# Approximate Graph Propagation


Hanzhi Wang[1], Mingguo He[1], Zhewei Wei[*1,2], Sibo Wang[3]
Ye Yuan[4], Xiaoyong Du[1], Ji-Rong Wen[1]

[1]Renmin University of China
[2]Pazhou Lab
[3]Beijing Institute of Technology
[4]The Chinese University of Hong Kong
[*] corresponding authors
Contact: zhewei@ruc.edu.cn

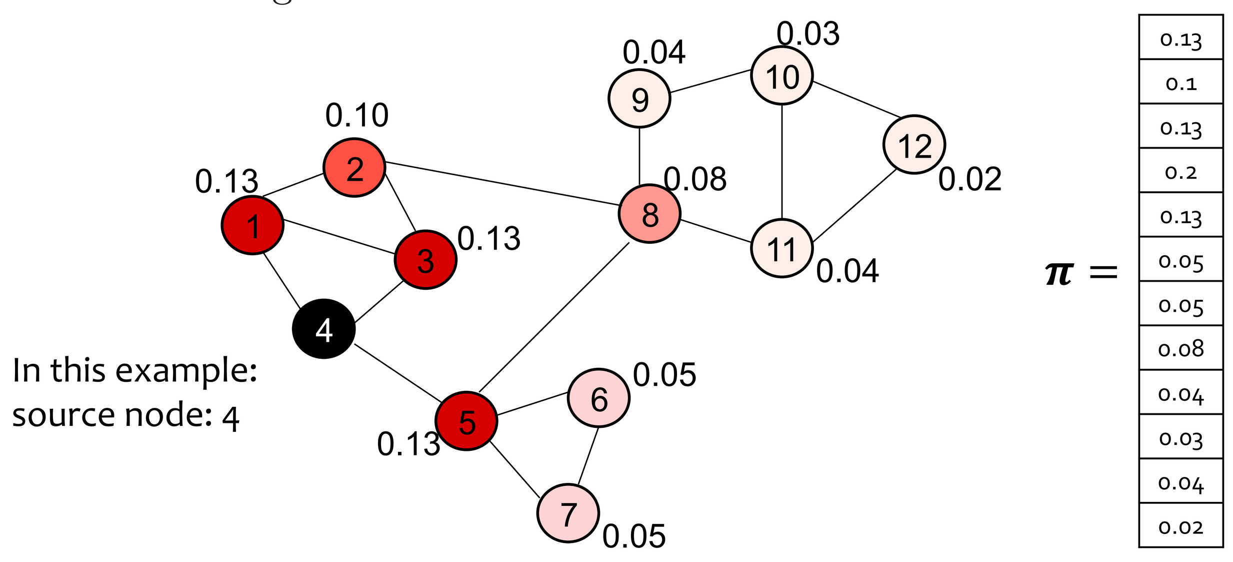


To appear at the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2021)


中國人民大學 RENMIN UNIVERSITY OF CHINA
琶洲实验室 PAZHOU LAB
香港中文大學 The Chinese University of Hong Kong
北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

## 1. Motivations

- **Node proximity:**
  - ◇ Node proximity measures the relative importance of nodes on the graph with respect to the given source node.
  - ◇ **Problem definition:** The node proximity query computes the node proximity vector $\boldsymbol{\pi}$, where $\boldsymbol{\pi}(v)$ denotes node $v$'s proximity w.r.t the given source node.




- **Popular node proximities**:

| Node proximity | Propagation equation |
|---|---|
| $L$-hop transition probability | $\boldsymbol{\pi} = \left(\mathbf{AD}^{-1}\right)^{L} \cdot \boldsymbol{e}_s$ |
| PageRank [1] | $\boldsymbol{\pi} = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i \cdot \left(\mathbf{AD}^{-1}\right)^i \cdot \frac{\mathbf{1}}{n}$ |
| Personalized PageRank [1] | $\boldsymbol{\pi} = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i \cdot \left(\mathbf{AD}^{-1}\right)^i \cdot \boldsymbol{e}_s$ |
| Single-target PPR [2] | $\boldsymbol{\pi} = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i \cdot \left(\mathbf{D}^{-1}\mathbf{A}\right)^i \cdot \boldsymbol{e}_t$ |
| Heat Kernel PageRank [3] | $\boldsymbol{\pi} = \sum_{i=0}^{\infty} e^{-t} \cdot \frac{t^i}{i!} \cdot \left(\mathbf{AD}^{-1}\right)^i \cdot \boldsymbol{e}_s$ |
| Katz [4] | $\boldsymbol{\pi} = \sum_{i=0}^{\infty} \beta^i \cdot \mathbf{A}^i \cdot \boldsymbol{e}_s$ |

- **Feature propagation in GNN**:

| GNN models | Propagation equation |
|---|---|
| SGC [5] | $\boldsymbol{\pi} = \left(\mathbf{D}^{-1/2}\mathbf{AD}^{-1/2}\right)^{L} \cdot \boldsymbol{x}$ |
| APPNP [6] | $\boldsymbol{\pi} = \sum_{i=0}^{L} \alpha(1-\alpha)^i \cdot \left(\mathbf{D}^{-1/2}\mathbf{AD}^{-1/2}\right)^i \cdot \boldsymbol{x}$ |
| GDC [7] | $\boldsymbol{\pi} = \sum_{i=0}^{L} e^{-t} \cdot \frac{t^i}{i!} \cdot \left(\mathbf{D}^{-1/2}\mathbf{AD}^{-1/2}\right)^i \cdot \boldsymbol{x}$ |

- **Two major questions:**
  - ◇ Is there a ***generalized graph propagation equation***?
  - ◇ Is there a ***universal algorithm*** that computes the approximate graph propagation with **near optimal cost**?

## 2. Contributions

- **We propose a Generalized Graph Propagation Equation:**

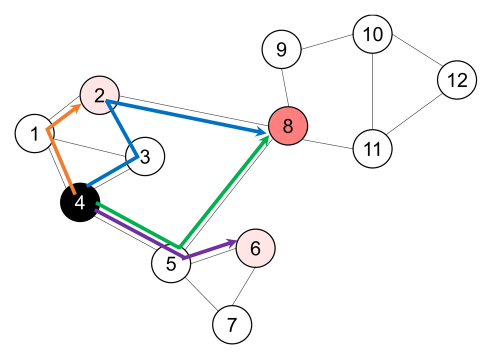
$$\boldsymbol{\pi} = \sum_{i=0}^{\infty} \boldsymbol{w_i} \cdot \left(\mathbf{D}^{-a}\boldsymbol{A}\mathbf{D}^{-b}\right)^{i} \cdot \boldsymbol{x}$$

| Node proximities / GNN models | weighted sequence $w_i$ | propagation matrix | graph signal $\boldsymbol{x}$ |
|---|---|---|---|
| transition probability/ SGC | $w_L = 1$, $w_i = 0\ (i \neq L)$ | $\mathrm{AD}^{-1}$/ $\mathrm{D}^{-1/2}\mathrm{AD}^{-1/2}$ | $\boldsymbol{e_s}$ / the feature vector $\boldsymbol{x}$ |
| PPR / APPNP | $\alpha(1-\alpha)^i$ | $\mathrm{AD}^{-1}$/ $\mathrm{D}^{-1/2}\mathrm{AD}^{-1/2}$ | $\boldsymbol{e_s}$ / the feature vector $\boldsymbol{x}$ |
| HKPR / GDC | $e^{-t} \cdot \frac{t^i}{i!}$ | $\mathrm{AD}^{-1}$/ $\mathrm{D}^{-1/2}\mathrm{AD}^{-1/2}$ | $\boldsymbol{e_s}$ / the feature vector $\boldsymbol{x}$ |
| Katz | $\beta^i$ | A | $\boldsymbol{e_s}$ |

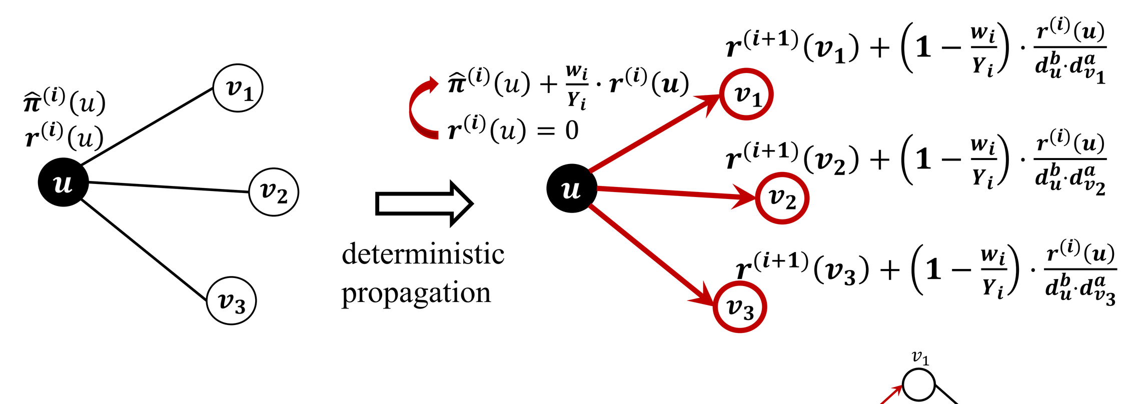
- We propose **Approximate Graph Propagation (AGP)**, **a unified randomized algorithm** that computes various node proximities and GNN models in **near optimal time**.

## 3. Previous Work

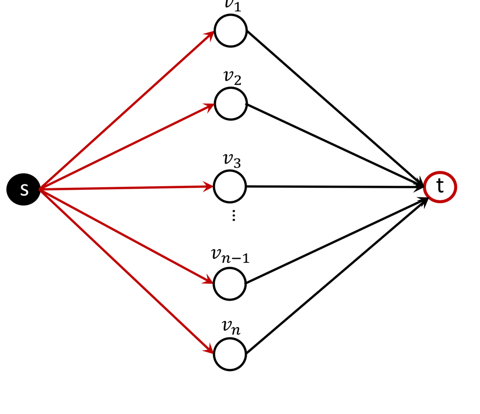
- **Monte-Carlo method:**
  - ◇ Generate multiple random walks from the source node
  - ◇ $\hat{\boldsymbol{\pi}}(v) = \frac{\#\text{ of walks terminate at } v}{\text{total }\#\text{ of walks}}$.
  - ◇ **Drawbacks:**
    - Multiple walks needed.
    - Don't support Katz ($a + b \neq 1$).



- **Deterministic propagation [8, 9]:**
  - ◇ residue $\boldsymbol{r}^{(i)}(u)$: the probability mass to be propagated to node $v$ in level $i$.
  - ◇ reserve $\hat{\boldsymbol{\pi}}^{(i)}(u)$: the probability mass stays at node $v$ in level $i$.




  - ◇ **Drawbacks:**
    In some bad cases, to derive $\hat{\boldsymbol{\pi}}(t)$:
    - Deterministic propagation: $O(n)$
    - Monte Carlo method: $O(1)$.

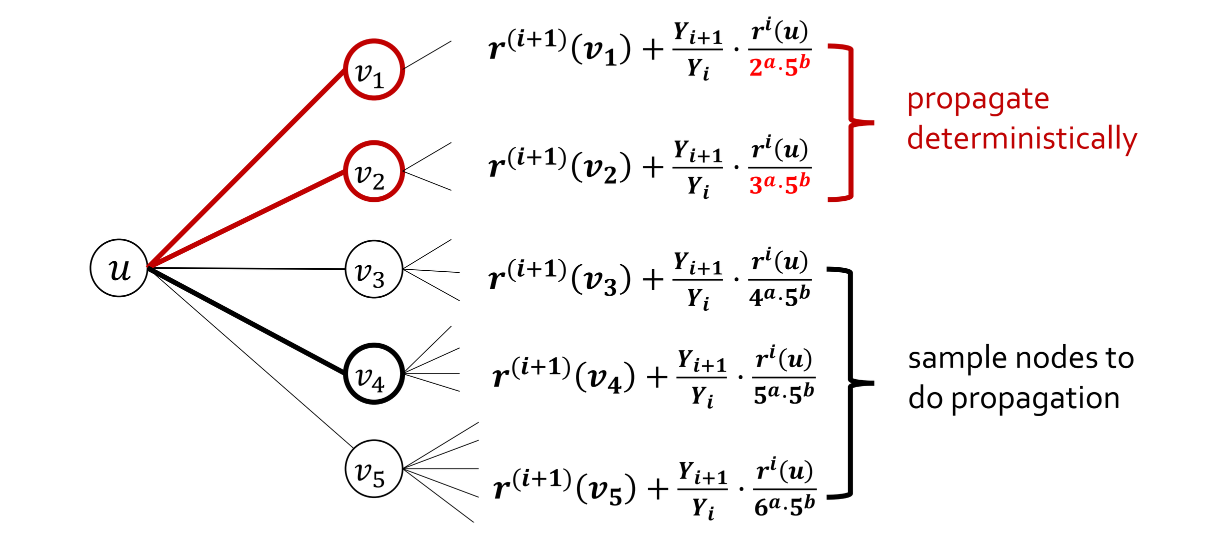


## 4. Algorithms

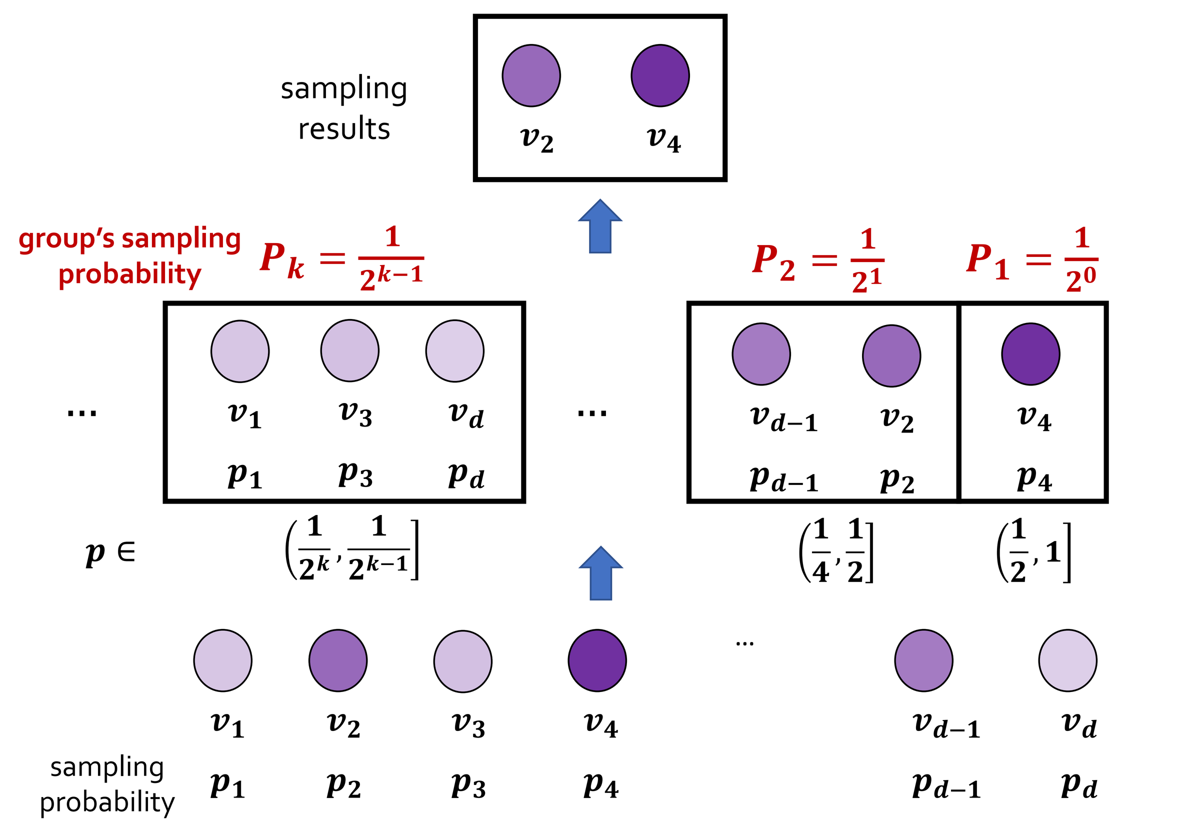
- **AGP:** combine the strengths of Monte-Carlo method and the deterministic propagation
  - ◇ In the propagation from node $u$ at level $i$ to $v$ at level $i+1$, the increment of $v$'s residue is:

$$\boldsymbol{r}^{i+1}(v) \leftarrow \boldsymbol{r}^{i+1}(v) + \frac{Y_{i+1}}{Y_i} \cdot \frac{\boldsymbol{r}^i(u)}{d_u^b \cdot d_v^a}.$$

  - ◇ **Pre-sorting adjacency list by degrees:**




  - ◇ **Randomized propagation with subset sampling:**
    - In each group, sample according to the binary distribution with the groups' sampling probability.
    - Reject the selected nodes w.p. $\frac{p_{individual}}{p_{group}}$ for unbiasedness.

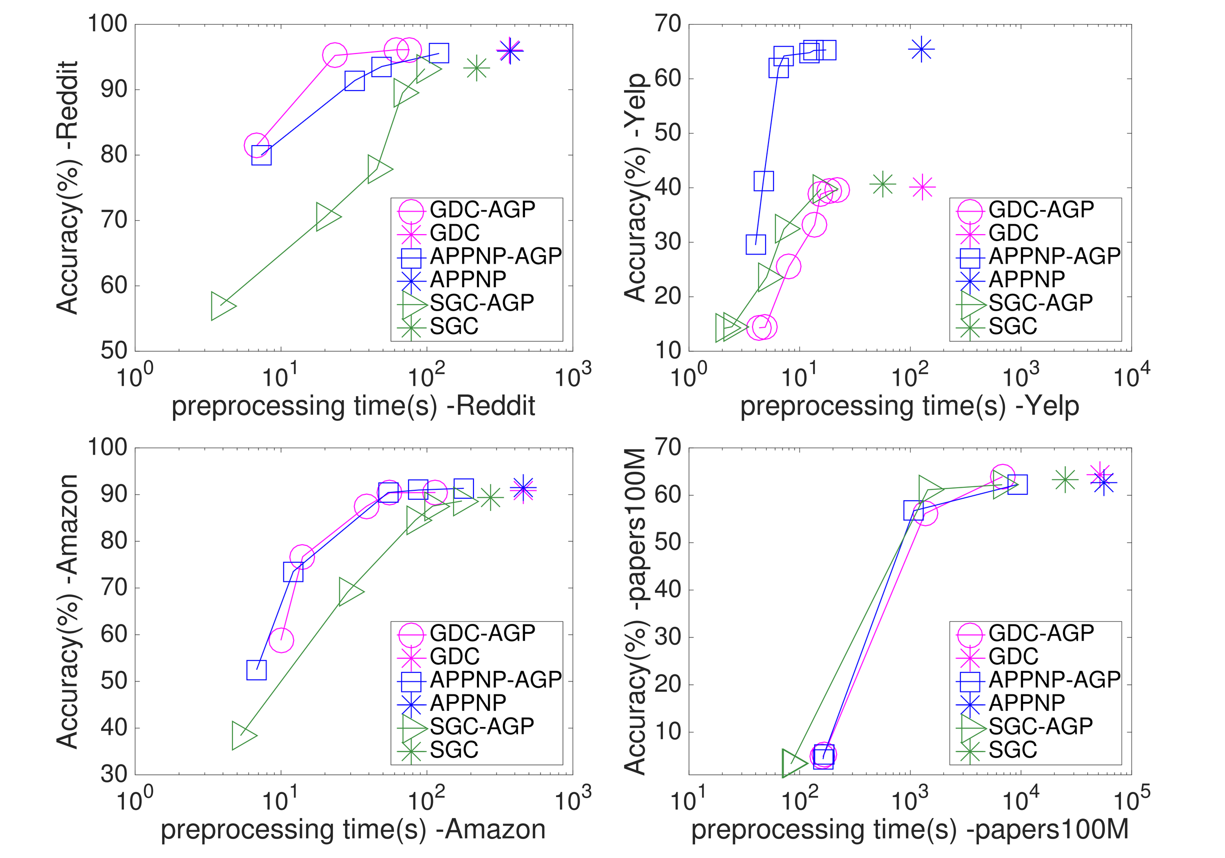



## 5. Experiments

- **Node Classification with GNN**

| Data Set | $n$ | $m$ | $d$ | Classes | Label % |
|---|---|---|---|---|---|
| Reddit | 232,965 | 114,615,892 | 602 | 41 | 0.0035 |
| Yelp | 716,847 | 6,977,410 | 300 | 100 | 0.7500 |
| Amazon | 2,449,029 | 61,859,140 | 100 | 47 | 0.7000 |
| Papers100M | 111,059,956 | 1,615,685,872 | 128 | 172 | 0.0109 |

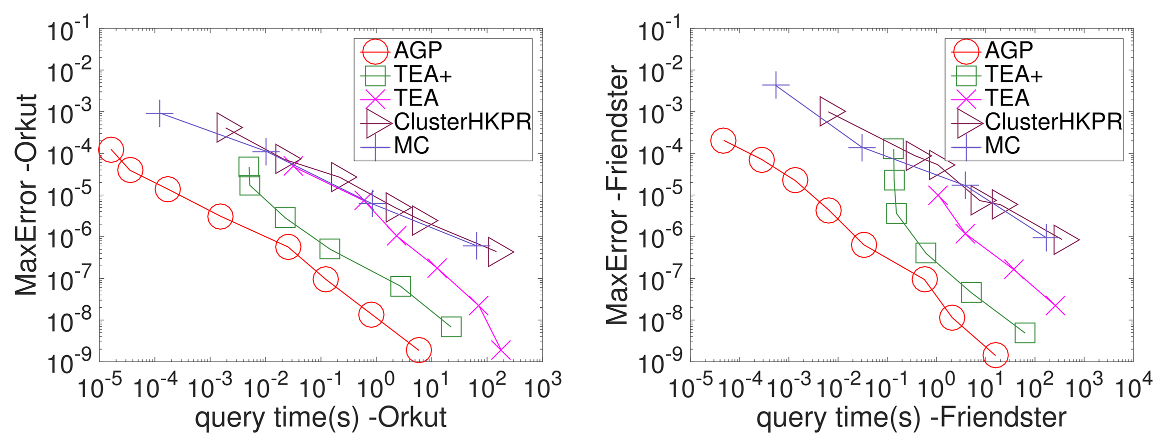
  - ◇ **Tradeoffs between** ***Accuracy*** **(%) and propagation time:**



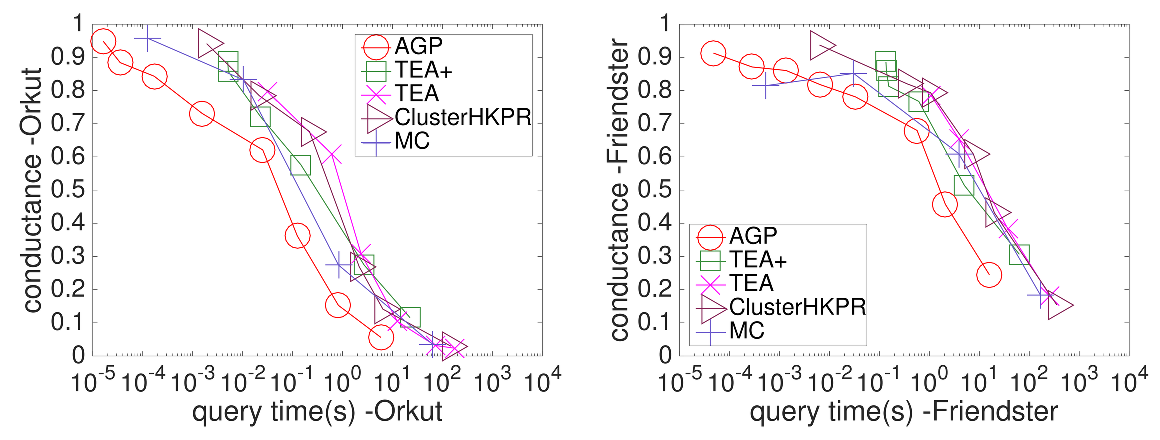

- **Local Clustering with HKPR**

| Dataset | Type | $n$ | $m$ |
|---|---|---|---|
| Orkut-Links | undirected | 3,072,441 | 234,369,798 |
| Friendster | undirected | 68,349,466 | 3,623,698,684 |

  - ◇ **Tradeoffs between** ***MaxError*** **and query time:**




  - ◇ **Tradeoffs between** ***conductance*** **and query time:**




## 6. Reference


[1] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[2] Peter Lofgren and Ashish Goel. Personalized pagerank to a target node. *arXiv preprint arXiv:1304.4658*, 2013.

[3] Fan Chung. The heat kernel as the pagerank of a graph. *PNAS*, 104(50):19735–19740, 2007.

[4] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[5] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *ICML*, pages 6861–6871. PMLR, 2019.

[6] Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2019.

[7] Johannes Klicpera, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *NeurIPS*, pages 13354–13366, 2019.

[8] Reid Andersen, Fan R. K. Chung, and Kevin J. Lang. Local graph partitioning using pagerank vectors. In *FOCS*, pages 475–486, 2006.

[9] Siddhartha Banerjee and Peter Lofgren. Fast bidirectional probability estimation in markov models. In *NIPS*, 2015.