High Quality SimRank-Based Similarity Search

Weiren Yu and Julie McCann

Department of Computing
Imperial College London

Outline

Overview

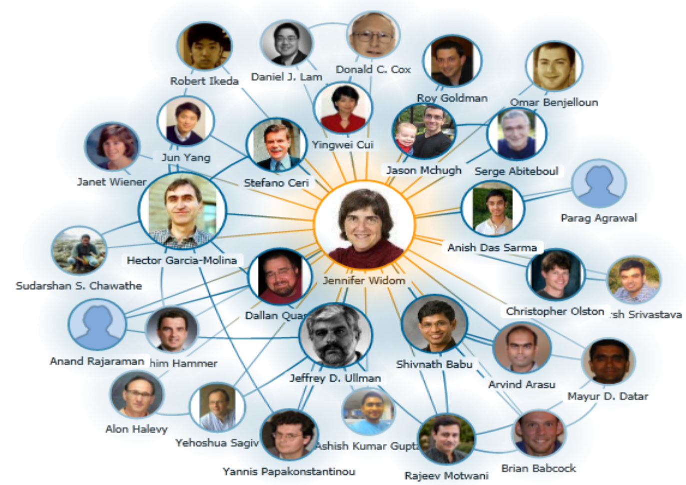
- The quality of SimRank search
  - Superfluous error
  - Connectivity trait
- Our solutions
  - A “varied-D” method to accurately evaluate SimRank
  - A “kernel-based” model to improve search quality
  - A semantic comparison of two SimRank models
- Experimental Results
- Conclusions

Overview

- SimRank in real-world applications:

Recommender System

Citation Graph



Collaboration Network

SimRank Overview

- SimRank
  - An appealing similarity measure based on graph structure
  - Central idea:

    Two nodes are similar if they are pointed to by similar nodes. (*recursion*)

    Each node is most similar to itself. (*base case*)

- Two formulations of SimRank
  - Jeh and Widom's form (*SIGKDD'02*)

similarity btw. nodes a and b

damping factor

$$s(a,b) = \begin{cases} 1 & (a = b) \\ \gamma \cdot \frac{\sum_{(i,j) \in N_a \times N_b} s(i,j)}{|N_a||N_b|} & (a \neq b) \end{cases}$$

in-neighbor set of node b

  - Kusumoto et al.'s form (*SIGMOD'14*)

$$S = \max\{\gamma P^\top S P,\ I\}$$

Kusumoto et al.'s linearization

- Linearized SimRank model:

$$S = \max\{\gamma P^\top S P,\ I\} \Leftrightarrow S = \gamma P^\top S P + D$$

- Single-pair score S(a,b) can be computed as

$$s(a,b) = e_a^\top D e_b + \gamma (P e_a)^\top D (P e_b) + \gamma^2 (P^2 e_a)^\top D (P^2 e_b) + \cdots$$

However, it is difficult to determine D in advance.

- Kusumoto et al.'s approximation

$$D \approx (1-\gamma) I$$

$$S = \max\{\gamma P^\top S P,\ I\} \not\Leftrightarrow \tilde{S} = \gamma P^\top \tilde{S} P + (1-\gamma) I$$

Prob 1: Superfluous Diag Error

- Two Types of Error:

Exact

$$s(a,b) = e_a^\top D e_b + \gamma (P e_a)^\top D (P e_b) + \gamma^2 (P^2 e_a)^\top D (P^2 e_b) + \cdots$$

Diag Err

$$\epsilon_{\text{diag}} := |s(a,b) - s_{\tilde{D}}(a,b)|$$

Approx. D

$$s_{\tilde{D}}(a,b) = e_a^\top \tilde{D} e_b + \gamma (P e_a)^\top \tilde{D} (P e_b) + \gamma^2 (P^2 e_a)^\top \tilde{D} (P^2 e_b) + \cdots$$

Iter Err

$$\epsilon_{\text{iter}} := |s_{\tilde{D}}(a,b) - s_{\tilde{D}}^{(k)}(a,b)| \leq \frac{\gamma^{k+1}}{1-\gamma}$$

K-th Partial Sums

$$s_{\tilde{D}}^{(k)}(a,b) = e_a^\top \tilde{D} e_b + \gamma (P e_a)^\top \tilde{D} (P e_b) + \cdots + \gamma^k (P^k e_a)^\top \tilde{D} (P^k e_b)$$

- “Iter Err” is convergent when k increases

  “Diag Err” is not convergent and sensitive to search quality

Our Method: Varied-D Iteration

- [Kusumoto et al. SIGMOD'14]
  - Hard to determine the exact D *in advance*

$$S_{\tilde{D}}^{(k)} = \tilde{D} + \gamma P^\top \tilde{D} P + \cdots + \gamma^k (P^k)^\top \tilde{D} P^k$$

- Our main idea : Varied-D Model
  - To iteratively compute D and S *at the same time*

$$S^{(k)} := D_k + \gamma P^\top D_{k-1} P + \cdots + \gamma^k (P^\top)^k D_0 P^k$$

When k increases

$$S_{\tilde{D}}^{(k)} \to S_{\tilde{D}} \ (\neq S) \quad (\text{since } \tilde{D} \neq D)$$

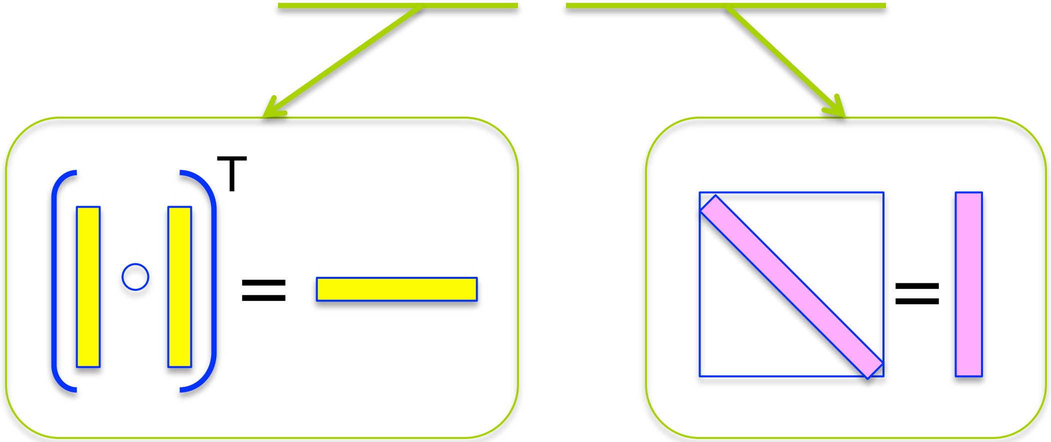
$$S^{(k)} \to S \quad (\text{since } D_k \to D)$$

How to iteratively find D_k ?

Iteratively Find D_k

- D_k can be obtained iteratively as follows:

$$(D_k)_{i,i} = 1 - \sum_{l=1}^{k} (h_l \circ h_l)^\top \overrightarrow{diag}(D_{k-l}) \quad with \quad D_0 = I$$



where
$$\begin{cases} h_0 = e_i \\ h_l = \sqrt{\gamma} P h_{l-1} \quad (l = 1, 2, \cdots, k) \end{cases}$$

- D_k is obtainable in linear memory, independent of $S^{(k)}$ *(scalability)*

Convergence of Varied-D Model

- Varied-D model to compute $S^{(k)}$:

$$S^{(k)} := D_k + \gamma P^\top D_{k-1} P + \cdots + \gamma^k (P^\top)^k D_0 P^k$$

$$S^{(k)} \to S \qquad\qquad (\text{since } D_k \to D)$$

- How close is $S^{(k)}$ to S ?
  - Our model: $\|S^{(k)} - S\|_{\max} \le \gamma^{k+1}$

    No Diag Error, with smaller Iter Error
  - Existing work [SIGMOD'14]:

Iter Error + Diag Error

$$\epsilon_{\text{iter}} := |s_{\tilde{D}}(a,b) - s_{\tilde{D}}^{(k)}(a,b)| \le \frac{\gamma^{k+1}}{1-\gamma}$$

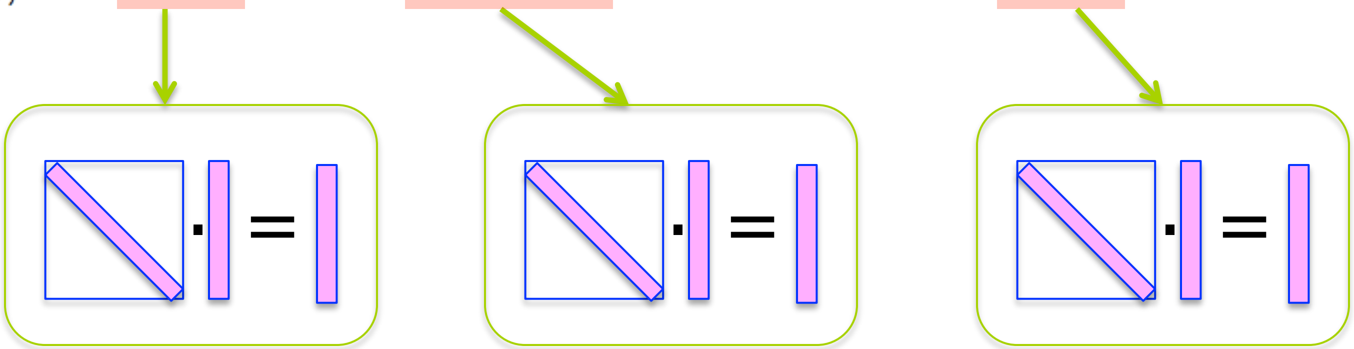
$$\epsilon_{\text{diag}} := |s(a,b) - s_{\tilde{D}}(a,b)|$$

Accelerate Computation for Each Column of SimRank

$$S^{(k)} := D_k + \gamma P^\top D_{k-1} P + \cdots + \gamma^k (P^\top)^k D_0 P^k$$

- Computing i-th column of $S^{(k)}$

$$(S^{(k)})_{i,*} = D_k x_0 + \gamma P^\top D_{k-1} x_1 + \cdots + \gamma^k (P^\top)^k D_0 x_k \quad \text{with} \quad x_l := P^l e_i$$

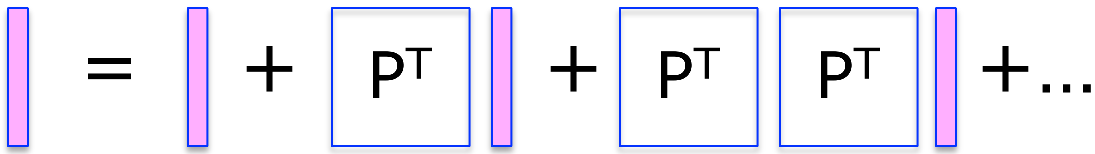


Accelerate Computation for Each Column of SimRank

$$S^{(k)} := D_k + \gamma P^\top D_{k-1} P + \cdots + \gamma^k (P^\top)^k D_0 P^k$$

- Computing i-th column of $S^{(k)}$

$$(S^{(k)})_{i,*} = D_k x_0 + \gamma P^\top D_{k-1} x_1 + \cdots + \gamma^k (P^\top)^k D_0 x_k \quad \text{with} \quad x_l := P^l e_i$$



Naïve Cost: $O(k^2|E|)$ time [SIGMOD'14]

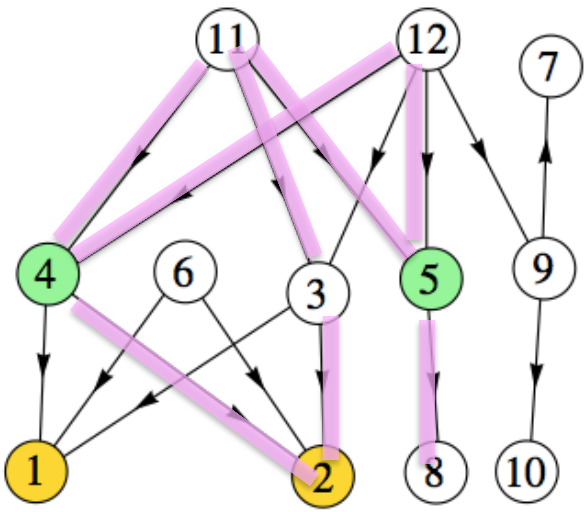
- Our approach
  - multiplying a matrix by a group of vectors added together

$$(S^{(k)})_{i,*} = D_k x_0 + \gamma P^\top (D_{k-1} x_1 + \gamma P^\top (D_{k-2} x_2 + \cdots + \gamma P^\top (D_1 x_{k-1} + \gamma P^\top (D_0 x_k))))$$

Our Cost: $O(k|E|)$ time

Prob 2: "Connectivity Trait" of SimRank

- "Connectivity Trait" Problem:
  - *increasing # of paths between two nodes, say a and b, would incur a decrease in SimRank s(a, b).*



	SR	SR^{++}	RS	$SR^{\#}$
$s(1,2) > s(4,5)$	✗	✗	✓	✓
$s(2,8) > s(8,10)$	✗	✗	✗	✓
$s(4,5) > s(3,9)$	✗	✓	✓	✓

Four paths between node pair (2,8):

$$2 \leftarrow 4 \leftarrow \boxed{11} \rightarrow 5 \rightarrow 8, \quad 2 \leftarrow 3 \leftarrow \boxed{11} \rightarrow 5 \rightarrow 8$$

$$2 \leftarrow 4 \leftarrow \boxed{12} \rightarrow 5 \rightarrow 8, \quad 2 \leftarrow 3 \leftarrow \boxed{12} \rightarrow 5 \rightarrow 8$$

SimRank ignores high connectivity between (2,8)

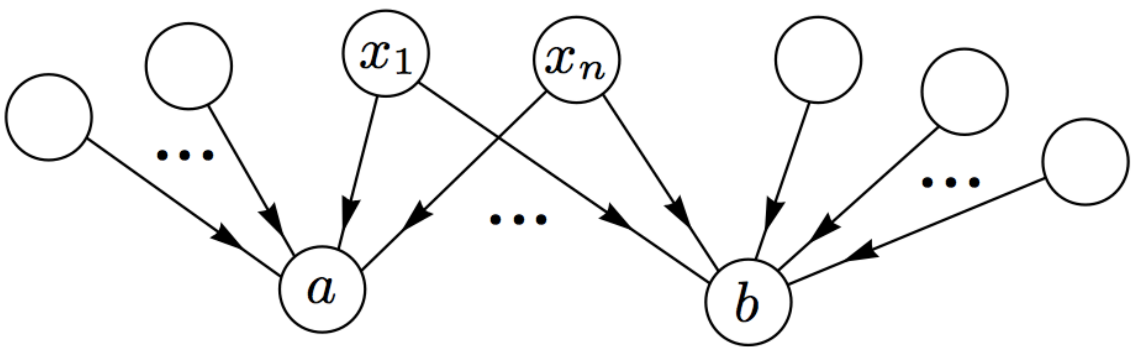
Only one path between node pair (8,10):

$$8 \leftarrow 5 \leftarrow \boxed{12} \rightarrow 9 \rightarrow 10$$

Root Cause of "Connectivity Trait" Issue

$$s(a,b) = \begin{cases} 1 & (a = b) \\ \gamma \cdot \frac{\sum_{(i,j)\in N_a \times N_b} s(i,j)}{|N_a||N_b|} & (a \neq b) \end{cases}$$

- The order of the normalized factor $\frac{1}{|N_a||N_b|}$ is too high.



After δ paths of $\{a \leftarrow x \rightarrow b\}$ are inserted into G:

$$s_\delta(a,b) = \gamma \cdot \frac{|N_a \cap N_b| + \delta}{(|N_a|+\delta)(|N_b|+\delta)} \sim \gamma \cdot \frac{\delta}{\delta^2} \rightarrow 0. \quad (\delta \rightarrow \infty)$$

Our Remedy

- "Cosine-based SimRank" model:

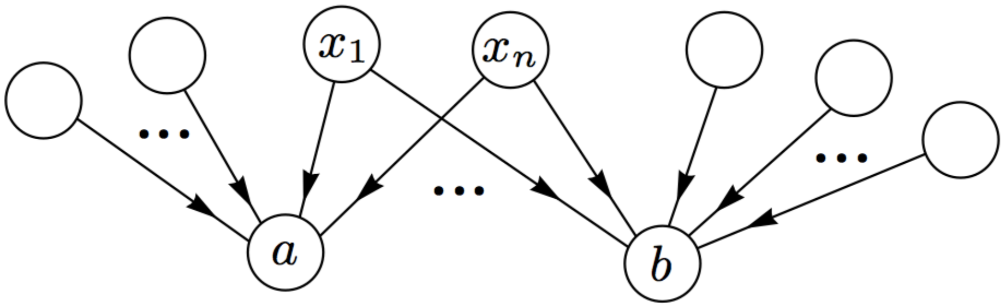
$$\hat{S}_{a,b} = (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \phi(A^k e_a, A^k e_b) \quad \text{with } \phi(x,y) := \frac{x^\top y}{\|x\|_2 \|y\|_2}$$

- Main idea:
  - Aggregates weighted cosine similarities between node a's and node b's multi-hop in-neighbor sets
- Advantage:
  - Provides a correct normalized factor for common multi-hop in-neighbors of a and b

$$\hat{S}_{a,b} = (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \frac{|\mathsf{hop}_k(a) \cap \mathsf{hop}_k(b)|}{\sqrt{|\mathsf{hop}_k(a)| \cdot |\mathsf{hop}_k(b)|}}$$

$$= (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \frac{e_a^\top (A^k)^\top A^k e_b}{\|A^k e_a\|_2 \|A^k e_b\|_2}$$

Fixing "Connectivity Trait" Issue



After δ paths of $\{a \leftarrow x \rightarrow b\}$ are inserted into G:

Cosine-Based SimRank

$$Ae_a = (\underbrace{1,1,\cdots,1}_{|N_a|},\underbrace{0,0,\cdots 0}_{|N_b - N_a|},\underbrace{1,1,\cdots,1}_{\delta})^\top \quad Ae_b = (\underbrace{0,0,\cdots 0}_{|N_a - N_b|},\underbrace{1,1,\cdots,1}_{|N_b|},\underbrace{1,1,\cdots,1}_{\delta})^\top$$

$$\hat{S}_{a,b}(\delta) = (1-\gamma)\gamma \cdot \frac{|N_a \cap N_b| + \delta}{\sqrt{|N_a|+\delta}\sqrt{|N_b|+\delta}} \rightarrow (1-\gamma)\gamma \quad (\delta \rightarrow \infty)$$

Naïve SimRank

$$s_\delta(a,b) = \gamma \cdot \frac{|N_a \cap N_b| + \delta}{(|N_a|+\delta)(|N_b|+\delta)} \sim \gamma \cdot \frac{\delta}{\delta^2} \rightarrow 0. \quad (\delta \rightarrow \infty)$$

Semantic Difference of Two SimRank models

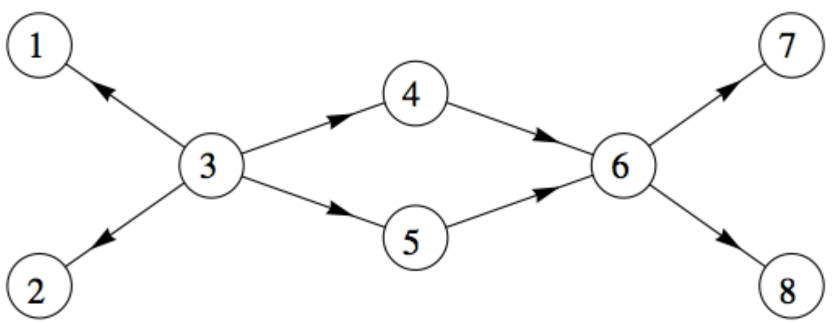
- Jeh and Widom' model:

$$S = \max\{\gamma P^{\top} S P,\ I\}$$

- Li et al.'s model:

$$\tilde{S} = \gamma P^{\top} \tilde{S} P + (1-\gamma) I$$

Any semantic relationship?



node pairs	(3, 3)	(6, 6)	...	(1, 2)	(7, 8)
rank by S	*1*	*1*	...	*9*	*9*
rank by $\tilde{S}$	*4*	*3*	...	*10*	*9*

These two models

1) **neither yield the same relative rankings, [SIGKDD'10]**

2) **nor have the same top-K rankings [SIGMOD'14]**

Their Semantic Relationship

- Jeh and Widom' model: $S = \max\{\gamma P^\top S P,\ I\}$

$\Leftrightarrow$

$$S = I + \gamma (P^\top P)_{off} + \gamma^2 (P^\top (P^\top P)_{off} P)_{off} + \cdots + \\ + \gamma^k \underbrace{(P^\top \cdots (P^\top (P^\top P)_{off} P)_{off} \cdots P)_{off}}_{k \ nested \ (*)_{off}} + \cdots$$

- Li et al.'s model: $\tilde{S} = \gamma P^\top \tilde{S} P + (1-\gamma) I$

$\Leftrightarrow$

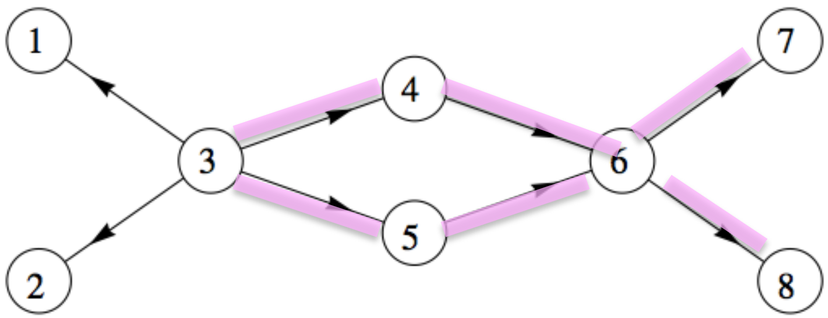
$$\frac{\tilde{S}}{1-\gamma} = I + \gamma P^\top P + \gamma^2 (P^2)^\top P^2 + \cdots + \gamma^k (P^k)^\top P^k + \cdots$$

$$\underbrace{(P^\top \cdots (P^\top (P^\top P)_{off} P)_{off} \cdots P)_{off}}_{k \text{ nested } (*)_{off}}$$

"off" ➔ "overlapping is disallowed"

$$x_0 \underbrace{\leftarrow x_1 \leftarrow \cdots \leftarrow x_{k-1} \leftarrow}_{k \ edges} \boxed{x_k} \underbrace{\rightarrow x_{k+1} \rightarrow \cdots \rightarrow x_{2k-1} \rightarrow}_{k \ edges} x_{2k}$$

Their Semantic Relationship



$$7 \leftarrow 6 \leftarrow 5 \leftarrow \boxed{3} \rightarrow 4 \rightarrow 6 \rightarrow 8$$

- can be tallied by $\left((P^3)^\top P^3\right)$
- but cannot be tallied by

$$(P^\top (P^\top (P^\top P)_{\mathit{off}} P)_{\mathit{off}} P)_{\mathit{off}}$$

$$\tilde{S} = \gamma P^\top \tilde{S} P + (1-\gamma) I$$

$$S = \max\{\gamma P^\top S P,\ I\}$$

	$k=0$	$k=1$		$k=2$				...
Li *et al.*'s SimRank Variation $\tilde{S}_k$	$i\ (j)$	① i j	② $i\ (j)$	③ i j	④ i j	⑤ $i\ (j)$	⑥ $i\ (j)$	...
Jeh and Widom's SimRank S_k	$i\ (j)$	i j		i j				...

Li et al.'s model can tally more paths with self-intersected nodes than Jeh and Widom's.

Experimental Settings

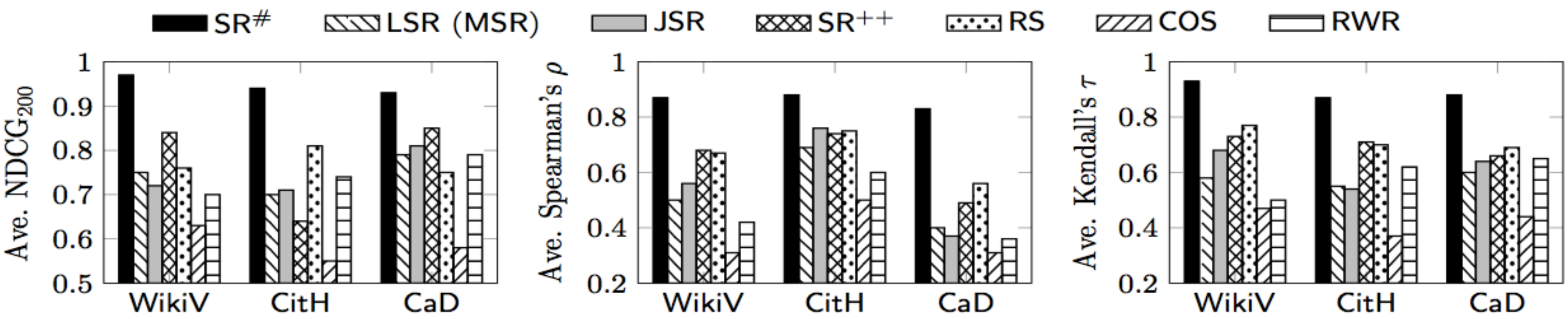
- Datasets
  - Real-life Data:

Dataset	$\|V\|$	$\|E\|$	$\|E\|/\|V\|$	Type
WikiV	7,115	103,689	14.57	Directed
CaD	15,683	55,064	5.31	Undirected
CitH	34,546	421,578	12.20	Directed
WebN	325,729	1,497,134	4.59	Directed
ComY	1,134,890	2,987,624	2.63	Undirected
SocL	4,847,571	68,993,773	14.23	Directed

  - Synthetic Data: GraphGen generator
- Compared Algorithms

Name	Description
SR#	our scheme ("cosine" kernel + computation sharing)
MSR	the state-of-the-art SimRank [7]
OIP	all-pairs SimRank (fine-grained clustering) [13]
PSUM	all-pairs SimRank (partial sums memoization) [12]
SMAT	single-source SimRank (matrix decomposition) [3]
JSR	Jeh and Widom's SimRank [5]
LSR	Li *et al.*'s SimRank [9]
SR++	SimRank++ (revised "evidence factor") [1]
RS	RoleSim (automorphism equivalence) [6]
RWR	Random Walk with Restart
COS	classic cosine similarity

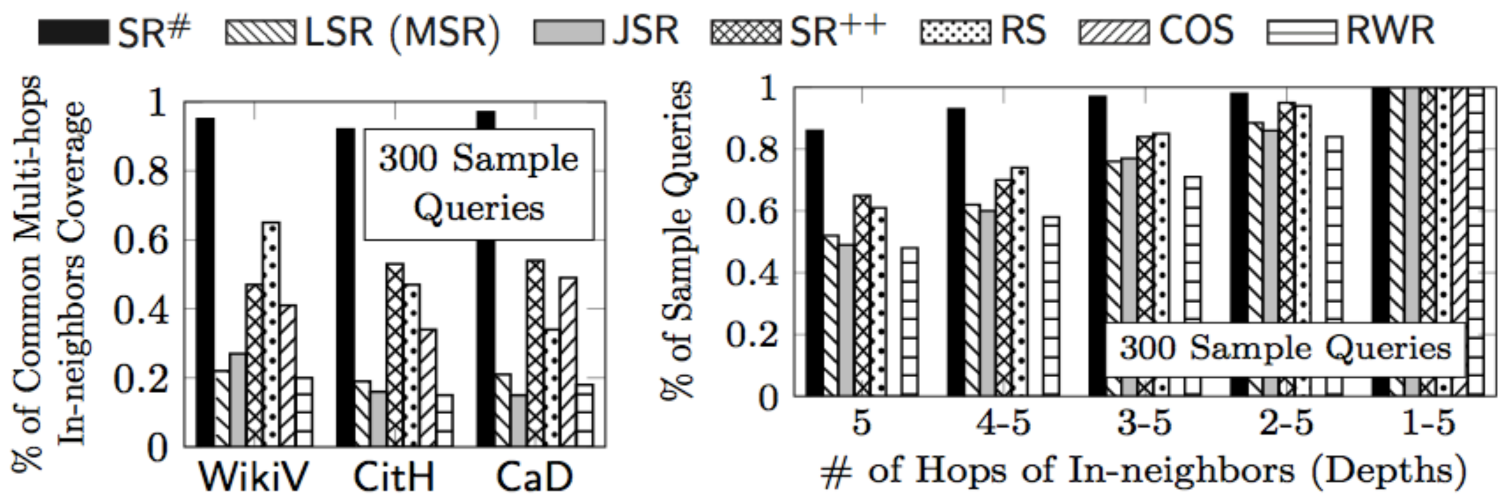
Exp 1: Semantic Quality



(a) Semantics on Real Data (Measured by NDCG, Spearman's ρ, Kendall's τ)

- SR# can avoid "connectivity trait" issue by using a "cosine" kernel.
- COS considers only direct overlapped in-neighbors.
- JSR and LSR both have a "connectivity trait" problem.

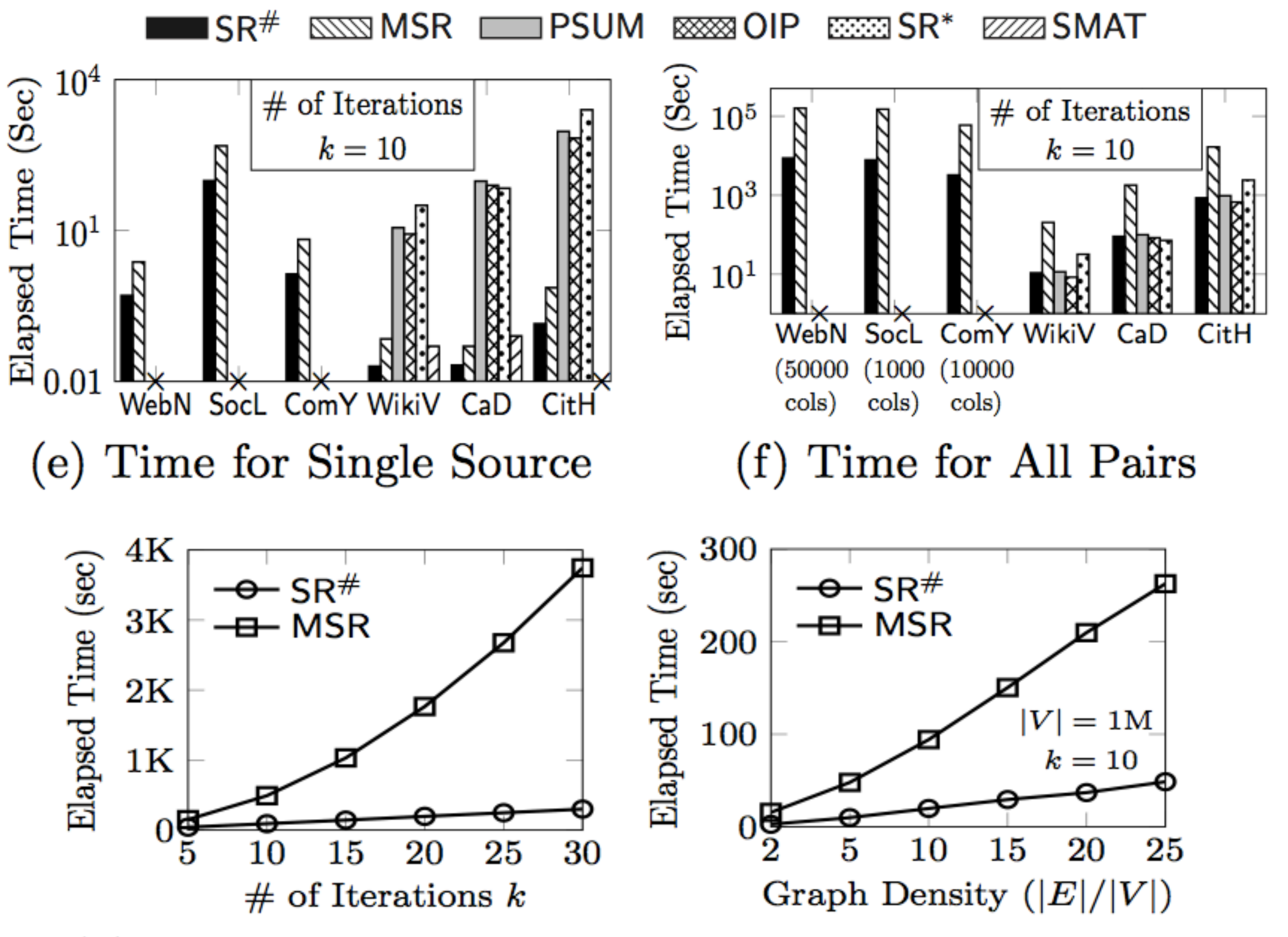
Exp 1: Semantic Quality



(c) Overlapping Coverage (d) Depth Coverage

- SR# achieves ~95% coverage of common multi-hop in-neighbors (due to its suitable normalized factor)
- COS (~0.41) consistently outperforms JSR/LSR (~0.20) since COS is not limited by the "connectivity trait" problem.
- The superiority of SR# is more pronounced in the groups with longer paths.

Exp 2: Speedup



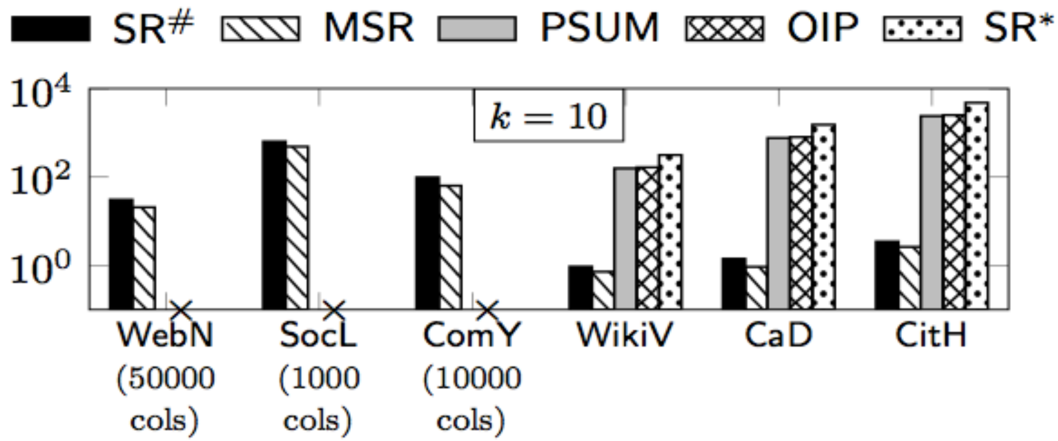
(e) Time for Single Source

(f) Time for All Pairs

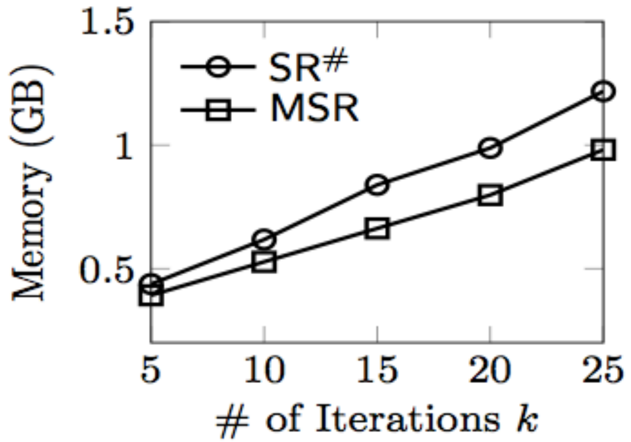
(g) Time *vs.* k on SocL

(h) Time *vs.* $\frac{|E|}{|V|}$ on SYN

Exp 3: Scalability



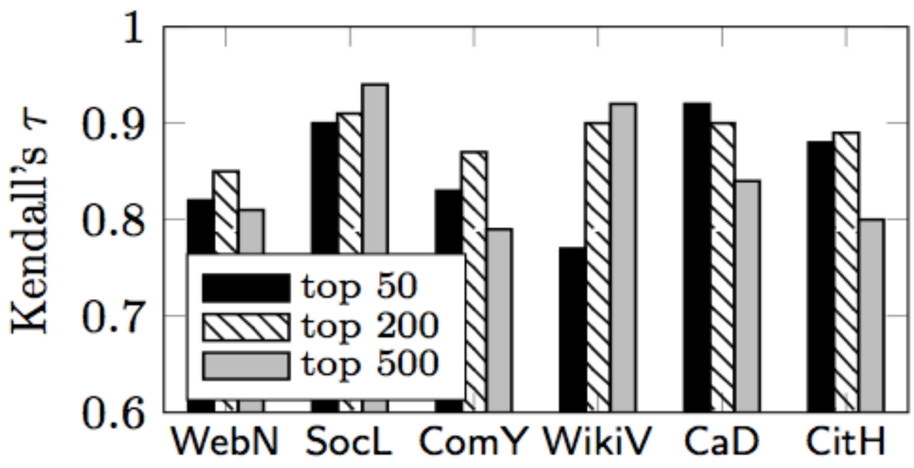
(i) Memory for Single Source/All Pairs



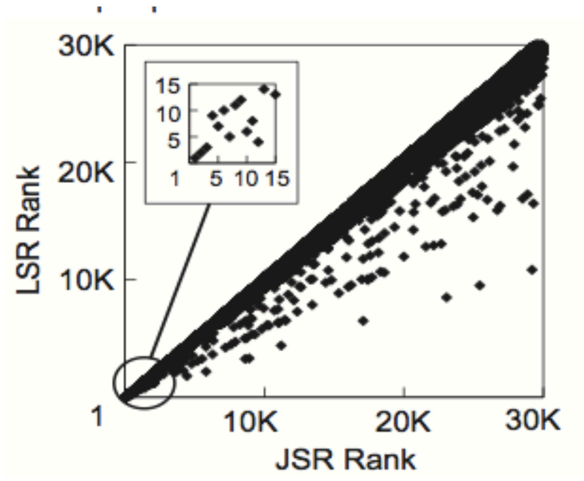
(j) Memory *vs.* k on SocL

- Only SR# and MSR survive on large datasets, highlighting their scalability.
- The disparity in the memory between SR# and MSR is comparatively small, due to SR# that stores the iterative diagonal correction matrix D_k.

Exp 4: Relative Ordering



(k) LSR and JSR Relative Ordering

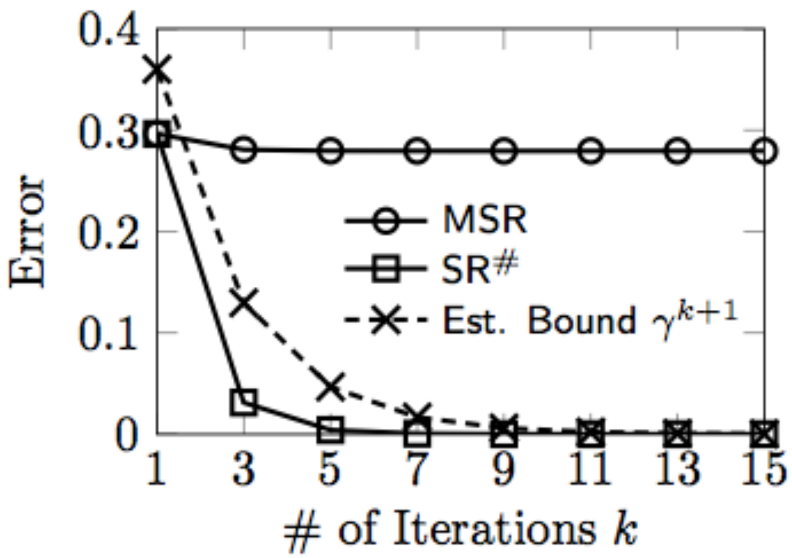


(l) Ranking on WikiV

- For different graphs, the quality of relative order is irrelevant to top K size.
- LSR does not maintain the relative rank of JSR, even for top 50.

- Many points below the diagonal imply that low-ranked node-pairs by JSR have greater likelihood to get promoted to a high rank of LSR.

Exp 5: Effect of Diag Error



(m) $(\epsilon_{\text{diag}} + \epsilon_{\text{iter}})$ $vs.$ k

- Our "varied-D" iterative model can guarantee the error to be small and convergent w.r.t k.
- The SR# curve is always below the Est. Bound curve, showing the correctness of our error estimation.

In Conclusion

- We have focused on high quality of SimRank search:
  - Devise a "varied-D" method to remove diagonal error of Kusumoto et al.'s SimRank model
  - Design a "kernel-based" model to resolve connectivity trait problem of SimRank
  - Semantically show the difference between Li et al.'s and Jeh et al.'s SimRank models

Thank you!

Q/A

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