# Fast Incremental SimRank on Link-Evolving Graphs

UNSW THE UNIVERSITY OF NEW SOUTH WALES

Imperial College London


**Weiren Yu [1,2], Xuemin Lin [1], Wenjie Zhang [1]**

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

[2] Department of Computing, Imperial College London, UK


## 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 $b$

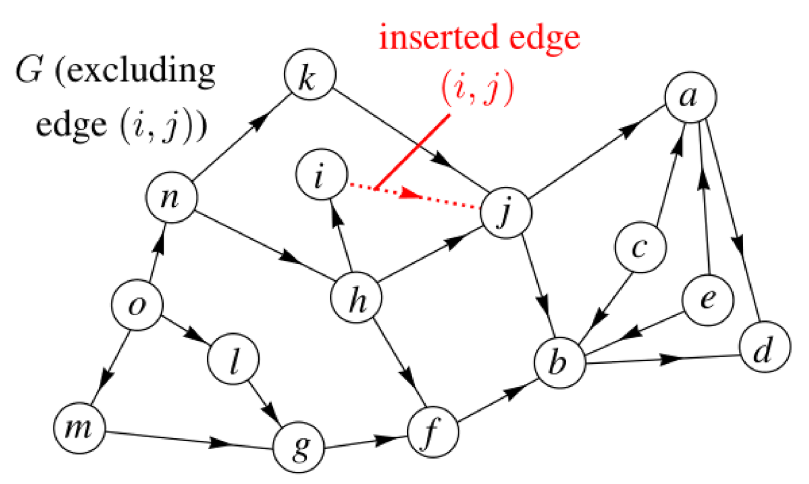
  - Matrix form (EDBT '10)

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

## Existing Work

- Batch Computations
  - All Pairs s(*,*)
  - Single Pair s(a,b)
  - Single Source s(*,q)
  - Similarity Join s(x,y) for all x in A, and y in B.
- Incremental Paradigms:
  - link-evolving:
    Li et. al. [EDBT 2010] needs $O(r^4 n^2)$ time for approximation.
  - node-evolving:
    He et al. [KDD 2010] --- GPU based

## Motivations




| Node-Pair | in $G$ | in $G \cup \Delta G$ | |
|---|---|---|---|
| | sim | $\text{sim}_{\text{true}}$ | $\text{sim}_{\text{Li et al.}}$ |
| $(a,b)$ | 0.075 | 0.062 | 0.073 |
| $(a,d)$ | 0.000 | 0.006 | 0.002 |
| $(i,f)$ | **0.246** | **0.246** | **0.246** |
| $(k,g)$ | **0.128** | **0.128** | **0.128** |
| $(k,h)$ | **0.288** | **0.288** | **0.288** |
| $(j,f)$ | 0.206 | 0.138 | 0.206 |
| $(m,l)$ | **0.160** | **0.160** | **0.160** |
| $(j,b)$ | 0.000 | 0.030 | 0.001 |

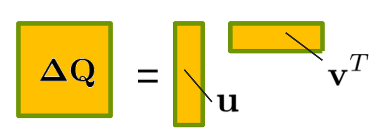
- Li et al. [EDBT 2010] using SVD for incremental SimRank is approximate.
- When ΔG is small, the "affected areas" of ΔS are also small.

**Problem** (INCREMENTAL SIMRANK COMPUTATION)
**Given:** G, S, ΔG, and C.
**Compute:** ΔS to S.

## Characterizing ΔS via a rank-one Sylvester

- Time complexity: $O(Kn^2)$

Step 1. Find u,v s.t. $\Delta \mathbf{Q} = \mathbf{u} \cdot \mathbf{v}^T$




$$\mathbf{u} = \begin{cases} \mathbf{e}_j & (d_j = 0) \\ \frac{1}{d_j+1}\mathbf{e}_j & (d_j > 0) \end{cases}, \quad \mathbf{v} = \begin{cases} \mathbf{e}_i & (d_j = 0) \\ \mathbf{e}_i - [\mathbf{Q}]_{j,\star}^T & (d_j > 0) \end{cases}$$

Can we further improve it?

Step 2. Find w s.t. $\mathbf{M} = C \cdot \tilde{\mathbf{Q}} \cdot \mathbf{M} \cdot \tilde{\mathbf{Q}}^T + C \cdot \mathbf{u} \cdot \mathbf{w}^T$

initialize $\xi_0 \leftarrow C \cdot \mathbf{u}, \quad \eta_0 \leftarrow \mathbf{w}, \quad \mathbf{M}_0 \leftarrow C \cdot \mathbf{u} \cdot \mathbf{w}^T$
for $k = 0, 1, 2, \cdots$
$\quad \xi_{k+1} \leftarrow C \cdot \tilde{\mathbf{Q}} \cdot \xi_k, \quad \eta_{k+1} \leftarrow \tilde{\mathbf{Q}} \cdot \eta_k$
$\quad \mathbf{M}_{k+1} \leftarrow \xi_{k+1} \cdot \eta_{k+1}^T + \mathbf{M}_k$

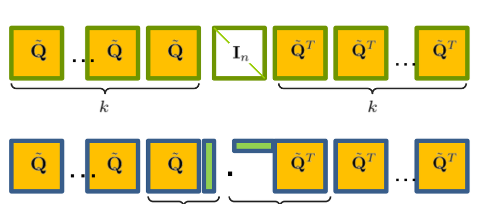
No mat-mat multiplications

Step 3. Compute ΔS as $\Delta \mathbf{S} = \mathbf{M} + \mathbf{M}^T$

## Finding w

$$\mathbf{M} = \sum_{k=0}^{\infty} C^{k+1} \cdot \tilde{\mathbf{Q}}^k \cdot \mathbf{u} \cdot \mathbf{w}^T \cdot (\tilde{\mathbf{Q}}^T)^k,$$

$$\tilde{\mathbf{S}} = (1-C) \cdot \sum_{k=0}^{\infty} C^k \cdot \tilde{\mathbf{Q}}^k \cdot \mathbf{I}_n \cdot (\tilde{\mathbf{Q}}^T)^k.$$



**Theorem** There exists $\mathbf{w} = \mathbf{y} + \frac{\lambda}{2}\mathbf{u}$ with

$$\mathbf{y} = \mathbf{Q} \cdot \mathbf{z}, \quad \lambda = \mathbf{v}^T \cdot \mathbf{z}, \quad \mathbf{z} = \mathbf{S} \cdot \mathbf{v}$$

s.t.

$$\mathbf{M} = C \cdot \tilde{\mathbf{Q}} \cdot \mathbf{M} \cdot \tilde{\mathbf{Q}}^T + C \cdot \mathbf{u} \cdot \mathbf{w}^T$$

is a rank-one Sylvester Equation w.r.t. M.

$\mathbf{z} = \mathbf{S} \cdot \mathbf{v} \quad \mathbf{y} = \mathbf{Q} \cdot \mathbf{z} \quad \lambda = \mathbf{v}^T \cdot \mathbf{z}$

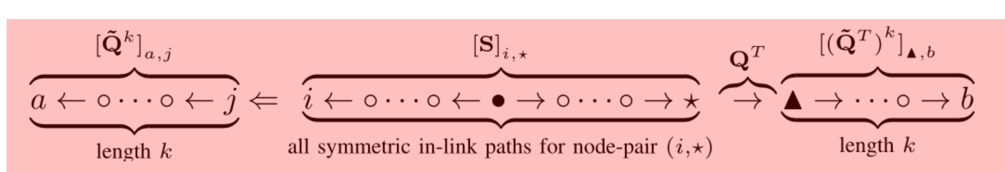
## Pruning "unaffected areas" of ΔS

- As M merely tallies these paths, node-pairs without having such paths can be pruned.

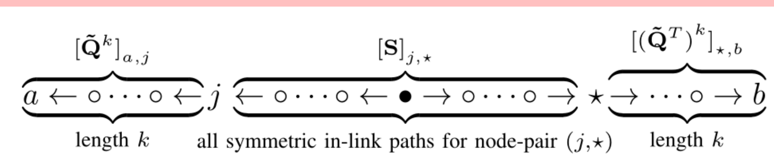
$$[\mathbf{M}]_{a,b} = \frac{1}{d_j+1}\Big(\underbrace{\sum_{k=0}^{\infty} C^{k+1} \cdot [\tilde{\mathbf{Q}}^k]_{a,j} [\mathbf{S}]_{i,\star} \mathbf{Q}^T \cdot [(\tilde{\mathbf{Q}}^T)^k]_{\star,b}}_{\text{Part 1}} - \underbrace{\sum_{k=0}^{\infty} C^k [\tilde{\mathbf{Q}}^k]_{a,j} [\mathbf{S}]_{j,\star} [(\tilde{\mathbf{Q}}^T)^k]_{\star,b}}_{\text{Part 2}}$$

$$+\mu \underbrace{\sum_{k=0}^{\infty} C^{k+1} [\tilde{\mathbf{Q}}^k]_{a,j} [(\tilde{\mathbf{Q}}^T)^k]_{j,b}}_{\text{Part 3}}\Big)$$

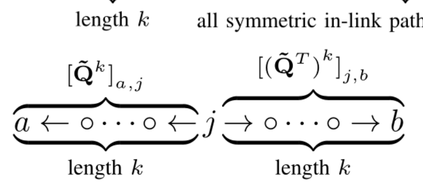
- Three types of paths captured by M
  - P1:




  - P2:




  - P3:




- Complexity: $O(K(nd+|AFF|))$ with $|AFF| := \text{avg}_{k \in [0,K]}(|\mathcal{A}_k| \cdot |\mathcal{B}_k|)$
- Iteratively Pruning:

Let $\mathcal{F}_1 := \{b \mid b \in \mathcal{O}(y),\ \exists y,\ s.t.\ [\mathbf{S}]_{i,y} \neq 0\}$

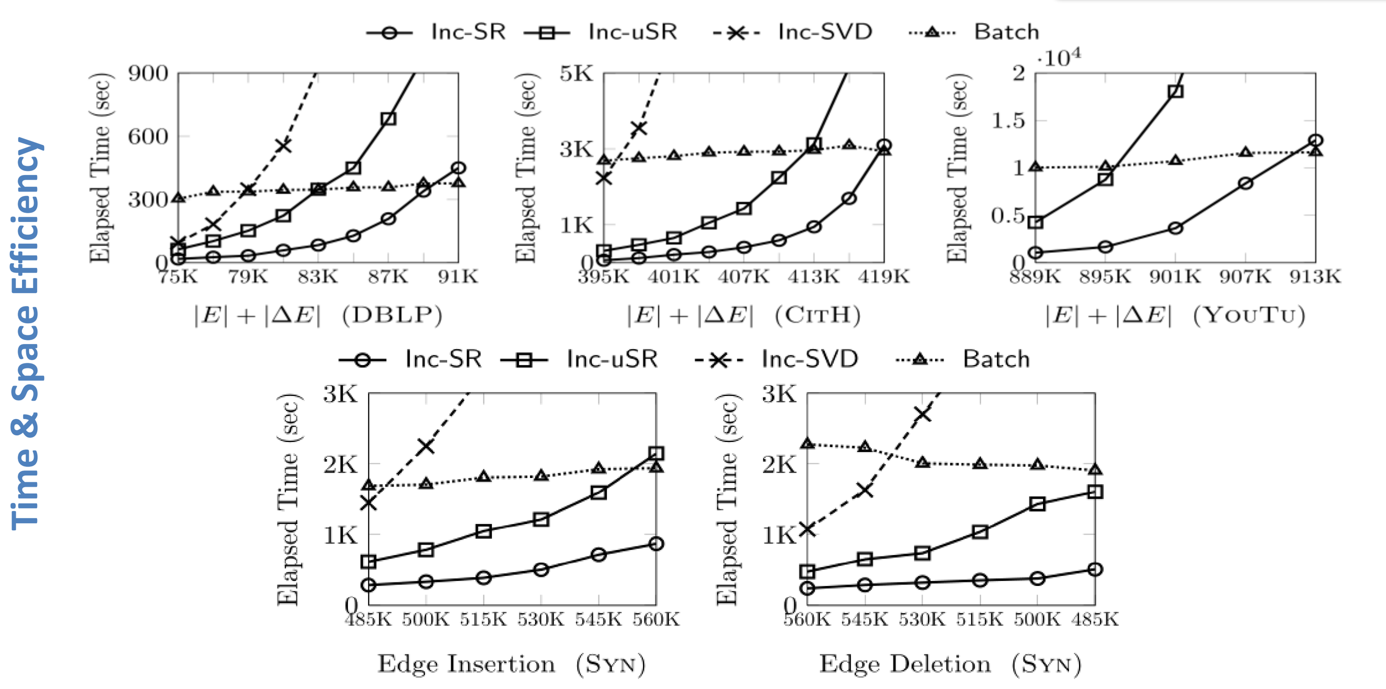
$$\mathcal{F}_2 := \begin{cases} \varnothing & (d_j = 0) \\ \{y \mid [\mathbf{S}]_{j,y} \neq 0\} & (d_j > 0) \end{cases}$$

$$\mathcal{A}_k \times \mathcal{B}_k := \begin{cases} \{j\} \times (\mathcal{F}_1 \cup \mathcal{F}_2 \cup \{j\}) & (k=0) \\ \{(a,b) \mid a \in \tilde{\mathcal{O}}(x),\ b \in \tilde{\mathcal{O}}(y),\ \exists x,\ \exists y,\ s.t.\ [\mathbf{M}_{k-1}]_{x,y} \neq 0\} & (k>0) \end{cases}$$

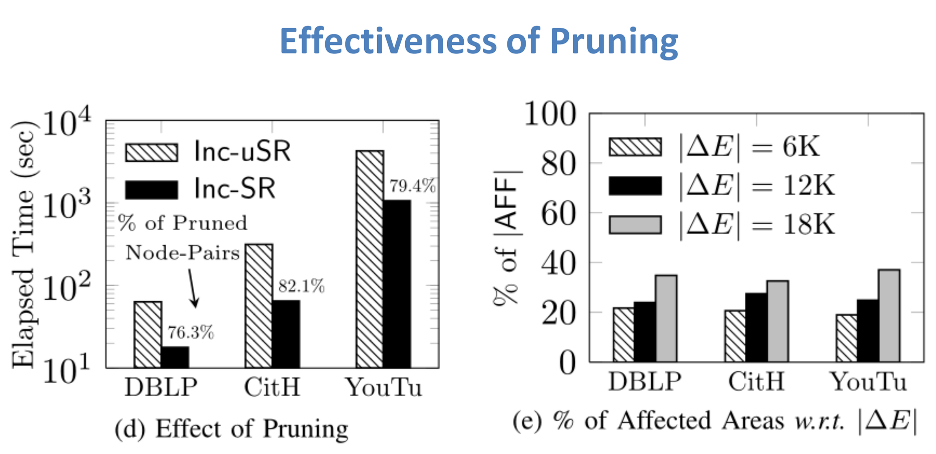
Then $[\mathbf{M}_k]_{a,b} = 0 \quad \textit{for all } (a,b) \notin (\mathcal{A}_k \times \mathcal{B}_k) \cup (\mathcal{A}_0 \times \mathcal{B}_0)$

## Experimental Evaluations

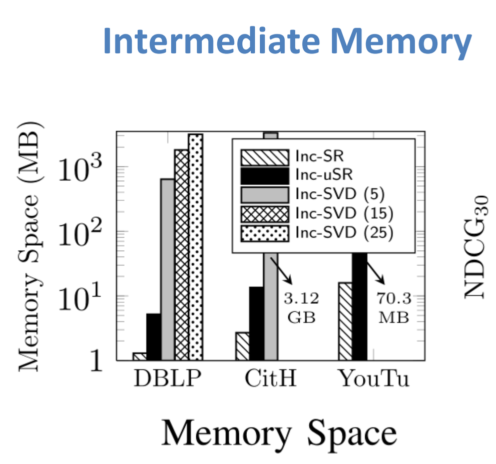
Time & Space Efficiency




Effectiveness of Pruning




Intermediate Memory




Exactness