Reading

More is Simpler: Effectively and Efficiently Assessing Node-Pair Similarities Based on Hyperlinks

Weiren Yu[†♯], Xuemin Lin[♭†], Wenjie Zhang[†], Lijun Chang[†], Jian Pei[♮]

[†]The University of New South Wales, Australia [♭]East China Normal University, China
[♯]NICTA, Australia [♮]Simon Fraser University, Canada

{weirenyu, lxue, zhangw, ljchang}@cse.unsw.edu.au jpei@cs.sfu.ca

Paper from PVLDB vol.7
(To appear in VLDB 2014)

ERATO Seminar 2013/11/28

Presenter: Kazuhiro Inaba

論文概要

- SimRank [Jeh and Widom 2002] を (similarity の評価尺度として) 改良した **SimRank*** を提案する
- 計算速度も速い

おさらい: SimRank

$$s(a,b) = \begin{cases} 1, & a = b \\ \frac{C}{|\mathcal{I}(a)||\mathcal{I}(b)|} \sum_{j \in \mathcal{I}(b)} \sum_{i \in \mathcal{I}(a)} s(i,j), & a \neq b \end{cases}$$

- $\mathcal{I}(v) = \{u \mid u \rightarrow v\}$
- C = 0.6 ～ 0.8
- 「似ているノードからリンクされているノードは似ている」

SimRank (別の書き方)

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

あるいは

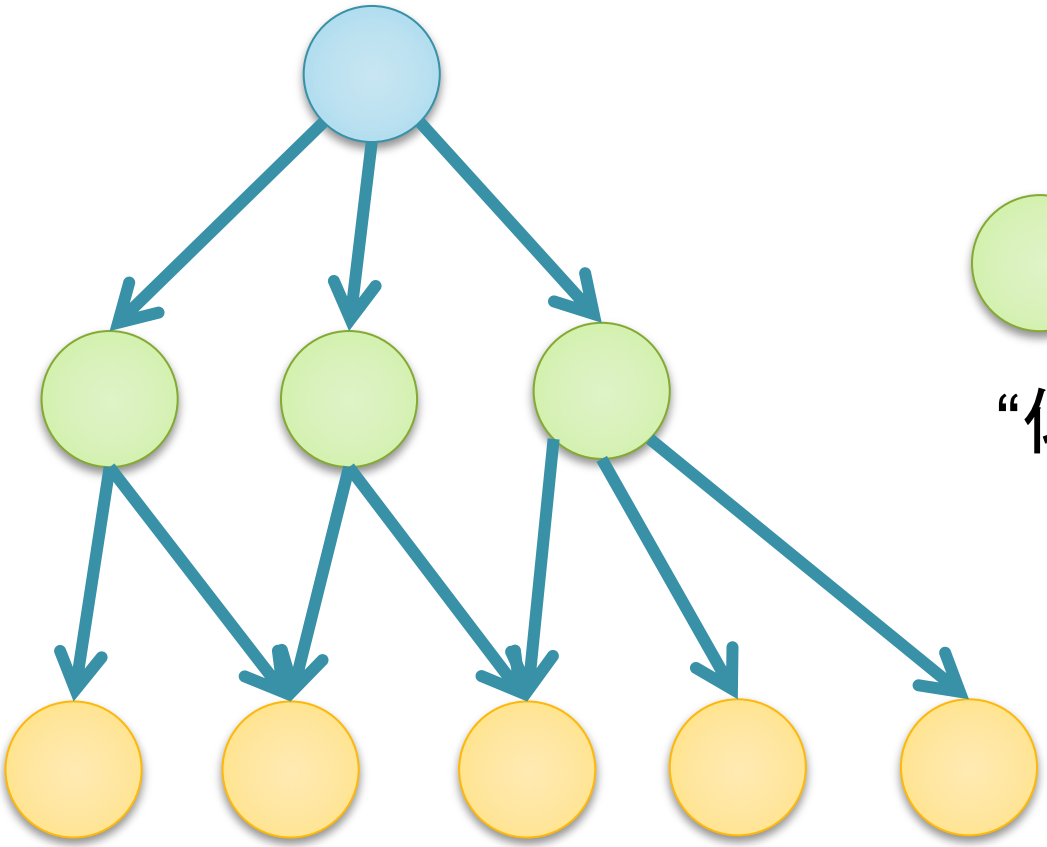
※発表中に、この式は会議論文レベルでもよく流布している間違い（(1-C)を一律に足すのでは対角成分が1に戻らない）という突込みがありました

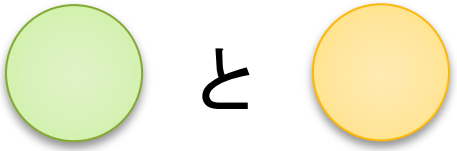
$$\mathbf{S} = (1 - C) \cdot \sum_{l=0}^{\infty} C^l \cdot \mathbf{Q}^l \cdot (\mathbf{Q}^T)^l$$

- $S_{ab} = s(a, b)$
- $Q_{ab} = 1 / |\mathcal{I}(a)|$ if b→a or 0 otherwise

SimRank の何が不満なのか？

$$s(a,b) = \begin{cases} 1, & a = b \\ \frac{C}{|\mathcal{I}(a)||\mathcal{I}(b)|} \sum_{j \in \mathcal{I}(b)} \sum_{i \in \mathcal{I}(a)} s(i,j), & a \neq b \end{cases}$$





と

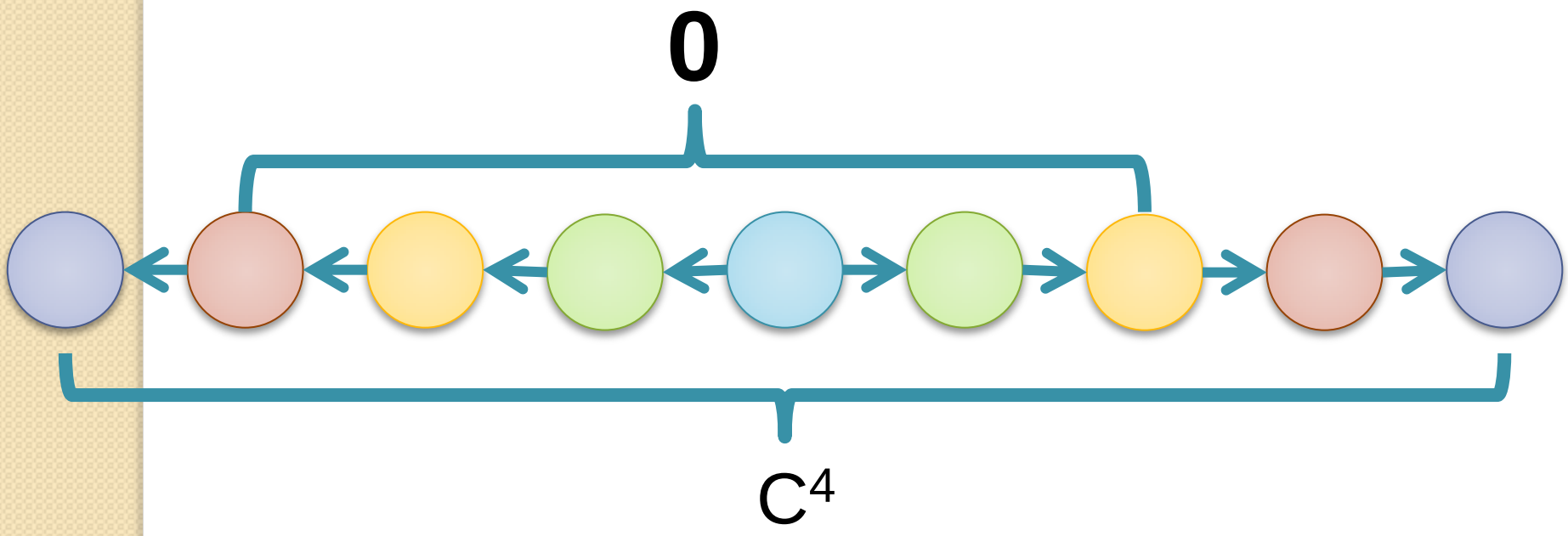
は

“似ていない”

つまり

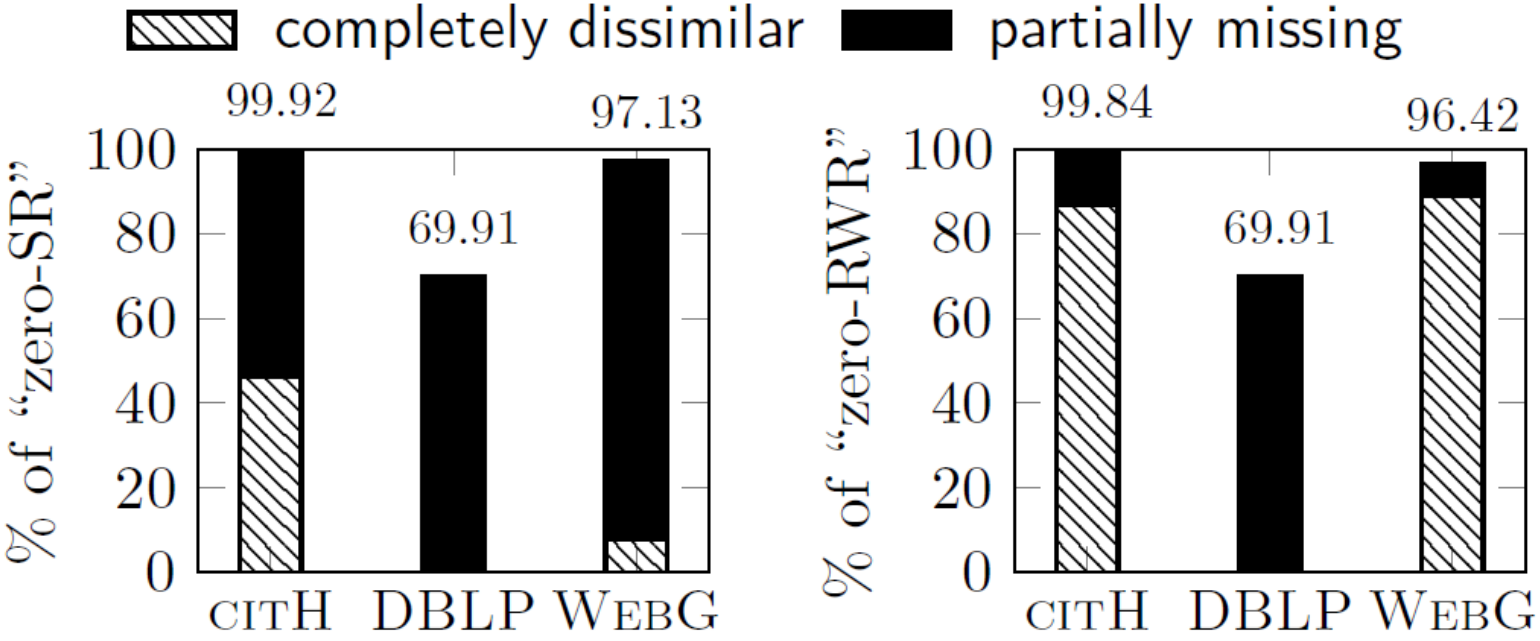
Theorem 1

For any two nodes a and b in G, $s(a, b) = 0$ if there does not exist a node c that has directed paths $c \rightarrow^k a$ and $c \rightarrow^k b$ of the same length.



“Zero-Similarity”

Node-Pair のうち dissymmetric source を持つものの割合



SimRank

$$\mathbf{S} = (1 - C) \cdot \sum_{l=0}^{\infty} C^l \cdot \mathbf{Q}^l \cdot (\mathbf{Q}^T)^l$$

g-SimRank* (提案手法1)

$$\hat{\mathbf{S}} = (1 - C) \cdot \sum_{l=0}^{\infty} \frac{C^l}{2^l} \cdot \sum_{\alpha=0}^{l} \binom{l}{\alpha} \cdot \mathbf{Q}^{\alpha} \cdot (\mathbf{Q}^T)^{l-\alpha}$$

e-SimRank* (提案手法2)

$$\hat{\mathbf{S}}' = e^{-C} \cdot \sum_{l=0}^{\infty} \frac{C^l}{l!} \cdot \frac{1}{2^l} \sum_{\alpha=0}^{l} \binom{l}{\alpha} \cdot \mathbf{Q}^{\alpha} \cdot (\mathbf{Q}^T)^{l-\alpha}$$

SimRank*

$$\hat{\mathbf{S}} = (1-C)\cdot\sum_{l=0}^{\infty}\frac{C^l}{2^l}\cdot\sum_{\alpha=0}^{l}\binom{l}{\alpha}\cdot\mathbf{Q}^{\alpha}\cdot(\mathbf{Q}^{T})^{l-\alpha}$$

- 長さの違うパス (α と l – α) にも非0重み
- 二項係数なので長さが近いほど重要

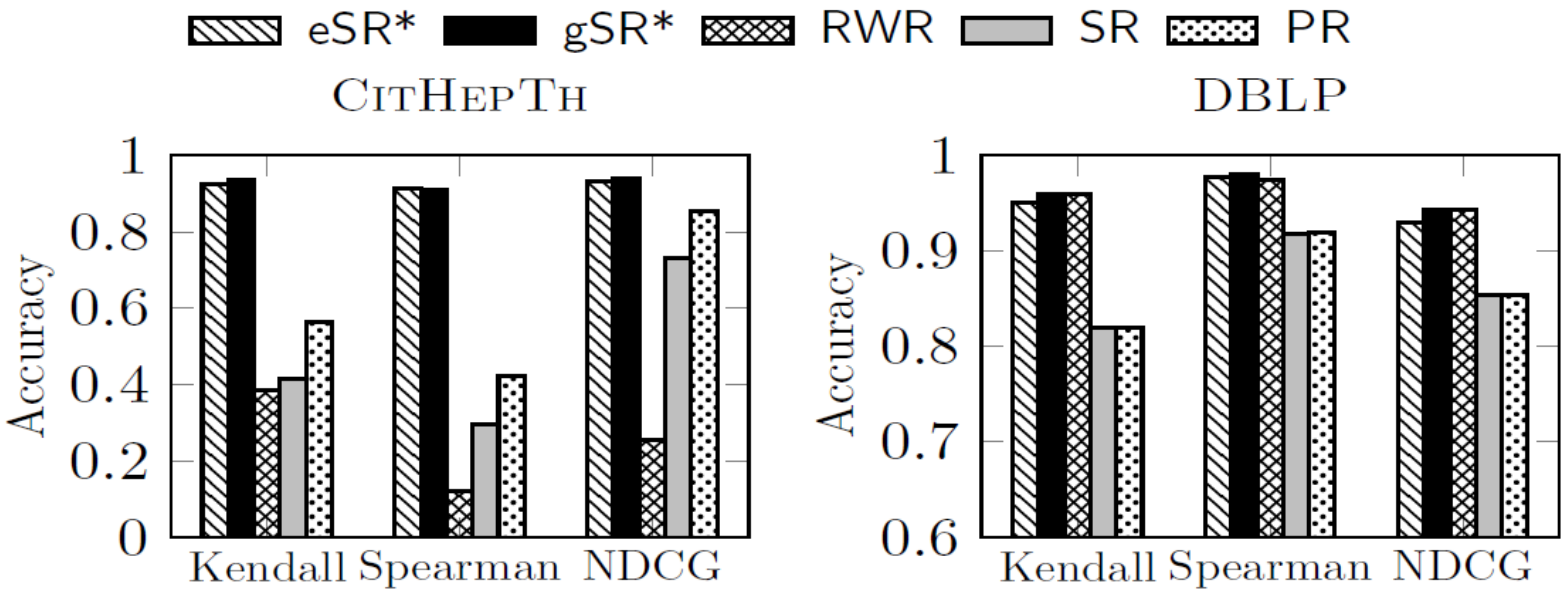
			1					/ 1
		1		1				/ 2
	1		2		1			/ 4
1		3		3		1		/ 8
1		4		6		4	1	/ 16

類似度の質に関する実験 (1)
“Ground Truth” との比較

Citation

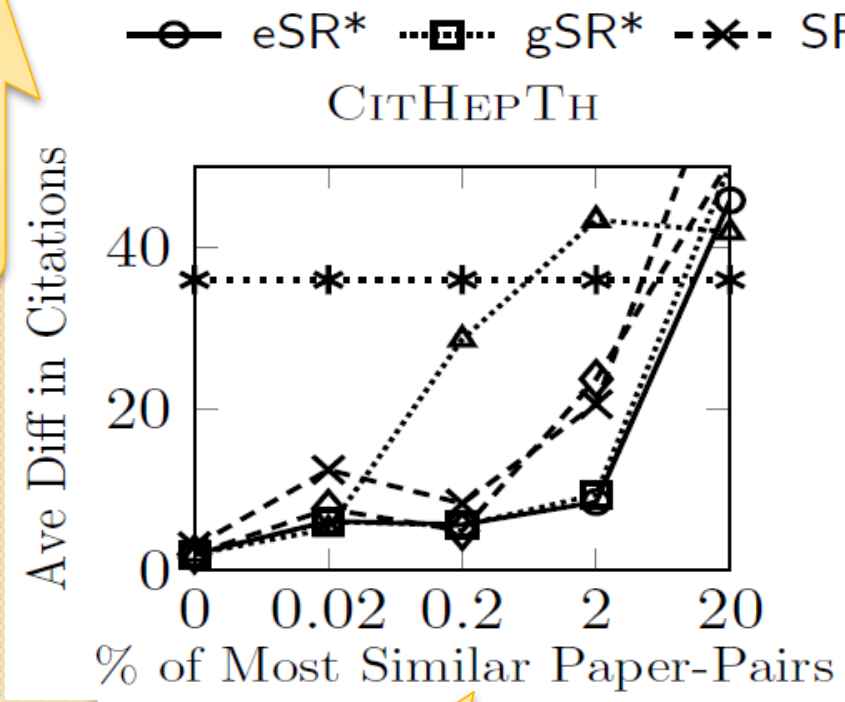
Dataset	$\|\mathcal{G}\|$ $(\|\mathcal{V}\|, \|\mathcal{E}\|)$	Density $(\|\mathcal{E}\|/\|\mathcal{V}\|)$
CitHepTh	451K (33K, 418K)	12.6
DBLP	102K (15K, 87K)	5.8

Coauthorship in
{SIGMOD,PODS,VLDB,ICDE,SIGKDD,SIGIR,WWW}



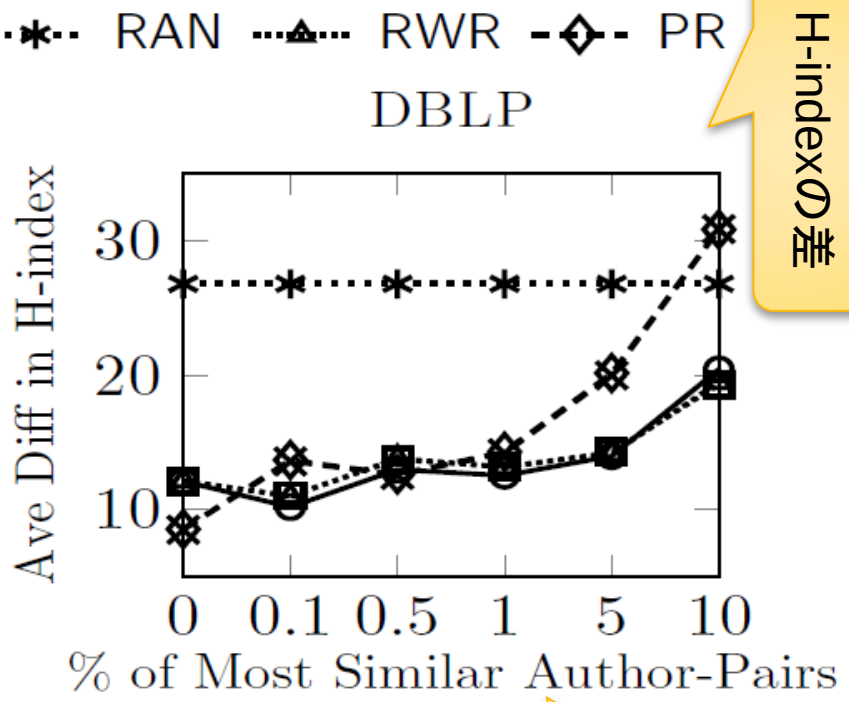
類似度の質に関する実験 (2)
類似ペアがどこまで「似ている」か

引用数の差



似ている論文ペア

H-indexの差

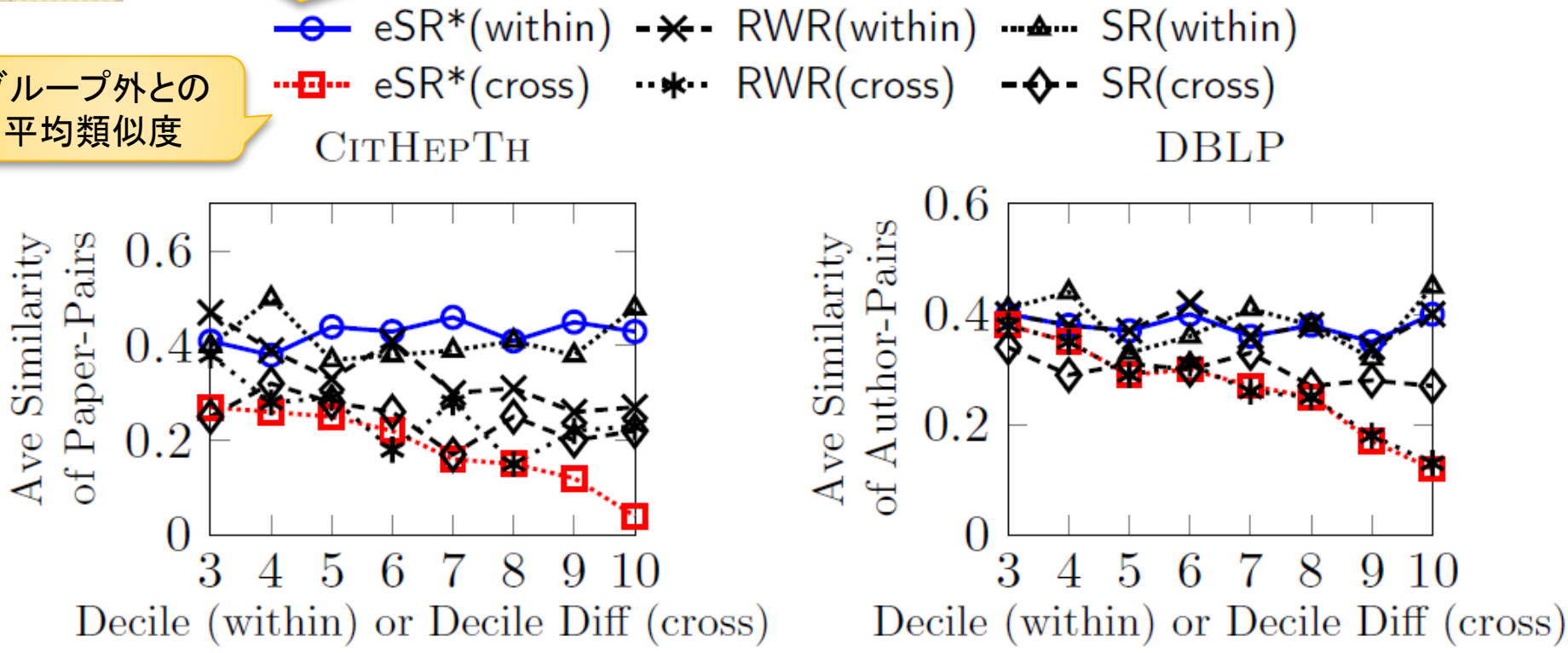


似ている著者ペア

類似度の質に関する実験 (3)

グループ間平均類似度

グループ外との平均類似度



引用数 / H-index の順でノードを10グループに分類

SIMRANK* の計算手法

定理 (g-SimRank*)

$$\hat{\mathbf{S}} = (1-C)\cdot\sum_{l=0}^{\infty}\frac{C^l}{2^l}\cdot\sum_{\alpha=0}^{l}\binom{l}{\alpha}\cdot\mathbf{Q}^{\alpha}\cdot(\mathbf{Q}^T)^{l-\alpha}$$

iff

$$\hat{\mathbf{S}} = \frac{C}{2}\cdot(\mathbf{Q}\cdot\hat{\mathbf{S}}+\hat{\mathbf{S}}\cdot\mathbf{Q}^T)+(1-C)\cdot\mathbf{I}_n$$

証明: 確認してみればわかる。

SimRank

$$\mathbf{S} = (1 - C) \cdot \sum_{l=0}^{\infty} C^l \cdot \mathbf{Q}^l \cdot (\mathbf{Q}^T)^l$$

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

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

g-SimRank*

$$\hat{\mathbf{S}} = (1 - C) \cdot \sum_{l=0}^{\infty} \frac{C^l}{2^l} \cdot \sum_{\alpha=0}^{l} \binom{l}{\alpha} \cdot \mathbf{Q}^\alpha \cdot (\mathbf{Q}^T)^{l-\alpha}$$

$$\hat{\mathbf{S}} = \frac{C}{2} \cdot (\mathbf{Q} \cdot \hat{\mathbf{S}} + \hat{\mathbf{S}} \cdot \mathbf{Q}^T) + (1 - C) \cdot \mathbf{I}_n$$

$$\hat{s}(a,b) = \frac{C}{2|\mathcal{I}(b)|} \sum_{y \in \mathcal{I}(b)} \hat{s}(a,y) + \frac{C}{2|\mathcal{I}(a)|} \sum_{x \in \mathcal{I}(a)} \hat{s}(x,b)$$

SimRank

行列乗算が2回（Q･S･Q^T）

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

g-SimRank*

行列乗算が1回
(S^ は対称なので一方は他方の転置)

$$\hat{\mathbf{S}} = \frac{C}{2} \cdot (\mathbf{Q} \cdot \hat{\mathbf{S}} + \hat{\mathbf{S}} \cdot \mathbf{Q}^T) + (1 - C) \cdot \mathbf{I}_n$$

さらに効率的な計算

$$\hat{s}_{k+1}(a,b) = \frac{C}{2|\mathcal{I}(b)|} \sum_{y \in \mathcal{I}(b)} \hat{s}_k(a,y) + \cdots$$

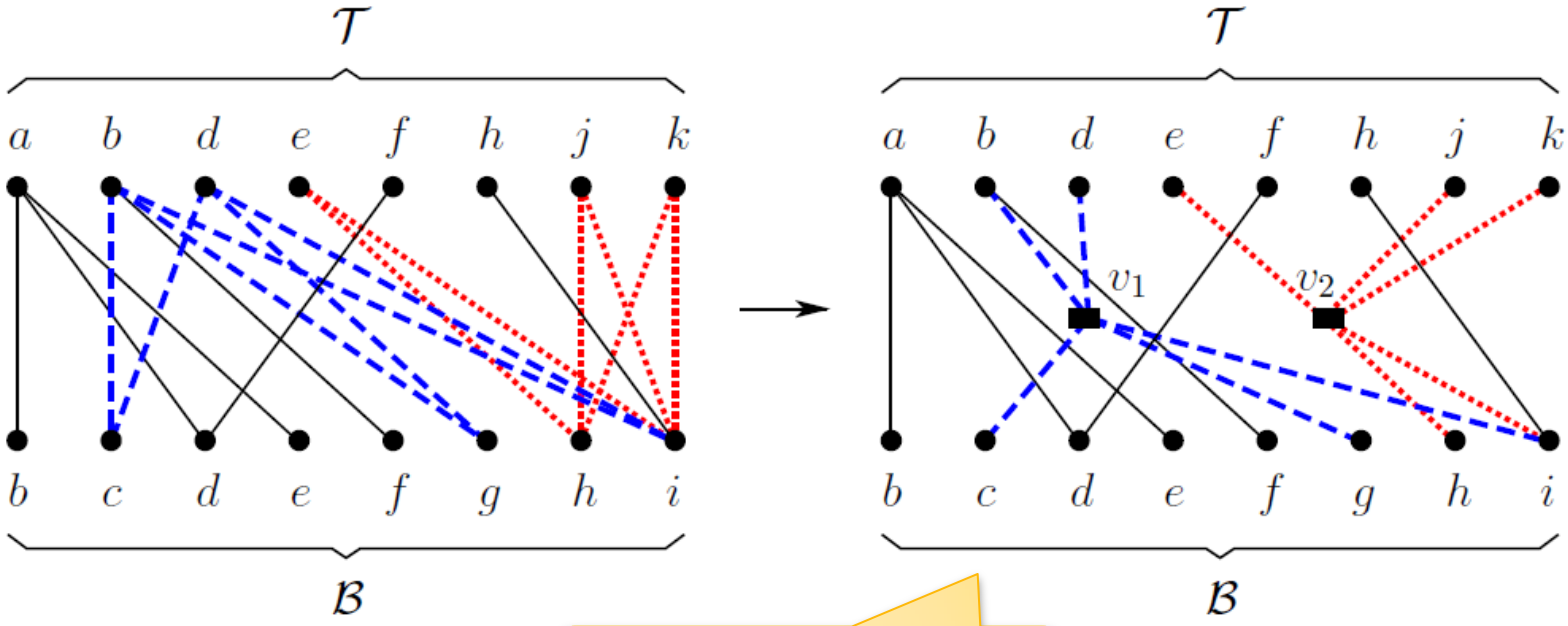
$$= \sum_{\biguplus \Delta = \mathcal{I}(b)} \sum_{y \in \Delta} \hat{s}_k(a,y)$$

$\Sigma_{\mathcal{I}(b)}$ を直接計算するのではなく、
部分集合 Δ に分割して計算。
$\mathcal{I}$(a), $\mathcal{I}$(b), … で共通する部分のΣは再利用

さらに効率的な計算

元のグラフの二部グラフ表現を圧縮する

- T = B = V (元のグラフのノード集合)
- $v_T \rightarrow u_B$ iff $v \rightarrow u \in E$



部分完全二部グラフを
見つけて、置き換える

$\Delta_i = I(v_i)$ として
Σの計算を再利用

速度実験

Subsets of DBLP

Web

Cite

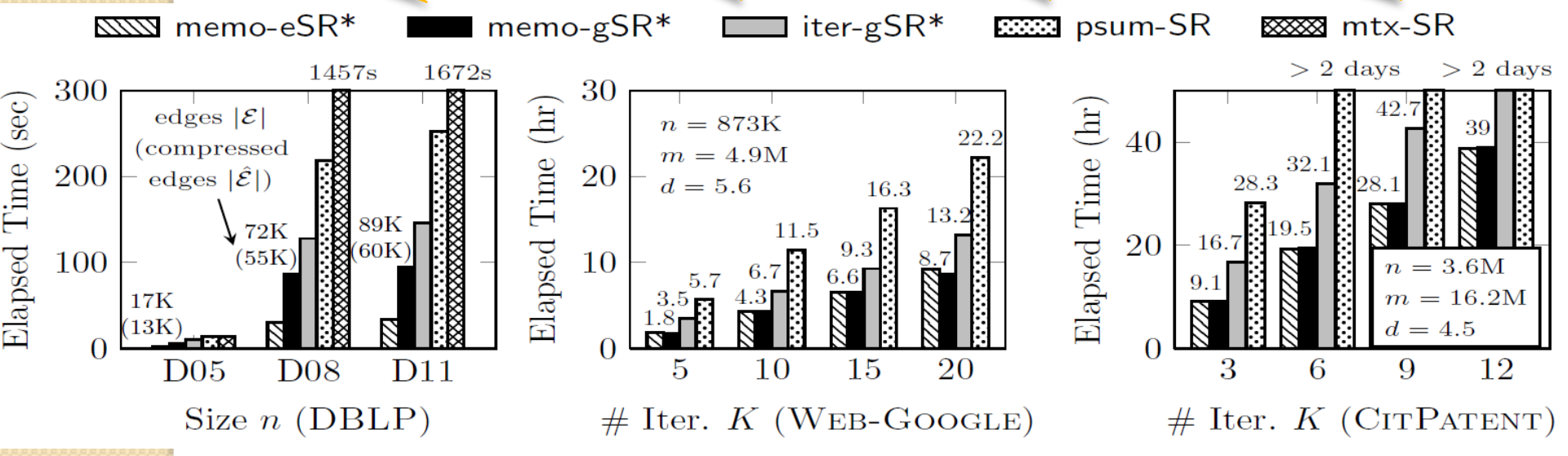
Dataset	$\|\mathcal{G}\|$ ($\|\mathcal{V}\|$, $\|\mathcal{E}\|$)	Density ($\|\mathcal{E}\|/\|\mathcal{V}\|$)
D05	21K (4K, 17K)	4.3
D08	85K (13K, 72K)	5.5
D11	103K (14K, 89K)	6.3
WEB-GOOGLE	5.8M (873K, 4.9M)	5.6
CITPATENT	19.8M (3.6M, 16.2M)	4.5

提案手法（メモ化）

提案手法（naive）

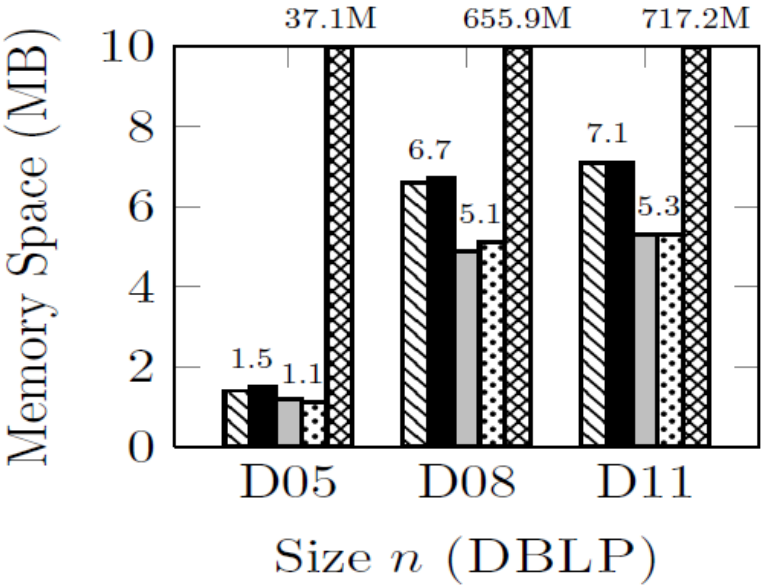
SimRank（メモ化）

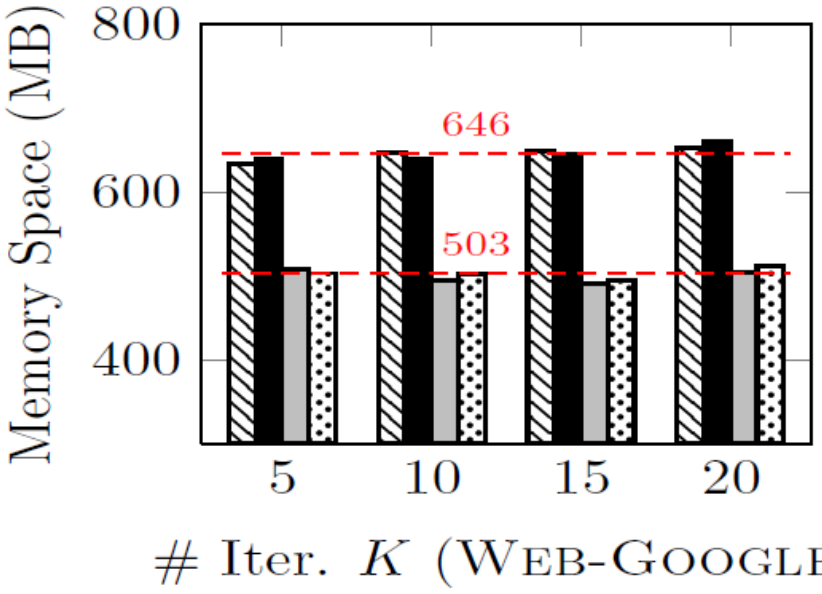
SimRank（特異値分解）

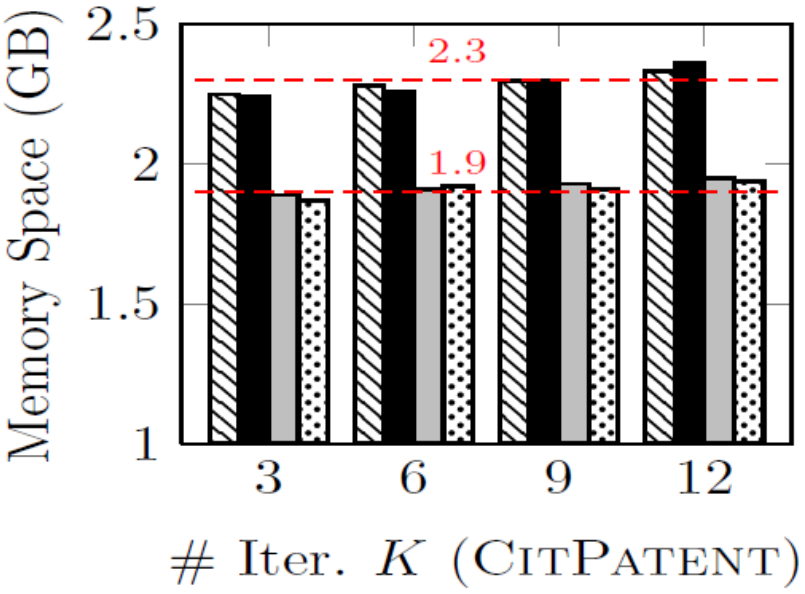


メモリ使用量の実験


memo-eSR*
memo-gSR*
iter-gSR*
psum-SR
mtx-SR


37.1M
655.9M
717.2M
Memory Space (MB)
10
8
6
4
2
0
1.5
1.1
6.7
5.1
7.1
5.3
D05
D08
D11
Size n (DBLP)


Memory Space (MB)
800
600
400
646
503
5
10
15
20
Iter. K (WEB-GOOGLE)


Memory Space (GB)
2.5
2
1.5
1
2.3
1.9
3
6
9
12
Iter. K (CITPATENT)

まとめ

SimRank* という類似度尺度を提案

$$\hat{\mathbf{S}} = (1-C) \cdot \sum_{l=0}^{\infty} \frac{C^l}{2^l} \cdot \sum_{\alpha=0}^{l} \binom{l}{\alpha} \cdot \mathbf{Q}^{\alpha} \cdot (\mathbf{Q}^T)^{l-\alpha}$$

$$\hat{\mathbf{S}} = \frac{C}{2} \cdot (\mathbf{Q} \cdot \hat{\mathbf{S}} + \hat{\mathbf{S}} \cdot \mathbf{Q}^T) + (1-C) \cdot \mathbf{I}_n$$

疑問点

- Out-Link を使った手法 (P-Rank 等) との組み合わせは可能か？
- H-index や 引用数ではなく、内容に関する類似度と比較しての評価は？
- SimRank と SimRank* のハイブリッドの用な計算は可能か？（リンクの種類による切替）