Efficient SimRank Computation on Large Graphs

Weiren Yu, Xuemin Lin, Wenjie Zhang

School of Computer Science & Engineering, University of New South Wales, Sydney, Australia

SimRank Overview

- SimRank
  - An appealing link-based similarity measure (KDD '02)
  - Basic philosophy

    Two vertices are similar if they are referenced by similar vertices.

- Two Forms
  - Original form (KDD '02)

$$s(a,a) = 1$$

$$s(a,b) = \frac{C}{|\mathcal{I}(a)|\,|\mathcal{I}(b)|} \sum_{j \in \mathcal{I}(b)} \sum_{i \in \mathcal{I}(a)} s(i,j)$$

similarity btw. nodes a and b

damping factor

in-neighbor set of node b

  - Matrix form (EDBT '10)

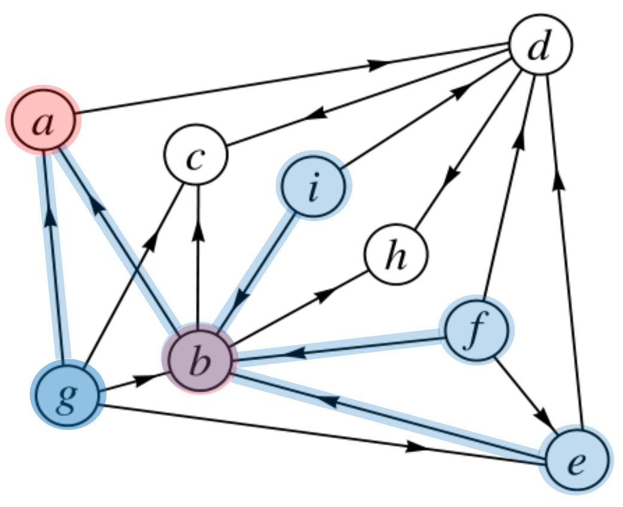
$$\mathbf{S} = C \cdot (\mathbf{Q} \cdot \mathbf{S} \cdot \mathbf{Q}^T) + (1-C) \cdot \mathbf{I}_n$$

Motivation

- Prior Work (VLDB J. '10)
  - High time complexity: O(Kdn²)
    - Duplicate computation among partial sums memoization
  - Slow (geometric) convergence rate
    - Require K = ⌈log$_C$ϵ⌉ iterations to guarantee accuracy ϵ
- Our Contributions
  - Propose an adaptive clustering strategy to reduce the time from O(Kdn²) to O(Kd'n²), where d'≤ d.
  - Introduce a new notion of SimRank to accelerate convergence from geometric to exponential rate.

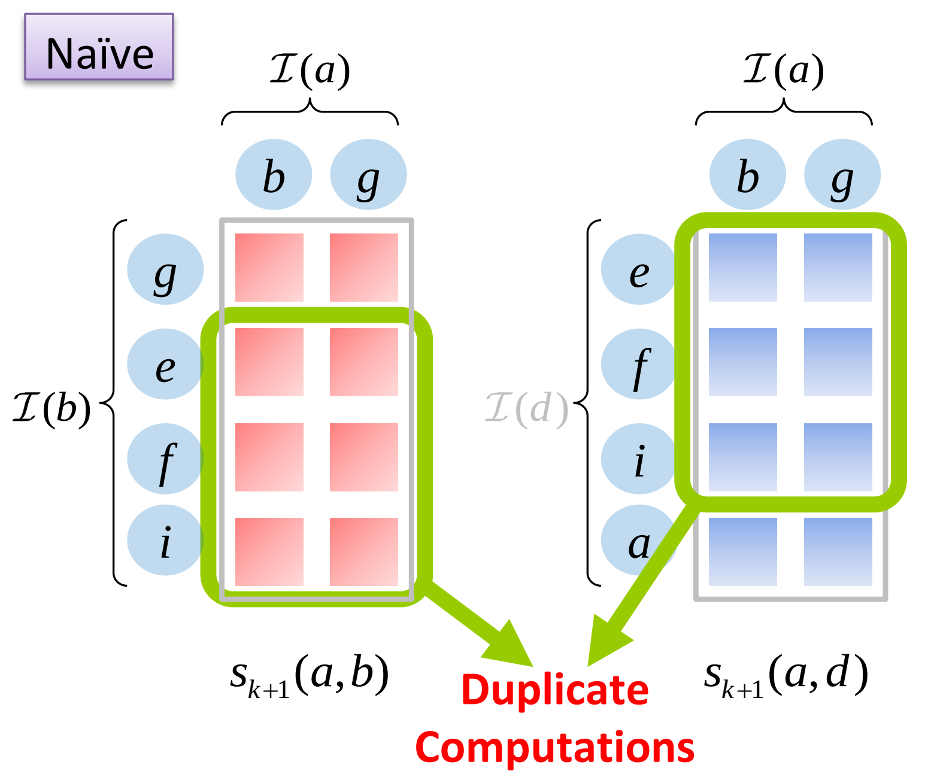
Duplicate Computations among Partial Sums

- Example:



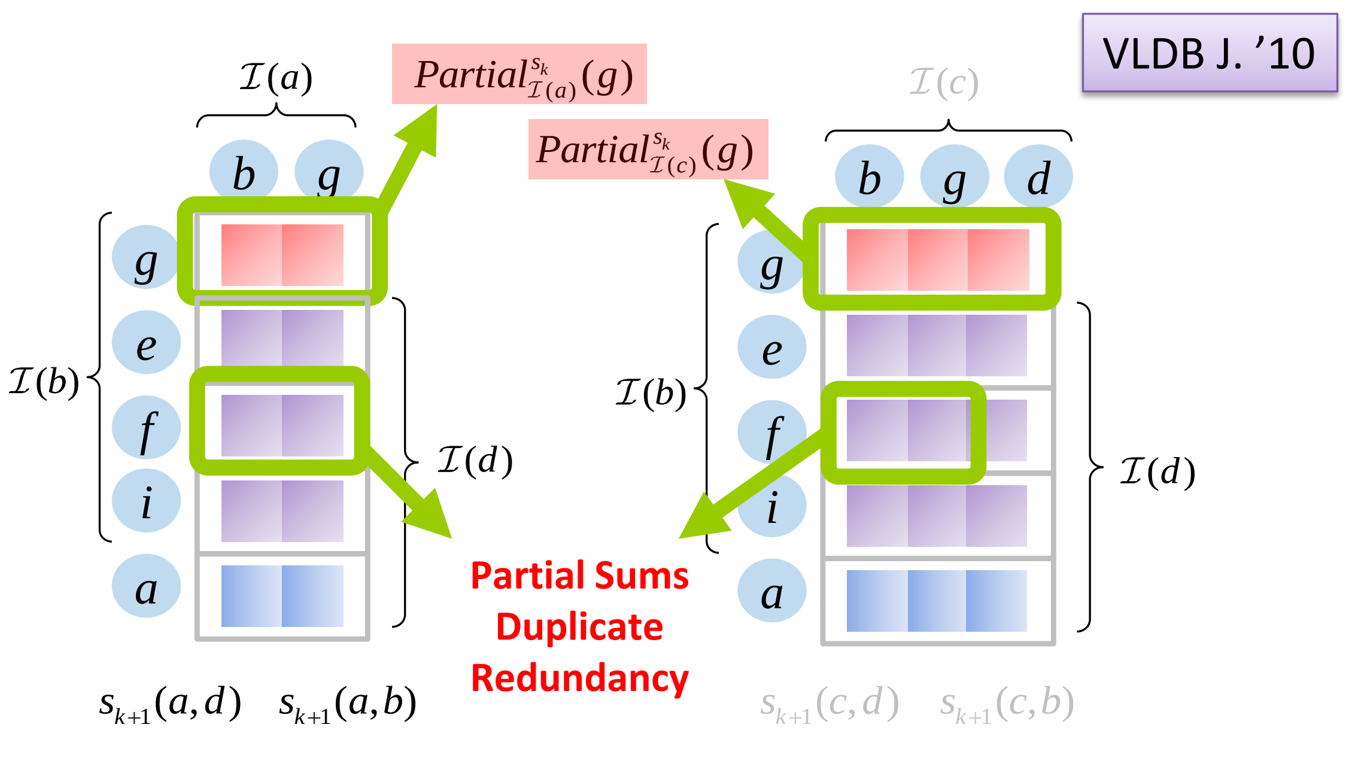
Compute s(a,b), s(a,d)

s(c,b), s(c,d)



$$s_{k+1}(a,b) = \frac{C}{|\mathcal{I}(a)||\mathcal{I}(b)|} \sum_{j \in \mathcal{I}(b)} \sum_{i \in \mathcal{I}(a)} s_k(i,j)$$

$$s_{k+1}(a,d) = \frac{C}{|\mathcal{I}(a)||\mathcal{I}(d)|} \sum_{j \in \mathcal{I}(d)} \sum_{i \in \mathcal{I}(a)} s_k(i,j)$$



$$\forall j,\quad Partial^{s_k}_{\mathcal{I}(a)}(j) = \sum_{i \in \mathcal{I}(a)} s_k(i,j) \qquad \forall j,\quad Partial^{s_k}_{\mathcal{I}(c)}(j) = \sum_{i \in \mathcal{I}(c)} s_k(i,j)$$

$$s_{k+1}(a,b) = \frac{C}{|\mathcal{I}(a)||\mathcal{I}(b)|} \sum_{j \in \mathcal{I}(b)} Partial^{s_k}_{\mathcal{I}(a)}(j) \qquad s_{k+1}(c,b) = \frac{C}{|\mathcal{I}(c)||\mathcal{I}(b)|} \sum_{j \in \mathcal{I}(b)} Partial^{s_k}_{\mathcal{I}(c)}(j)$$

Slow Convergence

- Existing Approach (VLDB J. '10)

$$\|\mathbf{S}_k - \mathbf{S}\|_{\max} \le C^{k+1}$$

Geometric Rate

For $C = 0.8$, $\epsilon = 0.0001$,

$K = \lceil \log_{0.8} 0.0001 \rceil = 41$ iterations.

- Our Approach

$$\|\hat{\mathbf{S}}_k - \hat{\mathbf{S}}\|_{\max} \le \frac{C^{k+1}}{(k+1)!}$$

Exponential Rate

For $C = 0.8$, $\epsilon = 0.0001$,
there are only 7 iterations.

Eliminating Partial Sums Redundancy

vertex	$\mathcal{I}(\star)$
a	$\{b,g\}$
e	$\{f,g\}$
h	$\{b,d\}$
c	$\{b,d,g\}$
b	$\{f,g,e,i\}$
d	$\{f,a,e,i\}$

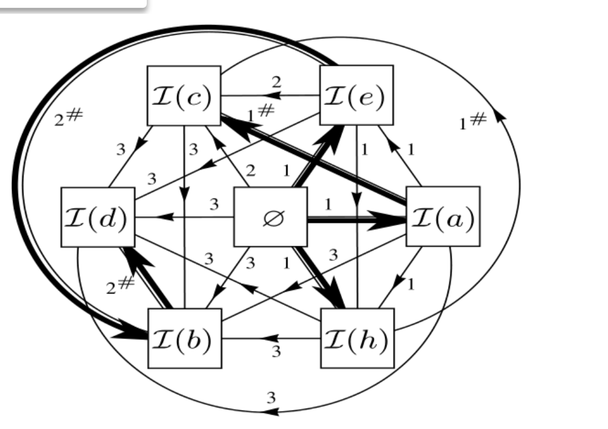
(a) In-neighbors in $\mathcal{G}$

(1)

	$\mathcal{I}(a)$	$\mathcal{I}(e)$	$\mathcal{I}(h)$	$\mathcal{I}(c)$	$\mathcal{I}(b)$	$\mathcal{I}(d)$
$\varnothing$	1	1	1	2	3	3
$\mathcal{I}(a)$		1	1	$1^{\#}$	3	3
$\mathcal{I}(e)$			1	2	$2^{\#}$	3
$\mathcal{I}(h)$				$1^{\#}$	3	3
$\mathcal{I}(c)$					3	3
$\mathcal{I}(b)$						$2^{\#}$

(b) Transition Costs (Edge Weights) in $\mathscr{G}$

(2)



(c) Minimum Spanning Tree $\mathscr{T}$ of $\mathscr{G}$

(3)

- (Inner) partial sums sharing

$$Partial^{s_k}_{\mathcal{I}(a)}(\star) = s_k(b,\star) + s_k(g,\star)$$

$$Partial^{s_k}_{\mathcal{I}(c)}(\star) = Partial^{s_k}_{\mathcal{I}(a)}(\star) + s_k(d,\star)$$

- Outer partial sums sharing

$$OuterPartial^{\mathcal{I}(\star),s_k}_{\mathcal{I}(a)} = \sum_{y \in \{b,g\}} Partial^{s_k}_{\mathcal{I}(\star)}(y)$$

$$OuterPartial^{\mathcal{I}(\star),s_k}_{\mathcal{I}(c)} = OuterPartial^{\mathcal{I}(\star),s_k}_{\mathcal{I}(a)} + Partial^{s_k}_{\mathcal{I}(\star)}(d)$$

$$s_{k+1}(a,\star) = \frac{C}{|\mathcal{I}(a)||\mathcal{I}(\star)|}\ OuterPartial^{\mathcal{I}(\star),s_k}_{\mathcal{I}(a)}$$

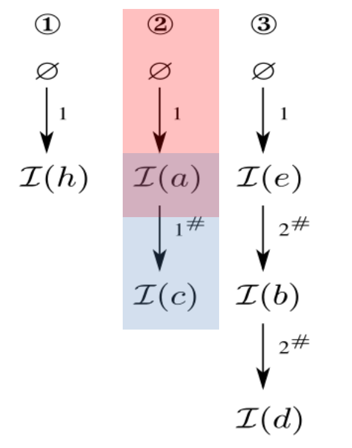
(5)

Partitions of $\mathcal{I}(\star)$ in $\mathcal{G}$

	$\mathscr{P}(\star)$
$\mathcal{I}(a)$	$\{\{b,g\}\}$
$\mathcal{I}(e)$	$\{\{f,g\}\}$
$\mathcal{I}(h)$	$\{\{b,d\}\}$
$\mathcal{I}(c)$	$\{\mathcal{I}(a),\{d\}\}$
$\mathcal{I}(b)$	$\{\mathcal{I}(e),\{e,i\}\}$
$\mathcal{I}(d)$	$\{\mathcal{I}(b)\backslash\{g\},\{a\}\}$

(4)

Partial Sums Order



From Geometric to Exponential Rate

$$\mathbf{S} = C \cdot (\mathbf{Q} \cdot \mathbf{S} \cdot \mathbf{Q}^T) + (1-C) \cdot \mathbf{I}_n \Rightarrow \mathbf{S} = (1-C) \cdot \sum_{i=0}^{\infty} C^i \cdot \mathbf{Q}^i \cdot (\mathbf{Q}^T)^i$$

Geometric Sum

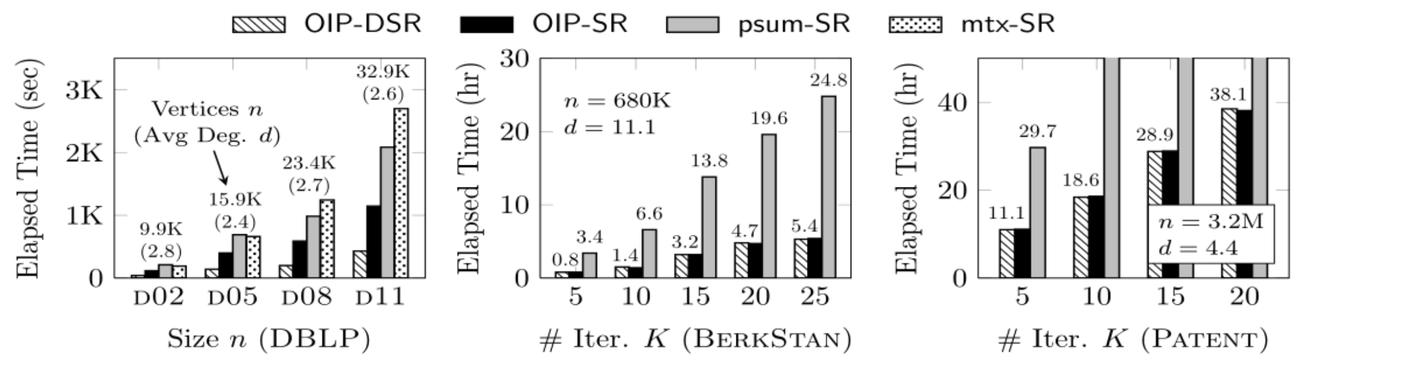
$$\frac{d\hat{\mathbf{S}}(t)}{dt} = \mathbf{Q} \cdot \hat{\mathbf{S}}(t) \cdot \mathbf{Q}^T, \quad \hat{\mathbf{S}}(0) = e^{-C} \cdot \mathbf{I}_n. \Leftarrow \hat{\mathbf{S}} = e^{-C} \cdot \sum_{i=0}^{\infty} \frac{C^i}{i!} \cdot \mathbf{Q}^i \cdot (\mathbf{Q}^T)^i$$

Exponential Sum

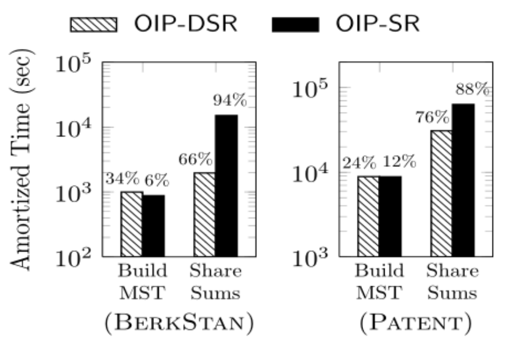
	(Geometric) SimRank	Exponential SimRank
Closed Form	$\mathbf{S} = C \cdot (\mathbf{Q} \cdot \mathbf{S} \cdot \mathbf{Q}^T) + (1-C) \cdot \mathbf{I}_n$	$\frac{d\hat{\mathbf{S}}(t)}{dt} = \mathbf{Q} \cdot \hat{\mathbf{S}}(t) \cdot \mathbf{Q}^T, \quad \hat{\mathbf{S}}(0) = e^{-C} \cdot \mathbf{I}_n.$
Series Form	$\mathbf{S} = (1-C) \cdot \sum_{i=0}^{\infty} C^i \cdot \mathbf{Q}^i \cdot (\mathbf{Q}^T)^i$	$\hat{\mathbf{S}} = e^{-C} \cdot \sum_{i=0}^{\infty} \frac{C^i}{i!} \cdot \mathbf{Q}^i \cdot (\mathbf{Q}^T)^i$
Iterative Form	$\mathbf{S}_0 = \mathbf{I}_n$ $\mathbf{S}_{k+1} = C \cdot (\mathbf{Q} \cdot \mathbf{S}_k \cdot \mathbf{Q}^T) + (1-C) \cdot \mathbf{I}_n$	$\begin{cases} \mathbf{T}_{k+1} = \mathbf{Q} \cdot \mathbf{T}_k \cdot \mathbf{Q}^T \\ \hat{\mathbf{S}}_{k+1} = \hat{\mathbf{S}}_k + e^{-C} \cdot \frac{C^{k+1}}{(k+1)!} \cdot \mathbf{T}_{k+1} \end{cases} \quad \begin{cases} \mathbf{T}_0 = \mathbf{I}_n \\ \hat{\mathbf{S}}_0 = e^{-C} \cdot \mathbf{I}_n \end{cases}$
Error	$\|\mathbf{S}_k - \mathbf{S}\|_{\max} \le C^{k+1}$	$\|\hat{\mathbf{S}}_k - \hat{\mathbf{S}}\|_{\max} \le \frac{C^{k+1}}{(k+1)!}$

Experimental Evaluations

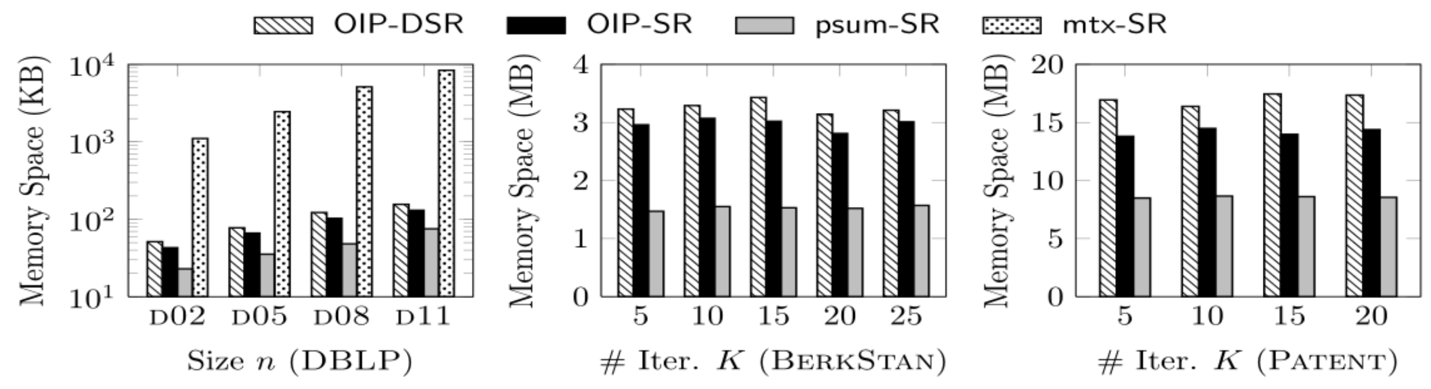
Time & Space Efficiency

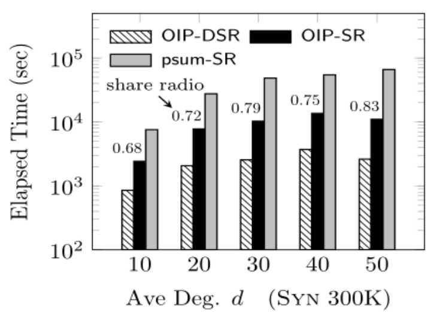


(a) Time Efficiency on Real Datasets



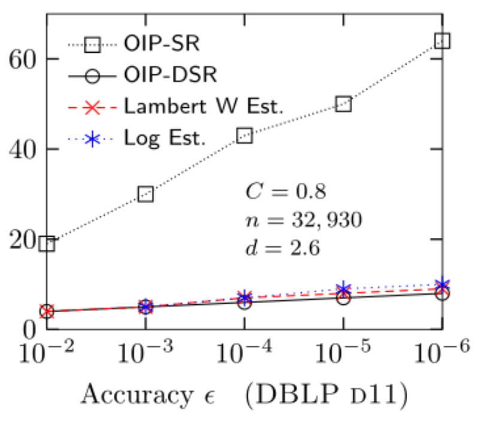
(b) Amortized Time on Real Data





(c) Effect of Density

Convergence Rate

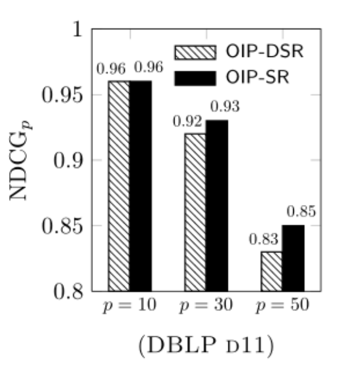


(e) Convergence Rate

Err ϵ	OIP-SR	OIP-DSR	LamW Est.	Log Est.
10^{-2}	19	4	4	-
10^{-3}	30	5	5	5
10^{-4}	43	6	7	7
10^{-5}	50	7	8	9
10^{-6}	64	8	9	10

(f) Lam W & Log Bound on K

Relative Order Preservation



(g) Relative Ordering

#	Co-authors	#	Co-authors
1	Hongjun Lu	16	Aoying Zhou
2	Lu Qin	17	Xiang Lian
3	Xuemin Lin	18	Cheqing Jin
4	Wei Wang	19	Baichen Chen
5	Lei Chen	20	Byron Choi
6	Lijun Chang	21	Wenfei Fan
7	Yiping Ke	22	Rong-Hua Li
8	Haifeng Jiang	23	**Hong Cheng ▼**
9	Philip S. Yu	24	**Jun Gao ▲**
10	Gabriel Pui Cheong Fung	25	Xiaofang Zhou
11	James Cheng	26	Ke Yi
12	Weifa Liang	27	Yufei Tao
13	Ying Zhang	28	Nan Tang
14	Bolin Ding	29	Jinsoo Lee
15	Haixun Wang	30	Kam-Fai Wong

(h) Top-30 Co-authors of "Jeffrey Xu Yu"