UNIVERSITY OF CALIFORNIA
SANTA BARBARA
THE UNIVERSITY of EDINBURGH
informatics

Resource-Bounded Graph Query Answering

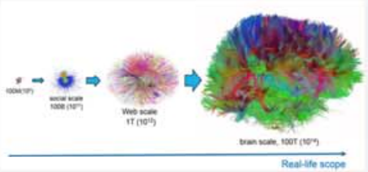
Wenfei Fan, Xin Wang, Yinghui Wu
{wenfei@inf, x.wang-36@sms}.ed.ac.uk yinghui@cs.ucsb.edu
西南交通大学 Southwest Jiaotong University

OVERVIEW

Motivation

Querying Big graphs

Given query Q, find answers Q(G) for Q from data graph G



Challenge

- Real graphs are Huge.
- Tractable methods could be in feasible!
- **Real-world applications require searching with limited resource**
- How to evaluate query using limited resource?

Contributions

Resource-bounded query answering

- Accessing small amount of data for accurate answers
- A tunable graph querying framework
- Balance computing resource and answer quality

Efficient resource bounded algorithms for

- Localized queries: strong simulation & subgraph queries
- Non-localized queries: reachability

Related Work: BlinkDB, budgeted search, graph indexing & compression, MapReduce, GraphLab...

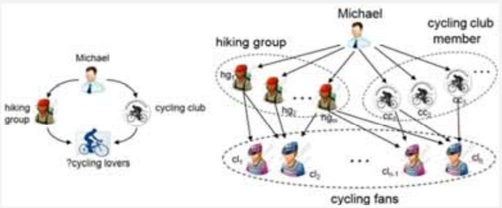
Graph Queries

Localized graph pattern queries

◊ A match can be determined by exploring its d_Q (diameter of Q) hops

◊ Simulation queries: strong simulation relation with a personalized node

◊ Subgraph isomorphism: injective mapping



Matching relation: (Michael, Michael), (hiking group, hg_m), (cycling club, cc_1),(cycling club, cc_3), (cycling lover, cl_{n-1}), (cycling lover, cl_n)

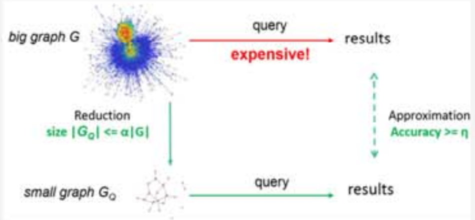
Non-localized queries

◊ We consider reachability queries: visit the entire G in the worst case

Question

How to effectively evaluate graph pattern queries and reachability with lim-

Resource-bounded Graph Querying



Resource-bounded algorithm for query class with resource bound α and accuracy guarantee η

- Access small (bounded) amount of G
- Guaranteed result quality
- Balance resource and answer quality

Resource-bounded graph query answering

Given: a query class L, α in (0,1] and η in (0,1], find algorithm with resource bound α and accuracy guarantee η

NP-hard for simulation and subgraph queries
Impossible for reachability queries

Resource-bounded Graph Simulation

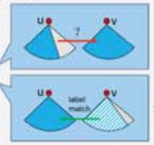
Auxiliary Information (Preprocessing)

Local information that benefits dynamic reduction

degree	|neighbor|	<label, frequency>	...

Dynamically updated during processing

Boolean guarded condition
Cost function c(u,v)
p(u,v), estimated probability that v matches u
bound b, determines an upper bound of the number of nodes to be visited



Dynamic Reduction

◊ Iteratively process each query edge

◊ Dynamically update cost and probability

◊ Select promising nodes

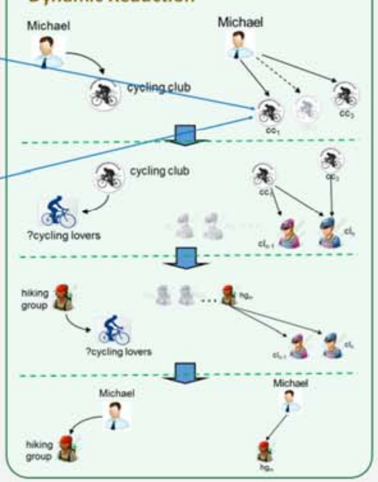
◊ Update G_Q until resource bound reached/no unvisited nodes

Approximate Querying

◊ apply graph simulation algorithms over G_Q to compute matches

Resource bounded graph simulation

Dynamic Reduction



Resource-bounded Reachability

Hierarchical Landmark Index

Landmarks are used to encode reachability.
Idea: select bounded number of landmarks for approximate reachability.

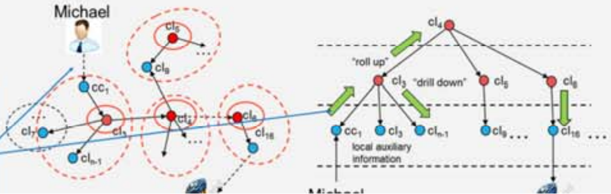
Dynamically updated auxiliary Information during processing

Boolean guarded condition (landmark labels)
Cover size (no. of node pairs covered)
Cost(v), size of subtree rooted at v
Potential p(v), total cover size of unvisited nodes in subtree rooted at v

Guided reachability search

◊ Bi-directed search with guided "roll-up"/"drill-down"

◊ Terminates if "yes" is determined or no unvisited nodes in the index



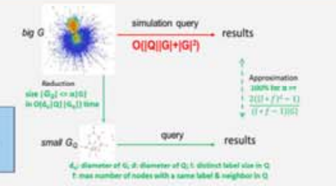
Resource-bounded Querying: Summary

Resource-bounded Simulation querying/ Subgraph isomorphism

accuracy measure

$$prec = \frac{|Y \cap Q(G)|}{|Y|} \quad recall = \frac{|Y \cap Q(G)|}{|Q(G)|}$$

$$acc = 2\,prec * recall/(prec + recall)$$



Resource-bounded reachability

accuracy measure

$$prec = \frac{tp}{tp + fp} \quad recall = \frac{tp}{tp + fn}$$

Other types of resource bounds, accuracy measures, graph query classes, applications...

RESULTS

Dataset

◊ Yahoo Web graph (http://webscope.sandbox.yahoo.com/catalog.php?datatype=g)

◊ Youtube (http://netsg.cs.sfu.ca/youtubedata)

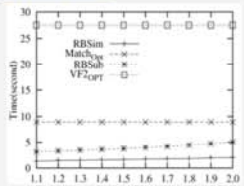
Baseline

◊ Resource bounded reachability **RBReach VS LM**: applying landmark vectors; **BFS** and **optimized BFS** over compressed graphs

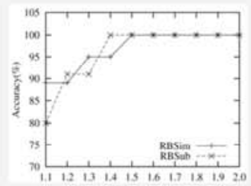
◊ Resource bounded simulation algorithm **RBSim VS** Optimized strong simulation **MatchOpt**

◊ Resource bounded subgraph isomorphism **RBSub VS** Optimized **VF2**

Evaluation



Yahoo, Efficiency, varying α (10^{-5})



Yahoo, Accuracy, varying α (10^{-5})

Big Picture

Application