

Discovering the Densest Subgraph in MapReduce for Assortative Big Natural Graphs



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

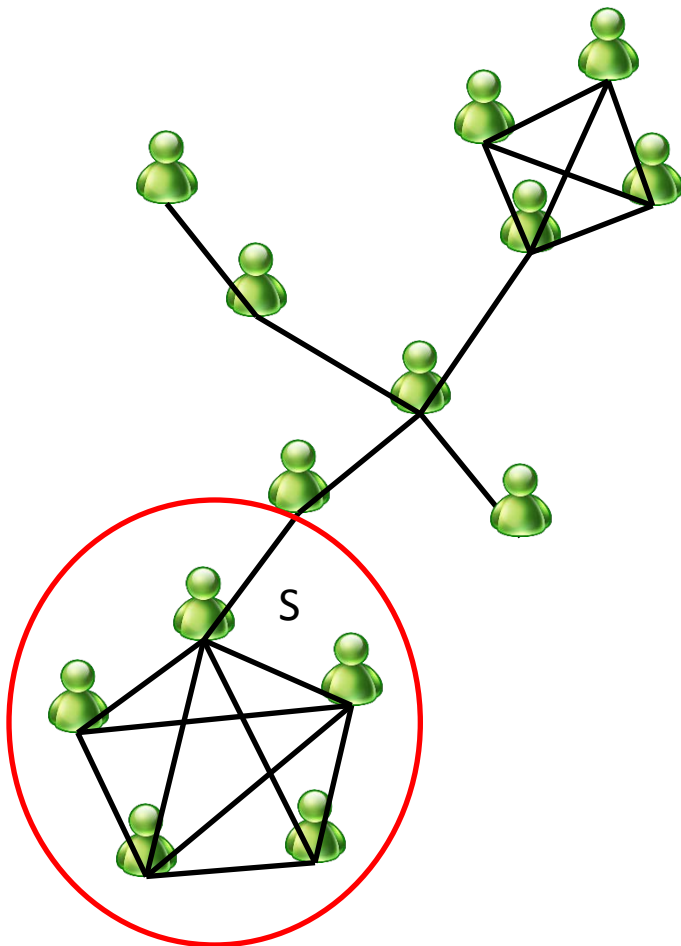
Outline

- Background
- Measurement
- Heuristic 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 $9/5=1.8$



Measurement

- Datasets [1]

ID	Description	$ V $
1	Collaboration network of Arxiv General Relativity	5,242
2	Social circles from Facebook (anonymized)	4,039
3	DBLP collaboration network	317,080
4	Amazon product network	334,863
5	LiveJournal online social network	3,997,962
6	Enron company email list	36,692
7	Wikipedia who votes on whom network	7,115
8	Slashdot social network from November 2008	77,360
9	Arxiv High Energy Physics paper citation network	34,546
10	Web graph of Notre Dame	325,729

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

Measurement

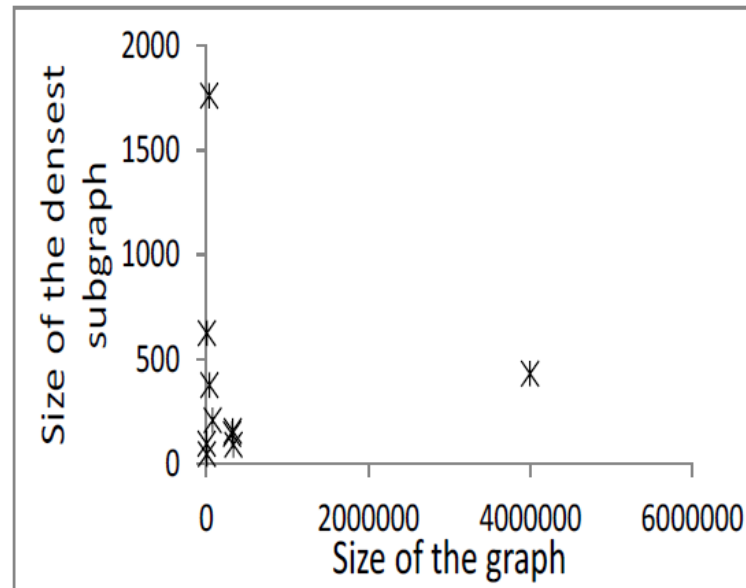
Phenomenon of small densest subgraph

Dataset	Diameter	Dataset	Diameter
Dataset 1	2	Dataset 6	2
Dataset 2	2	Dataset 7	2
Dataset 3	2	Dataset 8	2
Dataset 4	2	Dataset 9	2
Dataset 5	2	Dataset 10	2

- Observation:
 - The densest subgraphs are with extremely small diameters.

Measurement

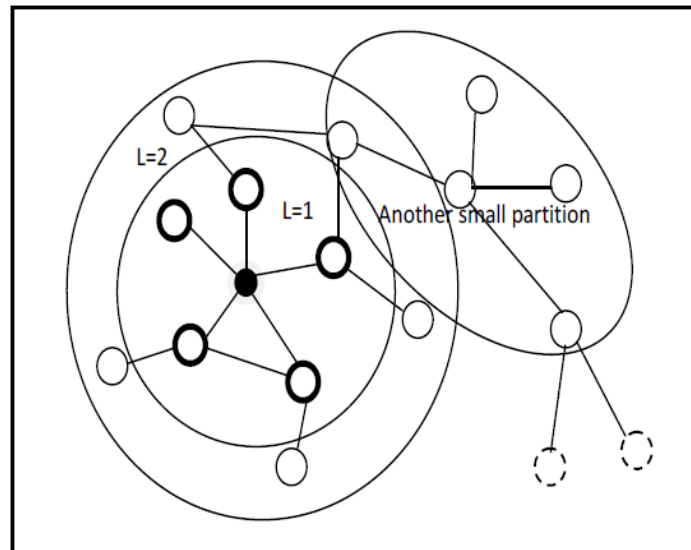
Phenomenon of small densest subgraph



- Observation:
 - The sizes of the densest subgraphs are extremely small.
 - The size of the densest subgraph has no correlation with the size of the whole graph.

Heuristic algorithm design

General idea:



- **Two steps:**
 - Sampling: Select several seeds; Get several small partitions by transversing from the seeds.
 - Find the densest subgraphs in the partitons.

Heuristic algorithm design

- Sampling:
 - How to determine the criteria of seeds.
 - Seeds with very large degrees can cause memory overload.
 - Estimate the degree of seeds based on the degree distribution of the initial graph.
 - How to determine the number of seeds selected.
 - How to determine the number of steps for transverse.

Heuristic algorithm design

- Find the densest subgraph in the partition:
 - Use max-flow min-cut technique to find the exact densest subgraph in each partition in polynomial time.
 - Select the densest subgraph discovered among partitions as our result.

Performance Evaluation

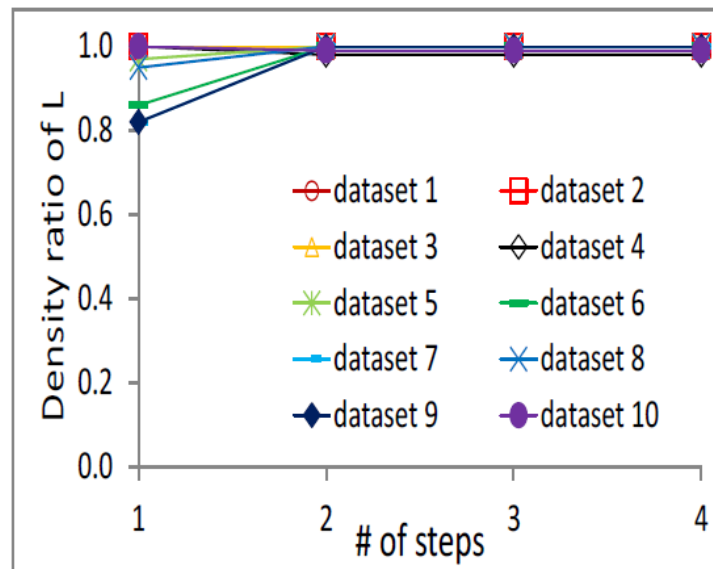
- Platform:
 - Hadoop MapReduce framework on 4 PCs; each PC is quipped with 2.1GHz Intel core i3 processor with 2 cores, and a 2GB memory.
- Data: Real datasets [1] & Simulated datasets [2]
- Metrics for the evaluation
 - **Density ratio:** The density ratio is the density of the subgraph found by our algorithm divides by the density of the exact densest subgraph.
 - **Running time**
- Performance vs. the following factors:
 - **The # of transverse steps**
 - **The # of seeds selected**

[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 (cont.)

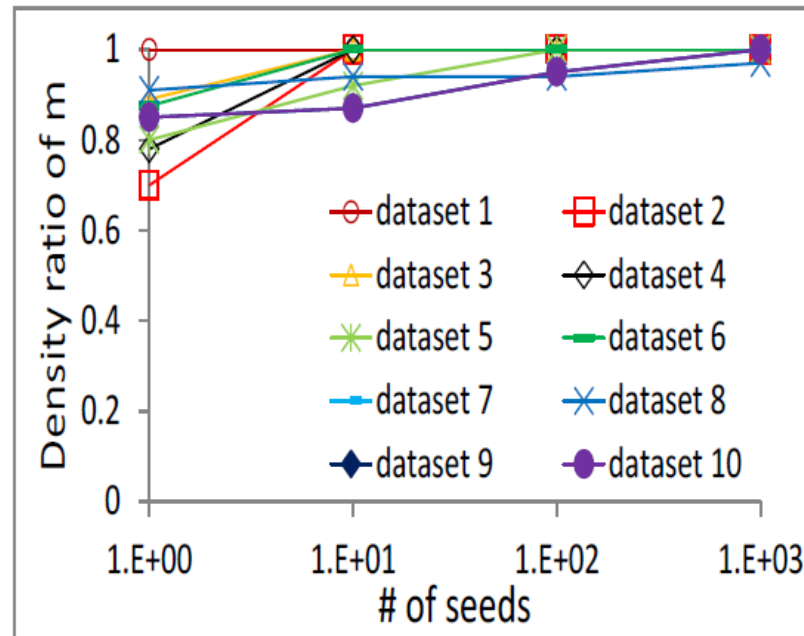
The influence of # of transverse steps on the density ratio of the result:



- Analysis:
 - We only need about two transverse steps to achieve the best performance.

Performance Evaluation (cont.)

The influence of # of seeds on the density ratio of the results:



- Analysis:
 - We only need about tens of seeds to achieve the best performance.

Performance Evaluation (cont.)

The final density of densest subgraph discovered by different algorithms:

MGreedy Greedy Heuristic

$G = (V, E)$	$\rho^*(G)$	$\rho^+(G)$	$\rho(G)$
Dataset 1	22.39	22.39	22.39
Dataset 2	69.97	77.35	77.35
Dataset 3	56.50	56.50	56.56
Dataset 4	4.08	3.76	4.79
Dataset 5	35.31	37.33	36.26
Dataset 6	35.31	37.33	32.09
Dataset 7	43.91	46.25	38.00
Dataset 8	38.65	42.27	40.74
Dataset 9	28.20	30.17	25.42
Dataset 10	78.43	78.43	78.66

Heuristic \approx Greedy $>$ MGreedy

Performance Evaluation (cont.)

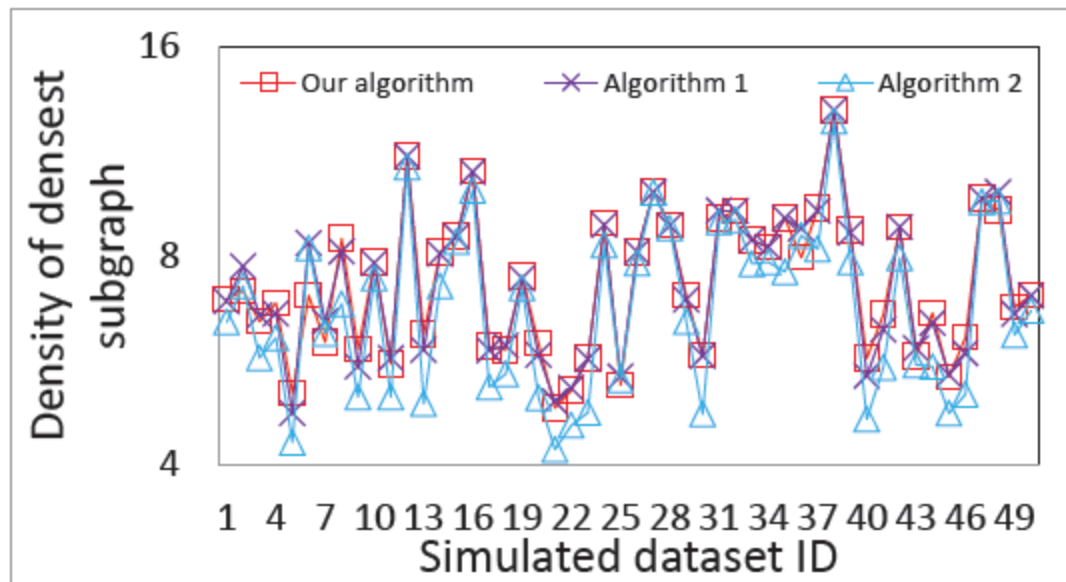
The running time of different algorithms:
MGreedy Heuristic

$G = (V, E)$	$t^*(G)$	$t(G)$
Dataset 1	160	81
Dataset 2	184	82
Dataset 3	257	105
Dataset 4	312	84
Dataset 5	2,175	905
Dataset 6	589	84
Dataset 7	184	82
Dataset 8	245	90
Dataset 9	238	88
Dataset 10	180	108

Heuristic > MGreedy

Performance Evaluation (cont.)

The results on simulated datasets:



Heuristic \approx Greedy $>$ MGreedy which is consistent with the results on real world datasets.

Conclusion

- Our heuristic algorithm can find densest subgraphs with very high density ratio in both real world datasets and simulated datasets. Our algorithm provides the possibility to build real application based on the densest subgraph problem.
- 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