3 Efficient graph mining techniques for big data

Fast algorithms for large-scale graphs

***Abstract*— Recent advances in information science have shown that linked data pervade our society and the natural world around us. Graphs have become increasingly important for representing complicated structures and schema-less data such as those generated by Wikipedia, Freebase, and various social networks. However, existing algorithms cannot handle large graphs efficiently, so fast algorithms are needed. We introduce two fast algorithms for identifying the top-k nodes of personalized PageRank and graph clustering. They outperform previous algorithms in terms of both speed and quality. Personalized PageRank and graph clustering are fundamental to many applications. Our algorithms allow many applications to be processed more efficiently and will help to improve the effectiveness of future applications.**

■ Fast algorithm for identifying top-k nodes of Personalized PageRank [1]

POINT
- **Identify important nodes for query nodes**
- **Up to 100x faster than original algorithm**

◎ Use matrices representation for non-iterative computation

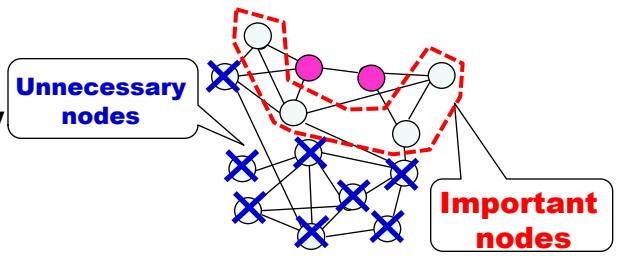
$$\begin{bmatrix} S \end{bmatrix} = (1-c)\begin{bmatrix} G \end{bmatrix}\begin{bmatrix} S \end{bmatrix} + c\begin{bmatrix} N \end{bmatrix}$$

***S*: Score vector N: Query vector**
***G*: Adjacency matrix *c*: Restart probability**

$$\begin{bmatrix} S \end{bmatrix} = C\begin{bmatrix} P^T \end{bmatrix}\left(\begin{bmatrix} Q \end{bmatrix}\begin{bmatrix} R \end{bmatrix}\right)^{-1}\begin{bmatrix} P \end{bmatrix}\begin{bmatrix} N \end{bmatrix}$$

P: Permutation matrix Q: Orthogonal matrix
R: Upper triangular matrix

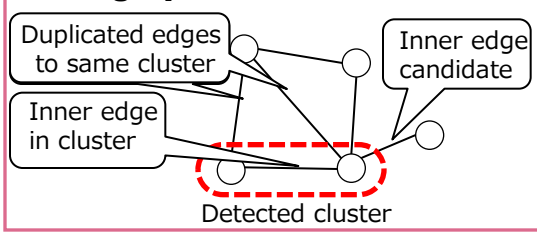
◎ Prune unnecessary nodes during top-k node search



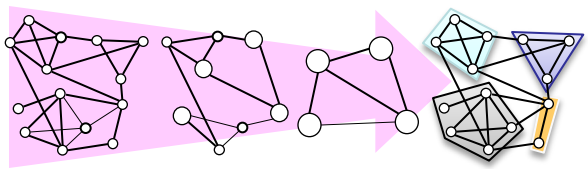
■ Fast algorithm for Graph Clustering [2]

POINT
- **Detect communities in graphs**
- **Graph with 100 million nodes is analyzed in 3 minutes**

◎ Prune unnecessary edges in graphs



◎ Incrementally aggregate nodes during hierarchical clustering



Related works

[1] Y. Fujiwara, M. Nakatsuji, T. Yamamuro, H. Shiokawa, M. Onizuka, "Efficient Personalized PageRank with Accuracy Assurance," in *Proc. the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.
[2] H. Shiokawa, Y. Fujiwara, M. Onizuka, "Fast Algorithm for Modularity-based Graph Clustering," in *Proc. the 27th AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

Contact

Makoto Onizuka Distributed Computing Technology Project, NTT Software Innovation Center
E-mail : onizuka.makoto{at}lab.ntt.co.jp (Please replace {at} with @)