Efficient and Scalable Subgraph Enumeration

David Oliveira Aparício (FCUP)

(**advisors:** Pedro Ribeiro and Fernando Silva)

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL

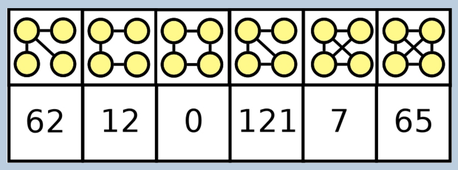
Subgraph Enumeration

Summarize/compare networks:

- Centrality measures (closeness, betweenness, PageRank, etc.).
- Degree Distributions/Power law exponent.
- **Subgraphs.**

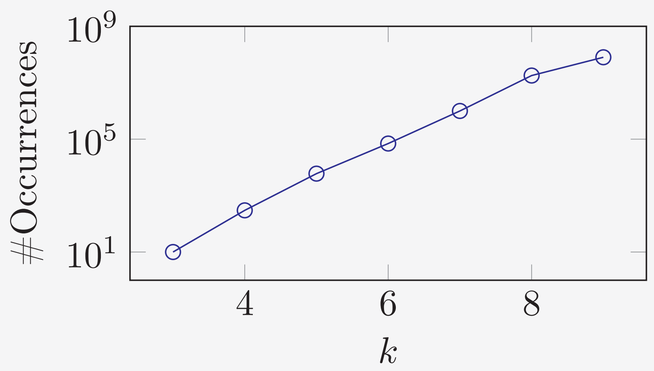
Enumerate *all* subgraphs of size k in network G.

Result ($k = 4$)**:**



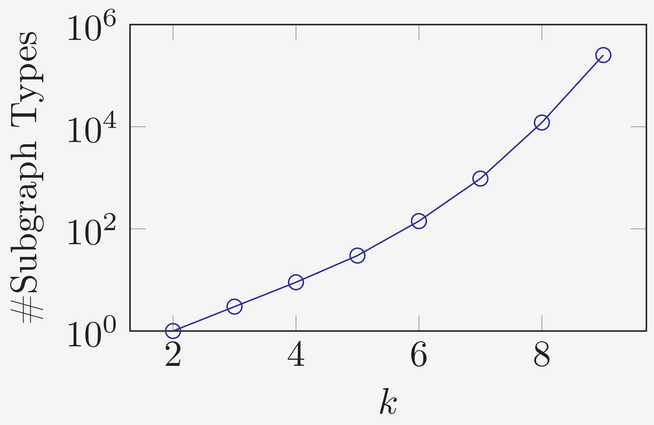
Computational Complexity

- **Millions/billions of occurrences** which **grows exponentially** with k.

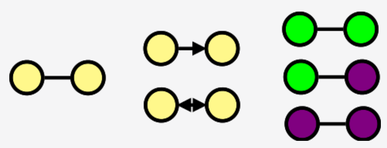


(Social Network – 51 nodes, 171 edges)

- Subgraph-types also grow exponentially. (Problematic if algorithm is subgraph-centric)



- Edge Direction, Node color, ..., increase the complexity.



Small networks ($< 10^6$ nodes)
Small subgraphs ($k < 10$)

Our Approach

1. Fast sequential algorithm (G-Tries).
2. Scalable parallel strategy (Dynamic workload balancing) [1,2].

References

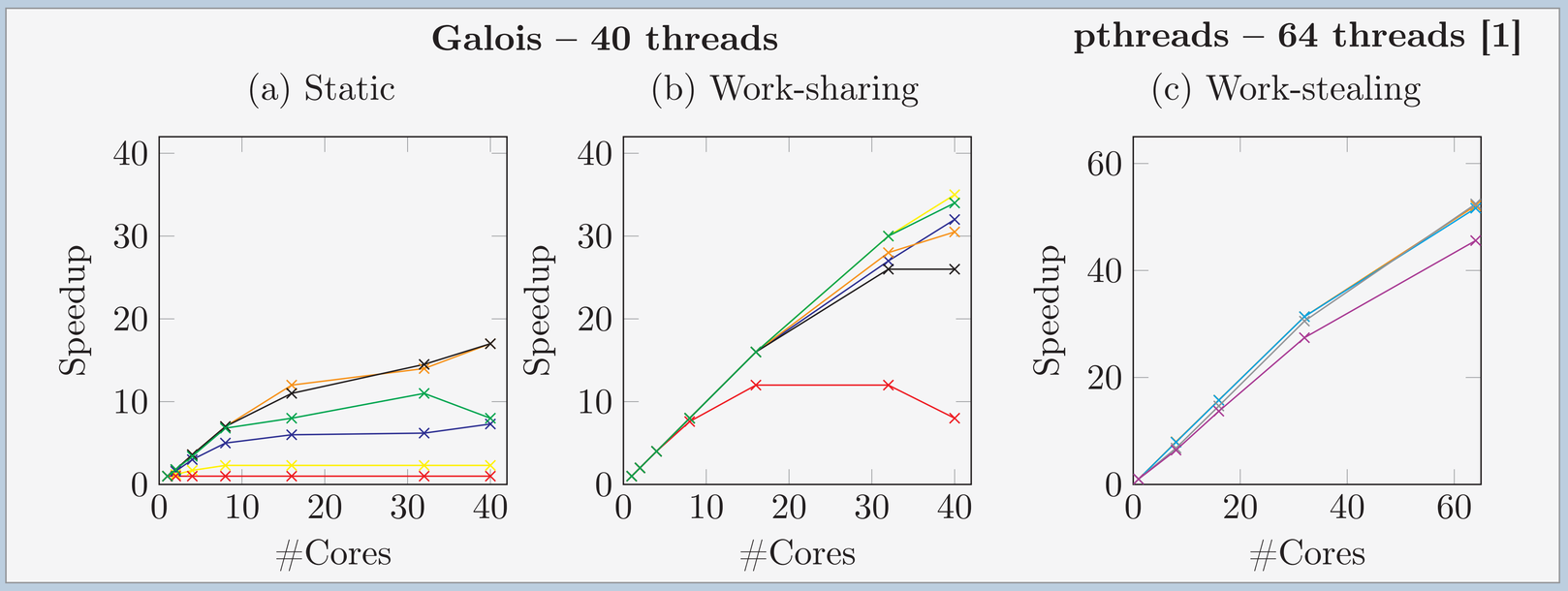
[1] D. Aparicio, P. Ribeiro, F. Silva. *Parallel subgraph counting for multicore architectures* in "Parallel and Distributed Processing with Applications", IEEE, 2014.

[2] D. Aparicio, P. Paredes, P. Ribeiro. *A Scalable Parallel Approach for Subgraph Census Computation* in "Euro-Par 2014: Parallel Processing Workshops", Springer, 2014.

[3] D. Aparicio, P. Ribeiro, F. Silva. "Network comparison using directed graphlets" in arXiv:1511.01964, 2015.

Parallel Subgraph Enumeration in Shared Memory

Networks: social (**jazz**, **blogs**, **e-mails**), co-authorships (**netscience**, **geometry**), biological (**neuronal**, **PPI**), linguistic (**dictionary**) and geometric (**routes**).



(a) Achieves some speedup but not for every network. **Work is not balanced**.
(b) Flexible but leads to overhead updating the work-sharing queue.
(c) **Near-linear speedup** for all tested networks.

Example: Size-6 census on **blogs** takes ≈ 2 days sequentially but only ≈ 1 hour using (c).

Distributed Memory, GPU, MapReduce

- **Distributed Memory:** near-linear speedup up to 128 processors.
- **GPU:** hard to distribute work dynamically and efficiently perform graph traversal.
- **MapReduce:** similar problems to the GPU *(ongoing work)*.

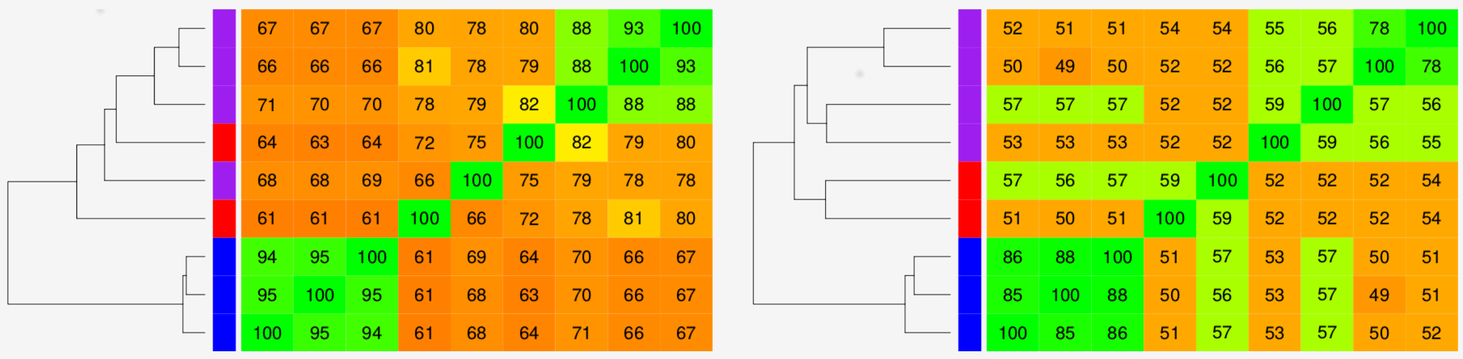
Application: Directed Graphlets [3]

- Node/Network comparison.
- Network Alignment.
- Subgraph-based metrics.

Usually **limited to small undirected networks**

k	2	3	4	5	6	7	8	9
$\lvert u\mathcal{G}_k \rvert$	10^0	10^0	10^1	10^1	10^2	10^3	10^4	10^5
$\lvert d\mathcal{G}_k \rvert$	10^0	10^1	10^2	10^3	10^6	10^8	10^{12}	10^{17}

Directed Biological Networks: cell signaling, **metabolic** and **transcription regulation**.



Undirected Graphlets ($k = 5$) | Directed Graphlets ($k = 4$)

Directed graphlets retrieve relevant topological information

Speedup:

$\mathcal{G}$	GraphCrunch	Orca	Kavosh	ESU
$u\mathcal{G}_5$	7.15 ± 2.56	2.04 ± 1.27	95.00 ± 30.97	80.11 ± 27.85
$u\mathcal{G}_6$	n/a	n/a	85.89 ± 24.07	70.31 ± 19.88
$d\mathcal{G}_4$	n/a	n/a	20.61 ± 3.80	18.73 ± 3.86
$d\mathcal{G}_5$	n/a	n/a	35.00 ± 9.77	31.75 ± 8.30

Our tool is more general and more efficient

Future Work

- **Temporal Graphlets:** compare/summarize temporal networks.
- **Subgraph Isomorphism on Streaming Graphs:** discover & report blacklisted patterns (colab w/UT-Austin)
- **Large Scale Subgraph Enumeration:** enumerate bigger subgraphs ($k >> 10$) on very large networks ($> 10^9$ nodes) → large scale parallel approach.

UT Austin | Portugal
INTERNATIONAL COLLABORATORY FOR EMERGING TECHNOLOGIES, COLAB