Paper ID : 053

Panther: Fast Top-k Similarity Search on Large Networks

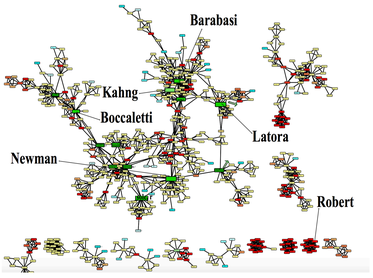
Jing Zhang[1], Jie Tang[1], Cong Ma[1], Hanghang Tong[2], Yu Jing[1], Juanzi Li[1]

[1]Department of Computer Science and Technology, Tsinghua University

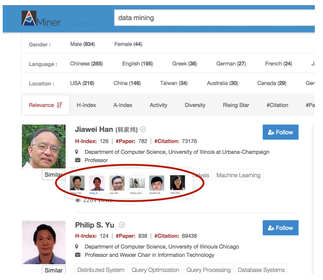
[2]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University

Goal: Develop a Fast Top-k Similarity Algorithm for Large Networks

- Who are similar with Barabási?

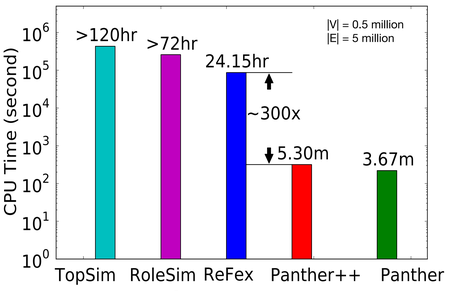


- Similar authors in Aminer



* Working with billions of edges.

- Efficiency Perforamnce



* T=5, $\varepsilon=\sqrt{1/|E|}$, δ=0.1, R=16,609,640

Our Approach: Panther

① Panther$_{ps}$

Path Similarity

Two vertices are similar if they frequently appear on the same paths.

$$S_{ps}(v_i, v_j) = \frac{\sum_{p \in P(v_i, v_j)} w(p)}{\sum_{p \in \Pi} w(p)}$$

A path is a T-length sequence of vertices p=(v_1, …, v_{T+1}).

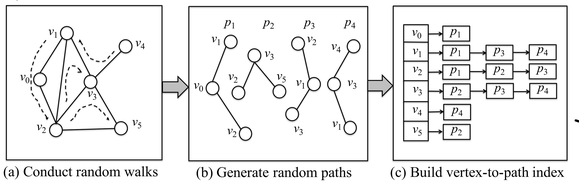
Π is all the T-paths in G.

Random Path

Uniformly sample a starting node.

Random walk according to the transition probability:

$$t_{ij} = \frac{w_{ij}}{\sum_{v_k \in \mathcal{N}(v_i)} w_{ik}}$$



If path weight is: $w(p) = \prod_{i=1, j=i+1}^{T} t_{ij}$.

Then path similarity is:

$$S_{ps}(v_i, v_j) = \frac{|P(v_i, v_j)|}{R}$$

Build index to improve O(RT) to O(dT) when finding all paths of a vertex.

How many random paths shall we sample?

Theorem 1

Let $\mathcal{R}$ be a range set on a domain $\mathcal{D}$ with $VC(\mathcal{R}) \le d$, and let ϕ be a distribution on $\mathcal{D}$. Given $\varepsilon, \delta \in (0,1)$, let S be a set of $|S|$ points sampled from $\mathcal{D}$ according to ϕ, with

$$|S| = \frac{c}{\varepsilon^2}(d + \ln\frac{1}{\delta}),$$

where c is a universal positive constant. Then S is a ε-approximation to $(\mathcal{R}, \phi)$ with probability of at least $1-\delta$.

① Domain is Π , range set is $\mathcal{R}_G = \{P(v_i, v_j) : v_i, v_j \in V\}$

② $VC(\mathcal{R}_G) \le \log_2 \binom{T}{2} + 1$

③ $\phi(p) = \text{prob}(p) = \frac{w(p)}{\sum_{p \in \Pi} w(p)}$, path similarity is $\phi(P(v_i, v_j))$

#Random paths: $R = \frac{c}{\varepsilon^2}(\log_2 \binom{T}{2} + 1 + \ln\frac{1}{\delta})$

② Panther$_{vs}$

Use top-D path similarities calculated by Panther$_{ps}$ to represent a vector:

$$\theta(v_i) = (S_{ps}(v_i, v_{(1)}), S_{ps}(v_i, v_{(2)}), \ldots, S_{ps}(v_i, v_{(D)}))$$

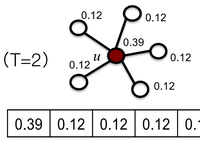
$$S_{vs}(v_i, v_j) = \frac{1}{\|\theta(v_i) - \theta(v_j)\|}$$

Build kd-tree based on the Euclidean distance between any vectors.

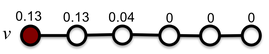
Vector Simlarity

The probability distributions of a vector linking to all other vertices are similar if their topology structures are similar.

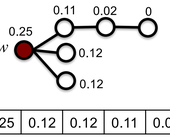
(T=2)



0.39	0.12	0.12	0.12	0.12	0.12



0.13	0.13	0.04	0	0	0



0.25	0.12	0.12	0.11	0.02	0

S_{vs}(u, w)=0.27 > S_{vs}(u, v)=0.16

Time complexity Panther$_{ps}$: O(RTc+NdT), Panther$_{vs}$: O(RTc+NdT+Nc)

Experiments

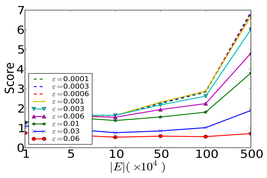
Data set:
- Tencent weibo: |V| = 0.3 billion, |E| = 6 billion.
- Extract 11 different Tencent networks.

Experimental Design

Efficiency Performance

Accuracy Performance

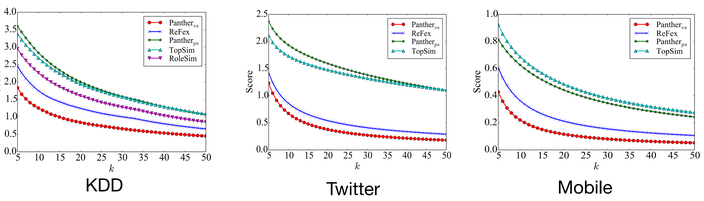
Parameter Analysis



- $(1/\varepsilon)^2$ is almost linearly proportional to |E|.
- Stable when T=5 and D=50. Refer to our paper for details.

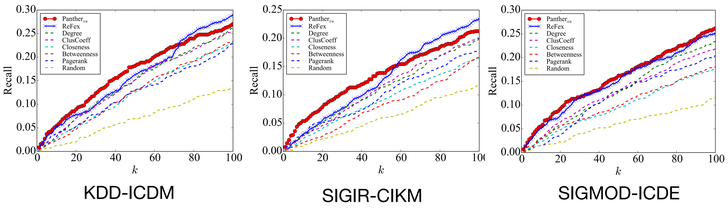
Accuracy of Panther$_{ps}$

- Evaluate how Panther$_{ps}$ can approximate common neighbors.



Accuracy of Panther$_{vs}$

- Evaluate how Panther$_{vs}$ can match the same identities between two networks.



Data sets:
- Co-author networks: |V|=3K, |E| = 7K.
- Twitter network: |V| = 100K, |E| = 500K.
- Mobile network: |V| = 200K, |E| = 200K