Discovering Maximal Motif Cliques in Large Heterogeneous Information Networks

Jiafeng Hu[1] Reynold Cheng[1] Kevin Chang[2] Aravind Sankar[2]
Yixiang Fang[1] and Brian Lam[3]

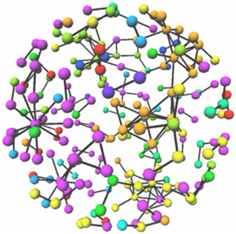
[1]Dept. of CS, The University of Hong Kong
[2]Dept. of CS, University of Illinois at Urbana-Champaign
[3]Metabolic Research Laboratories, University of Cambridge

April 10, 2019

Graphs are Everywhere

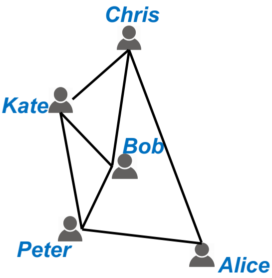
Social Network

E-Commerce



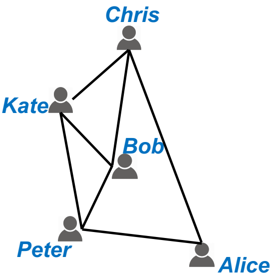
Protein-Protein interaction Network

Homogeneous VS Heterogeneous Graphs

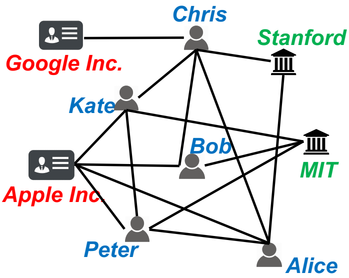


(a) Homogeneous

Homogeneous VS Heterogeneous Graphs

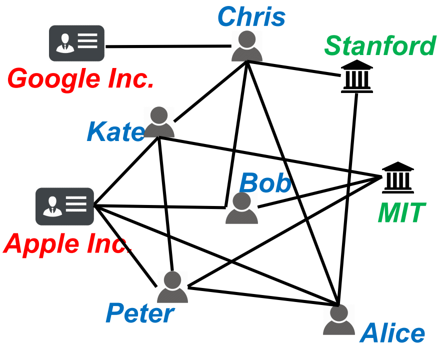


(a) Homogeneous



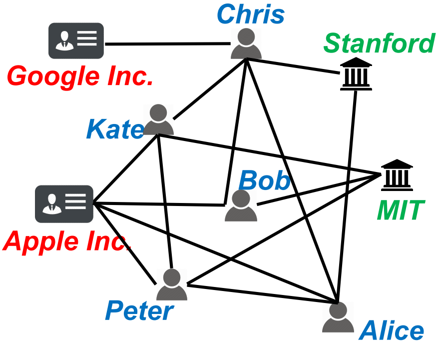
(b) Heterogeneous

Heterogeneous Information Networks (HINs)

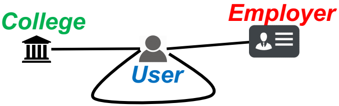


(a) An HIN

Heterogeneous Information Networks (HINs)



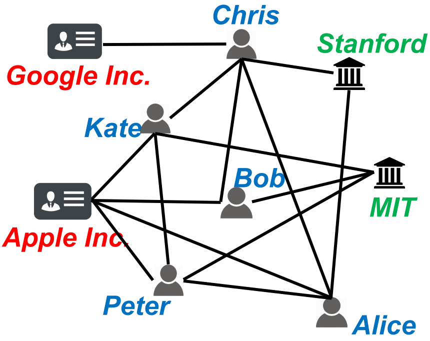
(a) An HIN



(b) Schema

How to Find Cliques for HINs?

- Clique: complete subgraph [E. Akkoyunlu, 1973]



How to Find Cliques for HINs?

- Clique: complete subgraph [E. Akkoyunlu, 1973]

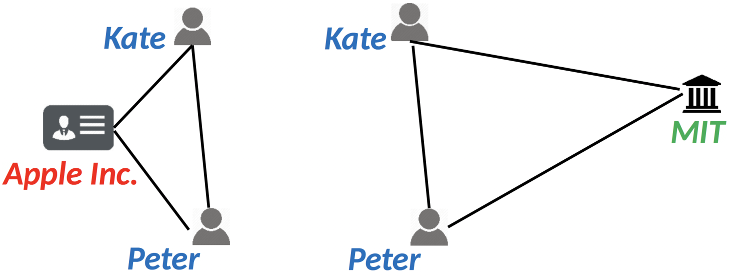
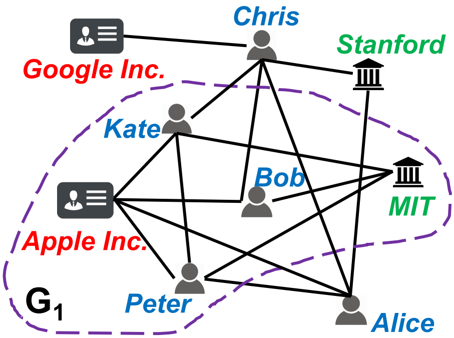


Figure: Traditional cliques

How to Find Cliques for HINs?

- Clique: complete subgraph [E. Akkoyunlu, 1973]

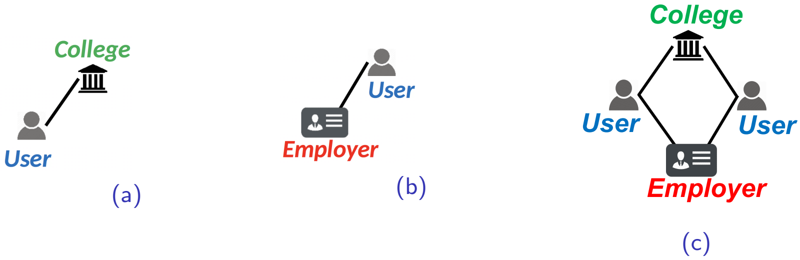


Motif

- Small pattern (higher-order structure)
- Building blocks of large and complex networks [Science'16]

Motif

- Small pattern (higher-order structure)
- Building blocks of large and complex networks [Science'16]



Motif

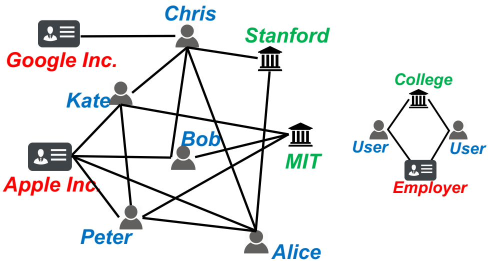
- Small pattern (higher-order structure)
- Building blocks of large and complex networks [Science'16]
- Existing works:
  - motif discovery [SIGMOD'15, DMKD'18]
  - graph node clustering [Science'16, KDD'17]
  - motif frequency estimation [WWW'15, WSDM'17, TKDD'17]

Motif clique (or m-clique)

$\boldsymbol{G}$: An HIN (each node has a **single** label)
$\boldsymbol{M}$: a small connected HIN which **follows the schema** of G

Motif clique (or m-clique)

G: An HIN (each node has a **single** label)
M: a small connected HIN which **follows the schema** of G



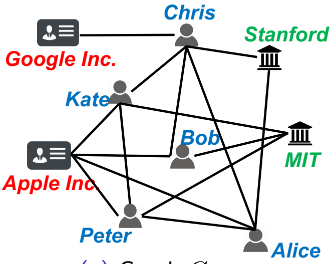
(a) Graph G (b) Motif M

Label-matched sets:

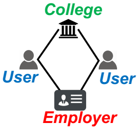
- {MIT, Kate, Peter, Apple}
- {MIT, Kate, Bob, Apple}
- ...

Motif clique (or m-clique)

G: An HIN (each node has a **single** label)
M: a small connected HIN which **follows the schema** of G



(a) Graph G



(b) Motif M

Label-matched sets:

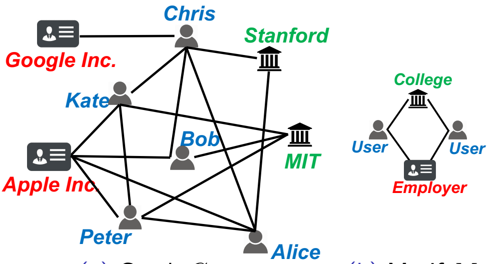
- {MIT, Kate, Peter, Apple}
- {MIT, Kate, Bob, Apple}
- ...

An induced subgraph G' of G is an m-clique of M:

1. G' has the same set of labels with M;
2. $\forall$label-matched set H in G', M is subgraph isomorphic to $G'[H]$.

Motif clique (or m-clique)

G: An HIN (each node has a **single** label)
M: a small connected HIN which **follows the schema** of G



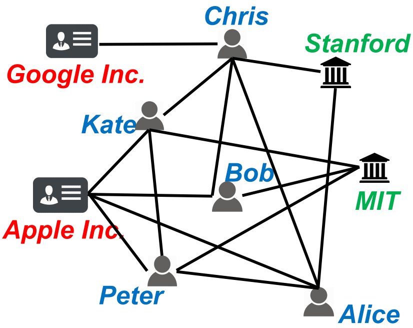
(a) Graph G (b) Motif M

Label-matched sets:

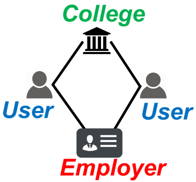
- {MIT, Kate, Peter, Apple}
- {MIT, Kate, Bob, Apple}
- ...

Maximal m-clique: m-clique & not contained in any other m-clique.
Maximal M-Clique **Enumeration** (**MMCE**): extract all maximal m-cliques in G.

Example

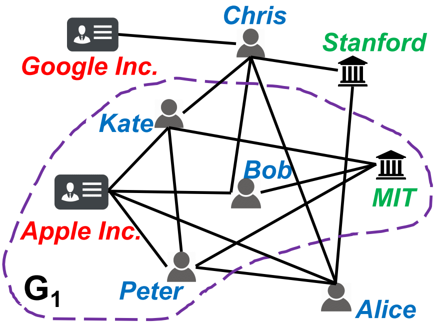


(a) Social network G

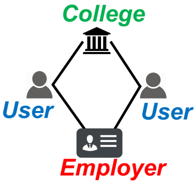


(b) Motif M

Example



(a) Social network G

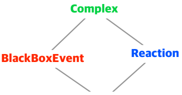


(b) Motif M

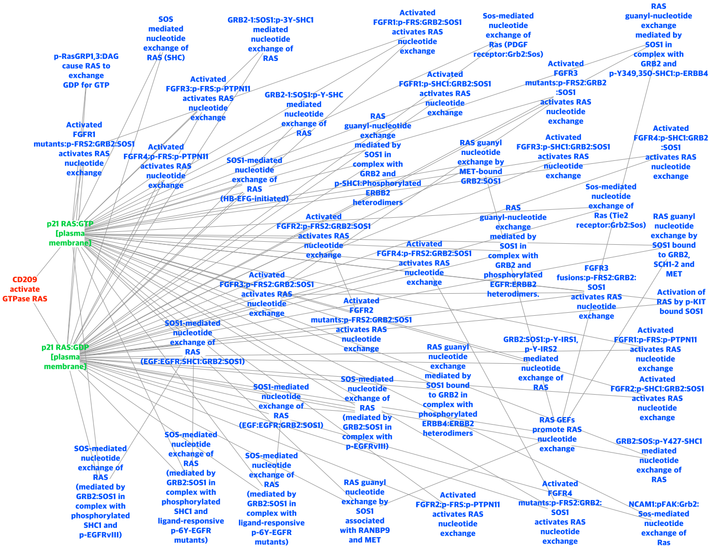
Apply m-clique in Biological Analysis

- Dataset: **Reactome**, a well known bioinformatic HIN [A. Fabregat et al., 2017]
- **Complex**: physical entities, e.g., proteins.
- **Reaction**: biochemical reactions which have balanced input and output entities.
- **BlackBoxEvent**: reactions or complex processes where details are not yet established.

Apply m-clique in Biological Analysis

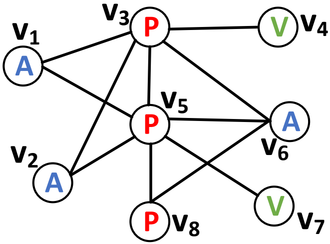


(a) motif

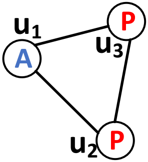


(b) maximal m-clique

Baseline Algorithm

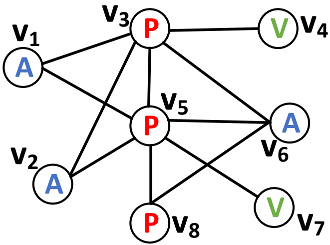


(a) G

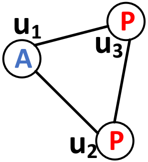


(b) M

Baseline Algorithm

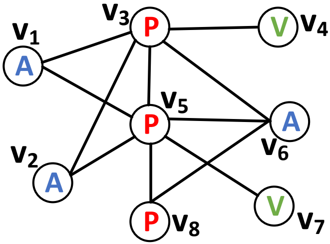


(a) G

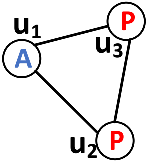


(b) M

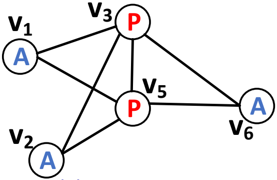
- Find all matched subgraphs
  - $S_1 = \{v_1, v_3, v_5\}$; $S_2 = \{v_2, v_3, v_5\}$; $S_3 = \{v_6, v_3, v_5\}$; $S_4 = \{v_6, v_5, v_8\}$
- For each matching, find all maximal m-cliques containing it.
  - $S_1 \rightarrow \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_2 \rightarrow \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_3 \rightarrow \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_4 \rightarrow \{v_6, v_5, v_8\}$



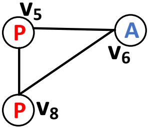
(a) G



(b) M

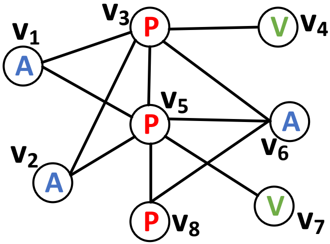


(c) maximal m-clique 1

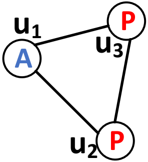


(d) maximal m-clique 2

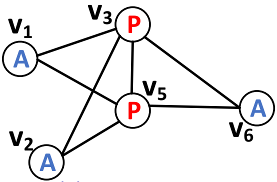
Baseline Algorithm



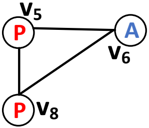
(a) G



(b) M



(c) maximal m-clique 1



(d) maximal m-clique 2

Time cost on Reactome: $>$ **1000 seconds/query**

Challenge 1: Node Expansion is NP-hard

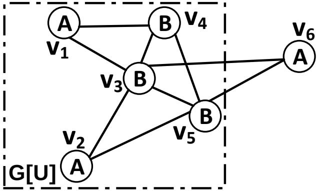


Figure: G

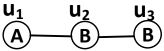


Figure: Motif

Possible label matched sets:

- $\{v_6, v_3, v_4\}$
- $\{v_6, v_3, v_5\}$
- $\{v_6, v_4, v_5\}$

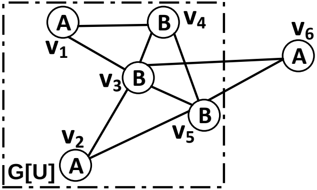


Figure: G

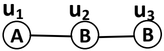


Figure: Motif

Dominance relationship:

- $\boldsymbol{v} = v_6; \quad \boldsymbol{u} = v_2$
- $L(v_6) = L(v_2) =$ "A"
- v_2 is dominated by v_6

Challenge 1: Node Expansion is NP-hard

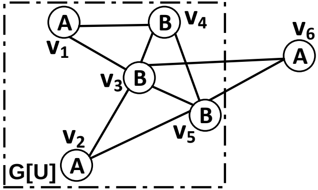


Figure: G

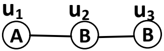


Figure: Motif

Dominance relationship:

- $\boldsymbol{v} = v_6; \quad \boldsymbol{u} = v_2$
- $L(v_6) = L(v_2) =$ "A"
- v_2 is dominated by v_6

Pruning strategies:

- Advanced node expansion
- Early stop pruning

Challenge 1: Node Expansion is NP-hard

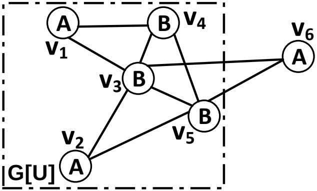


Figure: G

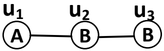


Figure: Motif

Dominance relationship:

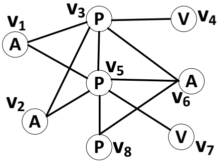
- $\boldsymbol{v} = v_6; \quad \boldsymbol{u} = v_2$
- $L(v_6) = L(v_2) =$ "A"
- v_2 is dominated by v_6

Pruning strategies:

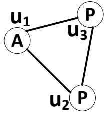
- Advanced node expansion
- Early stop pruning

Time cost on Reactome: around **10 seconds/query**

Challenge 2: Duplication Avoidance



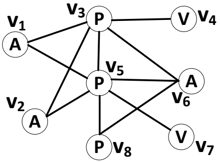
(a) G



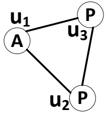
(b) M

- $S_1 = \{v_1, v_3, v_5\}; S_2 = \{v_2, v_3, v_5\}; S_3 = \{v_6, v_3, v_5\}; S_4 = \{v_6, v_5, v_8\}$
  - $S_1 \to \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_2 \to \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_3 \to \{v_1, v_2, v_3, v_5, v_6\}$

Challenge 2: Duplication Avoidance



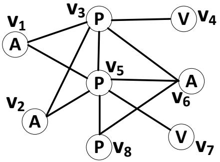
(a) G



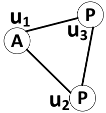
(b) M

- $S_1 = \{v_1, v_3, v_5\}; S_2 = \{v_2, v_3, v_5\}; S_3 = \{v_6, v_3, v_5\}; S_4 = \{v_6, v_5, v_8\}$
  - $S_1 \rightarrow \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_2 \rightarrow \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_3 \rightarrow \{v_1, v_2, v_3, v_5, v_6\}$
- Pruning strategy: **set-trie tree**
  - Dynamically build the set-trie tree and check candidates.

Challenge 2: Duplication Avoidance



(a) G



(b) M

- $S_1 = \{v_1, v_3, v_5\}$; $S_2 = \{v_2, v_3, v_5\}$; $S_3 = \{v_6, v_3, v_5\}$; $S_4 = \{v_6, v_5, v_8\}$
  - $S_1 \to \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_2 \to \{v_1, v_2, v_3, v_5, v_6\}$
  - $S_3 \to \{v_1, v_2, v_3, v_5, v_6\}$
- Pruning strategy: **set-trie tree**
  - Dynamically build the set-trie tree and check candidates.

Time cost on Reactome: around **0.1 seconds/query**

META: Maximal m-clique EnumeraTion Algorithm

Basic framework + the following pruning strategies:

- Dominance relationship between nodes
  - Advanced node expansion
  - Early stop pruning
- Duplication avoidance (set-trie tree)

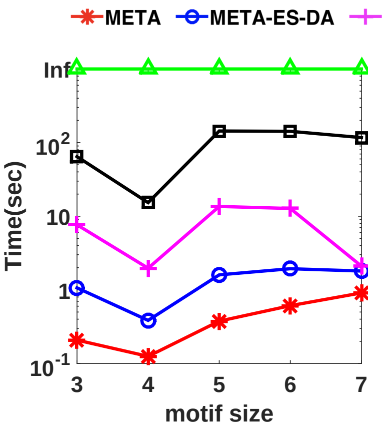
Experiments

Real graphs:

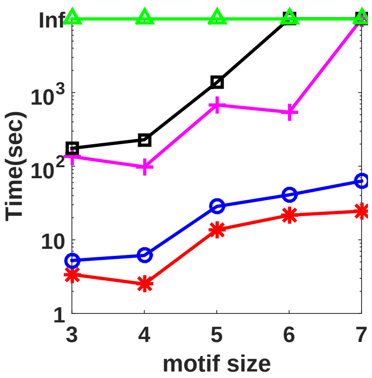
Dataset	#Nodes	#Edges	#Labels	Avg. Degree
DBLP	15K	51K	4	6.6
Amazon	548K	1.78M	4	6.5
Reactome	54K	98K	15	3.6
Yeast	3K	13K	71	8.1
Instacart	5K	13K	21	4.9

Time Cost by Varying Motif Size

The time of finding 10^3 maximal m-cliques.



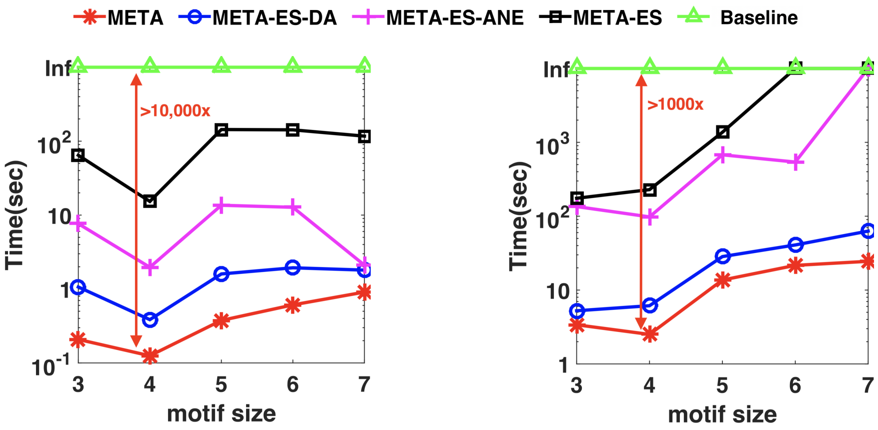
(a) Reactome



(b) Instacart

Time Cost by Varying Motif Size

The time of finding 10^3 maximal m-cliques.



(a) Reactome (b) Instacart

Conclusions

- m-clique for HINs
- The META algorithm
- Evaluation on both real and synthetic datasets

Future Work

- Extend META to handle more rich information on HINs
  - nodes with multiple labels
  - edges with directions and labels
  - ...
- Study other fundamental graph problems based on motif
  - motif-based path
  - motif-based connected component
  - ...

Thanks!

Q&A

Jiafeng Hu

acmhujiafeng@gmail.com