Personalized PageRank Estimation for Many Nodes: The Impact of Clustering on Complexity

Daniel Vial[1], Vijay Subramanian[1]

[1] EECS Department, University of Michigan, Ann Arbor, MI.

Introduction

Personalized PageRank (PPR): measure of importance of a node (called the *target*) from the perspective of another (called the *source*)

Example: provide Twitter user (source) personalized recommendations of who to follow (high PPR targets) [4]

Our goal: efficiently estimate PPR for many source/target pairs

Background

PPR: stationary distribution $\{\pi_s(v)\}_{v\in V}$ of random walk on $G = (V, E)$ with probability-α jumps to $s \in V$ at each step; satisfies (1) [3]

$$\pi_s(v) = \mathbb{P}[\text{random walk of length} \sim \text{geometric}(\alpha) \text{ from } s \text{ ends at } v] \tag{1}$$

`Bidirectional-PPR` [5] (Fig. 1): state-of-the-art $\pi_s(t)$ estimator (single source/target pair)

- Target stage: compute $p^t, r^t \in \mathbb{R}_+^{|V|}$ via `Approx-Contributions` [1], which satisfy

$$\pi_s(t) = p^t(s) + \sum_{v\in V} \pi_s(v) r^t(v) \; \forall \; s \in V$$

- Source stage: estimate $\sum_{v\in V} \pi_s(v) r^t(v) = \mathbb{E}_{V\sim\pi_s}[r^t(V)]$ with random walks via (1)

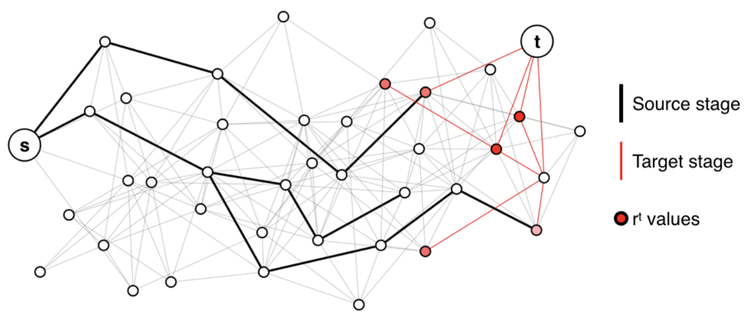

Figure 1: Single source/target pair PPR estimator `Bidirectional-PPR` [5]

Approach

Our goal: estimate $\pi_s(t) \; \forall \; (s, t) \in S \times T$ (many source/target pairs)

Naive strategy using `Bidir-PPR`: separate source stage $\forall \; s$, separate target stage $\forall \; t$

- Complexity linear in $|S|, |T|$

Proposed strategy: joint source and target stages

- Complexity scales with (2), (3), both interpretable as clustering measures

Joint source stage

Sample walks from intermediate nodes via following scheme (Fig. 2):

- Compute $p^s, r^s \in \mathbb{R}_+^{|V|}$ via `Approx-PageRank` [2], which satisfy

$$\pi_s(t) = p^t(s) + \langle p^s, r^t \rangle + \|r^s\|_1 \sum_{u\in V} \sum_{v\in V} \sigma_s(u) \pi_u(v) r^t(v), \quad \sigma_s := r^s / \|r^s\|_1$$

- Sample walks from $U \sim \sigma_s$ to estimate $\sum_{u,v} \sigma_s(u) \pi_u(v) r^t(v) = \mathbb{E}_{U\sim\sigma_s, V\sim\pi_U}[r^t(V)]$

This scheme allows sharing of walks across S (Fig. 3); complexity scales with (2)

$$c_S := \sum_{v\in V} \max_{s\in S} \sigma_s(v) \in [1, |S|] \tag{2}$$

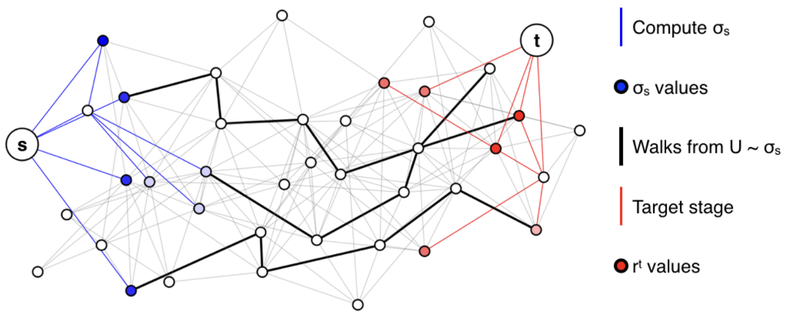

Figure 2: Scheme to sample walks from intermediate nodes

Joint target stage

Separate target stages separately traverse paths to each $t \in T$

Joint stage traverses overlapping paths once, avoiding duplicate computations (Fig. 4)

For $T = \{t_1, t_2, ..., t_{|T|}\}$, complexity reduction is (3), where $r^t_{\max} \in (0, 1)$ is algorithm input

$$c_T := \sum_{i=1}^{|T|} \left| \left\{ j \in \{1, 2, ..., i-1\} : \pi_{t_j}(t_i) > r^t_{\max} \right\} \right| \tag{3}$$

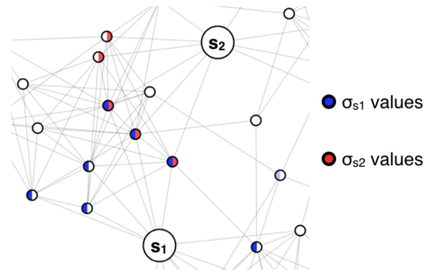

Figure 3: Walks from nodes with blue *and* red can be shared for both π_{s_1} and π_{s_2} estimates

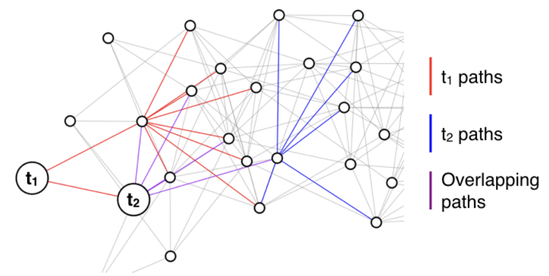

Figure 4: Purple paths traversed twice for separate target stages but only once for joint target stage

Experiments

Joint stage efficiency (Fig. 5): Walks, target stage iterations scale linearly using separate source/target stages but sublinearly using joint stages, corresponding to (2), (3).

Effect of clustering (Fig. 6): Complexity scale factors (2), (3) relate closely to the more common clustering measure *conductance*, defined for $U \subset V$ as

$$\Phi(U) = \frac{|\{u \to v \in E : u \in U, v \notin U\}|}{\min\{|\{u \to v \in E : u \in U\}|, |\{u \to v \in E : u \notin U\}|\}}.$$

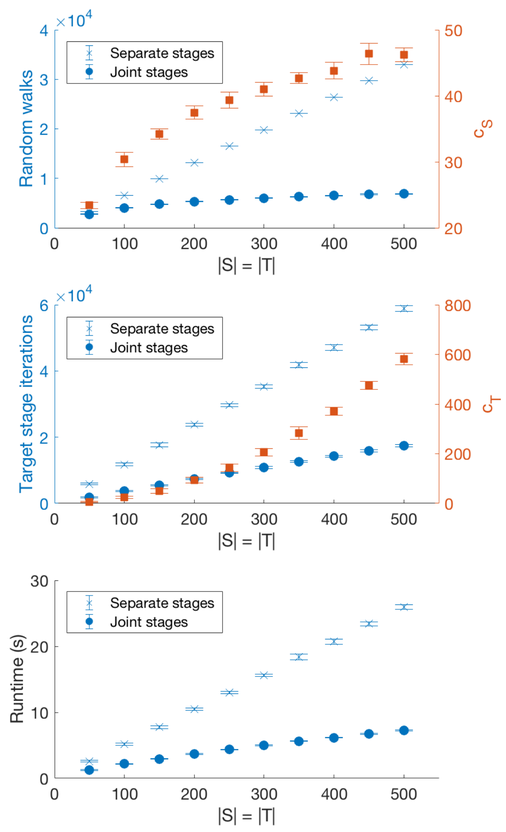

Figure 5: Joint stage efficiency, using directed Erdős-Rényi model with $|V| = 2000$, expected degree 10

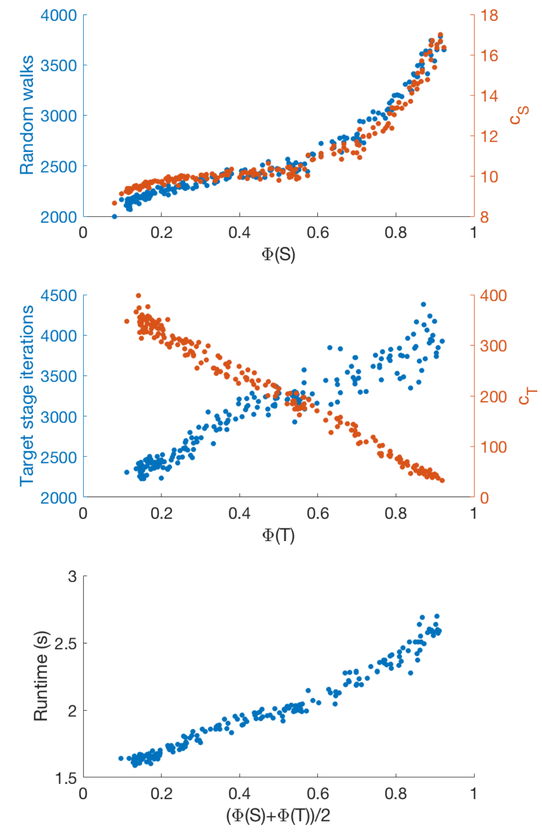

Figure 6: Effect of clustering, using directed stochastic block model with $|V| = 2000$, 20 communities, expected intra-community degree 9, expected inter-community degree 1

Conclusions

Key finding: PPR estimation is "easier" when sources and/or targets clustered

- Complexity (2) *low* when S clustered; complexity reduction (3) *high* when T clustered

Acknowledgements

Daniel Vial would like to acknowledge support from NSF via grant 1538827 and an ECE department fellowship. Vijay Subramanian would like to acknowledge support from NSF via grants 1343381, 1516075, 1538827 and 1608361.

References

[1] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Vahab Mirrokni, and Shang-Hua Teng. Local computation of pagerank contributions. *Internet Mathematics*, 5(1-2):23–45, 2008.

[2] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 475–486. IEEE, 2006.

[3] Konstantin Avrachenkov, Nelly Litvak, Danil Nemirovsky, and Natalia Osipova. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM Journal on Numerical Analysis*, 45(2):890–904, 2007.

[4] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514. ACM, 2013.

[5] Peter Lofgren, Siddhartha Banerjee, and Ashish Goel. Personalized pagerank estimation and search: A bidirectional approach. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 163–172. ACM, 2016.