GraphMineSuite

August 2021

Enabling High-Performance and Programmable Graph Mining Algorithms with Set Algebra

Onur Mutlu

Maciej Besta

Zur Vonarburg

Deniz Sert
April 6, 2023

Motivation: Not fast enough!

- Current solutions like GraphChallenge, SNAP, and Graph500 are slow when faced with repeated accesses
- CS scientists lack modern tool to evaluate and and construct high-performance graph mining algorithms

- Solution: Use set algebra to improve efficiency and programmability

Table of contents

Set Algebra

Mathematical Framework for manipulating sets

Identity :

- $A \cup \varnothing = A$
- $A \cap U = A$

Complement :

- $A \cup A^C = U$
- $A \cap A^C = \varnothing$

Commutative property:

- $A \cup B = B \cup A$
- $A \cap B = B \cap A$

Associative property:

- $(A \cup B) \cup C = A \cup (B \cup C)$
- $(A \cap B) \cap C = A \cap (B \cap C)$

Distributive property:

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

De Morgan's laws:

- $(A \cup B)^C = A^C \cap B^C$
- $(A \cap B)^C = A^C \cup B^C$

double complement or involution law:

- $(A^C)^C = A$

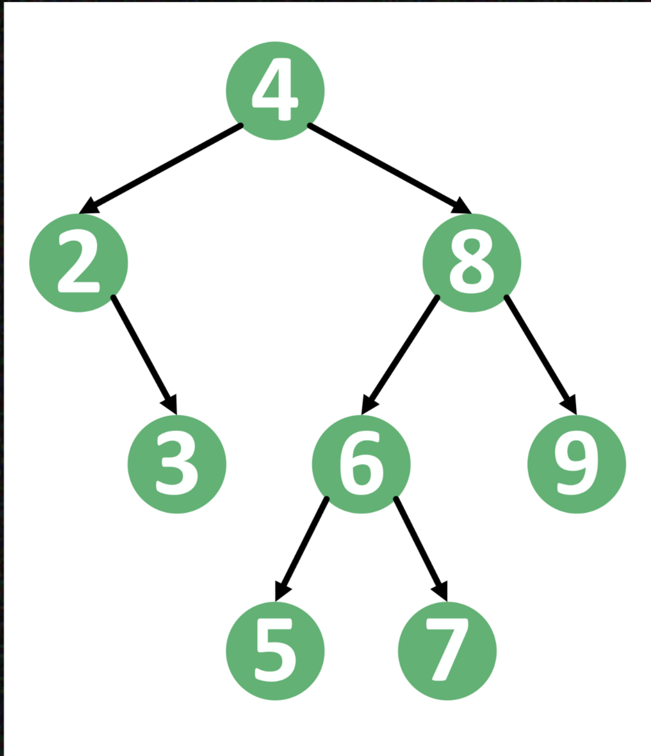
complement laws for the universe set and the empty set:

- $\varnothing^C = U$
- $U^C = \varnothing$

Graph Mining

Extracting useful information from graph-structured data

- This is important in various fields like social networks, biology, recommendation systems, fraud detection, and transportation networks
- Techniques include pattern mining, clustering, classification, similarity analysis, and visualization



SAPP Framework

Set Algebraic Parallel Processing

- SAPP is a new parallel processing framework for graph mining using set algebra
  It is the subroutine for GraphMineSuite (GMS), the first benchmarking suite for graph mining algorithms
- Properties include:
  - Fast: Speeds up modern graph mining algorithms by up to 9x
  - Flexible: easily extended to new algorithms and operations
  - Efficient: Built on top of set algebraic framework, allowing for efficient and expressive manipulation of graphs
  - Scalable: Designed to handle graphs with billions of vertices and edges

SAPP Framework

Asymptotic Bounds

	k-Clique Listing *Node Parallel* [41]	*k*-Clique Listing *Edge Parallel* [41]	★ ***k*-Clique Listing with ADG (§ 6)**	ADG (Section 6)	Max. Cliques Eppstein et al. [51]	Max. Cliques Das et al. [42]	★ **Max. Cliques with ADG (§ 7.3)**	Subgr. Isomorphism *Node Parallel* [26, 40]	Link Prediction†, JP Clustering
Work	$O\left(mk\left(\frac{d}{2}\right)^{k-2}\right)$	$O\left(mk\left(\frac{d}{2}\right)^{k-2}\right)$	$O\left(mk\left(d+\frac{\varepsilon}{2}\right)^{k-2}\right)$	$O(m)$	$O\left(dm3^{d/3}\right)$	$O\left(3^{n/3}\right)$	$O\left(dm3^{(2+\varepsilon)d/3}\right)$	$O\left(n\Delta^{k-1}\right)$	$O(m\Delta)$
Depth	$O\left(n+k\left(\frac{d}{2}\right)^{k-1}\right)$	$O\left(n+k\left(\frac{d}{2}\right)^{k-2}+d^2\right)$	$O\left(k\left(d+\frac{\varepsilon}{2}\right)^{k-2}+\log^2 n+d^2\right)$	$O\left(\log^2 n\right)$	$O\left(dm3^{d/3}\right)$	$O(d\log n)$	$O\left(\log^2 n+d\log n\right)$	$O\left(\Delta^{k-1}\right)$	$O(\Delta)$
Space	$O(nd^2+K)$	$O\left(md^2+K\right)$	$O\left(md^2+K\right)$	$O(m)$	$O(m+nd+K)$	$O(m+pd\Delta+K)$	$O(m+pd\Delta+K)$	$O(m+nk+K)$	$O(m\Delta)$

- d: graph degeneracy
- K: output size
- Delta: maximum degree
- p: number of processors
- k: number of vertices we're mining for
- n: number of vertices in the graph that we're mining
- m: number of edges in the graph

	Work
Chiba/Nishizeki [21]	$O\left(m\alpha^{k-2}\right)$

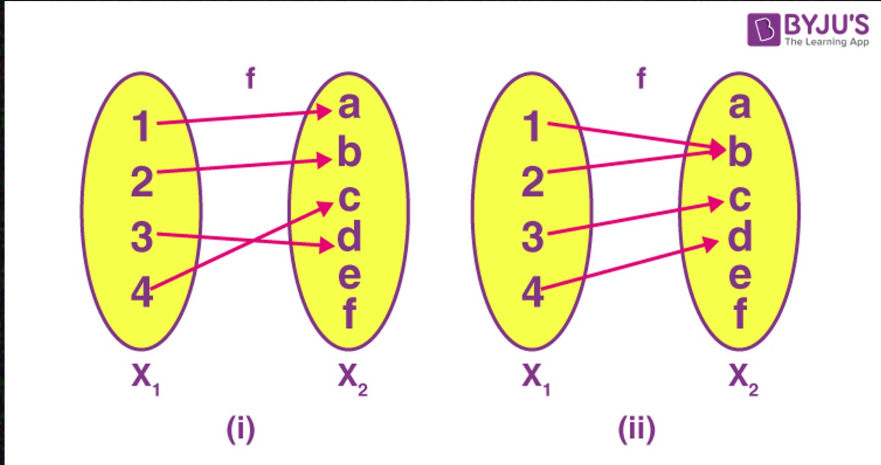
- Parallel Algorithms for Finding Large Cliques in Sparse Graphs
  - Gianinazzi, Besta, Shaffner
  - September 2021

Subgraph Isomorphism

Another Use Case for GMS

- Subgraph isomorphism is a problem in graph theory that involves determining whether a given pattern graph exists as a subgraph of a larger target graph

- Formally, given two graphs G and H, this problem asks whether there exists an injective function f from the vertices of H to the vertices in G s.t. if (u, v) is an edge in H, then (f(u), f(v)) is an edge in G.

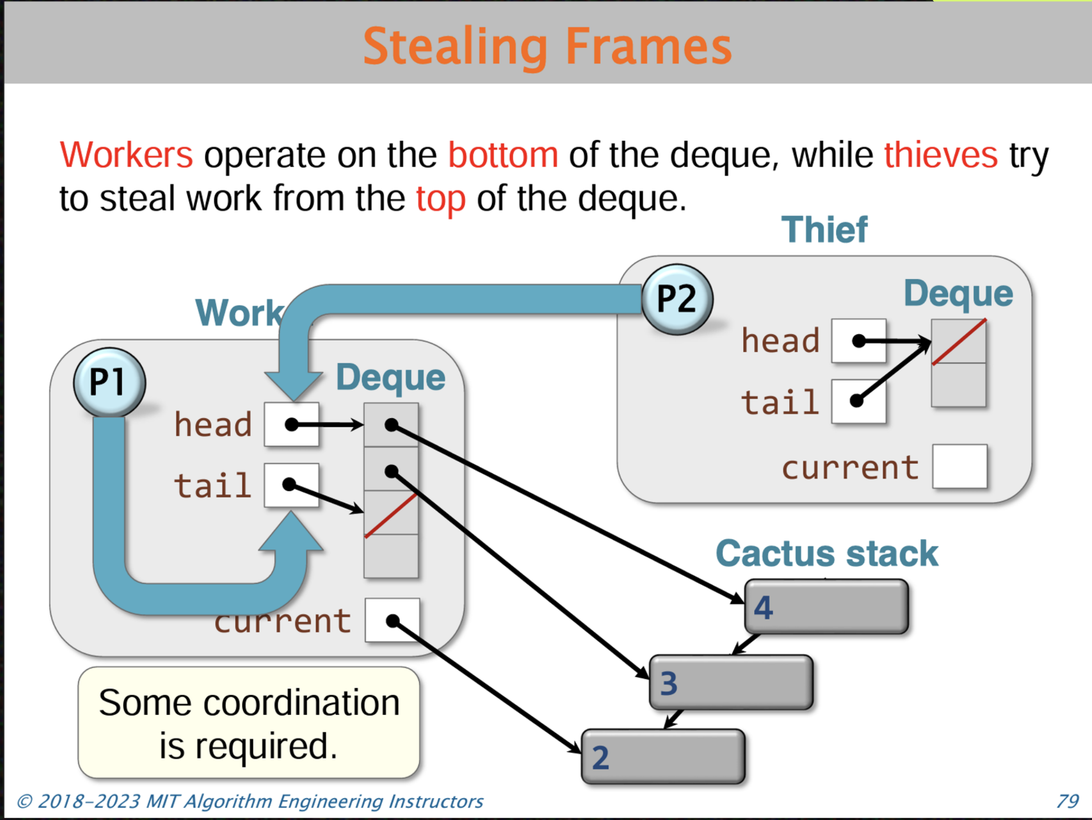


More Acceleration with Work-Stealing

Sometimes, crime does pay!

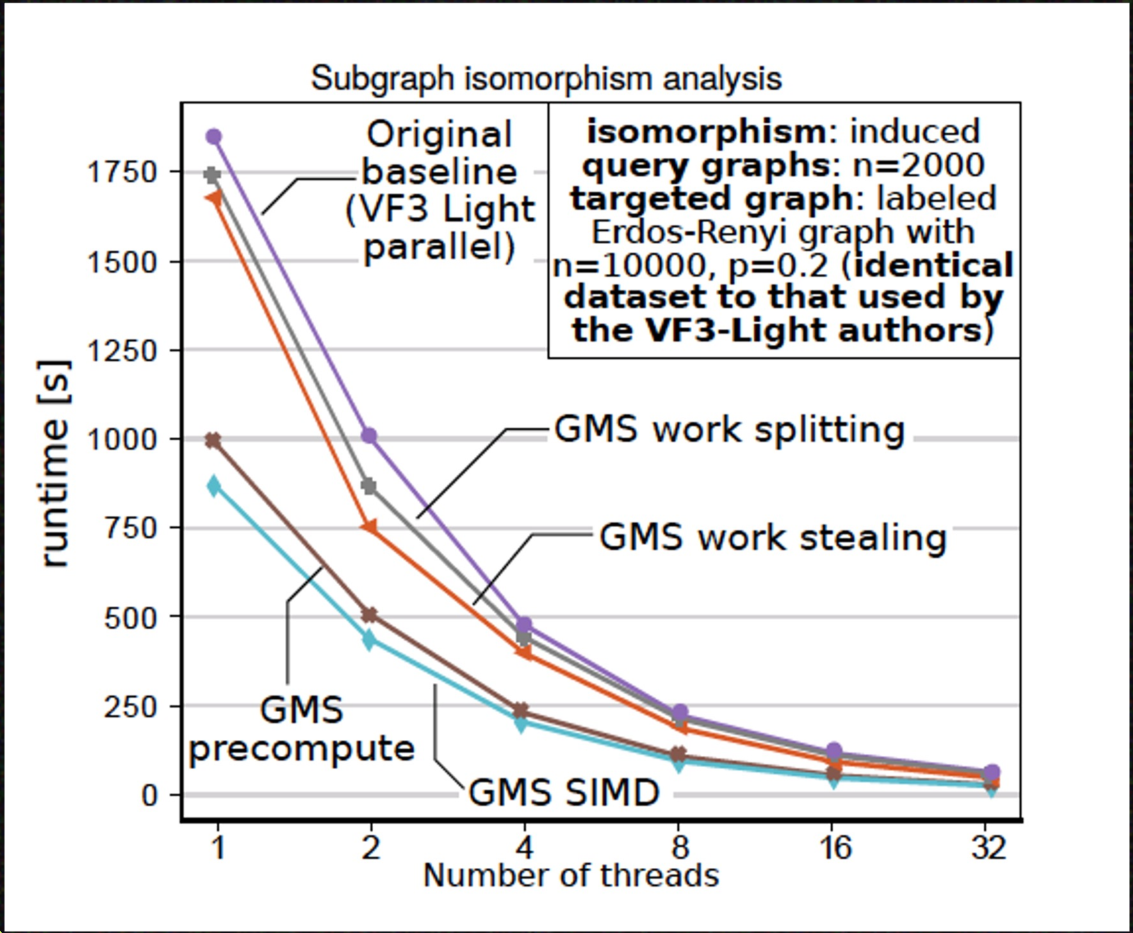
- Recall Cilk's work-stealing algorithm

- GMS combines this concept with Feb '19 paper regarding general purpose subgraph isomorphism algorithm to increase its performance by 2.5x!



- 6.506 Algorithm Engineering Lecture 4: The Cilk Runtime System
  Alexandros-Stavros Iliopoulous
  February 16, 2023

Compare purple (baseline), grey and red lines (work stealing)

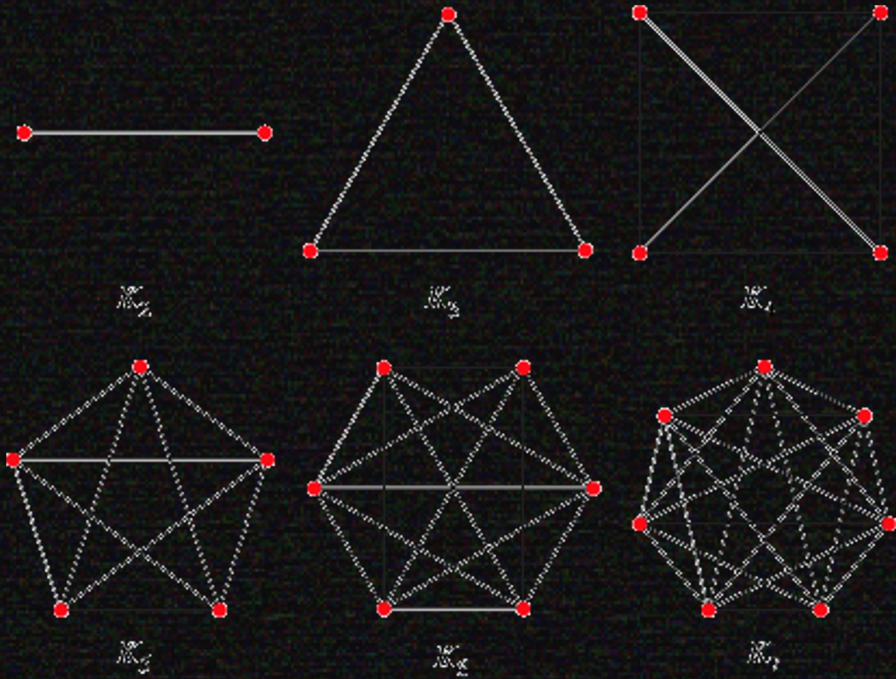


The 4-clique problem

GMS improvement in another popular graph problem

- The 4-clique problem is a graph problem that involves finding whether a graph contains a complete subgraph, or clique, or four nodes

- A clique is a subset of nodes in a graph s.t. every node in the subset is connected to every other node in the subset



Subtleties of Higher-Order Structure

Different types of graphs yield vastly different results

- Flickr, a photo sharing network, and Livemocha, a language-tutor matching app, have similar n, m, sparsity m/n, and degree distributions. But, 9 Billion and 4 Million 4-cliques, respectively!

- Why? In a social network of limited friendships, we expect 4-cliques to be only *relatively common*, whereas in photo sharing, we have metadata that *very often* link to other areas of the graph

Metadata Analysis

- Flickr: tags, descriptions, etc can be used to identify similar content, causing clusters of densely connected nodes

- Livemocha: language skills, proficiency levels, learning goals

flickr

Trade-Offs

Space complexity sometimes decreases for time increase

- Authors discuss the need to balance tradeoffs between work, depth, space, and approximation ratio
- Example: recursive clique-searching:
- Naive searching algorithm work / space: $\Theta(n\Delta^{k-1})$

After GMS and "Node Parallel" variant,

Work ↓ $\Theta(n + k(d/2)^{k-2} + d^2)$

Space ↑ $O(md^2)$

Closing Remarks

Big advancements to graph mining, but a prototype?

- Overviewed Set Algebra and technical details of GraphMineSuite
- Overviewed GMS
- Discussed several insightful use cases
- My thoughts:
  - Big speedups of up to 9x speeds
  - Better things to come?

It might soon be time to upgrade your iPhone (diy13 / Shutterstock)

iOS 17 could be leaving your old iPhone behind

In a couple of months we're expecting Apple to give us the lowdown on the iOS 17 update, and a reputable leaker says certain older devices won't be eligible for the update – devices including the iPhone 8, the iPhone 8 Plus and the iPhone X.

Read More

Thanks!

Questions?

Deniz Sert
dsert@mit.edu