Colorful *h*-star Core Decomposition

Sen Gao, Rong-Hua Li, Hongchao Qin, Hongzhi Chen,
Ye Yuan, Guoren Wang,

ICDE 2022

Presenter: Sen Gao

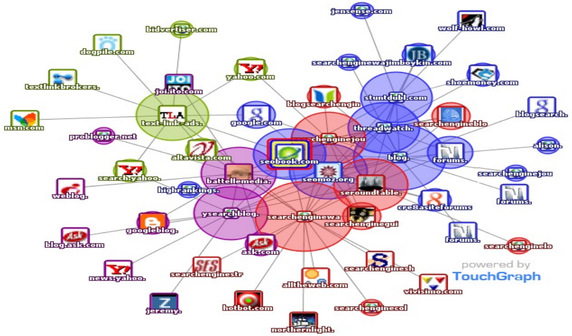
Contents

Background

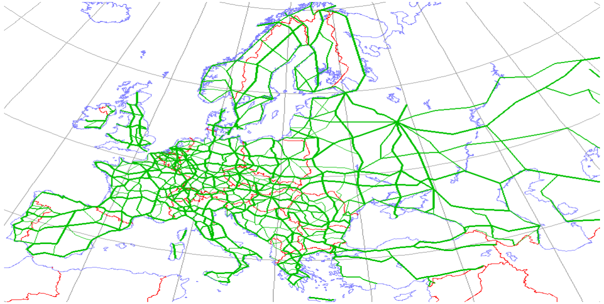
- Graph data are everywhere!

Social Network

Internet

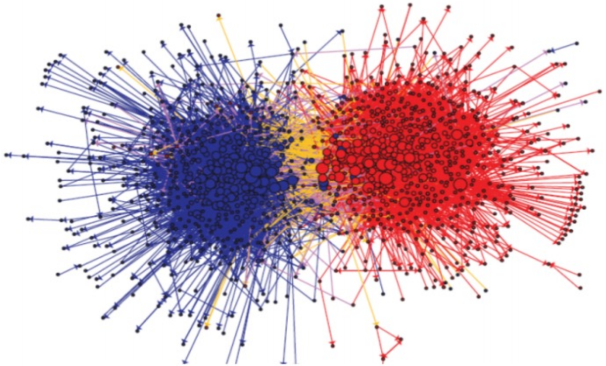


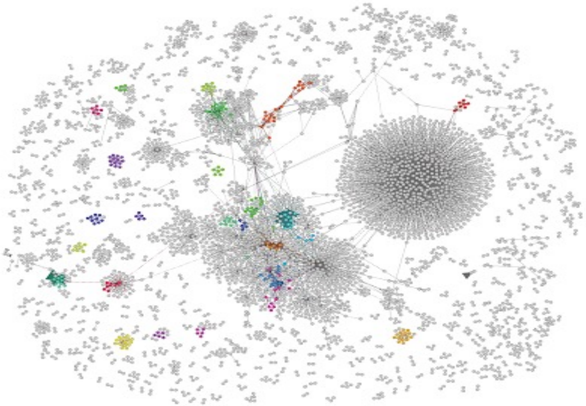
Road Network

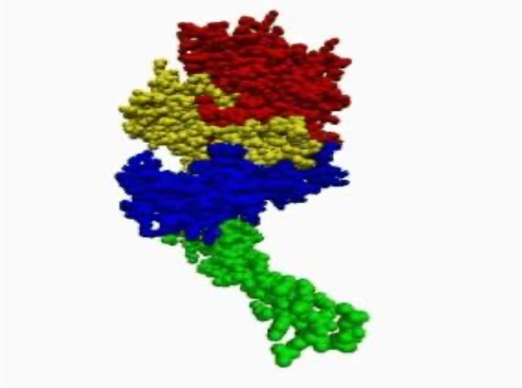


IoT Network

Cohesive subgraph Structure



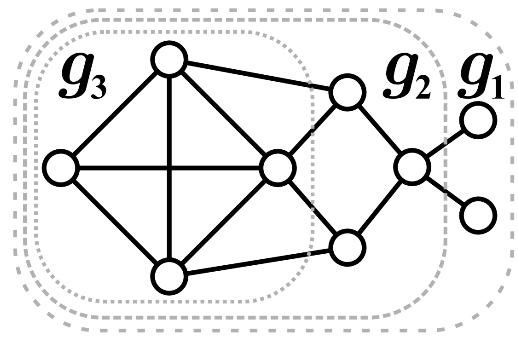


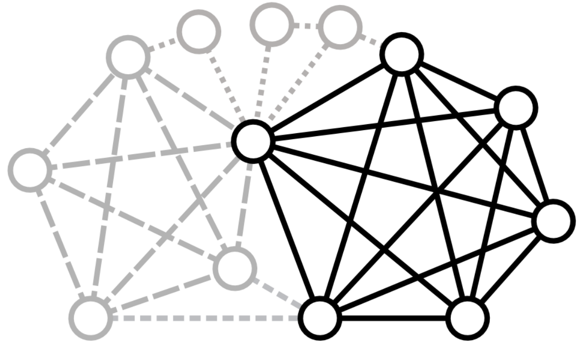


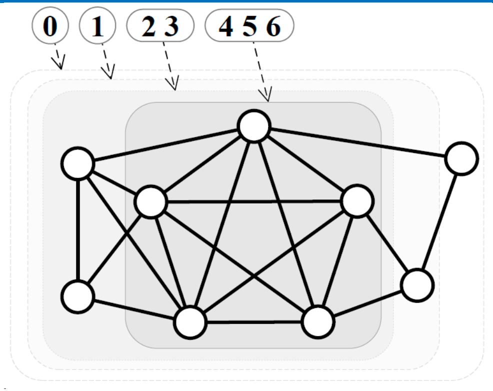
Blog Network **Twitter Social Network** **PPI Network**

Real-life networks often contains cohesive subgraph structures

Existing cohesive subgraph models







***k*-core**	***k*-truss**	***h*-clique *k* core** (*h*=3)
Minimum degree on Nodes	**Minimum number of triangles on Edges**	**Minimum number of *h*-clique on Nodes**

Existing cohesive subgraph models

Comparison between cohesive subgraph models

Models	cohesiveness measures	🙂	🙁
***k*-core**	Minimum degree	**Low complexity** **Really fast**	**Cannot locate denser regions**
***h*-clique *k* core**	Minimum number of motifs (*h*-clique)	**Higher-order graph analysis**	**High complexity** **Costly computation**
***h*-clique densest subgraph**	Average number of motifs (*h*-clique)		

Our contributions

New model

- The **colorful *h*-star core** model, which is a **relaxation** of the *h*-clique core model, can be applied to higher-order graph analysis.

Novel algorithms

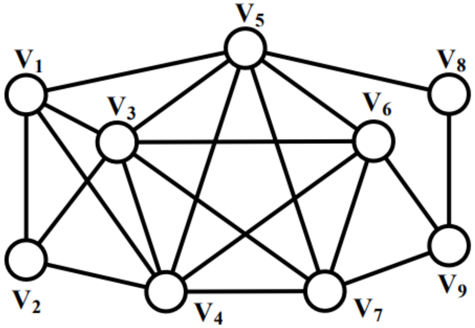
- **DP algorithm** to *compute* and **novel updating technique** to *update* the number of colorful h-stars for each node. (in **near-linear** time)

Extensive experiments

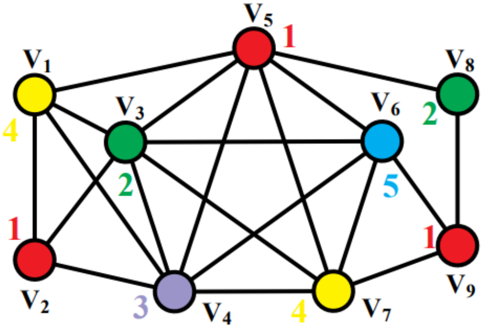
- We conduct **extensive experiments** on 11 large real-life datasets and **a case study** on DBLP, to evaluate our algorithms.

colorful *h*-star

Coloring the graph
(two adjacent nodes have different colors)

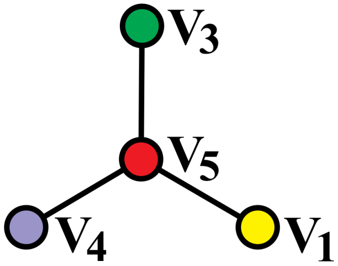


(a) An undirected graph

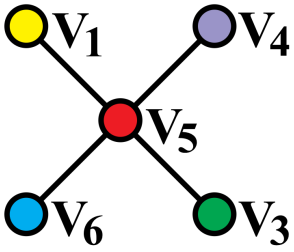


(b) The colored graph

Coloful *h*-star: **nodes in a star have different colors**

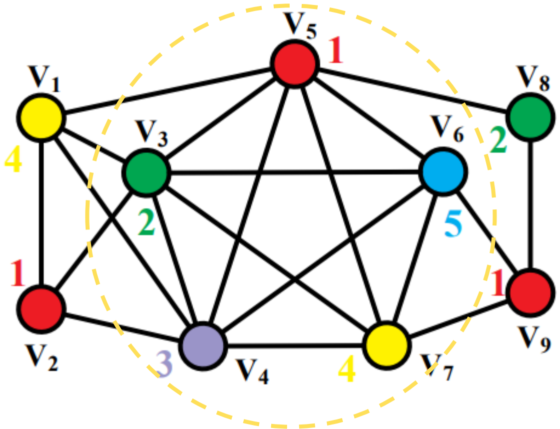


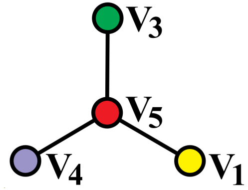
colorful 4-star



colorful 5-star

Motivation





The subgraph induced by $\{V_1, V_3, V_4, V_5\}$ is a 4-clique.

A colorful *h*-star is a *relaxed* definition of *h*-clique

Nodes within an *h*-clique have different colors

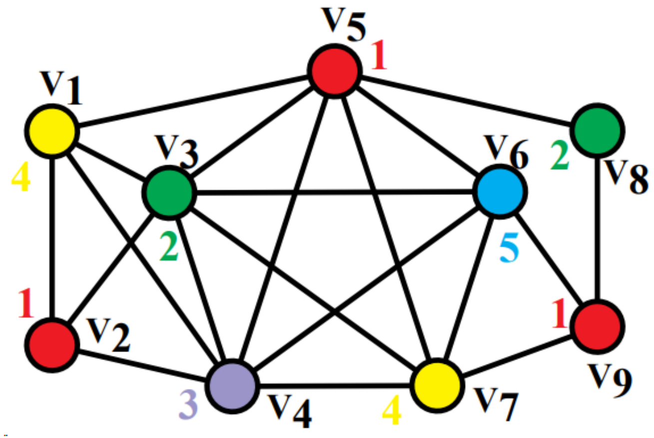
↓

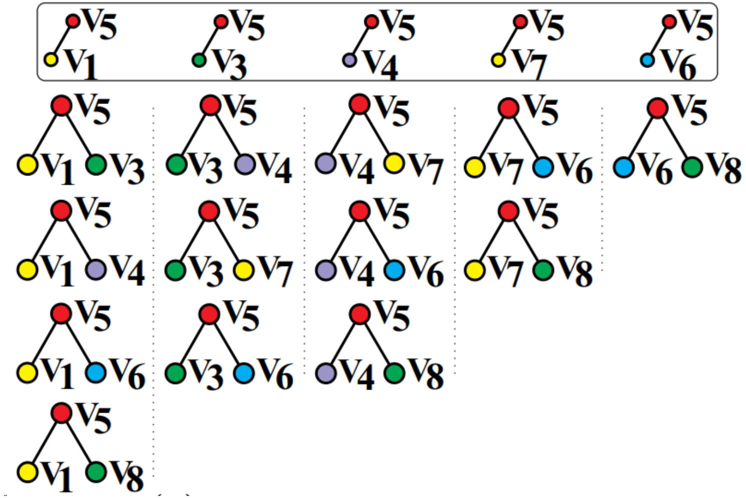
Any *h*-star in a *h*-clique must be colorful

A colorful *h*-star in *G* is more likely to be an *h*-clique
(induced subgraph)

Colorful *h*-star degree of *u*

the number of colorful *h*-stars centering on *u*

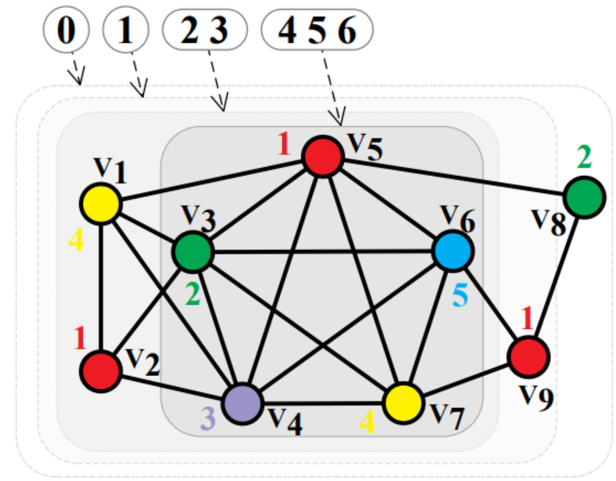




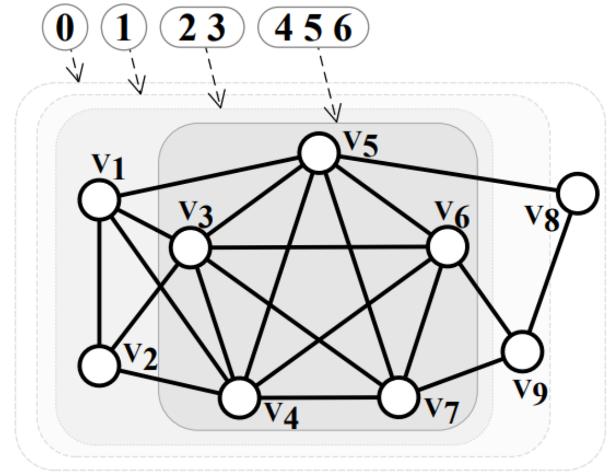
Colorful 3-star degree of V_5 = 13

Colorful *h*-star *k* core

The maximal subgraph in which the colorful *h*-star degree of each node is at least *k*



colorful *h*-star cores



***h*-clique cores**

Have the same cores in this example

The colorful *h*-star core is a good approximation of *h*-clique core

Colorful *h*-star core

The *h*-clique *K*max core has been proved to be a $\frac{1}{h}$-approximation solution to *h*-clique densest subgraph problem.

Therefore, our colorful *h*-star core can also provide a good approximation.

Challenge

We can use a peeling-style algorithm to compute each colorful h-star core, i.e. colorful *h*-star core decomposition, but

1. **how to compute the colorful *h*-star degree of each node? Using DP algorithm**
2. **how to update neighbors' colorful *h*-star degrees after removing a node? Using updating algorithm**

Colorful h-star core

Compute colorful h-star degree

Given a node u, divide u's neighbors into groups according to their colors

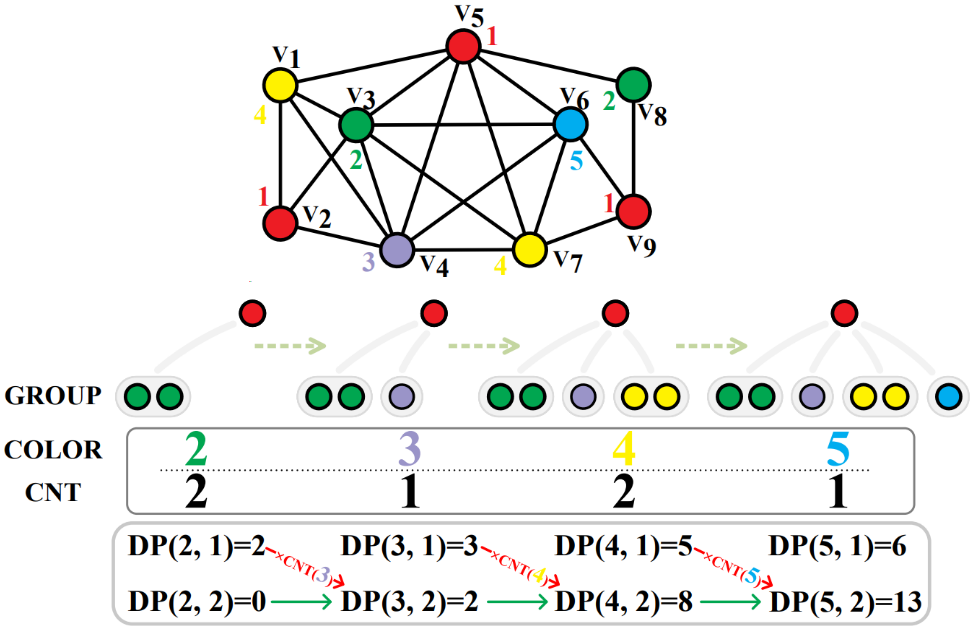


Fig. 4. Illustration of the computation for v_5's colorful 3-star degree.

DP(i, j): choose j neighbors with different colors from the first i groups

Case 1: only choose nodes from the first (i-1) groups

Case 2: choose a node from the i-th color group

$$\text{DP}(i, j) = \underbrace{\text{DP}(i-1, j)}_{\text{Case 1}} + \underbrace{\text{DP}(i-1, j-1) * cnt(i)}_{\text{Case 2}}$$

Time complexity

$$O(h \times \min\{\chi, d_u(G)\})$$

χ : the number of different colors used by the coloring algorithm

Updating colorful *h*-star degree

After removing a neighbor *v* of *u*, how to avoid recomputing the colorful *h*-star degree of *u* from scratch?

Observation

$$\mathsf{DP}(i,j) = \underbrace{\mathsf{DP}(i-1,j)}_{\text{Case 1}} + \underbrace{\mathsf{DP}(i-1,j-1) * cnt(i)}_{\text{Case 2}}$$

Removing *v* will reduce cnt(color(*v*)) only affects Case 2

Idea:

Decompose the DP equation into two different cases

Updating colorful h-star degree

Idea: Decompose the DP equation into two different cases

Given a node u, a color χ' $G(i)$ and $F(i)$ denotes the number of two types of colorful h-stars.

$G(i)$: has a node with color χ'

$F(i)$: not have a node with color χ'

$$DP(i) = F(i) + G(i)$$

After removing a neighbor, only $G(i)$ needs updating.

Colorful h-star core

Updating colorful h-star degree

Removing a neighbor v, three steps to update u

Procedure Updating($\mathcal{DP}$, v)

$\chi' \leftarrow \mathsf{color}(v)$, $\mathcal{F}(0) \leftarrow 1$;

for $i = 1$ **to** h **do**

  $\mathcal{G}(i) \leftarrow \mathcal{F}(i-1) \times \mathsf{cnt}(\mathsf{color}(v))$;

  $\mathcal{F}(i) \leftarrow \mathcal{DP}(i) - \mathcal{G}(i)$;

restore

$\mathsf{cnt}(\mathsf{color}(v)) \leftarrow \mathsf{cnt}(\mathsf{color}(v)) - 1$;

update

for $i = 1$ **to** h **do**

  $\mathcal{G}(i) \leftarrow \mathcal{F}(i-1) \times \mathsf{cnt}(\mathsf{color}(v))$;

  $\mathcal{DP}(i) \leftarrow \mathcal{F}(i) + \mathcal{G}(i)$;

regenerate

return $\mathcal{DP}(h-1)$;

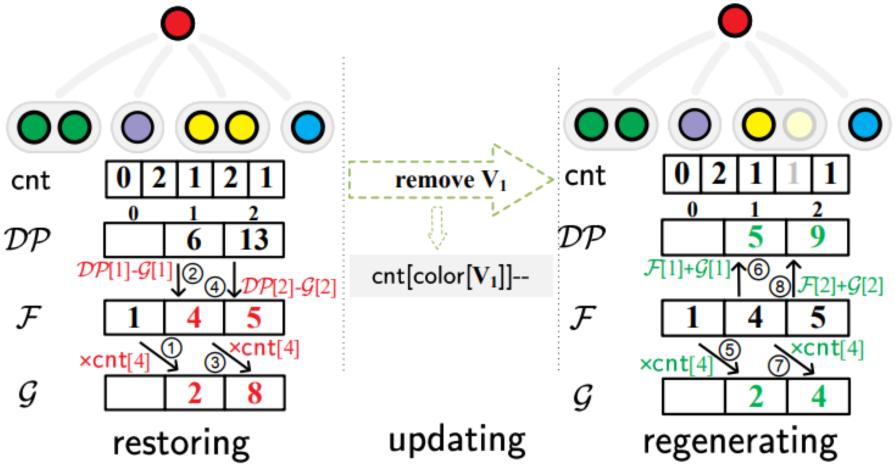


Fig. 5. Illustration of updating v_5's colorful 3-star degree.

Time complexity

$O(h)$

Experiments

Datasets

TABLE I
DATASETS

Dataset	$n = \|V\|$	$m = \|E\|$	χ	$d_{\max}$
Nasasrb	54,870	1,311,227	38	275
Pkustk	87,804	2,565,054	54	131
Buzznet	101,163	2,763,066	62	64,289
Pwtk	217,891	5,653,221	42	179
DBLP	317,080	1,049,866	114	343
MsDoor	404,785	9,378,650	42	76
Digg	770,799	5,907,132	66	17,643
LDoor	909,537	20,770,807	42	76
Skitter	1,694,616	11,094,209	71	35,455
Orkut	2,997,166	106,349,209	79	27,466
LiveJournal	4,847,572	42,851,237	324	20,333

various domains

1. **online social networks**
2. **collaboration networks**
3. **internet topology graphs**
4. **scientific computing networks**

1. Colorful *h*-star core decomposition

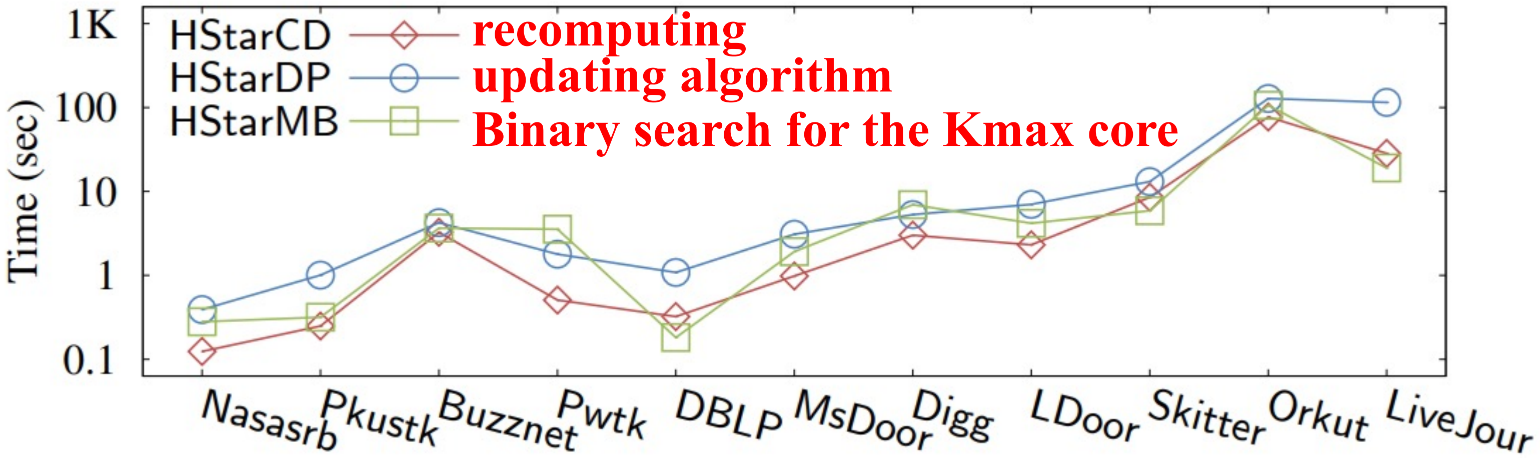
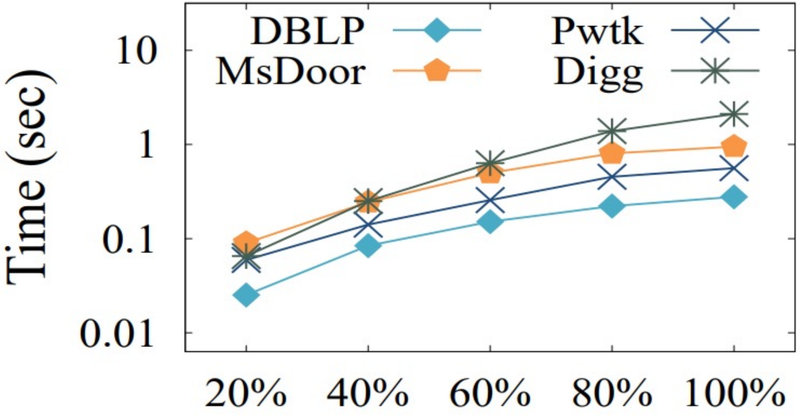


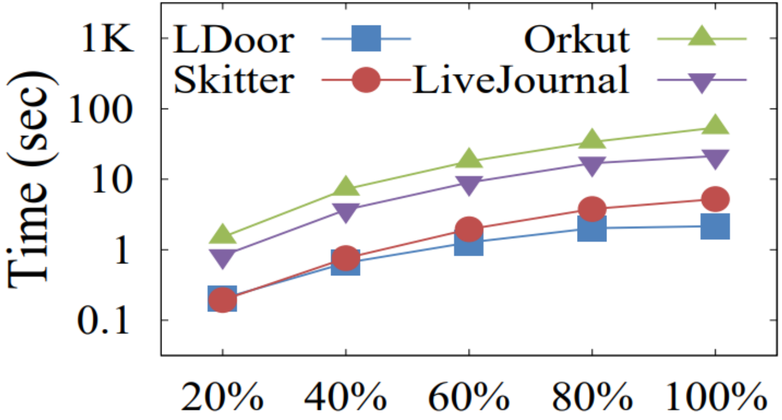
Fig. 7. Running time of HStarCD, HStarDP and HStarMB ($h = 6$)

Comparison between algorithms with and without the updating technique

1. Colorful *h*-star core decomposition



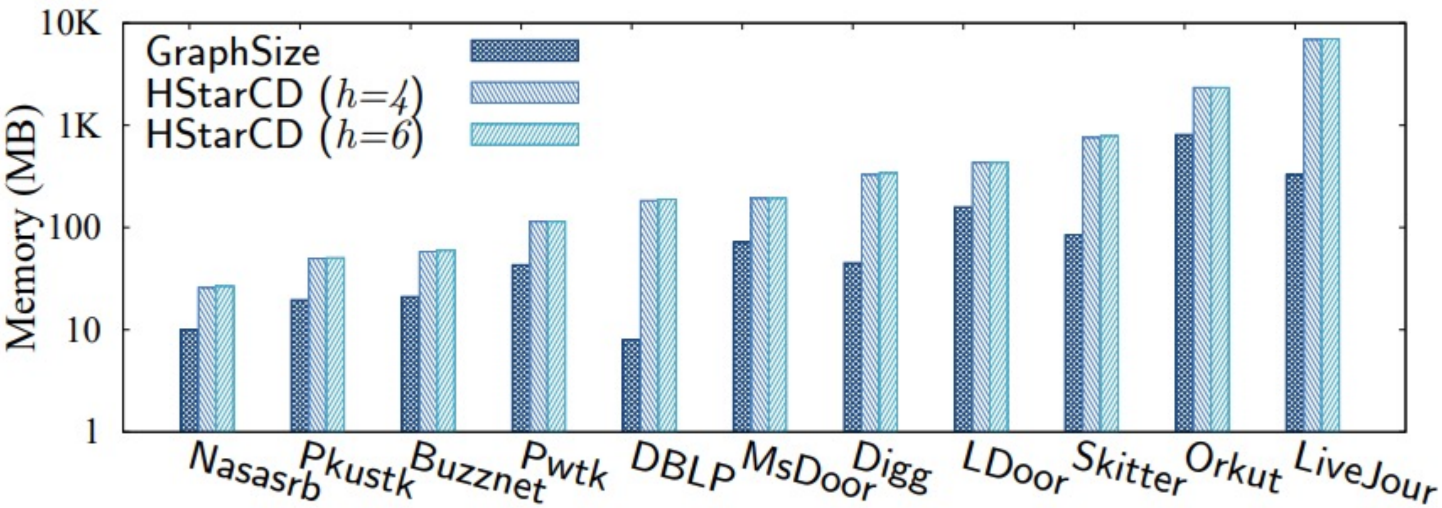
(b) Medium-sized graphs ($h = 6$)



(d) Massive graphs graphs ($h = 6$)

Scalability on medium-sized and massive graphs

1. Colorful *h*-star core decomposition



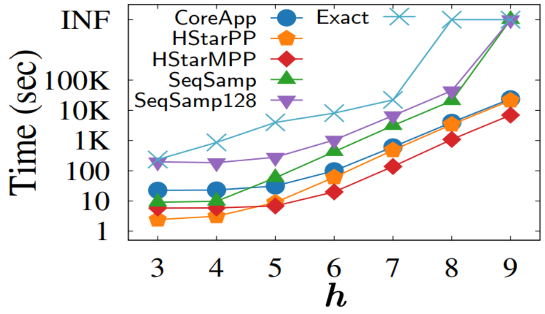
Memory overhead on different datasets (*h* = 4, 6)

2. *h*-clique densest subgraph problem

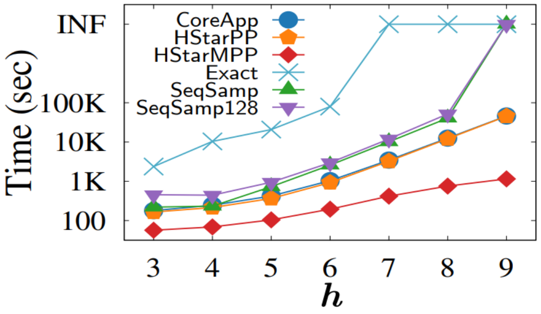
We proposed two algorithms

HStarPP	**a graph reduction technique based on the colorful h-star core model(with theoretical guarantee)**
HStarMPP	**a heuristic approach using the colorful h-star Kmax core as an approximation solution**

2. *h*-clique densest subgraph problem

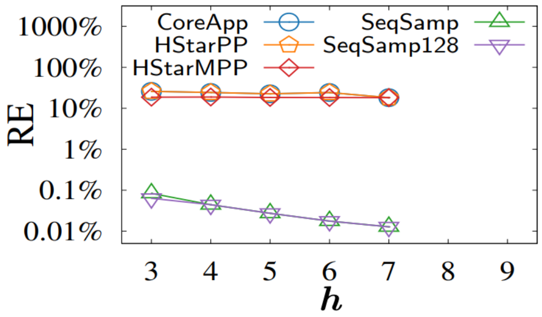


(b) Skitter

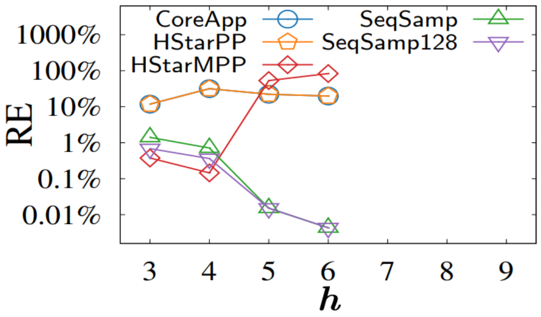


(c) Orkut

Runtime



(b) Skitter



(c) Orkut

Relative errors

2. *h*-clique densest subgraph problem

TABLE II
THE POWER OF PRUNING TECHNIQUES USED IN HStarPP ($h = 6$, 1K=1,000, 1M=1,000,000, 1G=1,000,000,000)

Dataset	$n = \|V\|$			$m = \|E\|$			#Density=m/n			Δ	σ_6: 6-clique density		
	G	C_θ^S	$C_{\delta\Psi}^\Psi$	G	C_θ^S	$C_{\delta\Psi}^\Psi$	G	C_θ^S	$C_{\delta\Psi}^\Psi$	C_θ^S	G	C_θ^S	$C_{\delta\Psi}^\Psi$
Nasasrb	54.9K	52.1K	1.6K	1.3M	1.3M	28.4K	23.90	24.13	17.50	5.13%	12.36K	12.99K	11.14K
Pkustk	87.8K	41.3K	396	2.6M	1.4M	9.1K	29.21	32.75	23.05	53.01%	25.90K	29.71K	95.19K
Buzznet	101K	33.8K	275	2.8M	2.2M	21.5K	27.31	65.04	78.23	66.63%	63.68K	191K	4.17M
DBLP	317K	114	114	1.0M	6.4K	6.4K	3.31	56.50	56.50	99.96%	13.31K	23.39M	23.39M
Digg	771K	23.4K	153	5.9M	2.9M	9.5K	7.66	125.26	62.29	96.96%	20.01K	658K	10.25M
Skitter	1.7M	3.0K	180	11.1M	222K	11.9K	6.55	73.87	66.24	99.82%	5.76K	2.66M	9.53M
Orkut	3.0M	693K	132	106M	50.4M	7.5K	35.48	72.70	56.77	76.87%	15.75K	64.80K	7.57M
LiveJournal	4.8M	483	385	42.9M	108K	73.7K	8.84	224.41	191.31	99.99%	1.70M	16.86G	10.72G

The power of graph reduction techniques

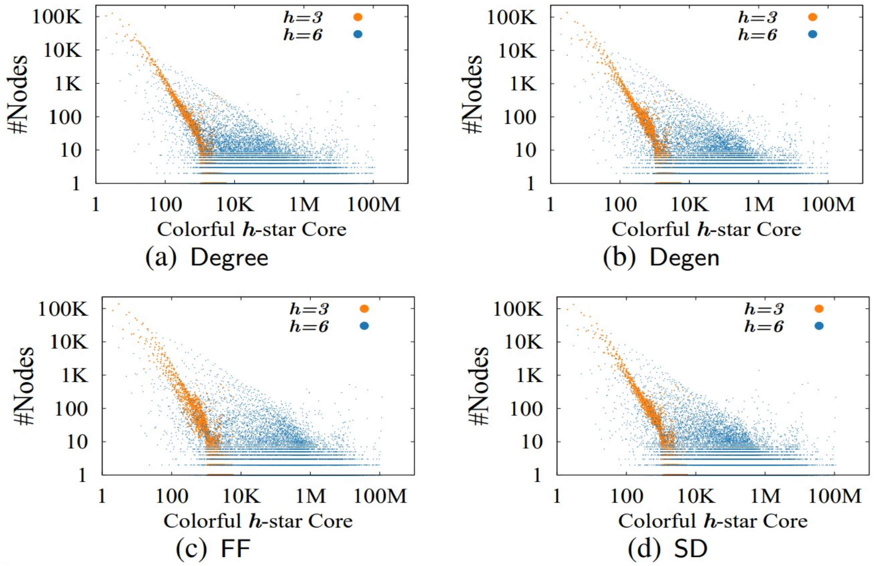
2. *h*-clique densest subgraph problem

The effects of graph colorings

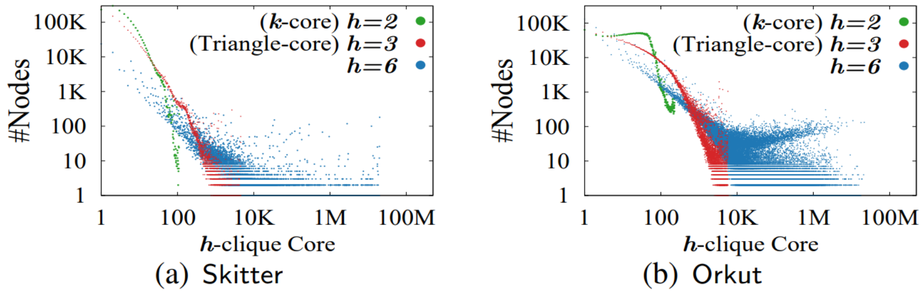
Degree	Coloring the nodes following a non-increasing ordering of degree.
Degen	Coloring the nodes following an inverse degeneracy ordering.
FF	Coloring the nodes in the order they appear in the input graph representation.
SD	Coloring an uncolored node whose colored neighbor nodes use the largest number of distinct colors.

2. *h*-clique densest subgraph problem

The effects of graph colorings



Nodes' distributions of colorful h-star cores



Nodes' distributions of h-clique cores

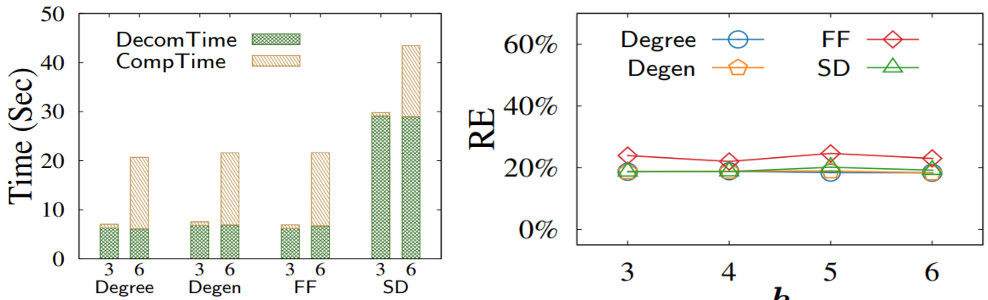
2. *h*-clique densest subgraph problem

The effects of graph colorings

PERFORMANCE OF DIFFERENT COLOR ALGORITHMS (Skitter, $h = 6$, H IS THE $(\delta^{\mathcal{S}}, \mathcal{S})$-CORE)

	$n = \|V_H\|$	$m = \|E_H\|$	χ	$\sigma_6(H)$
Degree	212	15,503	71	10.27M
Degen	213	15,609	75	10.28M
FF	233	15,145	101	10.08M
SD	213	15,600	68	10.31M

The performance of coloring algorithms



(a) Running Time ($h = 3, 6$)

(b) Relative Error

Performance of HStarMPP with various graph coloring tehcniques

Thank you!