On Analyzing Graphs with Motif-Paths

Supplemental Material

Xiaodong Li[†], Reynold Cheng[†], Kevin Chang[‡], Caihua Shan[†], Chenhao Ma[†], Hongtai Cao[‡]
[†]Department of Computer Science, University of Hong Kong, Hong Kong SAR
[‡]Department of Computer Science, University of Illinois at Urbana-Champaign, USA
{xdli,ckcheng,chshan,chma2}@cs.hku.hk;{kcchang,hongtai2}@illinois.edu

Table 1: Summary of algorithm complexities.

Alg.	Time	Space
BASE	$O(N_\tau^3)$	$O(N_\tau^2)$
MODC	$O(\sum_{s\in V}\sum_{\tau'\in B_k} D_{\tau'}^k)$	$O(I)$
MODCt	$O(\|V\| \times d_{\max}^2)$	$O(\|V\| \times \binom{d_{\max}}{2})$
MODQ	$O(D_\tau^{\phi_\tau - k})$	$O(I)$
SMP	$O(\|V\| \times D_\tau^{\phi_\tau - k} + N_\tau)$	$O(\|V\| + N_\tau + I)$
ESMP	$O(\|V\| \times D_\tau^{\phi_\tau - k} + \|E\| + N_\tau)$	$O(\|V\| + \|E\| + N_\tau + I)$
MGD	$O(\|V\| \times D_\tau^{\phi_\tau - k} + N_\tau)$	$O(\|V\| + N_\tau + I)$
MKI	$O((\|V\| \times D_\tau^{\phi_\tau - k})^{2L} + N_\tau)$	$O(\|V\|^{2L} + N_\tau + I)$
MLGC	$O(\|V\| \times \hat{k} \times D_\tau^{\phi_\tau - k} + N_\tau)$	$O(d_{\max} \times \hat{k} + \binom{\hat{k}}{\|V_\tau\|} + I)$
MBET	$O(\|V\| \times (D_\tau^{\phi_\tau - k} + \|V\| + \|E_W\|) + N_\tau)$	$O(\binom{d_{\max}^{\phi_\tau}}{\|V_\tau\|} + \|V\|^2 + I)$

Here $N_\tau = \binom{|V|}{|V_\tau|}$, $D_\tau = \binom{d_{\max}}{\tau.d_e}$ and $I = \sum_{s\in V}\sum_{\tau'\in B_k}\binom{d_{\max}^{\phi_{\tau'}}}{|V_{\tau'}|-1}$.

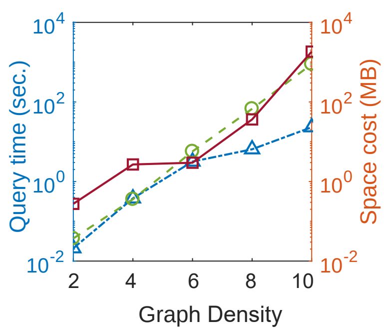

Figure 1: MOD-Indexing time and space cost as graph density grows, with shortest motif-path query time reported.

Table 2: MKI/MGD performance with AUC reported. Numbers of top-3 highest are marked bold.

Method	GAVI	EXTE	DBLP	AMAZ	YOUT
CN	0.72	0.56	0.77	0.62	0.54
JC	0.70	0.48	0.55	0.52	0.44
AA	0.75	0.57	**0.81**	0.65	0.52
PA	0.59	0.76	0.64	0.63	0.76
FM	0.65	0.65	0.59	0.64	0.69
HT	0.60	0.70	0.64	0.59	0.53
RPR	0.61	0.51	0.76	0.62	0.48
MCN	0.67	0.62	0.75	0.61	0.65
MLP+GB	**0.89**	**0.83**	**0.82**	**0.72**	**0.83**
KI	0.69	0.60	0.69	0.60	0.63
MKI	0.71	0.66	0.74	0.63	0.66
MKI-c	0.62	0.67	-	-	-
MKI-b	**0.76**	**0.87**	**0.75**	**0.73**	**0.76**
GD	0.50	0.50	0.50	0.50	0.50
MGD	0.67	0.63	0.65	0.66	0.55
MGD-c	0.52	0.51	-	-	-
MGD-b	**0.75**	**0.84**	0.72	**0.81**	**0.87**

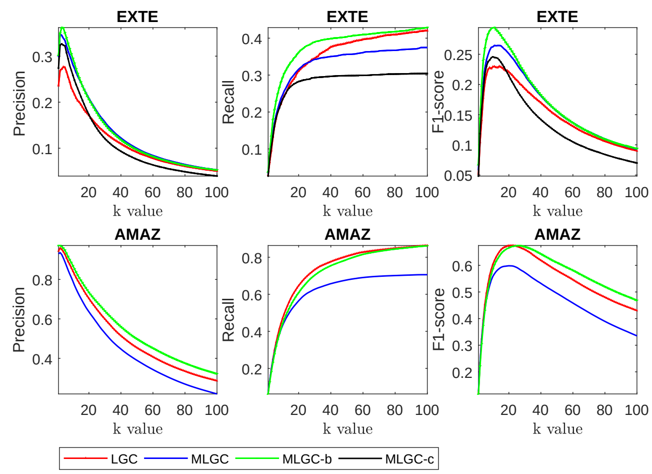

Figure 2: Local graph clustering results on EXTE and DBLP with Precision, Recall and F1-score reported.

Table 3: MLGC performance with F1-score reported. Numbers of top-3 highest F1-score are marked bold.

Method	GAVI	EXTE	DBLP	AMAZ	YOUT
TECTONIC	0.39	**0.44**	-	0.37	-
MAPPR	0.39	**0.42**	**0.34**	0.35	0.15
EdMot	0.33	0.38	-	-	-
LGC	**0.42**	0.36	0.33	**0.63**	**0.17**
MLGC	**0.41**	0.30	**0.35**	**0.59**	**0.16**
MLGC-c	0.39	0.29	-	-	-
MLGC-b	**0.42**	**0.38**	**0.35**	**0.65**	**0.23**

PVLDB Reference Format:
Xiaodong Li[†], Reynold Cheng[†], Kevin Chang[‡], Caihua Shan[†], Chenhao Ma[†], Hongtai Cao[‡]. On Analyzing Graphs with Motif-Paths. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

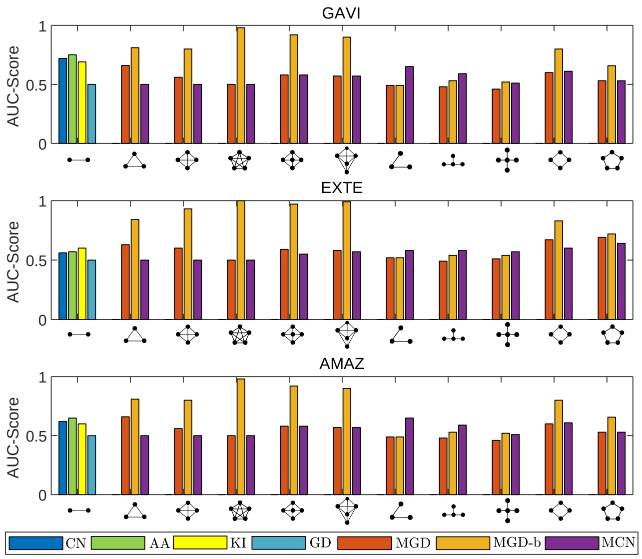

Figure 3: Evaluation of generic motifs (edge, cliques, quasi-cliques, stars and cycles) for link prediction on GAVI, EXTE and AMAZ.

1 PROOF OF THE LEMMAS

Proof of Lemma 1:

Proof. For c1, there is no need to add motif-instances containing a "searched" node since all motif-instances around node marked as "searched" have been found and added into candidates for $\mathcal{P}_{s,t}$. For c2, for the motif-instances in $\mathcal{P}_{s,t}$, there are only two status of the nodes: "searched" and "discovered". We only select "discovered" node as next seed because the "searched" nodes have been used as seed before and thus using them as next seed will find duplicates. For c3, in the incremental search manner, $\mathcal{P}_{s,v}$ is found for the "undiscovered" node v when v is covered by any motif-instance for the first time. Therefore, we only add motif-instances which contain at least one node marked as "undiscovered" to push the incremental search forward. □

Proof of Lemma 2:

Proof. Since $\mathcal{P}_{s,t}$ is the shortest sequence of motif-instances from m_s to m_t. For each motif-instance in the sequence, we pick the edge which links s, t and the nodes shared by the neighboring motif-instances, the path is the shortest one on W. Vise versa. □

Proof of Lemma 3:

Proof. Assume that $\exists(i,j) \in V \times V$, $W_{i,j} = 1$ but there is no motif-instance of $\bar{\tau}$, then there must be motif-instance m such that $m \simeq \bar{\tau}' \& (i,j) \in E_m$, where $\bar{\tau}'$ is another motif-orbit of τ with seed $s \in V_m$. By switching the seed node, $m \simeq \bar{\tau}$ with seed $s' \in V_m$, which is contradictory to the assumption. □

2 PROOF OF THE COMPLEXITIES

3 EFFICIENCY EVALUATIONS

4 EFFECTIVENESS EVALUATIONS

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX