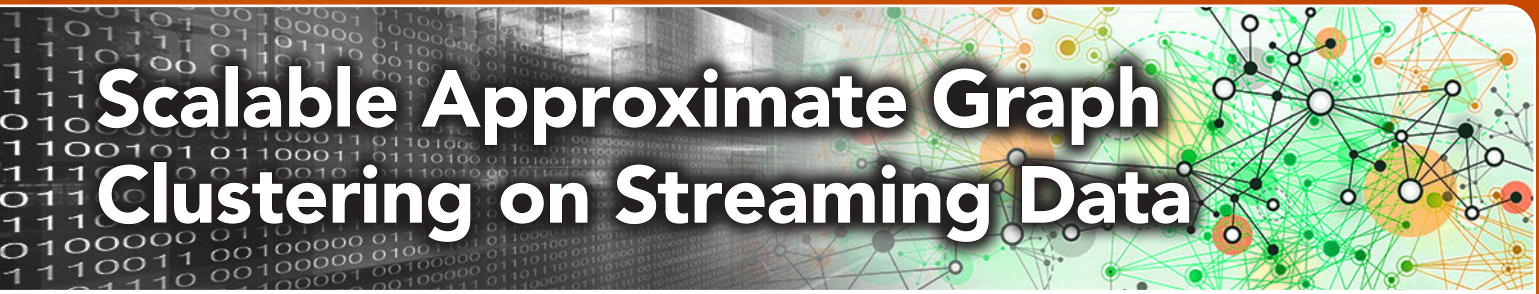
HIGH PERFORMANCE DATA ANALYTICS

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*



Scalable Approximate Graph Clustering on Streaming Data

CHALLENGE

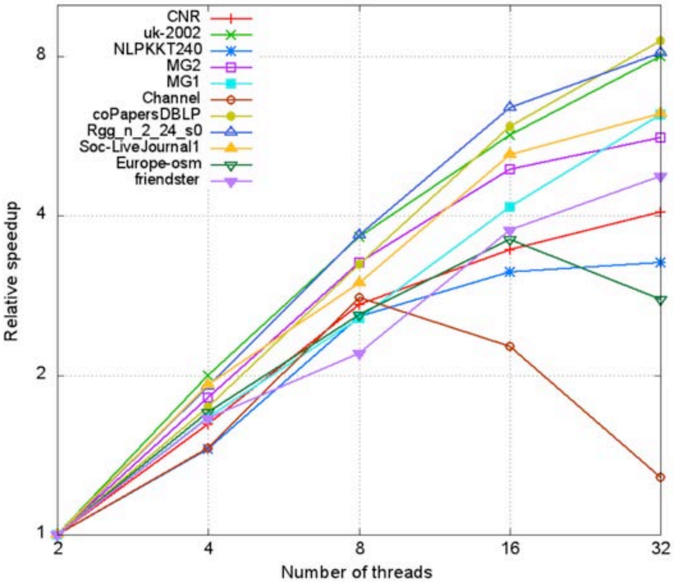
Graph clustering involves partitioning an input graph into clusters (or communities) of vertices that are closely related within and weakly related across. With numerous applications in life sciences, cybersecurity, and social network analysis, graph clustering has emerged as one of the most important discovery tools in the area of network analysis. The ability to handle dynamically evolving, temporal networks is an important extension of community detection. Dynamic (or temporal) community detection is the process of computing communities corresponding to multiple time steps in the evolution of a network. Similar to how temporal networks can be modeled as a sequence of time snapshots, a dynamic community can be seen as a sequence of static communities or a sequence of modifications to an initial static community.

CURRENT PRACTICE

The different approaches that have been proposed in the literature for temporal community detection can be summarized as follows:

1. Independent computation of static communities at each time step, followed by mapping the communities between consecutive time steps. This approach has a fundamental weakness arising from the stochasticity of community detection algorithms.
2. Systematic propagation of communities from the first time snapshot to subsequent snapshots. Limitations of this approach arise from the need to modify community detection algorithms and the inherently serial nature of dynamic algorithms. A further weakness is the lack of stability of static community detection algorithms.

> Developing clustering techniques for two fundamentally different formulations of the dynamic clustering problem to expose and exploit approximation strategies and achieve effective reduction in time-to-solution.



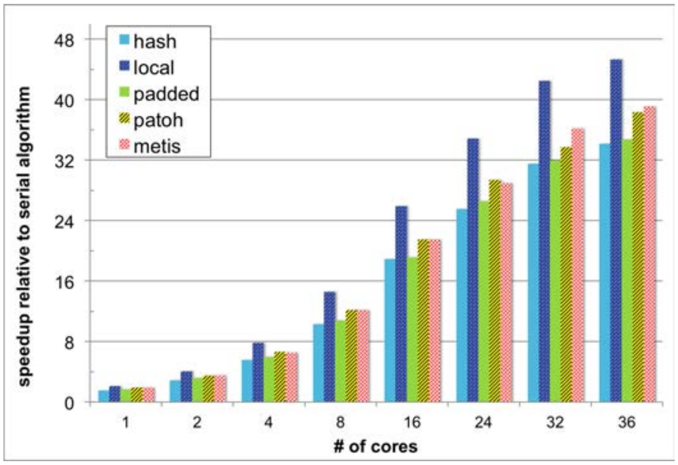
Strong scaling of Grappolo on Intel Xeon X7560[1].

3. Global community detection by simultaneous consideration of all time snapshots. While this approach addresses the problem of stability, a fundamental drawback is the high computational cost incurred by the need to incorporate the graphs from all time slices.

In this work, we focus on classes 1 and 3 by designing and developing novel algorithms and parallelization techniques.

TECHNICAL APPROACH

We consider two approaches for dynamic community detection: systematic propagation of communities and global community detection. The methods will be designed to overcome the stability and scalability (computational cost limitations) of current techniques that implement these two approaches. We will seek methods to overcome the stochastic nature of the underlying algorithms (e.g., by using weighted approximate matching to track communities and developing scalable heuristics for faster computation). We also will introduce the concept of end-to-end approximate computing that can enable scalability of dynamic community detection algorithms on graphs with billions to trillions of edges on modern computing platforms.



Strong scaling on Tilera Tile GX36 for input UK-2002[2].

We will build on our ongoing work, Grappolo, for static community detection. In Grappolo, we have developed several heuristics for parallelizing the Louvain method and demonstrated excellent scalability on several shared memory multi- and many-core architectures and distributed memory systems while also improving the quality of computed solutions. We will develop the proposed techniques as "Dynapolo." The novelty of this work is in the development of clustering techniques for two fundamentally different formulations of the dynamic clustering problem and for exposing and exploiting approximation strategies to achieve effective reduction in time-to-solution. We propose to apply the methods developed to data sets from cybersecurity, social media, and financial transaction networks.

IMPACT

Community detection has emerged as one of the most important kernels in network analysis. Because most networks evolve over a period of time, dynamic community detection has a greater significance with applications in multiple domains, including cybersecurity. However, current approaches are fundamentally inadequate to achieve scalability and are affected by stability issues. Consequently, the tools developed in this project can have a significant impact on complex network analysis.

Contacts

Mahantesh Halappanavar
Principal Investigator
(509) 372-5987
Mahantesh.Halappanavar@pnnl.gov

John Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

[1] Lu H, M Halappanavar, and A Kalyanaraman. 2015. "Parallel heuristics for scalable community detection." *Parallel Computing*, 47:19-37. DOI: 10.1016/j.parco.2015.03.003.

[2] Chavarría-Miranda D, M Halappanavar, and A Kalyanaraman. 2014. "Scaling Graph Community Detection on the Tilera Many-core Architecture." Presented at: *21st International Conference on High Performance Computing (HiPC)*, pp.1-11. December 17-20, 2014, Goa, India. IEEE Computer Society, Washington, D.C. DOI: 10.1109/HiPC.2014.7116708.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*