Scaling Graph Clustering with Distributed Sketches

Alec Dunton
Computing/CASC
Ben Priest and Geoff Sanders
Graph Clustering

• **Goal:** Identify “like” vertices as groups in a network.

• **Applications:** Contact Tracing!

• **Accepted Approach (Spectral):**
  • Embed graph matrix in low dimensional space spanned by subset of its eigenvectors.
  • Feed the embedding to a clustering algorithm.

• **Challenges:**
  • Distributed: bulk synchronous communication.
  • Streaming: spectral methods slow to update.

• **Proposed Solution:**

sciencedirect.com

github.com
Matrix Sketching: Overview and Results

- Randomly embed a graph matrix in a lower dimension, preserve important geometric structure.
  - Fast Walsh-Hadamard Transform (FWHT): theoretically sound, harder to implement.
  - CountSketch (CST): less rigorous, but faster and easily implemented in distributed setting.
- Use Stochastic Block Models (SBMs) in our numerical experiments.
  - SBMs function as surrogates for real world networks.
  - Randomly assign vertices to groups; high connectivity within groups, low between groups.

**Y-axes:** Required geometric fidelity of sketch. **X-axes:** (left) ratio of intra-group connectivity to inter-group connectivity; (center) number of true groups, fixed size; (right) size of graph, fixed number of groups.
Results, Next Steps, Questions?

- CST is faster than I/O, as accurate as FWHT!
- Next: Optimization of Clustering Algorithms