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

Weiren Yu [1], Xuemin Lin [1], Wenjie Zhang [1], Lijun Chang [1], Jian Pei [2]

[1] University of New South Wales, Australia
[2] Simon Fraser University, Canada

Outline

Overview

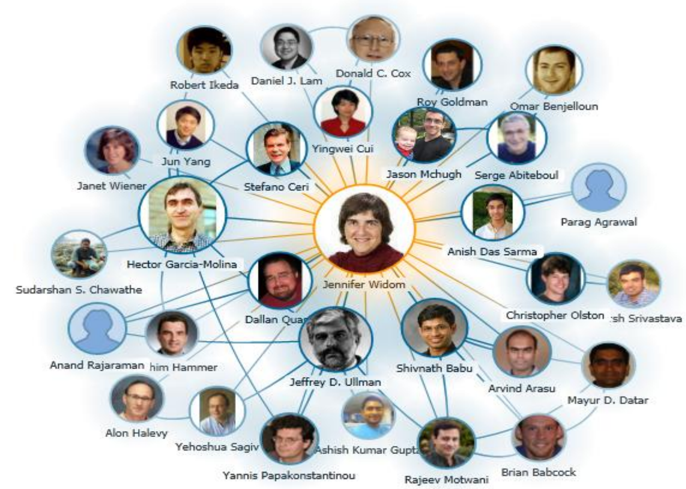
- The existing "zero-SimRank" problem
- Our approaches
  - SimRank*, a semantically-enhanced version
  - Two succinct closed forms of SimRank*
  - Edge concentration for speeding up computation
- Empirical evaluations
- Conclusions

Overview

- SimRank plays an important part in real applications.

Recommender System

Citation Graph



Collaboration Network

SimRank Overview

- SimRank
  - An attractive similarity measure based on hyperlinks, (proposed by Jeh and Widom in KDD '02)
  - Basic philosophy

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

- Two SimRank models
  - Basic 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 model (KDD '10)

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

Existing Link-based Measure

- PageRank

$$\mathbf{p} = C \cdot \mathbf{W}^T \cdot \mathbf{p} + (1-C) \cdot \mathbf{1}$$

vector of all 1s

- Personalized PageRank

$$\mathbf{p} = C \cdot \mathbf{W}^T \cdot \mathbf{p} + (1-C) \cdot \mathbf{q}$$

personalized vector

- Random Walk with Restart

$$\mathbf{p} = C \cdot \mathbf{W}^T \cdot \mathbf{p} + (1-C) \cdot \mathbf{e}_i$$

unit vector

- SimRank

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

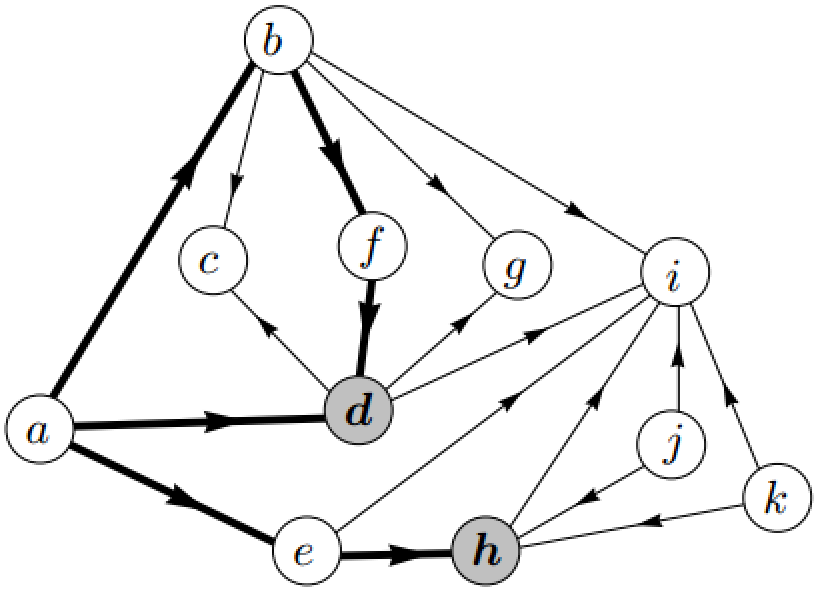
identity matrix

$$\mathbf{S} = C \cdot \mathbf{Q} \cdot \mathbf{S} \cdot \mathbf{Q}^T + \mathbf{D}$$

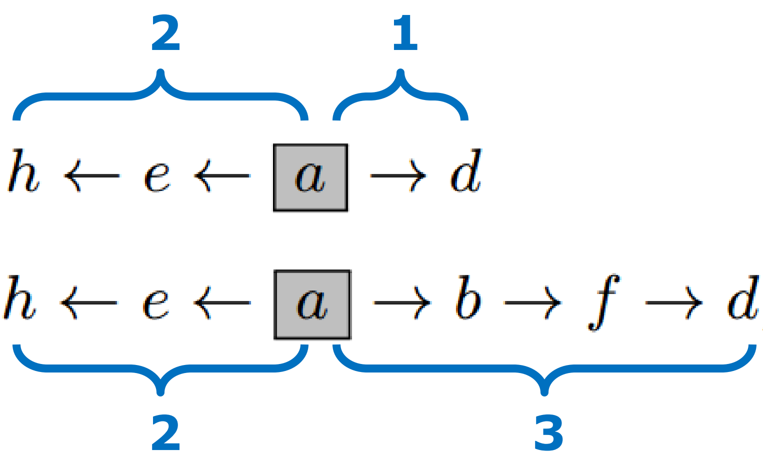
diagonal matrix

Motivation

- "Zero-Similarity" Problem



Node-Pairs	SR	PR	SR*	RWR
(h, d)	0	.049	.010	0
(a, f)	0	.075	.032	.032
(a, c)	0	0	.025	.024
(g, a)	0	0	.025	0
(g, b)	0	0	.075	0
(i, a)	0	0	.015	0
(i, h)	.044	.041	.031	0



Simrank (h,d) =0 !!

There are no nodes **with equal distance** to nodes h and d

Motivation

- Zero-Similarity" Problem

$$a_{-n} \leftarrow \cdots \leftarrow a_{-1} \leftarrow a_0 \rightarrow a_1 \rightarrow \cdots \rightarrow a_n$$

$$s(a_i, a_j) = 0, \text{ for all } |i| \neq |j|$$

Simrank (a_i, a_j) =0 (for all |i| !=|j|)

There are no nodes **with equal distance** to nodes a_i and a_j

Zero-Similarity” Problem

- Power of Adjacency Matrix A
  - The (x, j)-entry of $\mathbf{A}^l$ counts # of paths: $\underbrace{x \to \circ \to \circ \cdots \circ \to j}_{l \text{ length}}$
  - The (i, x)-entry of $(\mathbf{A}^T)^l$ counts # of paths: $\underbrace{i \leftarrow \circ \leftarrow \circ \cdots \circ \leftarrow x}_{l \text{ length}}$
  - The value of $\sum_{k=1}^{\infty} [(\mathbf{A}^T)^k \cdot \mathbf{A}^k]_{i,j}$ counts # of paths:

$$\underbrace{i \leftarrow \circ \leftarrow \circ \cdots \circ \leftarrow}_{k \text{ length}} \boxed{\circ} \underbrace{\to \circ \cdots \circ \to \circ \to j}_{k \text{ length}}$$

$\mathbf{Q} = rowNorm(\mathbf{A}^T)$

- SimRank series form:

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

Sim (i, j) = 0 if there are no nodes with equal length to (i, j)

A Remedy for SimRank

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

Length	SimRank	RWR / PPR	α	SimRank*
1	N/A	$\boxed{i} \to j$	0	$\boxed{i} \to j$
			1	$i \leftarrow \boxed{j}$
2	$i \leftarrow \boxed{\bullet} \to j$	$\boxed{i} \to \circ \to j$	0	$\boxed{i} \to \circ \to j$
			1	$i \leftarrow \boxed{\bullet} \to j$
			2	$i \leftarrow \circ \leftarrow \boxed{j}$
3	N/A	$\boxed{i} \to \circ \to \circ \to j$	0	$\boxed{i} \to \circ \to \circ \to j$
			1	$i \leftarrow \boxed{\bullet} \to \circ \to j$
			2	$i \leftarrow \circ \leftarrow \boxed{\bullet} \to j$
			3	$i \leftarrow \circ \leftarrow \circ \leftarrow \boxed{j}$
4	$i \leftarrow \circ \leftarrow \boxed{\bullet} \to \circ \to j$	$\boxed{i} \to \circ \to \circ \to \circ \to j$	0	$\boxed{i} \to \circ \to \circ \to \circ \to j$
			1	$i \leftarrow \boxed{\bullet} \to \circ \to \circ \to j$
			2	$i \leftarrow \circ \leftarrow \boxed{\bullet} \to \circ \to j$
			3	$i \leftarrow \circ \leftarrow \circ \leftarrow \boxed{\bullet} \to j$
			4	$i \leftarrow \circ \leftarrow \circ \leftarrow \circ \leftarrow \boxed{j}$

$\circ$ – any node in $\mathcal{G}$ $\boxed{i}$, $\boxed{\bullet}$, $\boxed{j}$ – in-link "source"

Two Kinds of Weighted Coefficients

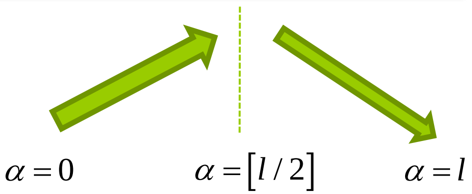
- SimRank* :

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

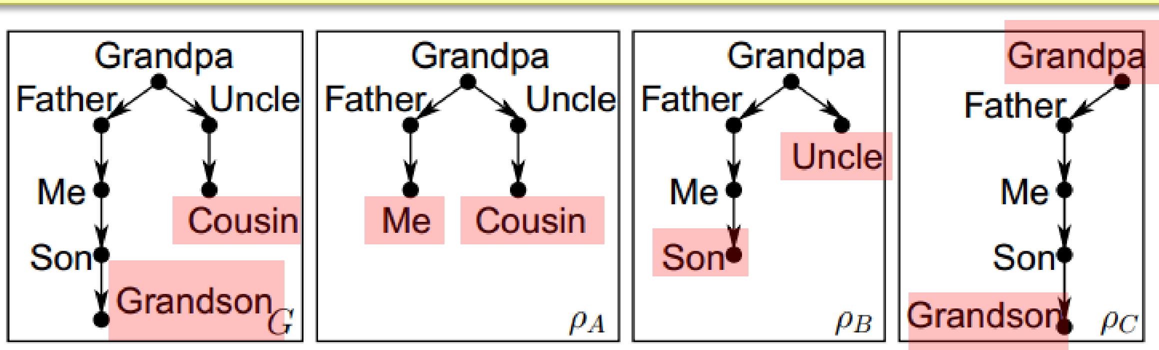
  - Length weights: $\{C^l\}_{l=0}^{\infty}$ is decreasing w.r.t. l (0<C<1)

Longer **paths should have a** ***smaller*** **contribution to S**

  - Symmetry weights: $\{\binom{l}{\alpha}\}_{\alpha=0}^{l}$ (binomial)



More symmetric **paths should have a** ***larger*** **contribution to S**



Variations of SimRank*

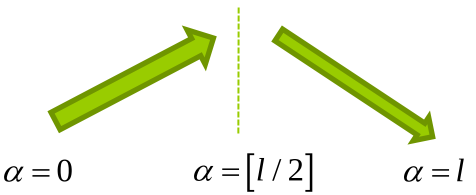
- SimRank* : $$\hat{\mathbf{S}} = (1-C) \cdot \sum_{l=0}^{\infty} \frac{C^l}{2^l} \cdot \sum_{\alpha=0}^{l} \binom{l}{\alpha} \cdot \mathbf{Q}^{\alpha} \cdot (\mathbf{Q}^T)^{l-\alpha}$$

$$\|\mathbf{Q}^{l_1} \cdot (\mathbf{Q}^T)^{l_2}\|_{\max} \leq 1, \text{ for } \forall l_1, l_2$$

  - Length weights: $\{C^l\}_{l=0}^{\infty}$

$$\sum_{l=0}^{\infty} C^l = \frac{1}{1-C}$$

  - Symmetry weights: $\{\binom{l}{\alpha}\}_{\alpha=0}^{l}$ (binomial)



$$\sum_{\alpha=0}^{l} \binom{l}{\alpha} = 2^l$$

Why not use $e^{-(l-\frac{\alpha}{2})^2}$?

$\sum_{\alpha=0}^{l} e^{-(l-\frac{\alpha}{2})^2}$ is not a simple form for normalization

Variations of SimRank*

- SimRank* :

Geometric version

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

- Length weights: C^l

$$\sum_{l=0}^{\infty} C^l = \frac{1}{1-C}$$

Can we use $\frac{C^l}{l!}$?

$$\sum_{l=0}^{\infty} \frac{C^l}{l!} = e^C$$

is a simple form for normalization

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

Exponential version

Convergence of SimRank*

- The first k-th partial sums:

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

- (Geometric) convergence:

$$\|\hat{\mathbf{S}} - \hat{\mathbf{S}}_k\|_{\max} \leq C^{k+1}. \quad (\forall k = 0, 1, \cdots)$$

- (Exponential) convergence:

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

$$\|\hat{\mathbf{S}}' - \hat{\mathbf{S}}'_k\|_{\max} \leq \frac{C^{k+1}}{(k+1)!}. \quad (\forall k = 0, 1, \cdots)$$

Recursive Form of SimRank*

SimRank recursion

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

SimRank series form

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

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

SimRank* series form

SimRank* resursion ?

- Geometric SimRank* has the following recursive form:

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

Closed Form of Exponential SimRank*

SimRank closed form

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

SimRank series form

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

SimRank* series form

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

SimRank* closed form?

- Exponential SimRank* has the following closed form:

$$\hat{\mathbf{S}}' = e^{-C}\cdot e^{\frac{C}{2}\mathbf{Q}}\cdot e^{\frac{C}{2}\mathbf{Q}^T}$$

where $e^{\mathbf{X}} = \sum_{k=0}^{\infty}\frac{\mathbf{X}^k}{k!}$

SimRank* Computation

- Iterative Model:

$$\begin{cases} \hat{\mathbf{S}}_0 = (1-C)\cdot \mathbf{I}_n, \\ \hat{\mathbf{S}}_{k+1} = \frac{C}{2}\cdot(\mathbf{Q}\cdot\hat{\mathbf{S}}_k + \hat{\mathbf{S}}_k\cdot\mathbf{Q}^T) + (1-C)\cdot\mathbf{I}_n \end{cases}$$

- Entry-wise Form:

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

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

If $\mathcal{I}(b)\cap\mathcal{I}(\star)\neq\varnothing$ **, then**

$$\text{Partial}_{\Delta}^{\hat{s}_k}(a) \triangleq \sum_{y\in\Delta}\hat{s}_k(a,y) \text{ with } \Delta\subseteq\mathcal{I}(\star)\cap\mathcal{I}(b)$$

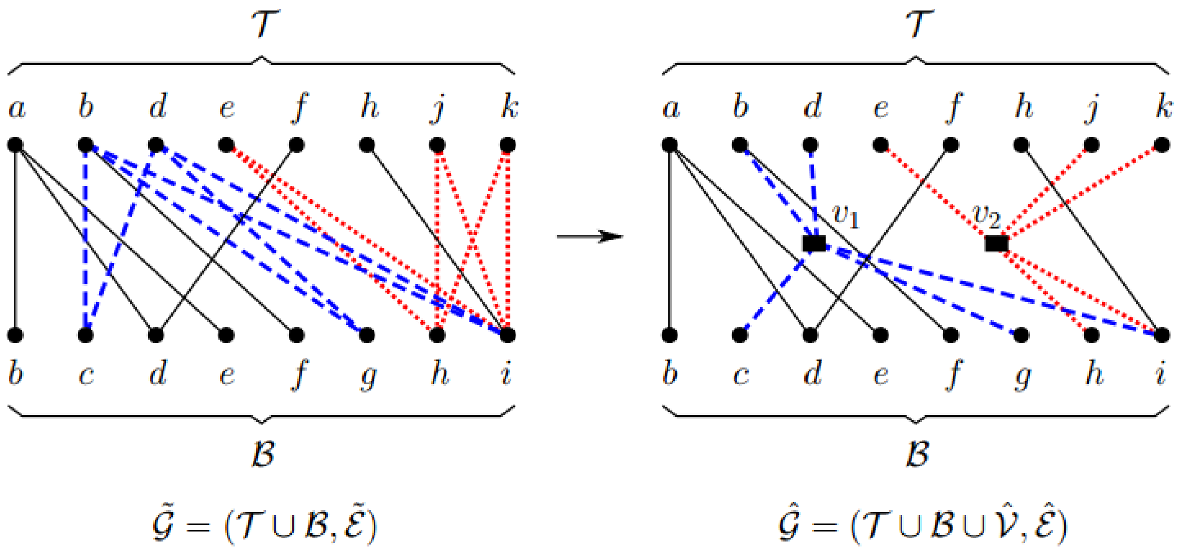
can be memoized for subsequent reuse.

Fine-grained Memoization

- How to find Δ in general case for maximal sharing?

$$\text{Partial}^{\hat{s}_k}_{\Delta}(a) \triangleq \sum_{y\in\Delta} \hat{s}_k(a,y) \text{ with } \Delta \subseteq \mathcal{I}(\star) \cap \mathcal{I}(b)$$

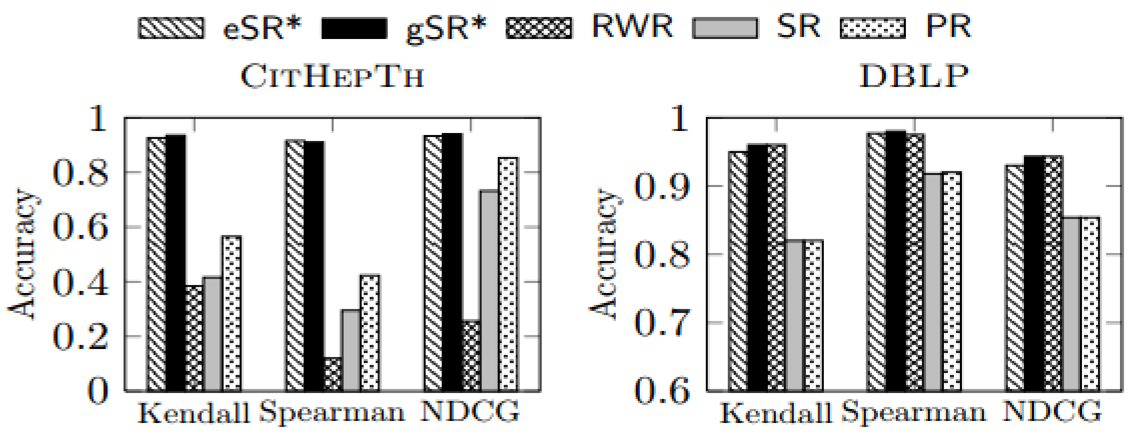
- Edge Concentration
  - Replace bicliques with stars: |X|*|Y| → |X|+|Y|
  - Apply Buehrer and Chellapilla's heuristic



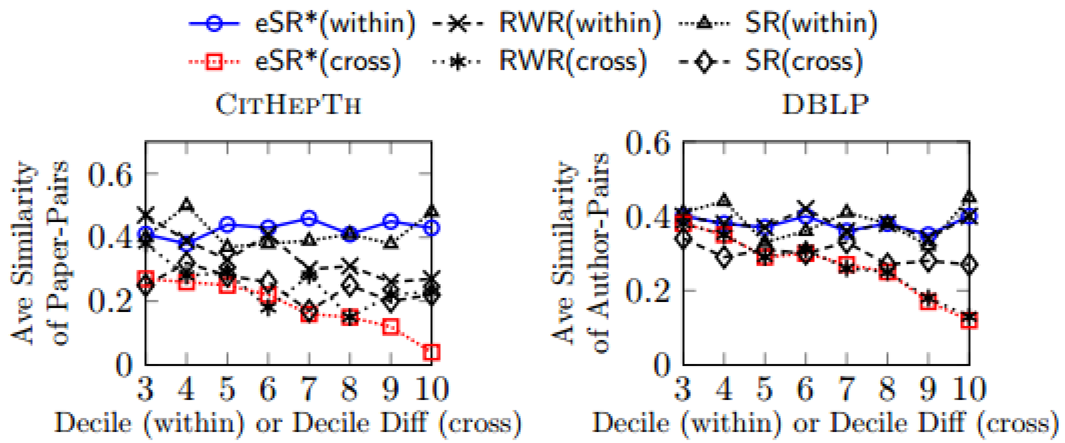
Experimental Settings

- Datasets
  - Real: CitHepTh, DBLP (D05, D08, D11), WebG
  - Synthetic: GraphGen generator
- Compared Algorithms
  - memo-gSR* : our geometric SimRank* + fine-grained memoization
  - memo-eSR*: our exponential SimRank* + fine-grained memoization
  - iter-gSR*: our geometric SimRank* + conventional iteration
  - psum-SR : best-known SimRank
  - mtx-SR: SimRank + singular value decomposition
- Evaluations
  - Semantics & Relative Ordering
  - Computational Efficiency

Semantic Effectiveness

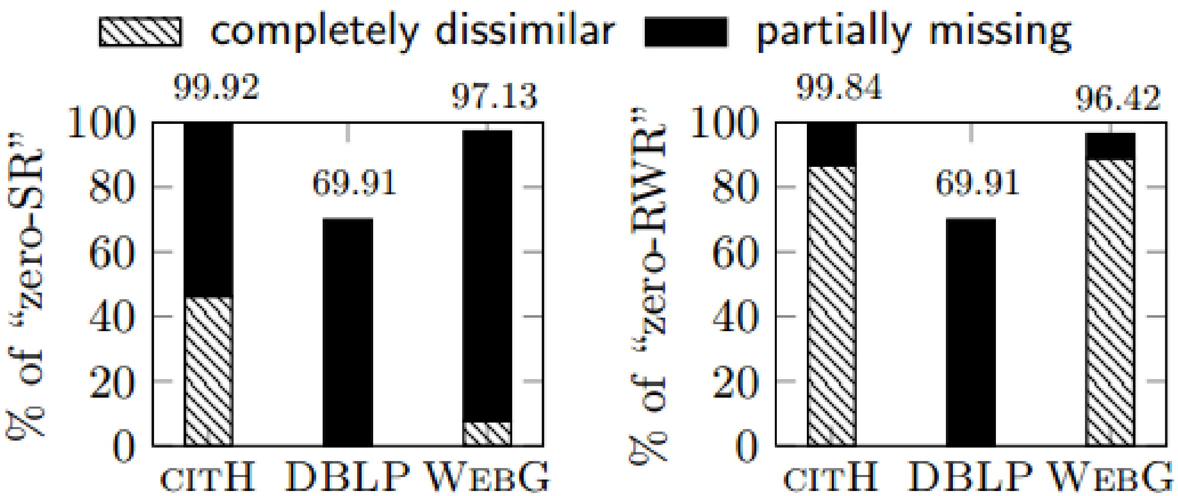


(a) Semantic Effectiveness on Real Data



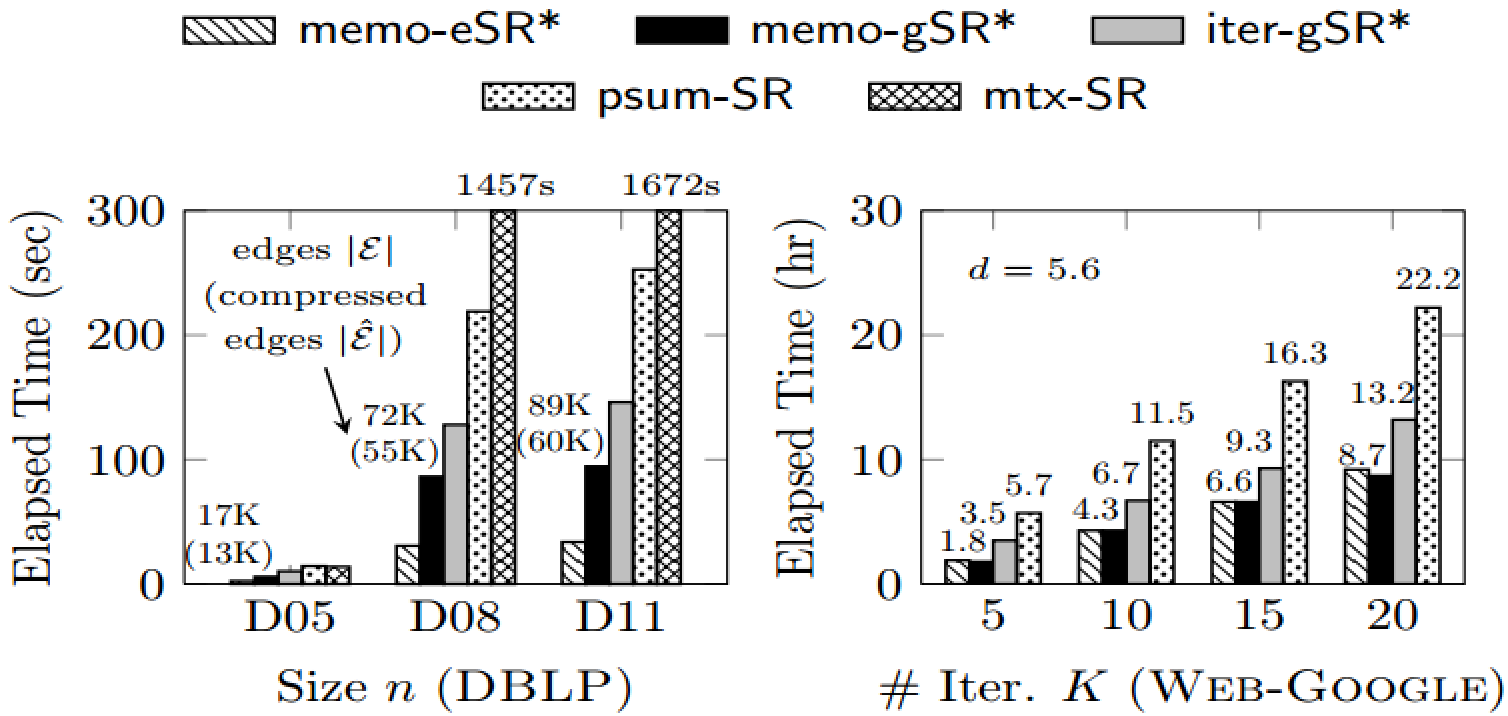
(c) Average Similarity of Grouped Node-Pairs

Existence of "Zero-Similarity" Problem



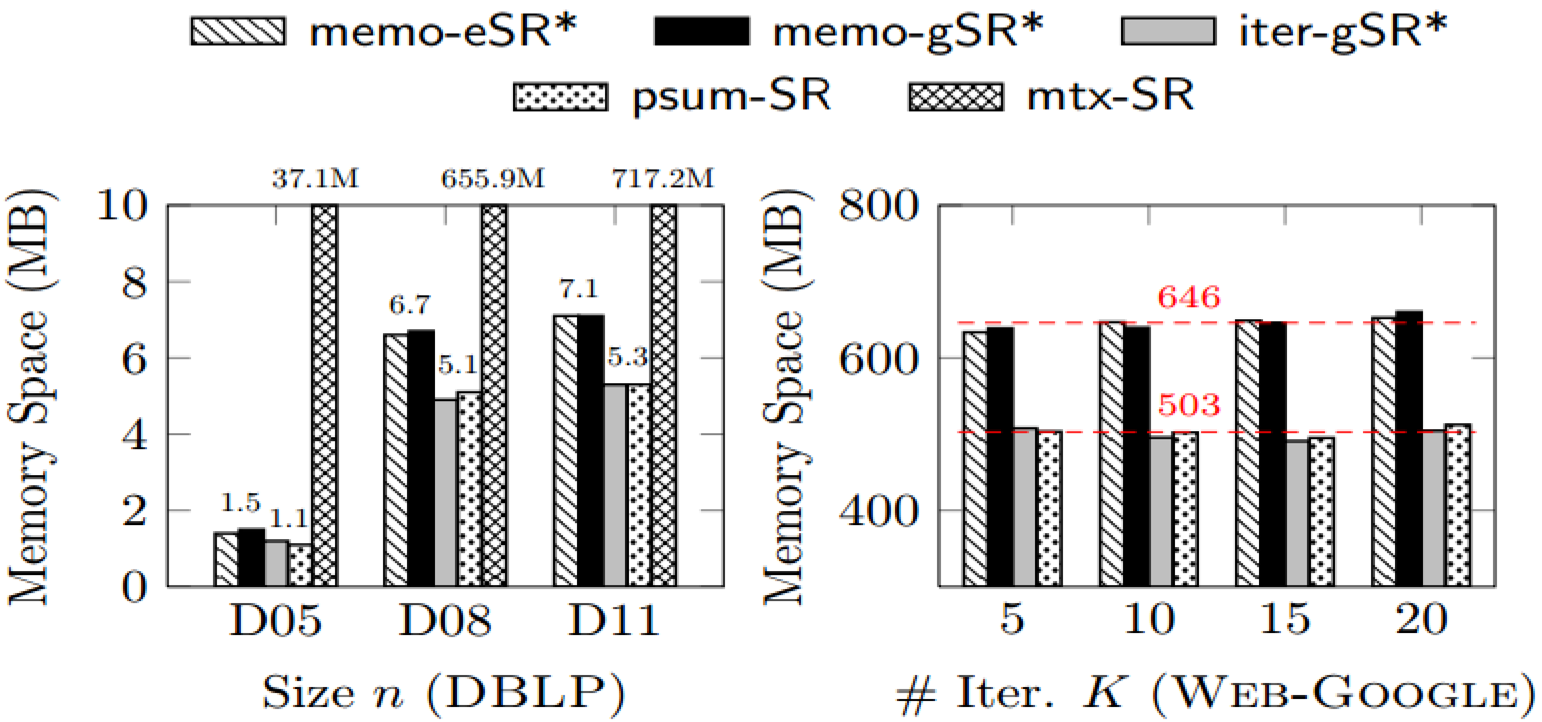
(d) % of "Zero-Similarity" Node-Pairs on Real Data

Time Efficiency



(e) Time Efficiency on Real Datasets

Memory Requirement



(h) Memory Space on Real Datasets

Conclusions

- We have proposed SimRank*, a refinement of SimRank.
  - Resolve “Zero SimRank” issue for semantic richness
  - Geometric & Exponential SimRank*
  - Derive the closed forms and recursive forms of SimRank*
  - Fine-grained memoization for speeding up its computation
- Empirical evaluations to show richer semantics and higher computation efficiency of SimRank*.

Thank you!

Q/A