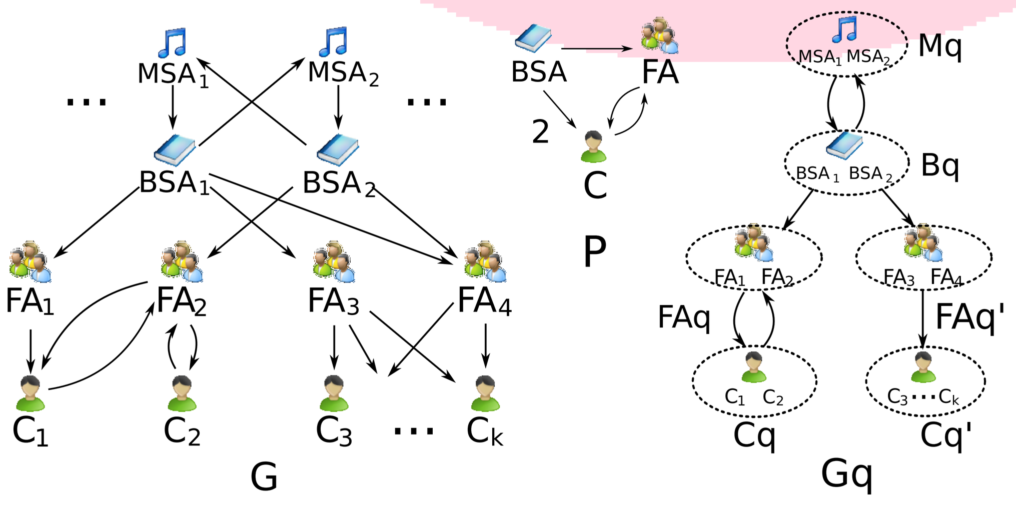
Query Preserving Graph Compression

Wenfei Fan [1,2], Jianzhong Li [2], Xin Wang [1], Yinghui Wu [1,3]
[1] University of Edinburgh [2] Harbin Institute of Technology
[3] UC Santa Barbara

{wenfei@inf.,x.wang-36@sms.,y.wu-18@sms.}ed.ac.uk,lijzh@hit.edu.cn

Introduction

- Queries over large real-life graphs are prohibitively expensive.
  - Reachability queries: $O(|V|+|E|)$ for $G(V,E)$
  - bounded simulation (**relation-based, edge-path matching**): $O(|E_p||V|^2)$ for $Q(V_p, E_p)$
- Indexing methods with construction and maintenance cost
- *unlikely* to lower the computational complexity
- Graph compression: construct compressed graphs which preserve information *only* related to a class of queries of users' choice



Querying Recommendation Network

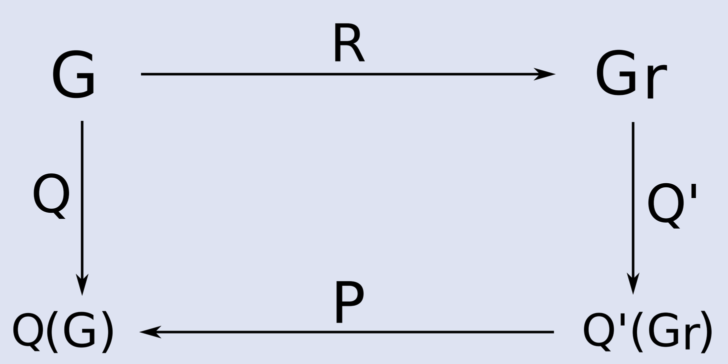
Query Preserving Graph Compression

Query Preserving Graph Compression. A triple $<R,F,P>$ where

- R: a *compression function*,
- $F \subseteq L_q \times L_q$: a *query rewriting function* for a class of graph queries L_q, and
- P: a *post-processing* function.

For any data graph G, $Q \in L_q$ and $G_r = R(G)$,

- $Q(G) = P(Q'(G_r))$,
- *any* query evaluation algorithm can be directly applied on G_r, without decompression,
- indexing and optimization can be directly applied on G_r.



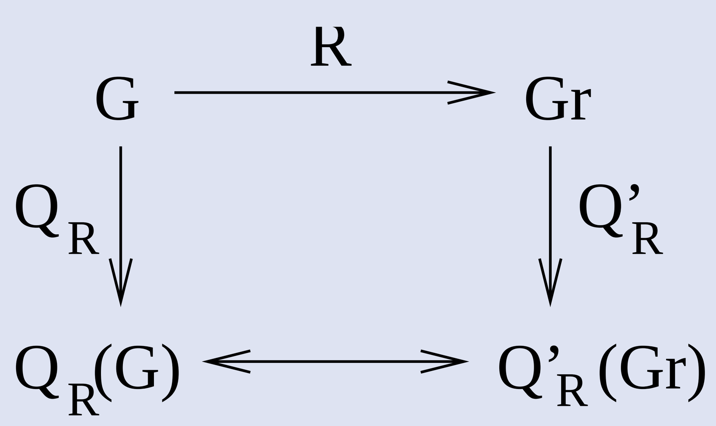
Query Preserving Graph Compression

Reachability Preserving Compression

Reachability equivalence relation. A node pair $(u,v) \in R_e$ if they have the same set of ancestors and descendants in G.

Theorem: There is a reachability preserving compression $<R, F>$ for G where

- R maps each node in G to its reachability equivalence class
- F maps each node in Q to its reachability equivalence class



Reachability Preserving Compression

Graph Pattern Preserving Compression

Bisimulation relation. A binary relation B over V of G, s.t. for each $(u,v) \in B$,

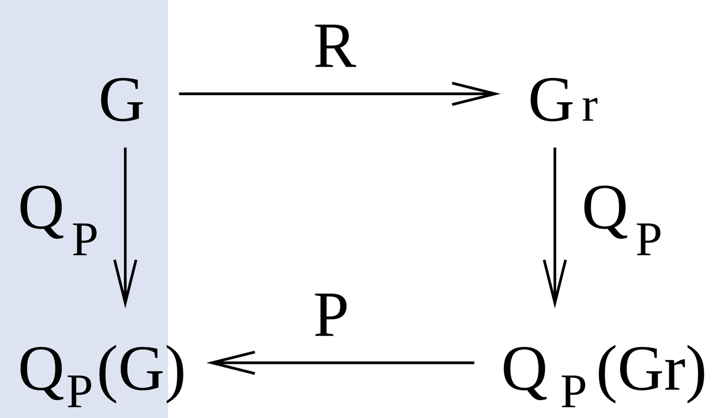
- the label of u and v are equivalent, and
- for each u's (resp. v's) child u' (resp. v'), v (resp. u) has a child v' (resp. u') that $(u',v') \in B$.

Theorem: There is a graph pattern preserving compression $<R, F, P>$ for G where

- R maps each node v in G to its bisimulation equivalence class $[v]$($O(|E|\log|V|)$)
- F is the identity mapping
- P maps each query node u and its match (as an equivalence class $[v]$) to node pairs (u,v') for each $v' \in [v]$ (linear time in the size of query result)

Algorithm.

- Compute the *unique maximum* bisimulation relation by iteratively refine the equivalence classes (initialized as V);
- Construct the compression graph G_r where each node denotes a bisimulation equivalence class, and each edge connects two nodes $[v_1]$ and $[v_2]$ if (v_1, v_2) is an edge in G.



Graph Pattern Preserving Compression

Incremental Query Preserving Compression

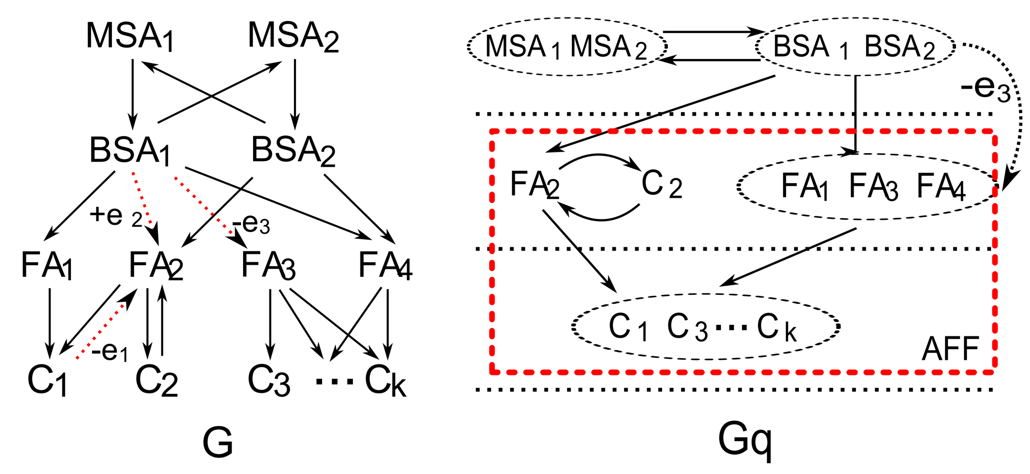
- Real-life graphs are changing. To compute the compressed graph from scratch is expensive.
- *Incremental graph compression*: given a data graph G, its changes ΔG, and a compressed graph G_r, compute ΔG_r, *i.e.,changes* to G_r, such that $G_r \oplus \Delta G_r = R(G \oplus \Delta G)$

Affected area: the total changes in the data graph ΔG and the compressed graph ΔG_r. Unbounded, bounded, optimal ...

- Incremental reachability preserving compression is unbounded even for unit updates, and is in $O(|AFF||G_r|)$ time
- Incremental pattern preserving compression is unbounded for unit updates, and is in $O(|AFF|^2 + |G_r|)$ time

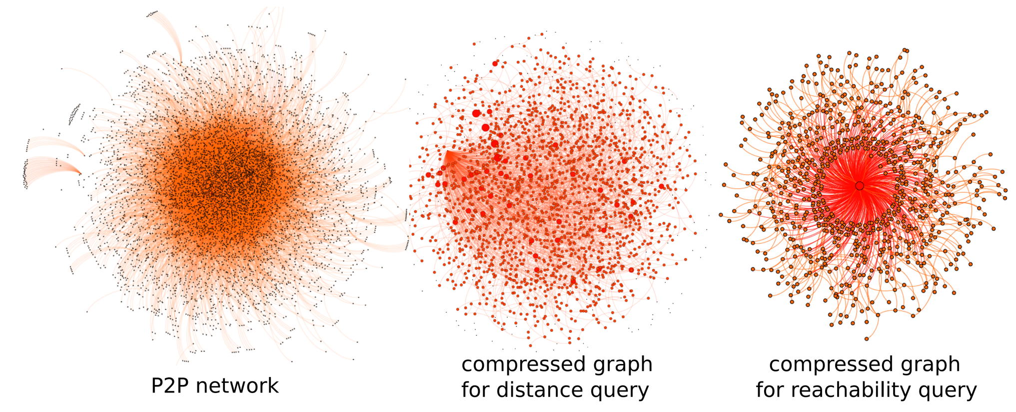
Algorithms.

- Update the ranks of the nodes (blocks), identify initial affected area
- Split-merge the blocks and propagate the affected area, until a fixpoint is reached



Incremental Compression

Experimental Study

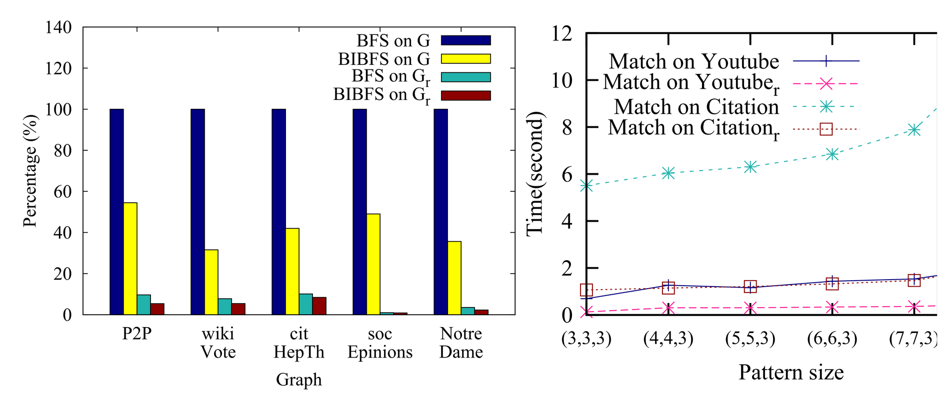


Compressing P2P network

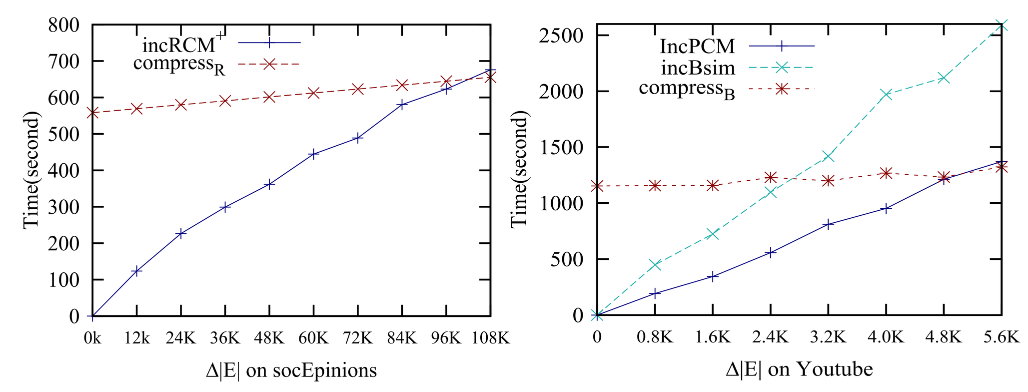
dataset	$\|G\|(\|V\|,\|E\|)$	RC$_{aho}$	RC$_{scc}$	RC$_r$
facebook	$1.6M$ $(64K, 1.5M)$	13.19%	5.89%	0.028%
amazon	$1.5M$ $(262K, 1.2M)$	35.09%	18.94%	0.18%
Youtube	$931K(155K, 796K)$	41.60%	17.02%	1.77%
wikiVote	$111K(7K, 104K)$	65.56%	8.33%	1.91%
wikiTalk	$7.4M$ $(2.4M, 5.0M)$	48.21%	16.82%	3.27%
socEpinions	$585K$ $(76K, 509K)$	29.53%	19.59%	2.88%
NotreDame	$1.8M$ $(326K, 1.5M)$	43.27%	10.75%	2.61%
P2P	$27K$ $(6K, 21K)$	73.24%	17.02%	5.97%
Internet	$155K$ $(52K, 103K)$	88.32%	28.89%	16.08%
citHepTh	$381K$ $(28K, 353K)$	71.32%	37.15%	14.70%

dataset	$\|G\|(\|V\|,\|E\|,\|L\|)$	PC$_r$
California	$26K$ $(10K, 16K, 95)$	45.9%
Internet	$155K$ $(52K, 103K, 247)$	29.8%
Youtube	$951K$ $(155K, 796K, 16)$	41.3%
Citation	$1.2M$ $(630K, 633K, 67)$	48.2%
P2P	$27K$ $(6K, 21K, 1)$	49.3%

Compression Ratio: Reachability (in average 5%) and Pattern Preserving (in average 43%)



Query efficiency: Reachability (in average 2%) and Pattern Preserving (in average 30%)



Incremental maintenance

Conclusion

- construct compressed graphs that can be directly queried without decompression
- Reachability and pattern preserving compression are efficient, and can be maintained without accessing original graphs