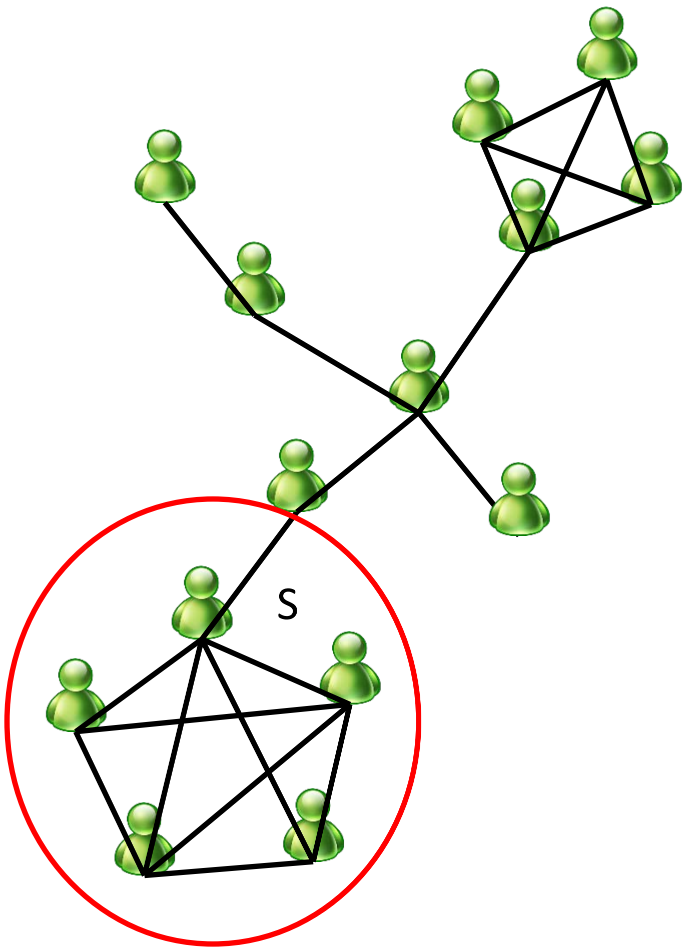
A Time-Efficient Connected Densest Subgraph Discovery Algorithm for Big Data

Bo Wu and Haiying Shen
Dept. of Electrical and Computer Engineering
Clemson University, SC, USA

Outline

- Background
- Algorithm design
- Evaluation
- Conclusion

Background

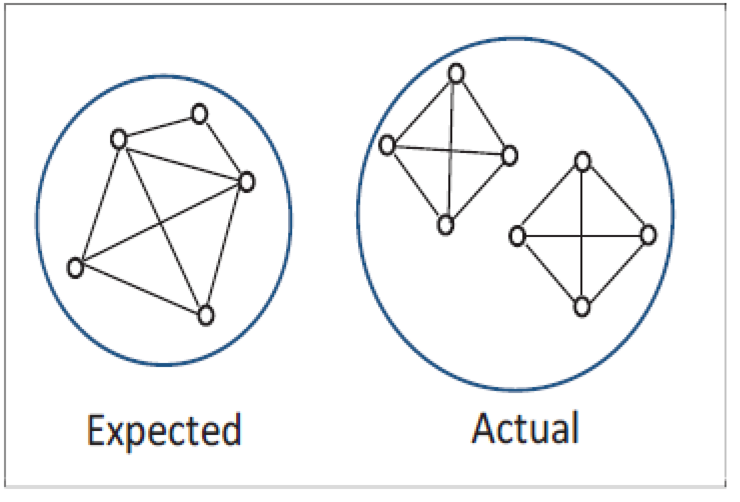


Densest subgraph problem

- Motivation: find the main community in a social network.
  - denotes different person.
  - The link between denotes friendship.
- Definition: densest subgraph is a subgraph with largest average degree.
  - e.g. the main community S is with a density 9/5=1.8

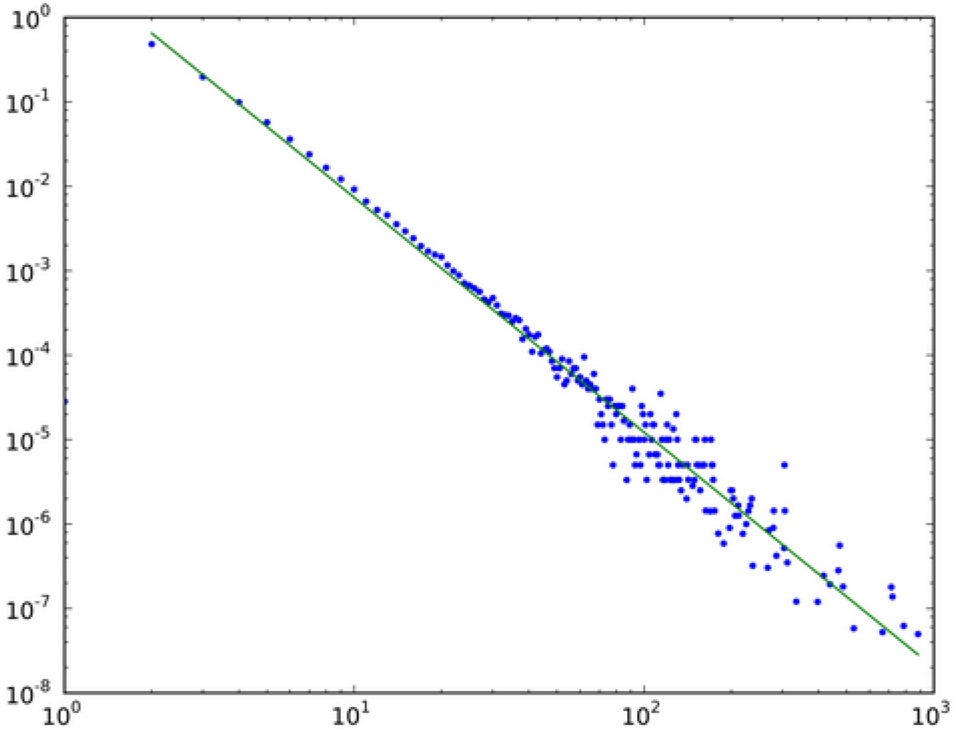
Background (cont.)

- Exact algorithm[Goldberg'84]
  - In memory
- Approximate algorithm [APPROX'00]
  - Connectivity problem
- Can we find an exact algorithm for big datasets?



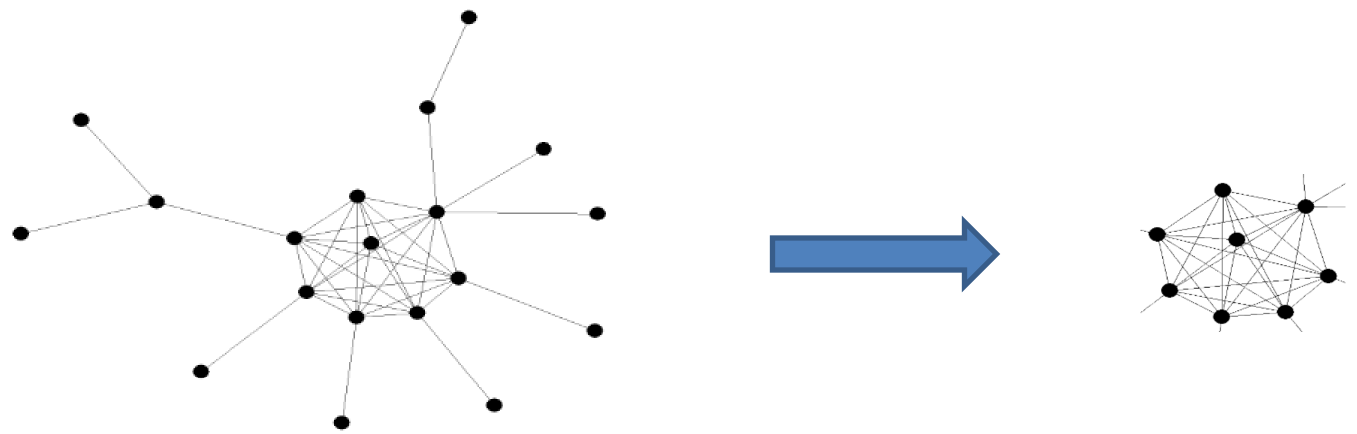
Background (cont.)

- Degree distribution of natural graphs



Algorithm design

- General idea
  - **Reduction:** delete all the nodes with very small degrees.
  - **Solution:** use exact algorithm to find the densest subgraph.



Algorithm design (cont.)

- Challenges
  - Correctness.
    - We need to be careful enough so that no nodes in the densest subgraph will be deleted.
    - We need to make sure the exact algorithm is suitable for the reduced graph.
  - Suitability
    - We need to make sure the reduced graph can be handled in memory.
  - Efficiency
    - We need to make sure the reduction is not time consuming.

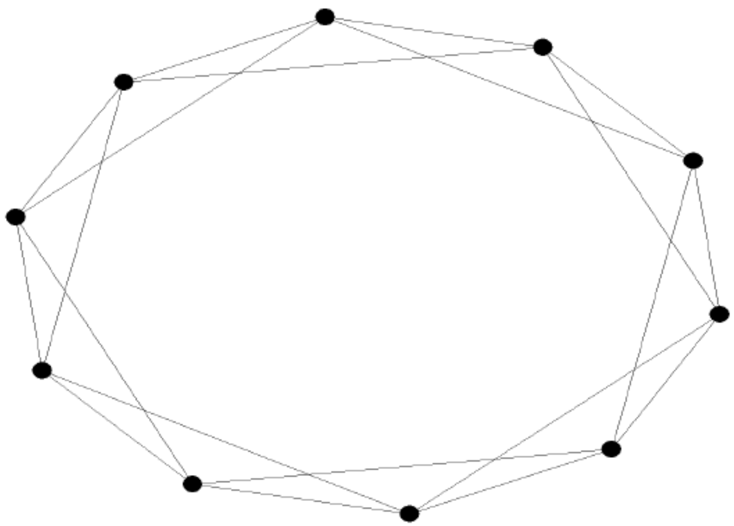
Algorithm design (cont.)

- Correctness
  - After we recursively delete all the nodes with degrees smaller than or equal to the density of remaining graph, the densest subgraph is still in the remaining graph.
  - No matter the remaining graph is connected or disconnected, we can find the connected densest subgraph by applying min-cut max-flow technique.

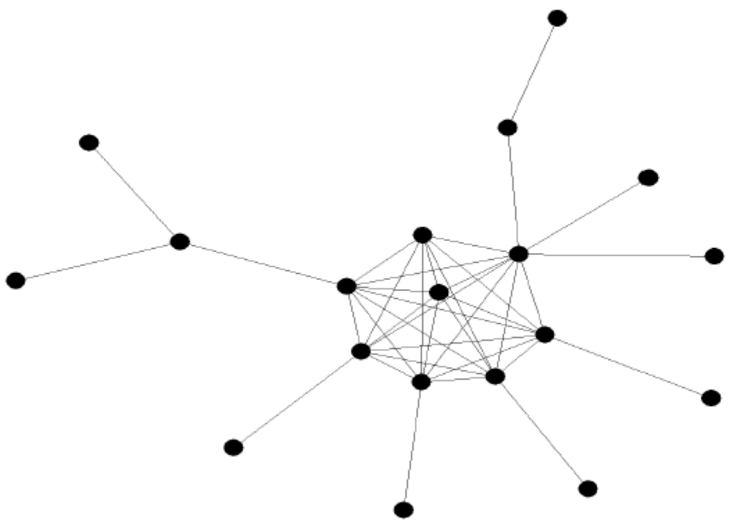
Algorithm design (cont.)

- Suitability and efficiency

Unsuitable

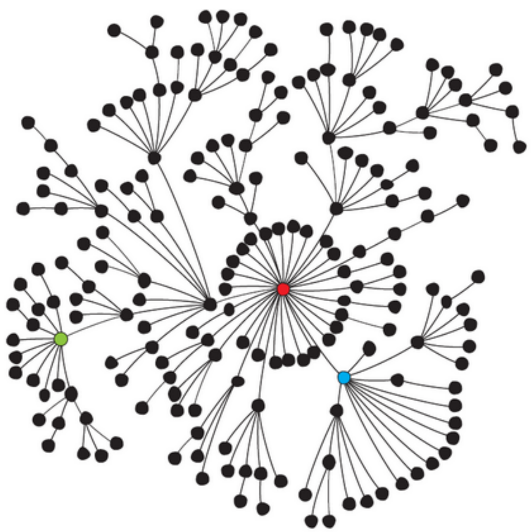


Suitable



Algorithm design (cont.)

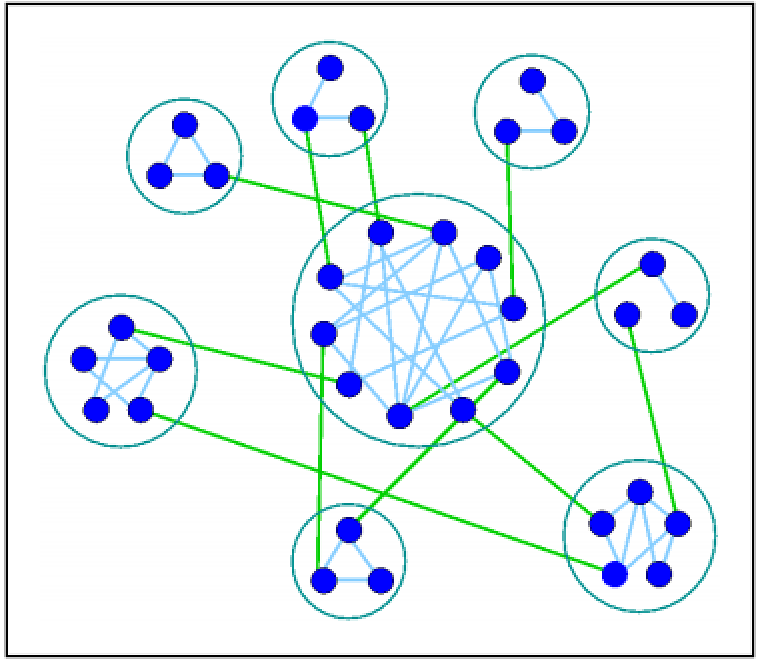
- Suitability and efficiency
  - Scale free network (without community) [1]
    - The density of the whole network equals the density of the densest sub-network. Therefore, no nodes can be deleted from the network.



[1] A. L. Barabasi and R. Albert, “Emergence of scaling in random networks,” Science, 1999.

Algorithm design (cont.)

- Suitability and efficiency
  - BTER network (with community) [1]
    - More than 90% of the nodes can be deleted in first few rounds.



[1] C. Seshadhri, T. G. Kolda, and A. Pinar, “Community structure and scale-free collections of er graphs,” CoRR, vol. abs/1112.3644, 2011.

Performance Evaluation

- Platform:
  - Hadoop MapReduce framework on 4 PCs; each PC is quipped with 2.1GHz Intel core i3 processor with 2 cores, and a 2GB memory.
- Metrics for the evaluation
  - **Percentage of data reduced (suitability)**
  - **The number of rounds needed for the reduction (efficiency)**

[1] “Stanford network analysis project.” https://snap.stanford.edu/.
[2] C. Seshadhri, T. G. Kolda, and A. Pinar, “Community structure and scale-free collections of er graphs,” CoRR, 2011.

Performance Evaluation

- Datasets [1]

Name	Description	$\|V\|$	$\|E\|$	Type
Wiki-Vote	Wikipedia who votes on whom network	7,115	207,378	small
CA-GrQc	Collaboration network of Arxiv General Relativity	12,008	237,010	small
Email-Enron	Enron company email list	36,692	367,662	small
CA-HepPh	Arxiv High Energy Physics paper citation network	34,546	421,578	small
slash	Slashdot social network from November 2008	77,360	905,468	small
com-youtube	Youtube online social network	1,134,890	2,987,624	large
com-lj	LiveJournal online social network	3,997,962	34,681,189	large
com-orkut	Orkut online social network	3,072,441	117,185,083	large

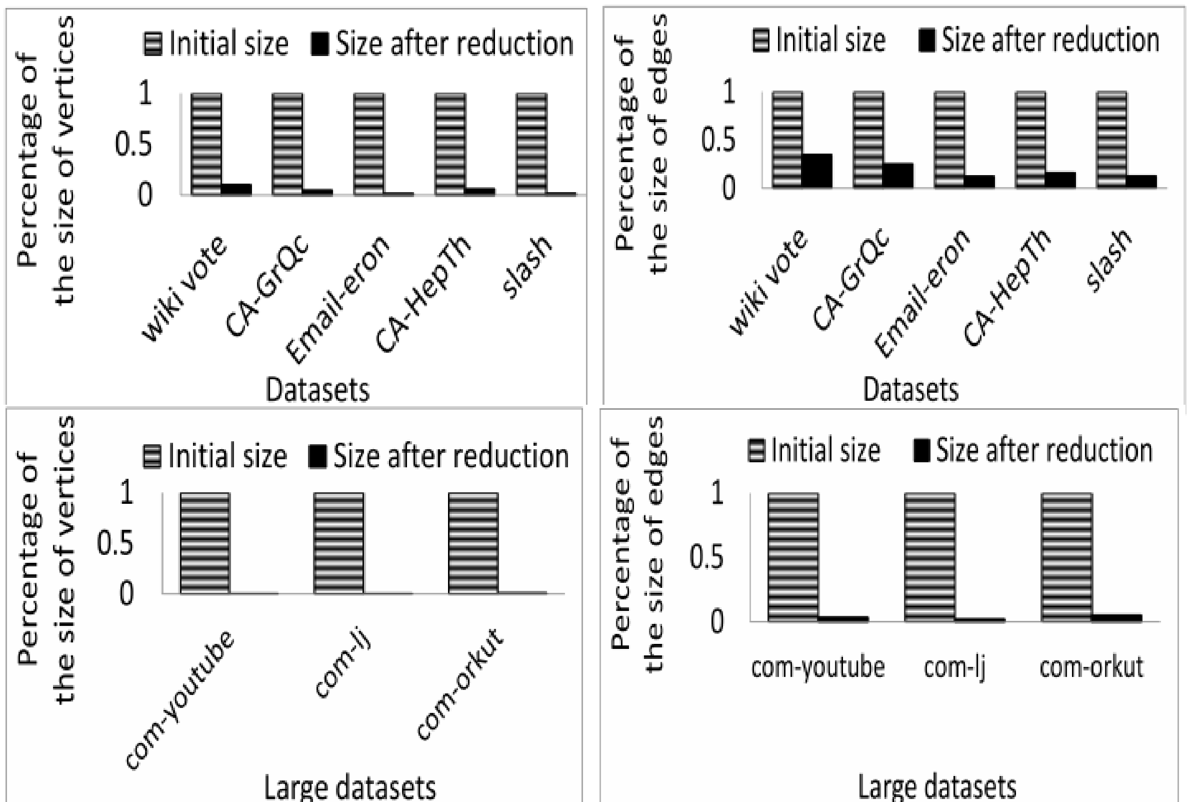
[1] "Stanford network analysis project." https://snap.stanford.edu/.

Performance Evaluation (cont.)

- Performance of reduction

Small

Large

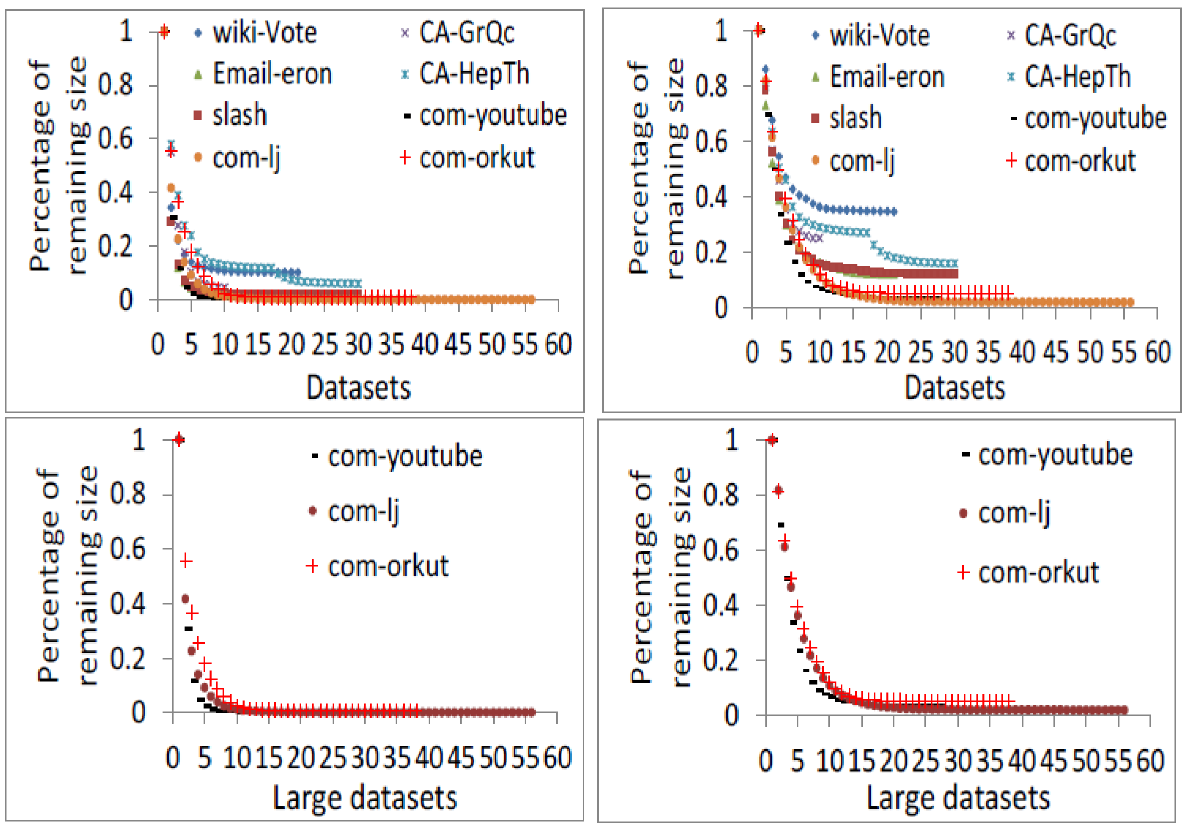


Performance Evaluation (cont.)

- Number of rounds

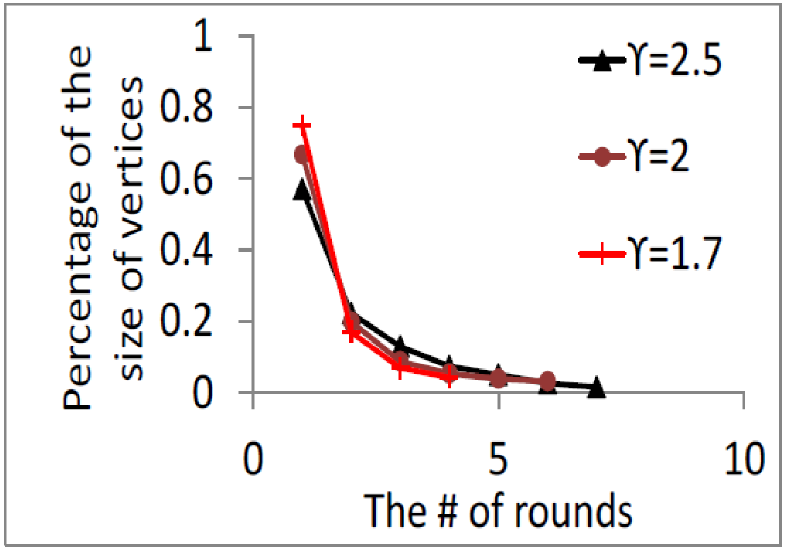
Small

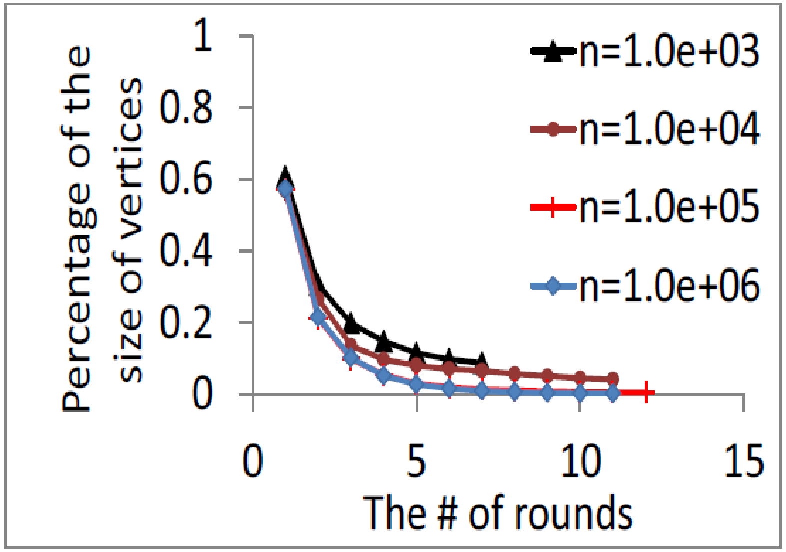
Large



Performance Evaluation (cont.)

- Simulation
  - The simulation is consistent with the experiment on real datasets.





Conclusion

- Our algorithm perform better on big datasets than small datasets.

- In the future, we will exploit to implement real application based on our algorithm.

Thank you!

Questions & Comments?

Bo Wu, PhD Candidate

bwu2@clemson.edu

Pervasive Communication Laboratory

Clemson University