value due to the existence of the triangular structure where user 1 and user 2 both appear. The premise of this paper is that it is important in social network mining to consider higher-order structures involving multiple nodes.

In this paper, we propose a novel framework, called motif-based PageRank (MPR), to incorporate such higher-order structures into PageRank computation for user ranking in networks. The higher-order structures can be represented as *network motifs* (or subgraphs or graphlets) (Milo et al. 2002; Benson, Gleich, and Leskovec 2016a). We show some typical 3-node motifs in Figure 2, where M_6 characterizes the triangular structure in Figure 1. In this work, we prove that motif-based and traditional edge-based relations are complementary to each other in computing the authority of a node in complex networks. We propose to first compute the motif-based adjacency matrix, which captures the pairwise relations between two nodes appearing in a specific motif.

[1] Equal contribution. This work was done when Xiaogang Xu was an intern at HKUST.

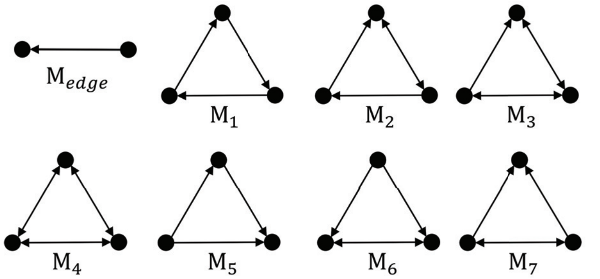


Figure 2: All triangular motifs in a directed unweighted graph. Note that we use M_{edge} to represent the edge-based relation.

Then we design a method to combine the motif-based adjacency matrix with the edge-based adjacency matrix. In this way, we re-weigh the links based on motifs in complex networks. In other words, we incorporate higher-order relations into conventional authority computation. We conduct experiments on an academic citation dataset, DBLP, and two trustworthiness datasets, Epinions and Ciao, to extract influential or trustworthy users. The results show that our proposed method significantly outperforms conventional PageRank and other baselines. Moreover, we give detailed analysis on different motifs, providing insights into authority computation with higher-order relations. The code of this work is available at https://github.com/HKUST-KnowComp/Motif-based-PageRank.

The rest of the paper is organized as follows. We first review the related work on PageRank and motif analysis in graphs. Then we introduce in detail MPR. Further we present our experimental results as well as the analysis. Finally, we conclude our work and discuss some future directions.

Background and Related Work

In this section, we introduce related work on authority computation that uses PageRank and motif in complex networks.

PageRank

PageRank was first introduced to rank Web pages on the Internet (Page et al. 1999). Apart from ranking Web pages, PageRank has been used in many other domains (Gleich 2015), such as citation network analysis (Ding 2011) and link prediction (Liben-Nowell and Kleinberg 2007). In (Xiang et al. 2013), Xiang et.al. explicitly connected PageRank with social influence model and showed that authority is equivalent to influence under their framework. Thus, PageRank can also help to select influential nodes in networks. Moreover, PageRank has been used to identify opinion leaders (Song et al. 2007) and find trustworthy users (Wang et al. 2015) in social networks. Compared to our work, all of the previous studies only considered direct edges only in PageRank computation and ignored higher-order structures among multiple nodes.

Motif in Complex Networks

Motif characterizes higher-order network structures and is also associated with other names such as graphlets or subgraphs. Network motif was first introduced in (Milo et al. 2002). It has been shown to be useful in many applications such as social networks (Ugander, Backstrom, and Kleinberg 2013; Granovetter 1973; Rotabi et al. 2017), scholar networks (Wang, Lü, and Yu 2014), biology (Pržulj 2007), neuroscience (Sporns and Kötter 2004), and temporal networks (Paranjape, Benson, and Leskovec 2017). Besides, most of the previous work focused on how to efficiently count the number of motifs in complex networks (Ahmed et al. 2015; Jha, Seshadhri, and Pinar 2015; Wang et al. 2016; Han and Sethu 2016; Stefani et al. 2017; Pinar, Seshadhri, and Vishal 2017). Recently, it was proven that motifs can also be used for graph clustering or community detection (Benson, Gleich, and Leskovec 2016a; Yin et al. 2017). In (Wang, Lü, and Yu 2014), Wang et.al. proposed to measure the importance of a node in a network by its participation in different motifs. In (Zhang et al. 2017), Zhang et.al. proposed to predict users' behaviors based on structural influence, i.e., the influence from a specific structure he/she appears. Compared to these previous studies, we incorporate motif to explore the higher-order relations in pairs of nodes and then use the relations to compute the authority of nodes using PageRank. In other words, we consider motifs from a global perspective while previous works only consider local motif structures of the nodes.

Motif-based PageRank

In this section, we introduce in detail our framework and algorithm.

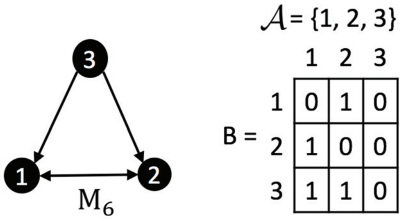
Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V} = \{v_i | i = 1, , ..., n\}$ is the node set, and $\mathcal{E} = \{e_{ij} | i, j = 1, ..., n\}$ is the edge set, where e_{ij} is an edge from v_i to v_j. $\mathbf{W}$ is the adjacency matrix, where $\mathbf{W}_{ij}$ represents the weight of e_{ij}. For a directed unweighted graph, $\mathbf{W}_{ij} = 1$ if e_{ij} exists and $\mathbf{W}_{ij} = 0$ otherwise. Then we normalize the adjacency matrix to obtain the transition probability matrix $\mathbf{P}$, where $\mathbf{P}_{ij} = \mathbf{W}_{ij} / \sum_i \mathbf{W}_{ij}$. Then, the PageRank over the graph can be defined as follows.

$$\mathbf{x} = d\mathbf{P}^T\mathbf{x} + \frac{1-d}{N}\mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{x}_i$ is the PageRank value of the i-th node in $\mathcal{G}$, $\mathbf{e} \in \mathbb{R}^N$ is a vector with every entry equal to 1, and $d \in (0, 1)$ is a damping factor. In (Bianchini, Gori, and Scarselli 2005), Bianchini et.al. proved that this iterative computation always converges.

Given a social network $\mathcal{G}$, we can generalize this definition as follows. If there is an edge e_{ij} from node v_i to v_j, then we use $\mathbf{W}_{ij}$ to represent the strength of endorsement v_i gives to v_j or the strength of influence of v_j exerts on v_i. The weight can be computed based on binary link relation, the interaction frequency, content similarity of users' posts, etc. However, all the above weights are still first-order proximity between two users. As shown in Figure 1, for user 1,



$\mathcal{A} = \{1, 2, 3\}$

$$B = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & 1 & 0 \end{array}$$

Figure 3: Motif example. This is the example of M_6. $\mathbf{B}$ is the binary matrix recording the edge pattern for M_6 in the left side. $\mathcal{A}$ is the anchor set, including all nodes in M_6. Therefore, it is a simple motif.

$\mathbf{W}_{12}$ should be larger than $\mathbf{W}_{14}$ because node 1 and node 2 participate in a triangular relation in trustworthiness evaluation. Therefore, it is desirable if the weight can also incorporate information about such higher-order structures.

In the rest of this section, we will first introduce the formal definition of motif to characterize higher-order structures, and then introduce our algorithm to incorporate such information into PageRank.

Motif Definitions

We first introduce the definition of motif as follows.

Definition 1. Network Motif. A motif M is defined on k nodes by a tuple $(\mathbf{B}, \mathcal{A})$, where $\mathbf{B}$ is a $k \times k$ binary matrix and $\mathcal{A} \subset \{1, 2, ..., k\}$ is a set of *anchor nodes*.

Anchor nodes represent the nodes we are interested in. Usually, anchor nodes are all of the k nodes. In this case it is called simple motif. Otherwise, it is called anchored motif (Benson, Gleich, and Leskovec 2016b). In this paper, we focus on simple motifs. An example is given in Figure 3 to illustrate the definition of motif.

Given a motif definition, we can define the set of motif instances as follows.

Definition 2. Motif Set. The motif set in an unweighted directed graph $\mathcal{G}$ with an adjacency matrix $\mathbf{W}$, denoted as $\mathcal{M}(\mathbf{B}, \mathcal{A})$, is defined by

$$\mathcal{M}(\mathbf{B}, \mathcal{A}) = \{(set(\mathbf{v}), set(\chi_{\mathcal{A}}(\mathbf{v})))|\mathbf{v} \in V^k, \\ v_1, ..., v_k, \text{distinct}, \mathbf{W}_{\mathbf{v}} = \mathbf{B}\}.$$

where $\chi_{\mathcal{A}}$ is a selection function that takes the subset of a k-tuple indexed by $\mathcal{A}$, and set$(\cdot)$ is an operator that transforms an ordered tuple to an unordered set, $set((v_1, v_2, ..., v_k)) = \{v_1, v_2, ..., v_k\}$. $\mathbf{v}$ is an ordered vector representing the k nodes, and $\mathbf{W}_{\mathbf{v}}$ is the $k \times k$ adjacency matrix of the subgraph induced by $\mathbf{v}$.

The set operator is used to avoid duplicates when $\mathcal{M}(\mathbf{B}, \mathcal{A})$ is defined for motifs exhibiting symmetries. Therefore, we will just use $(\mathbf{v}, \chi_{\mathcal{A}}(\mathbf{v}))$ to denote $(set(\mathbf{v}), set(\chi_{\mathcal{A}}(\mathbf{v}))$ when we discuss elements of $\mathcal{M}(\mathbf{B}, \mathcal{A})$. When $\mathbf{B}$ and $\mathcal{A}$ are arbitrary or clear from context, we will simply denote the motif set by $\mathcal{M}$. Then any $(\mathbf{v}, \chi_{\mathcal{A}}(\mathbf{v})) \in \mathcal{M}$ is called a *motif instance*.

Motif-based Adjacency Matrix

In this work, we propose to use motif to capture higher-order relations between nodes in a graph. When given a motif set $\mathcal{M}$, we use the co-occurrence of two nodes in $\mathcal{M}$ to capture this relation. Given a motif $\mathcal{M}$, the definition of the motif-based adjacency matrix or co-occurrence matrix is defined by:

$$(\mathbf{W}_M)_{ij} = \sum_{(\mathbf{v}, \chi_{\mathcal{A}}(\mathbf{v})) \in \mathcal{M}} \mathbf{1}(\{i, j\} \subset \chi_{\mathcal{A}}(\mathbf{v})), \quad (2)$$

where $i \neq j$, and $\mathbf{1}(s)$ is the truth-value indicator function, i.e., $\mathbf{1}(s) = 1$ if the statement s is true and 0 otherwise. Note that the weight is added to $(\mathbf{W}_M)_{ij}$ only if i and j appear in the anchor set. For simple motifs, it requires i and j to be members of set($\mathbf{v}$).

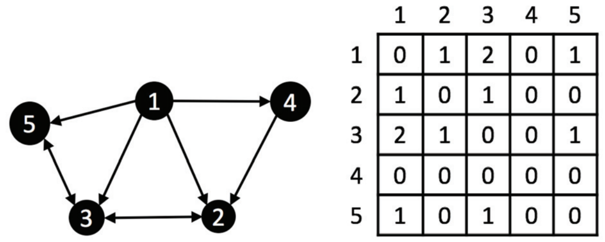
The motif-based adjacency matrix represents the frequency of two nodes appearing in a given motif, i.e., one type of higher-order structure. The larger $(\mathbf{W}_M)_{ij}$ is, the more significant the relation between i and j is within the motif. Then given a motif M_k, if we want to capture the high-order relations, we need to construct the motif-based adjacency matrix $\mathbf{W}_{M_k}$. The procedure is related to subgraph counting in larger graphs, which has been extensively explored in the literature (Ahmed et al. 2015; Jha, Seshadhri, and Pinar 2015; Wang et al. 2016; Han and Sethu 2016; Stefani et al. 2017; Pinar, Seshadhri, and Vishal 2017). In this paper, we focus only on triangular motifs because of triadic closure in social networks (Simmel 1908). We show that it can be computed based on simple matrix computation (Benson, Gleich, and Leskovec 2016b).

Let $\mathbf{W}$ be the adjacency matrix for $\mathcal{G}$, and let $\mathbf{U}$ and $\mathbf{B}$, respectively, be the adjacency matrix of the unidirectional and bidirectional links of $\mathcal{G}$. Here we focus on unweighted graphs where elements in $\mathbf{W}$ are either ones or zeros. For example, in Figure 4, e_{23} is a bidirectional edge while e_{12} is unidirectional. Then we have $\mathbf{B} = \mathbf{W} \odot \mathbf{W}^T$ and $\mathbf{U} = \mathbf{W} - \mathbf{B}$, where $\odot$ denotes the Hadamard (entry-wise) product. Note that $\mathbf{B}$ is a binary matrix representing the existence of bidirectional edges between two nodes in a directed graph. The computation of adjacency matrices based on all seven motifs is summarized in Table 1.

We use an example to illustrate the computing process for M_6 (shown in Figure 3). Taking two arbitrary nodes, v_i and v_j intermediated with v_k, there are six different cases for them to participate in M_6 because the graph is directed. We use $1, 2, 3$ or $3, 2, 1$ to denote the positions of three users v_i, v_k, v_j shown in M_6. We have six cases for the three nodes, i.e., $\{(3, 1, 2), (2, 1, 3)\}, \{(1, 2, 3), (3, 2, 1)\}$, $\{(1, 3, 2), (2, 3, 1)\}$, where we put them in three groups according to the intermediate node v_k. As shown in Figure 3, e_{12} is a bidirectional edge, while e_{13} and e_{23} are unidirectional edges. To compute the frequency of v_i and v_j participating M_6, we need to add up all their frequencies in the six cases. In $(3, 1, 2)$, where v_i is in position 3 and v_j is in position 2, the frequency can be obtained by $(\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{U}$, where $\mathbf{U} \cdot \mathbf{B}$ is for path $3 - 1 - 2$ without edge e_{32}, and $\odot\mathbf{U}$ just complements the motif with edge e_{32}. In this way, we get the frequency of v_i and v_j appearing together in a specific pattern $(3, 1, 2)$. Similarly, we can obtain the other cases. Note

Table 1: Computation of motif-based adjacency matrices for M_1 to M_7.

Motif	Matrix Computation	$\mathbf{W}_{M_i} =$
M_1	$\mathbf{C} = (\mathbf{U} \cdot \mathbf{U}) \odot \mathbf{U}^T$	$\mathbf{C} + \mathbf{C}^T$
M_2	$\mathbf{C} = (\mathbf{B} \cdot \mathbf{U}) \odot \mathbf{U}^T + (\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{U}^T + (\mathbf{U} \cdot \mathbf{U}) \odot \mathbf{B}$	$\mathbf{C} + \mathbf{C}^T$
M_3	$\mathbf{C} = (\mathbf{B} \cdot \mathbf{B}) \odot \mathbf{U} + (\mathbf{B} \cdot \mathbf{U}) \odot \mathbf{B} + (\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{B}$	$\mathbf{C} + \mathbf{C}^T$
M_4	$\mathbf{C} = (\mathbf{B} \cdot \mathbf{B}) \odot \mathbf{B}$	$\mathbf{C}$
M_5	$\mathbf{C} = (\mathbf{U} \cdot \mathbf{U}) \odot \mathbf{U} + (\mathbf{U} \cdot \mathbf{U}^T) \odot \mathbf{U} + (\mathbf{U}^T \cdot \mathbf{U}) \odot \mathbf{U}$	$\mathbf{C} + \mathbf{C}^T$
M_6	$\mathbf{C} = (\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{U} + (\mathbf{B} \cdot \mathbf{U}^T) \odot \mathbf{U}^T + (\mathbf{U}^T \cdot \mathbf{U}) \odot \mathbf{B}$	$\mathbf{C}$
M_7	$\mathbf{C} = (\mathbf{U}^T \cdot \mathbf{B}) \odot \mathbf{U}^T + (\mathbf{B} \cdot \mathbf{U}) \odot \mathbf{U} + (\mathbf{U} \cdot \mathbf{U}^T) \odot \mathbf{B}$	$\mathbf{C}$



	1	2	3	4	5
1	0	1	2	0	1
2	1	0	1	0	0
3	2	1	0	0	1
4	0	0	0	0	0
5	1	0	1	0	0

(a) An example graph with five nodes.

(b) Corresponding M_6-based adjacency matrix.

Figure 4: An example for computing M_6 based adjacency matrix according to the equation in Table 1. For example, $\mathbf{W}_{13} = 2$ because node 1 and 3 appear in two instances of M_6, i.e., $\{1, 2, 3\}$ and $\{1, 3, 5\}$.

that, due to the symmetric structure of $(3, 2, 1)$ and $(3, 1, 2)$, we only need to add up three equations as shown in Table 1. Figure 4 shows an example for the motif-based adjacency matrix based on motif M_6.

Computation Analysis. In all computation formulas of $\mathbf{W}_{M_i}$ in Table 1, the core computation kernel is $(\mathbf{X} \cdot \mathbf{Y}) \odot \mathbf{Z}$, which can be efficiently computed with sparse matrices. The computational cost is proportional to the numbers of columns and rows, as well as the number of non-zero elements in the sparse matrices.

Higher-Order PageRank

After computing the motif-based adjacency matrices, we incorporate them into the ranking model for ranking users. Since the non-zero elements in the motif-based adjacency matrices will be no more than the non-zero elements of the original edge-based adjacency matrix, we combine the motif-based relations with the edge-based relations instead of replacing the edge-based relations. In this way, motif-based edges can be regarded as supplementary to conventional authority computation with PageRank. We propose to use a linear combination to fuse tje edge-based and motif-based adjacency matrices. Specifically, for a given motif M_k, we generate a new matrix as follows.

$$\mathbf{H}_{M_k} = \alpha \cdot \mathbf{W} + (1 - \alpha) \cdot \mathbf{W}_{M_k}, \tag{3}$$

where $\alpha \in [0, 1]$ balances the original edge-based relations and higher-order relations provided by motifs. Then we can compute the transition probability matrix $(\mathbf{P}_{M_k})_{ij} = (\mathbf{H}_{M_k})_{ij} / \sum_i (\mathbf{H}_{M_k})_{ij}$ and substitute $\mathbf{P}_{M_k}$ for the transition probability matrix $\mathbf{P}$ in Eq. (1).

Table 2: Statistics of the three datasets: DBLP, Epinions, Ciao. The density is computed by $\frac{\#edges}{\#nodes \times (\#nodes - 1)}$.

	Nodes	Edges	Density(%)
DBLP	35,315	941,936	0.076
Epinions	18,089	355,217	0.109
Ciao	2,342	57,544	1.049

Experiments

In this section, we present the experimental results to demonstrate the effectiveness of MPR.

Datasets and Settings

Our experiments are conducted on three real-world networks. The first is a scholar network, DBLP, which is provided by ArnetMiner (Tang et al. 2008). The other two are trust networks, Epinions and Ciao, which are provided by (Tang, Gao, and Liu 2012; Tang et al. 2012). More information about the datasets is given below.

DBLP. We use the DBLP dataset (version V8) in the AMiner Website.[2] DBLP is an academic dataset which records the publications of authors. We can extract the citation networks from all publications. We focus on a sub-network from six research domains: “Artificial Intelligence,” “Computer Vision,” “Database,” “Data Mining,” “Information Retrieval,” and “Machine Learning.” After extracting authors and papers from these domains, the citation network is constructed to evaluate the social influence of authors. When constructing the citation network, we add one edge from i to j if author i cites at least one paper of author j.

Epinions and Ciao. They are two review websites where users can write reviews on products as well as rate the reviews of other users, indicated by the *review helpfulness rating*. Moreover, users can add other users as trustworthy users if they like their reviews. Then it is intuitive that the higher helpfulness rating obtained by a user’s reviews, the more likely it is for him to be added as a trustworthy user. When constructing the trust network, an edge e_{ij} will be added when user v_i trusts user v_j.

The statistics of the datasets are listed in Table 2.

[2] https://cn.aminer.org/billboard/citation

Table 3: NDCG for top10, top50, top500 users from DBLP, Epinions, and Ciao datasets. The best performance in each column is emphasized with boldface.

	DBLP			Epinions			Ciao		
TopK	10	50	500	10	50	500	10	50	500
IND	0.9879	0.9639	0.9400	0.9476	0.9563	0.9343	0.9218	0.8651	0.9120
BET	0.9796	0.9710	0.9559	0.9566	0.9559	0.9403	0.9421	0.8961	0.8911
CLO	0.9875	0.9614	0.9285	0.9308	0.9346	0.9382	0.9021	0.9225	0.9251
BPR	0.9464	0.9414	0.9527	0.9777	0.9543	0.9365	0.8332	0.8599	0.8932
WPR	0.9154	0.8871	0.9350	0.9777	0.9543	0.9365	0.8332	0.8599	0.8932
M_1	0.9753	0.9590	0.9623	0.9777	**0.9656**	0.9406	0.9802	0.9347	0.9392
M_2	0.9890	0.9424	0.9585	0.9777	0.9581	0.9417	**0.9905**	0.9453	0.9401
M_3	0.9895	0.9508	0.9586	0.9788	0.9568	0.9378	0.9768	0.9576	0.9441
M_4	0.9809	0.9477	0.9528	0.9827	0.9557	0.9395	0.9719	0.9357	0.9401
M_5	0.9877	0.9513	0.9574	0.9777	0.9551	0.9454	0.9792	**0.9792**	0.9401
M_6	0.9634	0.9525	0.9588	**0.9957**	0.9596	0.9382	0.9459	0.9459	**0.9427**
M_7	**0.9920**	**0.9766**	**0.9640**	0.9780	0.9614	**0.9442**	0.9514	0.9500	0.9418

Table 4: Results of Z-scores for M_1 to M_7 on DBLP, Epinions and Ciao datasets.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7
DBLP	8.83	190.78	368.29	125.15	300.67	689.21	348.71
Epinions	-6.82	39.09	162.74	30.553	145.47	167.94	182.01
Ciao	7.83	105.58	205.93	25.50	133.97	114.44	206.71

Evaluation Metrics. To evaluate the effectiveness of the proposed framework, we compare the quality of the topK users ranked by different algorithms. Specifically, we extract K users with the largest PageRank values and then compute the Normalized Discounted cumulative Gain (NDCG), which is a popular metric for ranking quality in Information Retrieval (IR) (Järvelin and Kekäläinen 2002). For topK results, DCG@K is defined as:

$$DCG_K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)}, \tag{4}$$

where rel_i represents the relevance score of a document for a given query. $\log_2(i+1)$ is used to penalize the algorithm if it ranks higher relevant items in lower positions. To normalized the results, normalized DCG_K is computed as:

$$NDCG_K = \frac{DCG_K}{IDCG_K}, \tag{5}$$

where $IDCG_K$ is the ideal ranking for the results, i.e. the results are sorted according to their relevance scores. In this way, it measures how good a ranking list is compared to an ideal one.

In our experiments, for the DBLP dataset, we use the H-index of authors as the relevance scores. In the research community, H-index is a metric to measure the influence of an author by considering both the quality and quantity of the author's published papers based on the citation network. The larger H-index an author has, the more influential s/he is. We crawled the H-index of all authors in our dataset from the AMiner website[3] before July 2017. For Epinions and Ciao, we use the average helpfulness rating of a user's reviews as the relevance score, which means that the larger it is, the more trustworthy the user is.

Baselines. We compare our proposed framework with the following methods:

- **IND**: It selects influential nodes based on the incoming degree, i.e., those whose works are cited by most authors in DBLP, or users who are trusted by most people in Epinions or Ciao.
- **BET**: It selects influential nodes based on *betweenness score*. Betweenness score is a centrality measure of a node in a graph which quantifies the number of times a node acts as a bridge along the shortest path between two other nodes (Freeman 1977).
- **CLO**: It selects influential nodes based on *closeness score*. Closeness score is a centrality measure of a node in a graph which is the average length of the shortest paths between the node and all other nodes in a graph (Sabidussi 1966).
- **BPR**: This method runs PageRank in a binary network, where the weights of all edges are set to 1.
- **WPR**: This method runs PageRank in a weighted network, where the weight of an edge from i to j is set to the frequency of i citing the work of j or i trusting j.

For MPR, we work on all of the triangular motifs separately, and then show the results for each of them separately. Note that for performance comparison, we show the best results of each motif we can obtain among different values of α. Then we show how the parameter α affects the performance of MPR.

[3] https://aminer.org/

Performance Comparison

The overall results are shown in Table 3. We show the performance of top10, top50, top500 ranking results with different algorithms. From Table 3, we can see that the best performance for all three datasets with different K's is achieved by our proposed framework. This demonstrates the effectiveness of incorporating higher-order relations into PageRank computation.

We emphasize the following three observations in Table 3. First, we can see that on all three datasets, the larger K is, the smaller NDCG is. This makes sense because more users means more difficulty in achieving an ideal ranking list.

Second, when $K = 10$, we can see that the NDCG of BPR on DBLP and Epinions are greater than 0.94, which is actually very strong in practice. However, our proposed framework can further improve the NDCG from 0.9464 to 0.9920 on DBLP, and from 0.9777 to 0.9957 on Epinions. On Ciao dataset, the performance is improved from 0.8332 to 0.9905, which is even more significant.

Third, when comparing the performance of the five baseline methods, we found that the performance of BPR and WPR are similar, which means that the use of frequencies as weights cannot bring much additional benefit.

When comparing IND, BET, CLO, and BPR, BPR wins only on the Epinions dataset, since different centrality measures can characterize different local or global topological structures of the network. By introducing higher-order structures, MPR can still outperform all of these baselines. Obviously not all motifs can outperform all baselines on all datasets. In the next section, we will present a possible explanation to this phenomenon using Z-score analysis.

Analysis of Z-score

To have more quantitative analysis of MPR, we introduce Z-score (Milo et al. 2002), which is a qualitative measure of statistical significance of a motif in a particular network. To give a formal definition of Z-score, we first introduce randomized networks. Corresponding to a given complex network, a randomized network has the same single-node characteristic as in the real network: each node in the randomized networks has the same number of incoming and outgoing edges as that of the corresponding node in the real network. A random local rewiring algorithm that preserves the degrees of the nodes can be used to generate such randomized network (Milo et al. 2002). For each motif, the number of appearances in the real network and randomized networks are denoted as N_{real} and N_{rand}, respectively. By constructing a number of randomized networks, e.g., 1,000, we compute the standard deviation of N_{rand}, denoted as SD. Therefore, the Z-score is defined as follows.

$$\text{Z-score} = \frac{N_{real} - N_{rand}}{SD}, \tag{6}$$

which measures the significance of the frequency that a motif appears in a given network comparing to the randomized networks. The larger Z-score a motif is, the more frequent it occurs. We then compute the Z-scores of motifs M_1 to M_7 for all of the three datasets with the software FANMOD (Wernicke and Rasche 2006). The results are shown in Table 4. From the table, we can see that the Z-scores of M_1 to M_7 on DBLP are all larger than those on Epinions and Ciao. Therefore, it can help explain the phenomenon that the performance of M_1 to M_7 are all better than BPR on DBLP.

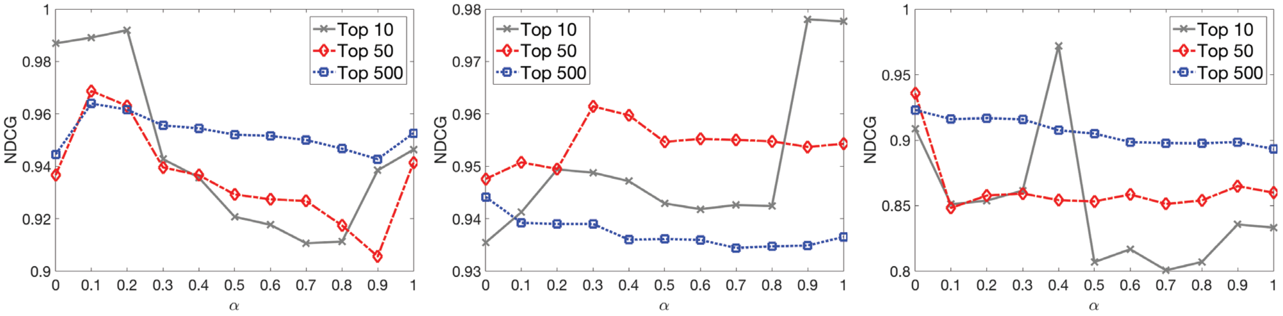
Analysis of α

In Eq. (3), we have a parameter α to control the balance between edge-based and motif-based higher-order relations. When $\alpha = 0$, it means we only use higher-order relations for authority computation. When $\alpha = 1$, it means we only use the original edge-based relations for authority computation. In this section, we show how this parameter affects NDCG performance. For simplicity, we only show the results of one motif with better performance on each dataset, which is M_7 for DBLP and Epinions, and M_4 for Ciao. The results are shown in Figure 5. We can see that the trends are consistent in most cases and the best performance is achieved at some value in [0,1]. It means that the best performance on the three datasets is achieved by combining the edge-based and motif-based relations. The best performance of top500 ranking results on Epinion and Ciao and top50 ranking results on Ciao is achieved at $\alpha = 0$, i.e., using only the higher-order relations. This again demonstrates that higher-order relations can provide useful information for ranking users in social networks. Top50 and top500 ranking results are more consistent and the trends of the curves are also similar, whereas top10 results are more diverse. This may be because the top ten users in the social networks are very prominent and may have different behaviors than the other users.

Case Study

In this section, we analyze some real cases to show the effectiveness of MPR. Specifically, we extract the top ten users and compute the average of their relevance scores, namely, H-index of top ten authors on DBLP and trustworthiness scores of top ten users on Epinions and Ciao. The results are shown in Table 5. Due to space limitation, we only show the results of one motif achieving the best performance for $NDCG_{10}$, i.e., M_7 for DBLP, $M6$ for Epinions, and M_2 for Ciao.

From Table 5, we can see that, on DBLP and Epinions, the average relevance scores of MPR are larger than those of BPR, which means that MPR can select more influential authors or trustworthy users. This demonstrates the effectiveness of incorporating higher-order relations when ranking authors in the research community or users in social networks. On Ciao dataset, the average trustworthiness of M_2 is smaller than that of BPR. The main reason is that the trustworthiness of the user at the 3-rd position of BPR is very large. This may be an outlier considering other users. We support this observation with Figure 6. From the figure, we can see that with increase of K, the trend of relevance of MPR on all three datasets is decreasing despite some small fluctuation. However, the trends of BPR on all three datasets are more unstable. Thus, it demonstrates that MPR can rank influential users in a more reasonable order. It again shows the advantage of incorporating higher-order relations when ranking users in social networks.

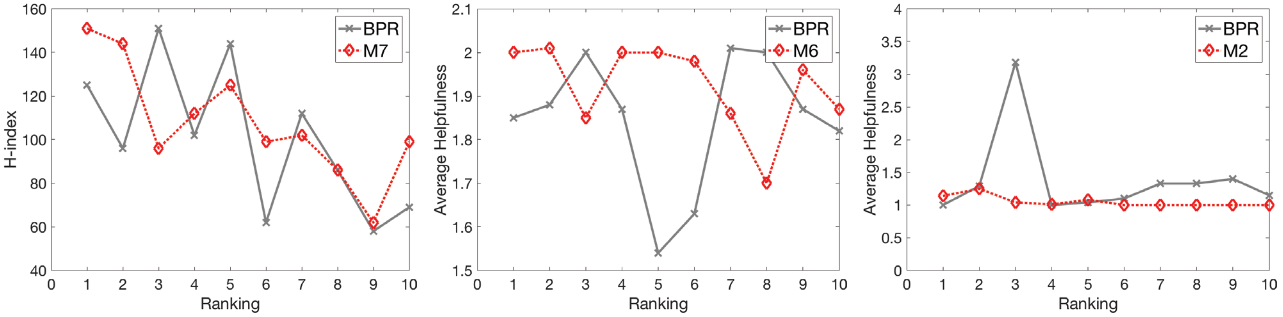


(a) M_7 on DBLP

(b) M_7 on Epinions

(c) M_4 on Ciao

Figure 5: Parameter analysis of α on the three datasets. We show Top10, Top50 and Top500 ranking results. $\alpha = 0$ means we use motif-based relations alone to perform PageRank while $\alpha = 1$ means we use the original edge based relations alone to perform PageRank.



(a) BPR vs. M_7 on DBLP.

(b) BPR vs. M_6 on Epinions.

(c) BPR vs. M_2 on Ciao.

Figure 6: The trend of relevance scores of top ten users on all three datasets.

Table 5: The average relevance scores of top10 users, i.e., H-index on DBLP and trustworthiness scores on Epinions and Ciao.

	DBLP		Epinions		Ciao	
	BPR	M_7	BPR	M_6	BPR	M_2
Average	100.5	107.6	1.85	1.92	1.38	1.05

Conclusion and Future Work

In this paper, we propose MPR to incorporate higher-order relations into conventional authority computation with PageRank. The higher-order relations are captured by small subgraphs, which are called network motifs. We propose to use the motif-based adjacency matrix to re-weigh the links of the edges in social networks. Then, the higher-order and edge-based relations are combined together to perform PageRank. We conduct experiments on an academic network, DBLP, and two trust networks, Epinions and Ciao. The experimental results demonstrate that MPR can significantly improve the performance of ranking users in social networks, and thus the effectiveness of higher-order relations. Overall, we believe our work is a fundamental framework that can be applied to many user ranking problems in social networks.

For future work, we point out here three potential directions. First, linear combination of higher-order and edge-based relations may not be the best way. Some non-linear combination methods are worth exploring. Second, in MPR, we incorporate only one motif at each time. It would be very interesting to incorporate multiple motifs simultaneously and automatically select the weighting parameters for different motifs. Third, to perform better social user ranking it may involve other features, such as content posted by users in the social networks. It will be also interesting if we can combine higher-order relations with other features to perform better user ranking. However, second and third directions would involve supervised information, while our current work is fully unsupervised and only based on social connections.

Acknowledgment

This paper was supported by HKUST-WeChat WHAT Lab, China 973 Fundamental R&D Program (No. 2014CB340304), and the Research Grants Council HKSAR GRF (No.615113 and No. 26206717). We also thank the anonymous reviewers for their valuable comments and suggestions that help improve the quality of this manuscript.

References

Ahmed, N. K.; Neville, J.; Rossi, R. A.; and Duffield, N. 2015. Efficient graphlet counting for large networks. In *ICDM*, 1–10.

Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016a. Higher-order organization of complex networks. *Science* 353(6295):163–166.

Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016b. Supplementary materials for higher-order organization of complex networks. *Science*.

Bianchini, M.; Gori, M.; and Scarselli, F. 2005. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)* 5(1):92–128.

Ding, Y. 2011. Topic-based pagerank on author cocitation networks. *Journal of the American Society for Information Science and Technology (JASIST)* 62(3):449–466.

Freeman, L. C. 1977. A set of measures of centrality based on betweenness. *Sociometry* 35–41.

Gleich, D. F. 2015. Pagerank beyond the web. *SIAM Review* 57(3):321–363.

Granovetter, M. S. 1973. The strength of weak ties. *American journal of sociology (AJS)* 78(6):1360–1380.

Han, G., and Sethu, H. 2016. Waddling random walk: Fast and accurate mining of motif statistics in large graphs. In *ICDM*, 181–190.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)* 20(4):422–446.

Jha, M.; Seshadhri, C.; and Pinar, A. 2015. Path sampling: A fast and provable method for estimating 4-vertex subgraph counts. In *WWW*, 495–505.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology (JASIST)* 58(7):1019–1031.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: Simple building blocks of complex networks. *Science* 298(5594):824–827.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Paranjape, A.; Benson, A. R.; and Leskovec, J. 2017. Motifs in temporal networks. In *WSDM*, 601–610.

Pinar, A.; Seshadhri, C.; and Vishal, V. 2017. Escape: Efficiently counting all 5-vertex subgraphs. In *WWW*, 1431–1440.

Pržulj, N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* 23(2):e177–e183.

Rotabi, R.; Kamath, K.; Kleinberg, J.; and Sharma, A. 2017. Detecting strong ties using network motifs. In *WWW Companion*, 983–992.

Sabidussi, G. 1966. The centrality index of a graph. *Psychometrika* 31(4):581–603.

Simmel, G. 1908. Sociology: investigations on the forms of sociation. *Duncker & Humblot, Berlin Germany*.

Song, X.; Chi, Y.; Hino, K.; and Tseng, B. L. 2007. Identifying opinion leaders in the blogosphere. In *CIKM*, 971–974.

Sporns, O., and Kötter, R. 2004. Motifs in brain networks. *PLoS biology* 2(11):e369.

Stefani, L. D.; Epasto, A.; Riondato, M.; and Upfal, E. 2017. Trièst: Counting local and global triangles in fully dynamic streams with fixed memory size. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11(4):43.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. ArnetMiner: Extraction and mining of academic social networks. In *KDD*, 990–998.

Tang, J.; Sun, J.; Wang, C.; and Yang, Z. 2009. Social influence analysis in large-scale networks. In *KDD*, 807–816.

Tang, J.; Gao, H.; Liu, H.; and Das Sarma, A. 2012. eTrust: Understanding trust evolution in an online world. In *KDD*, 253–261.

Tang, J.; Gao, H.; and Liu, H. 2012. mTrust: Discerning multi-faceted trust in a connected world. In *WSDM*, 93–102.

Ugander, J.; Backstrom, L.; and Kleinberg, J. 2013. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *WWW*, 1307–1318.

Wang, Y.; Wang, X.; Tang, J.; Zuo, W.; and Cai, G. 2015. Modeling status theory in trust prediction. In *AAAI*, 1875–1881.

Wang, P.; Lui, J. C.; Towsley, D.; and Zhao, J. 2016. Minfer: A method of inferring motif statistics from sampled edges. In *ICDE*, 1050–1061.

Wang, P.; Lü, J.; and Yu, X. 2014. Identification of important nodes in directed biological networks: A network motif approach. *PloS one* 9(8):e106132.

Wernicke, S., and Rasche, F. 2006. Fanmod: a tool for fast network motif detection. *Bioinformatics* 22(9):1152–1153.

Xiang, B.; Liu, Q.; Chen, E.; Xiong, H.; Zheng, Y.; and Yang, Y. 2013. Pagerank with priors: An influence propagation perspective. In *IJCAI*, 2740–2746.

Yin, H.; Benson, A. R.; Leskovec, J.; and Gleich, D. F. 2017. Local higher-order graph clustering. In *KDD*, 555–564.

Zhang, J.; Tang, J.; Zhong, Y.; Mo, Y.; Li, J.; Song, G.; Hall, W.; and Sun, J. 2017. StructInf: Mining structural influence from social streams. In *AAAI*, 73–80.