SAND2012-5985P

Exceptional service in the national interest

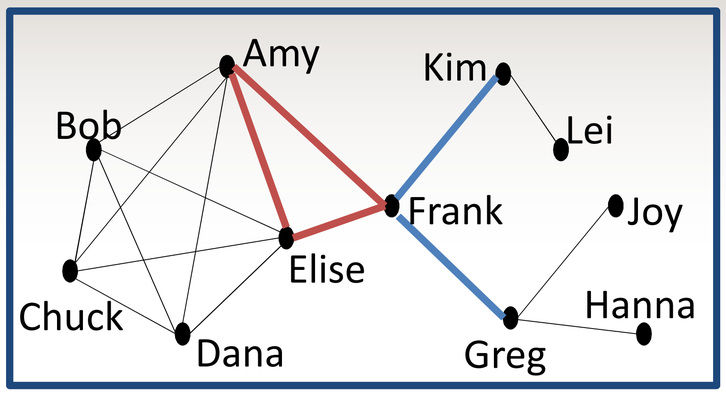
Sandia National Laboratories

Parallel Triangle Sampling in Social Networks

Christine Task*

Mentors: Tammy Kolda, Ali Pinar

Triangles Indicate Social Cohesion:
"Are my friends friends with each other? Or Not?"
This computational measure of graph structure gives insight into the nature of the community, and is used in network modeling and comparison.



This Graph Has:
11 Triangles and 15 Open Wedges

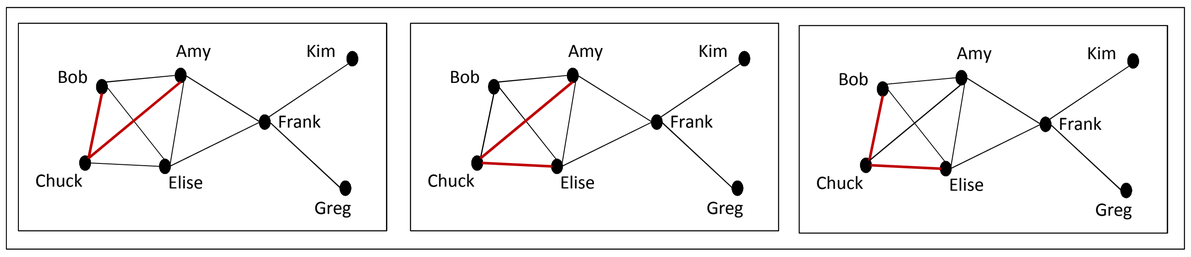
Real data can have more than:
10^7 triangles and ***10^{10} wedges!***

"How should a computer count triangles?"

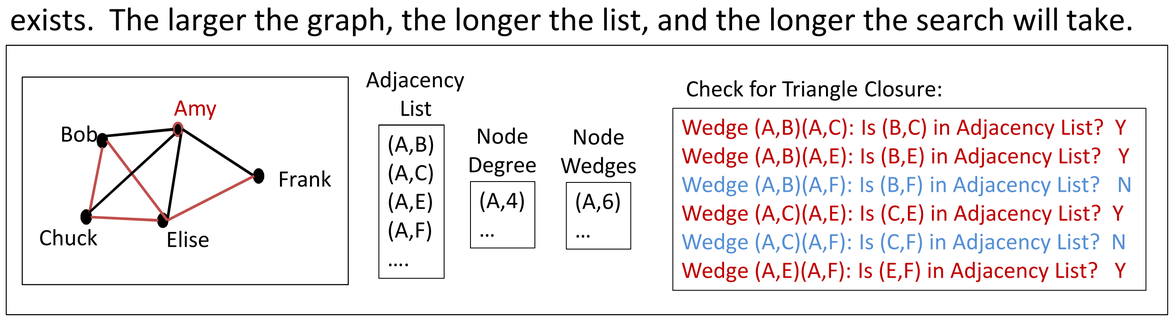
Sequential Triangle Counting:

Given a graph, stored as an adjacency (edge) list with **m** edges, how long does it take to check *every* wedge in the graph to see if it's a triangle?

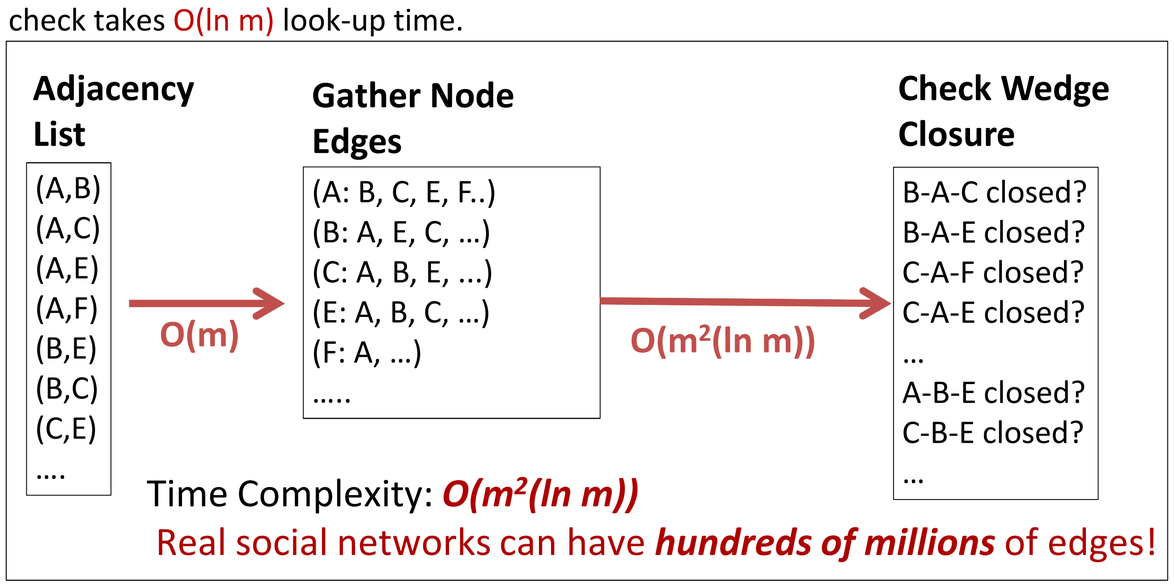
The total number of wedges at a node with *d* friends is *(d choose 2) = d(d-1)/2*



Even for a node with only four friends (we say "degree 4"), we have to ask a lot of wedge closure questions. Four is a small degree. How many friends do *you* have on facebook? Each closure check requires searching the adjacency list to see if an edge exists. The larger the graph, the longer the list, and the longer the search will take.



If we want to count triangles over the whole graph, first, we read through adjacency list and compile list of edges at each node. Then, for each wedge, check to see if it's closed to make a triangle. There are at maximum m^2 possible wedges, and each check takes O(ln m) look-up time.



Parallel Triangle Sampling:

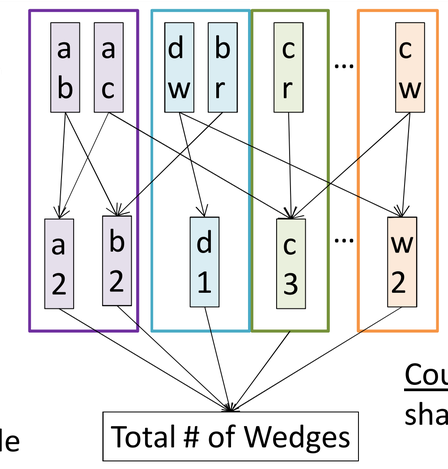
Instead of enumerating *all* triangles, we uniformly randomly sample wedges and check for closure to *estimate* triangle/wedge ratio. Then we multiply by total wedge count to get a close estimate of the triangle count.

Using Hadoop MapReduce we can distribute our graph data across many processors and do *all our reads of the adjacency list* ***simultaneously!***

Adjacency List
Stored Across Multiple Processors

Node Degrees List
Stored across Multiple Processors

Totals Wedge Counter
Stored in Shared Variable



Reduce/ Global Sorting:
Gathers data by node ID. All information about node x goes to same processor

Counter: Processors update shared counter in Parallel

Pass 1:
REDUCE: We can access the adjacency list in parallel and find the degree of each node (as described above).

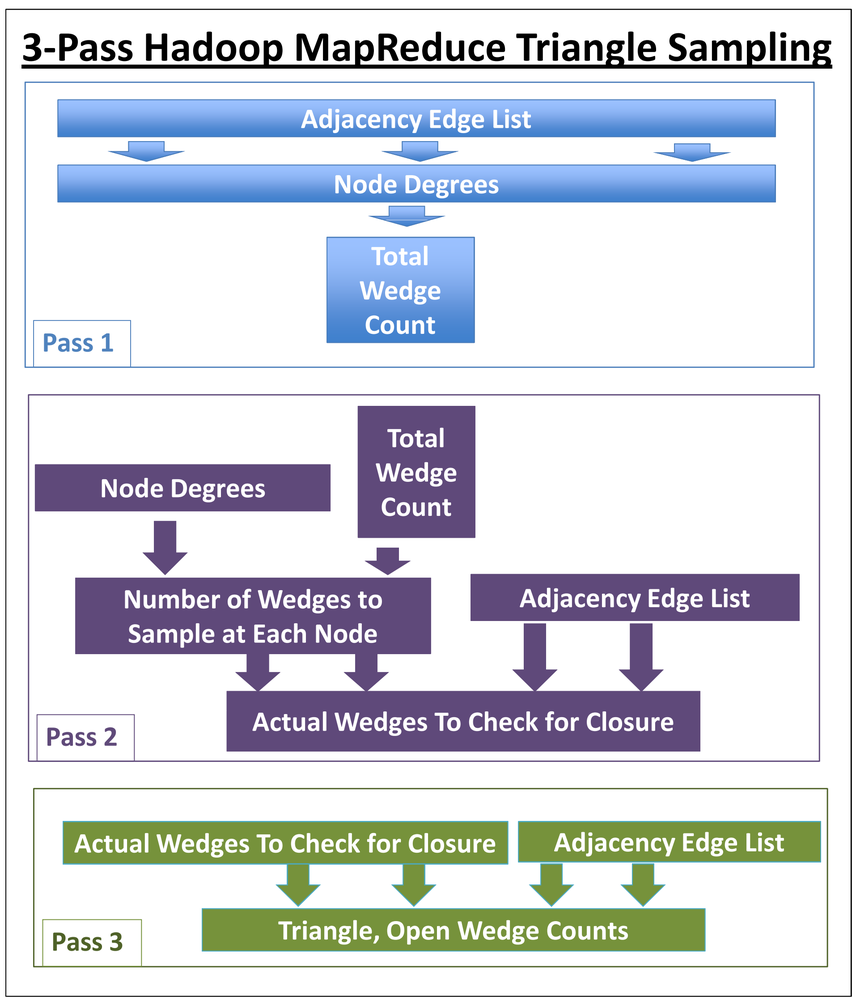
COUNTER: When we know the degree of a node, we know how many wedges it has, and we can easily keep a global running tally of the total wedges in the graph.

Pass 2:
MAP: If we know the degree of a node, and the total number of wedges, then we can compute how many wedges we should sample from that node. This can be done completely in parallel on each node.

REDUCE: If we know we need w wedges from node x, and we know x's edges, we can sample the wedges for x.

Pass 3:
REDUCE: We can check whether wedge closure edges exist in the graph by grouping together real edges from the adjacency list with hypothetical wedge-closure edges.

3-Pass Hadoop MapReduce Triangle Sampling



U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

*PhD Candidate, Computer Science Department, Purdue University