Co-Simmate: Quick Retrieving All Pairwise Co-Simrank Scores

Weiren Yu, McCann A. Julie
Department of Computing, Imperial College London, UK

Co-Simrank Overview

- Co-Simrank
  - A graph-based similarity measure (Rothe and Schutze, ACL'14)
  - integrates both features of Simrank and Pagerank
  - Intuition
    More similar nodes are likely to be pointed to by other similar nodes.
  - Formulation

$$\mathbf{S}_k(a,b) = c^k \langle \mathbf{p}_k(a), \mathbf{p}_k(b) \rangle + \mathbf{S}_{k-1}(a,b)$$

where $\mathbf{p}_k(a) = \mathbf{A}^T \mathbf{p}_{k-1}(a)$ with $\mathbf{p}_0(a) = \mathbf{I}(*,a)$

Existing Work

- Only suitable for computing a single pair s(a,b) on G(V,E)
  - O(k|E|) time : PR vector $p_k(a)$
  - O(|V|) time : dot product of two PR vectors $<p_k(a), p_k(b)>$
- Inefficient for computing all pairs scores s(*,*) ($|V|^2$ pairs)
  - Equivalent to solving

$$\mathbf{S}_k = c\mathbf{A}^T\mathbf{S}_{k-1}\mathbf{A} + \mathbf{I} \text{ with } \mathbf{S}_0 = \mathbf{I}$$

  - O(k|V|^3) time in total

Observation

- The exact Co-Simrank solution S can be expressed as

$$\mathbf{S} = \mathbf{I} + c\mathbf{A}^T\mathbf{A} + c^2(\mathbf{A}^T)^2\mathbf{A}^2 + c^3(\mathbf{A}^T)^3\mathbf{A}^3 + c^4(\mathbf{A}^T)^4\mathbf{A}^4 + \cdots$$

- The existing iterative method adopts the following association:

$$\mathbf{S} = \left(c\mathbf{A}^T\left(c\mathbf{A}^T\underbrace{\left(c\mathbf{A}^T\mathbf{A}+\mathbf{I}\right)}_{=\mathbf{S}_1}\mathbf{A}+\mathbf{I}\right)\mathbf{A}+\mathbf{I}\right)+\cdots \quad (=\mathbf{S}_2)$$

Our Association Approach

- Our method reorganizes S as follows:

$$\mathbf{S} = \left(\mathbf{I}+c\mathbf{A}^T\mathbf{A}\right) + \left(c^2(\mathbf{A}^T)^2\mathbf{A}^2 + c^3(\mathbf{A}^T)^3\mathbf{A}^3\right) + \left(c^4(\mathbf{A}^T)^4\mathbf{A}^4 + \cdots + c^7(\mathbf{A}^T)^7\mathbf{A}^7\right) + \cdots$$

- Computation sharing + Repeated Squaring

$$\mathbf{S} = \underbrace{\left(\overbrace{\left(\mathbf{I}+c\mathbf{A}^T\mathbf{A}\right)}^{=\mathbf{R}_1} + (c\mathbf{A}^T)^2\overbrace{\left(\mathbf{I}+c\mathbf{A}^T\mathbf{A}\right)}^{=\mathbf{R}_1}\mathbf{A}^2\right)}_{=\mathbf{R}_2} + (c\mathbf{A}^T)^4\underbrace{\left(\left(\mathbf{I}+c\mathbf{A}^T\mathbf{A}\right) + (c\mathbf{A}^T)^2\left(\mathbf{I}+c\mathbf{A}^T\mathbf{A}\right)\mathbf{A}^2\right)}_{=\mathbf{R}_2}\mathbf{A}^4 + \cdots$$

Is there an iterative formulation?

Co-Simmate Model

- Our iterative method:

$$\begin{cases} \mathbf{R}_0 = \mathbf{I}, \quad \mathbf{A}_0 = \mathbf{A} \\ \mathbf{R}_{k+1} = \mathbf{R}_k + c^{2^k}(\mathbf{A}_k{}^T\mathbf{R}_k\mathbf{A}_k) \\ \mathbf{A}_{k+1} = \mathbf{A}_k{}^2 \end{cases}$$

- Convergence rate:

$$\mathbf{R}_k = \mathbf{S}_{2^k-1}$$

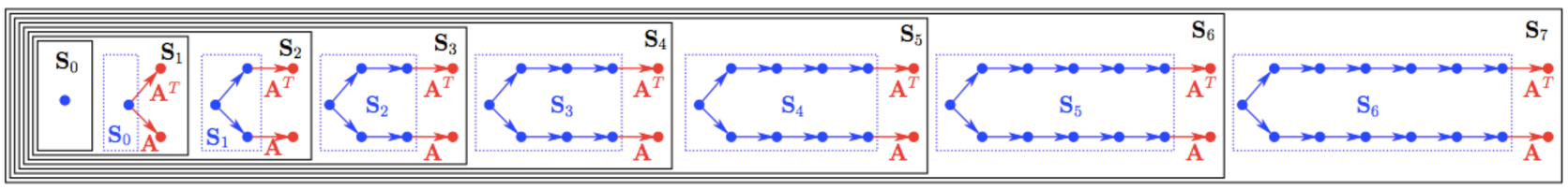
(total # of operations for $\mathbf{R}_k$) $= 3k|\mathfrak{M}|$

(total # of operations for $\mathbf{S}_k$) $= 2(2^k-1)|\mathfrak{M}|$

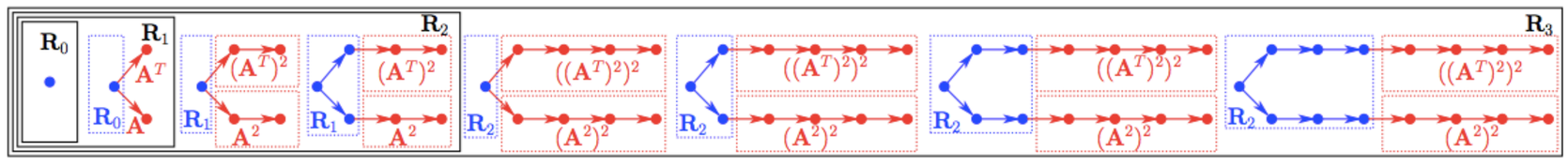
$\mathcal{O}(\log(1/\epsilon)n^3) \rightarrow \mathcal{O}(\log_2(\log(1/\epsilon))n^3)$

Pictorial Comparision

- Co-Simrank

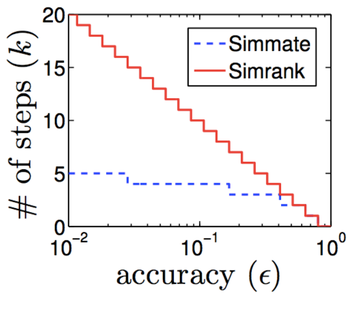


- Co-Simmate

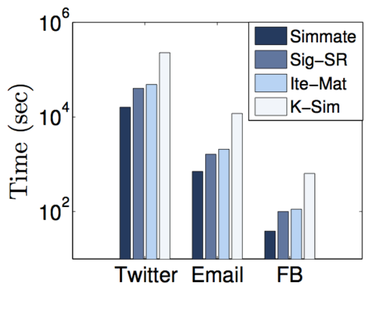


It speeds up Co-Simrank by aggregating more first terms of S at each step

Experimental Evaluations



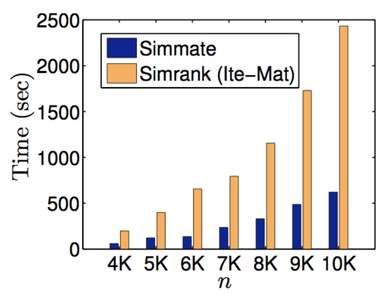
(a) Rate of Convergence (on FB dataset, $c = 0.8$)



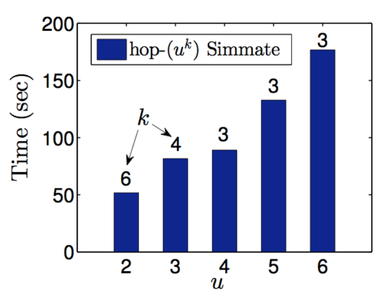
(b) Total Computational Time (on three real datasets, $c = 0.8$)

ϵ	$c=0.6$		$c=0.7$		$c=0.8$	
	SM	SR	SM	SR	SM	SR
0.1	3	4	3	6	4	10
0.01	4	9	4	12	5	20
0.001	4	13	5	19	5	30
0.0001	5	18	5	25	6	41
0.00001	5	22	6	32	6	51

(c) Effect of Damping Factor c on Iterations k (on FB)



(d) Scalability *w.r.t.* # nodes (on 7 synthetic datasets)



(e) Effect of Hop-(u^k) (on FB dataset, $c = 0.8$)