香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

Keyword-Centric Community Search

PI : **Dr. ZHANG Zhiwei**

Funding Scheme: **General Research Fund**
Project Ref. No.: **12258116**
Amount Awarded (to HKBU): **HK$ 712,325**
Project Period: **Jan 2017 - Dec 2019**

OBJECTIVES

1. To formally analyze the differences and benefits of a new framework based on graph contraction compared to existing frameworks.
2. To design new algorithms for basic graph operations, including node centric graph operations and edge centric graph operations.
3. To efficiently integrate these developed basic graph algorithms for graph applications.

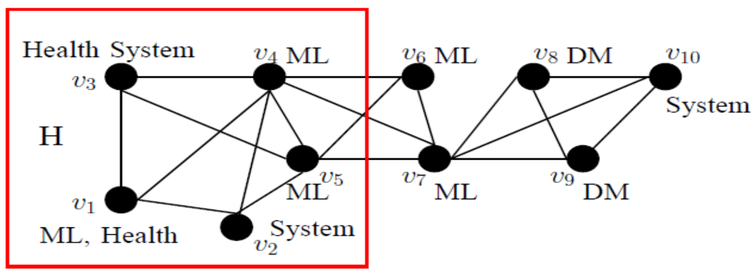
HIGHLIGHTS

Problem Statement

- **Community Search Query with Keywords in Large Graphs**
  - *ACQ* query aim to the number of keywords in the query shared by all vertices.
  - *ATC* query aim to minimize the distance from each node to the query vertex.
  - Existing solutions consider the keyword relationships and graph structure separately.
  - The time cost is high.

Keyword-Centric Community Search

- **Keyword-Structure Measurement**
  - Given a graph G, a set of keywords W and an integer k:
    - The community to find is a k-core,
    - *Keyword closeness*: there does not exist any subgraph H', such that H is minimal among all the subgraphs $H' \nsubseteq H$, and $KwdC(W, H') < KwdC(W, H)$.
    - $KwdC(W, H)$: largest value of the shortest distance from each vertex in H to each keyword in W.

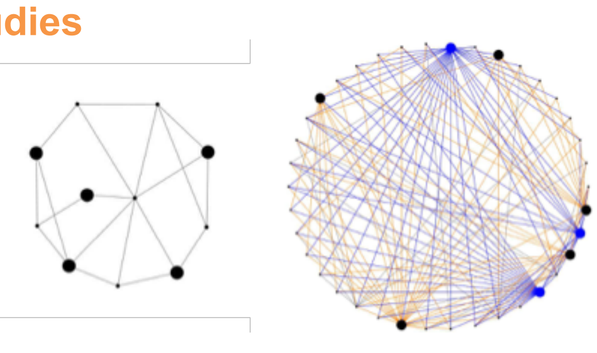


Our Approach

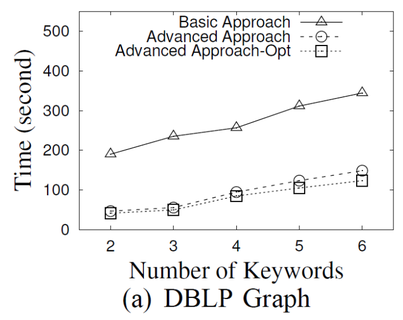
- **Idea: iteratively update $cd^L(u)$ according to the keyword-closeness and structural cohesiveness conditions.**
  - Given $d' = cd^L(u)$: u is only possible belong to $G = KC_k^d$ with $d \geq d'$:
  - Theorem: if for each node with $cd^L(u) \leq d$, $cd^L(u)$ cannot be updated, then $G = KC_k^d$ is the induced graph of the nodes with $cd^L \leq d$.
  - Algorithm
    - (recursively) compute the k-core of the induced graph of the node with $cd^L \leq d$.
    - update cd^L
    - return the induced subgraph of the nodes with the minimum d at the condition that the nodes with $cd^L \leq d$ will not change their cd^L value.

Performance Studies

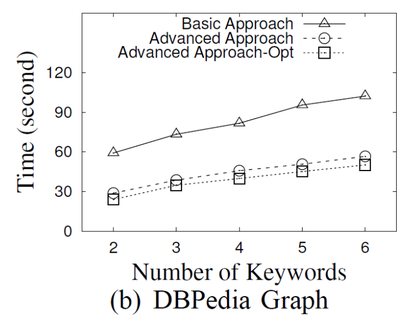
- **Effectiveness**



- **Efficiency**



(a) DBLP Graph



(b) DBPedia Graph

SELECTED PUBLICATIONS

1. Z. Zhang, X. Huang, J. Xu, B. Choi, and Z. Shang, “Keyword-Centric Community Search,” in *Proceedings of the 35th IEEE International Conference on Data Engineering* (ICDE ’19), Macau SAR, China, 2019.
2. Z. Shang, J. X. Yu, and Z. Zhang, "TuFast: A Lightweight Parallelization Library for Graph Analytics." *in Proceedings of the 35th IEEE International Conference on Data Engineering (*ICDE '19), Macau SAR, China, 2019.
3. Z. Zhang, H. Wei, J. Xu, and B. Choi, “GScan: Exploiting Sequential Scans for Subgraph Matching,” in *Proceedings of the 24th International Conference on Database Systems and Advanced Applications.* (DASFAA ’19), Chiang Mai, Thailand, 2019.