1785

The University of Georgia

Department of Statistics

big data analytics Lab

Subsampling in Large Graphs Using Ricci Curvature

Shushan Wu

Huimin Cheng

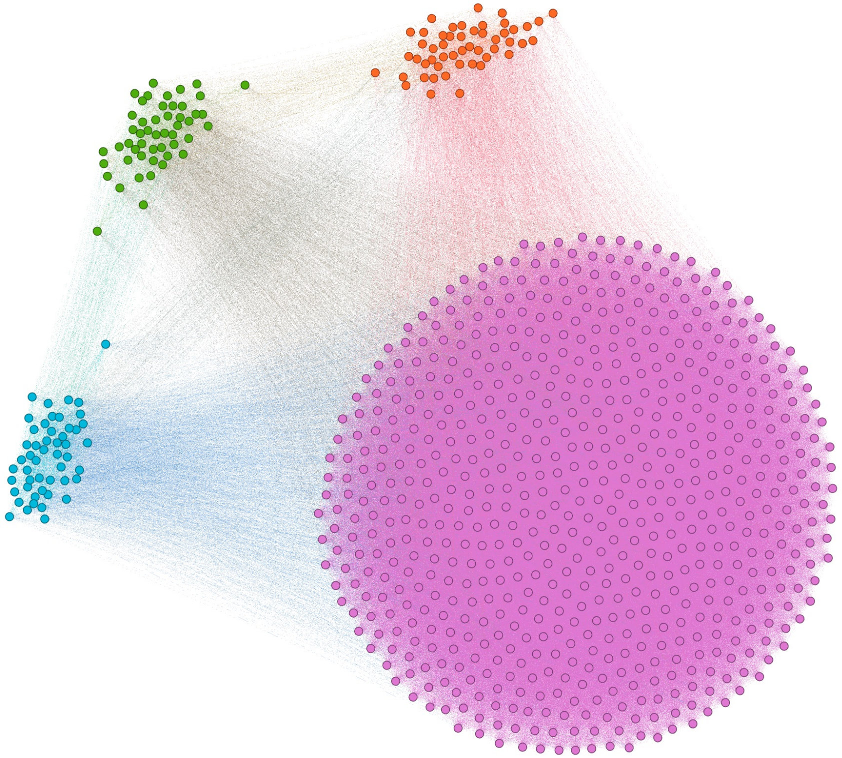
Jiazhang Cai

Ping Ma

Wenxuan Zhong

Challenges

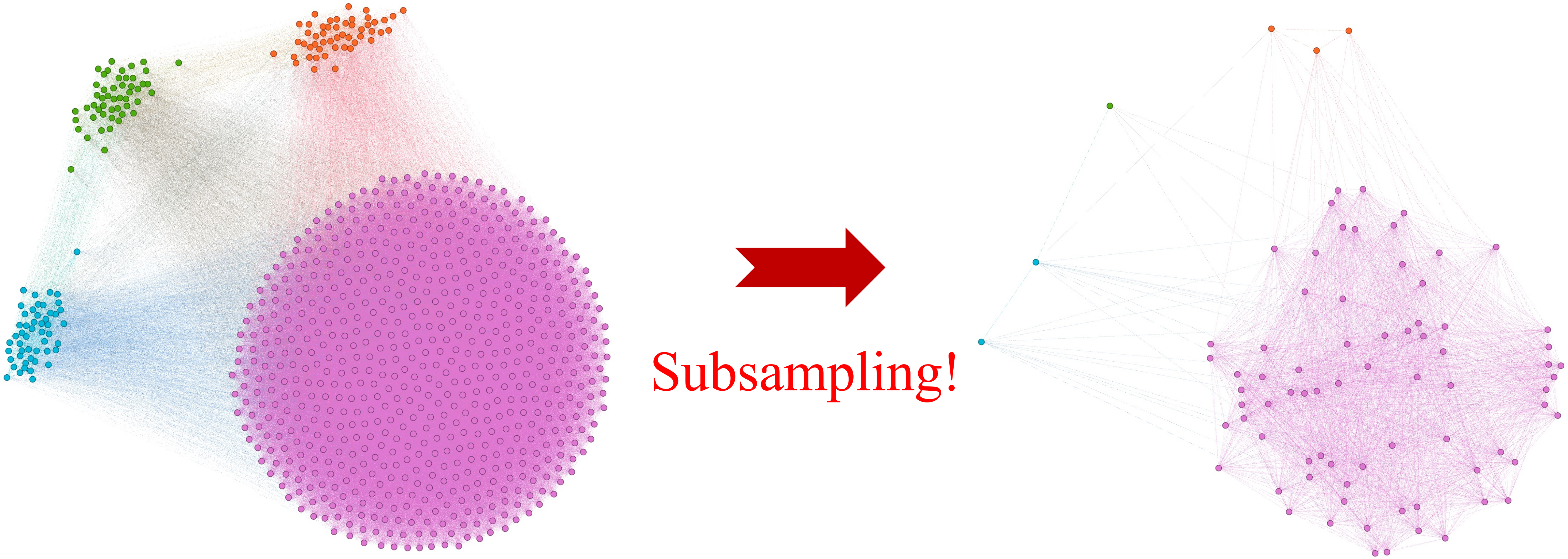
How to tackle huge networks?



Computational cost of network cross validation:

$$O(n^3)$$

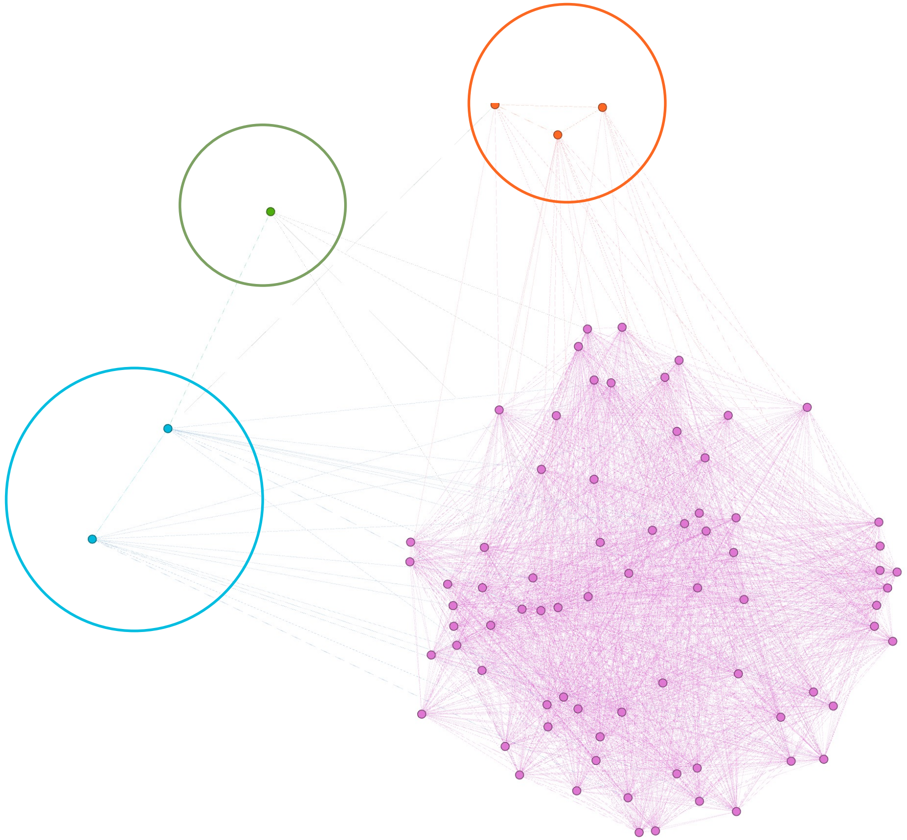
Solution



Subsampling!

Related Work

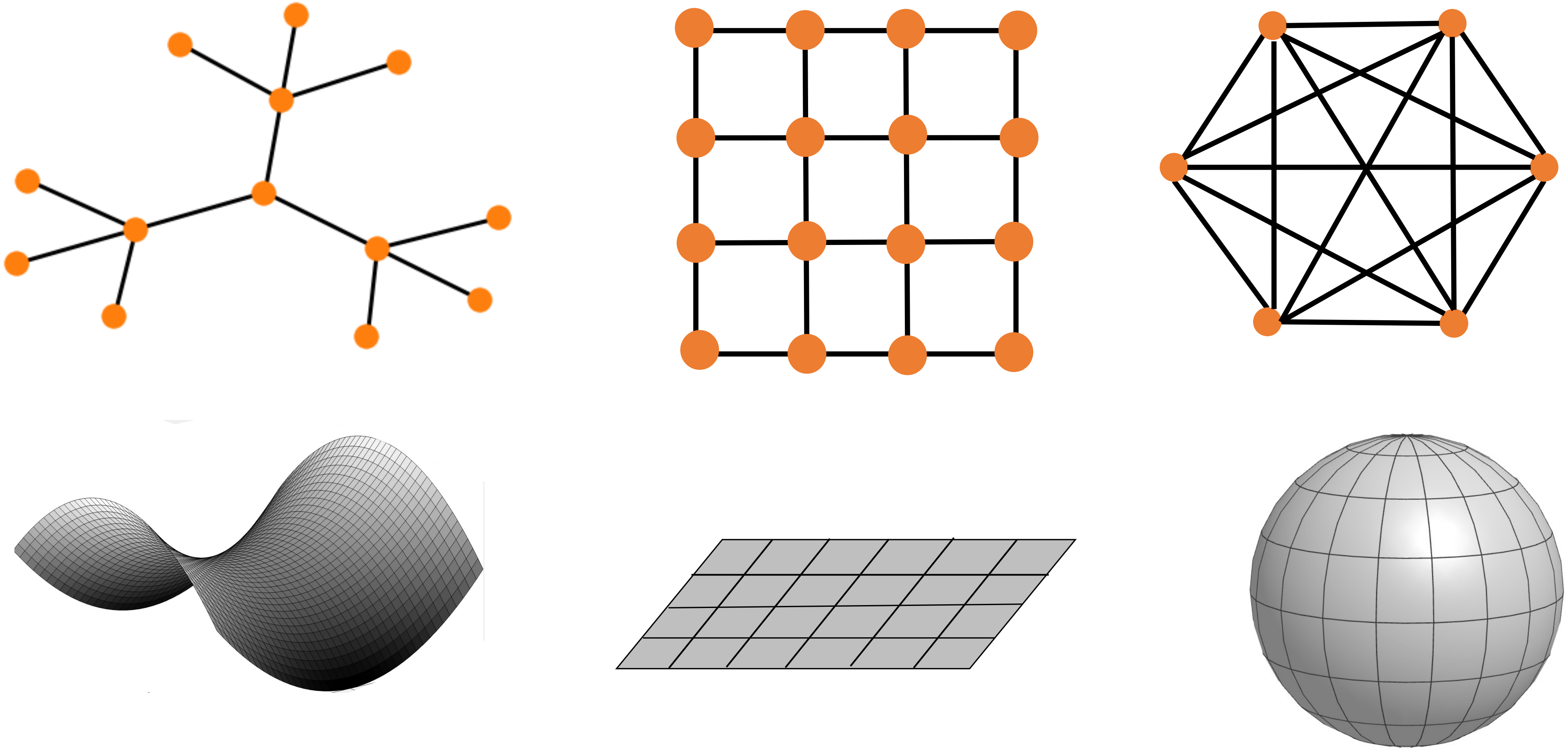
- ➢Node sampling
- ➢Edge sampling
- ➢Exploration sampling



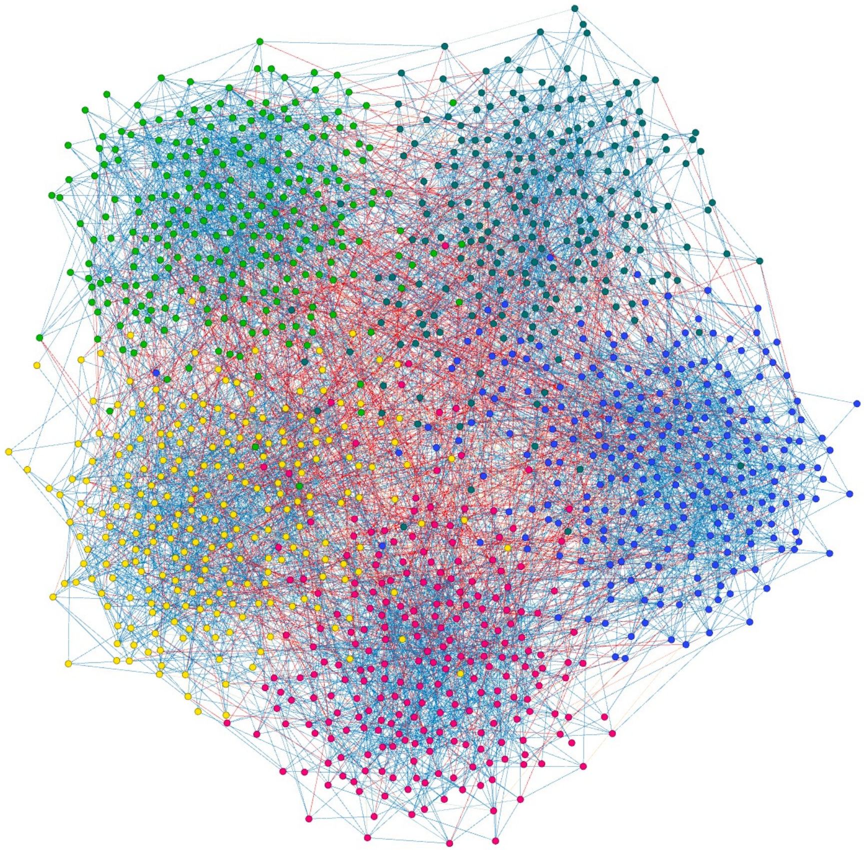
Underestimate *K* !

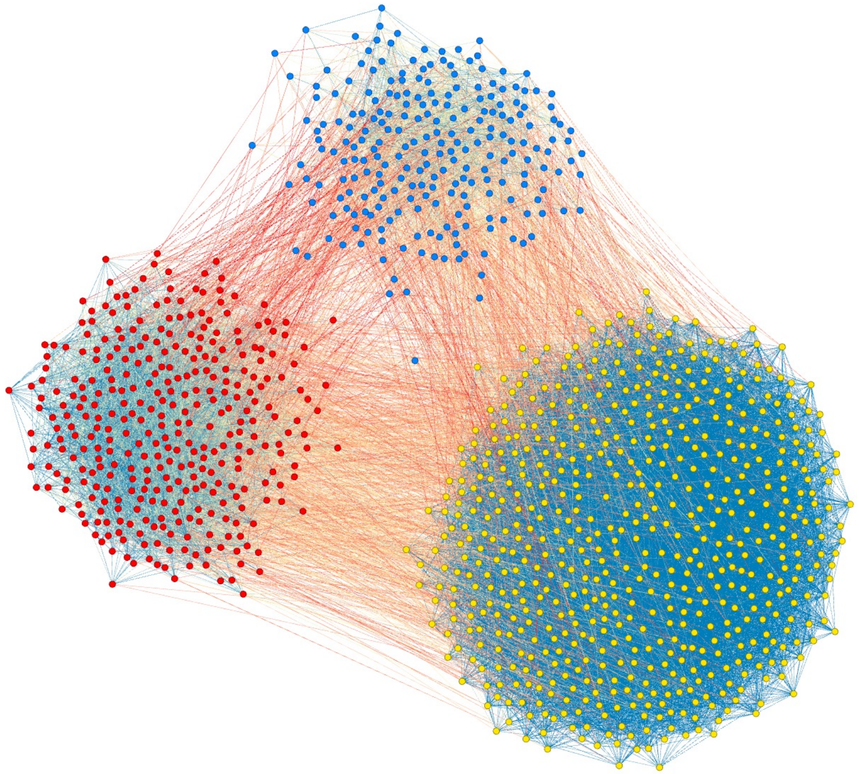
How to exploit community information?

Underling Manifold of Graphs



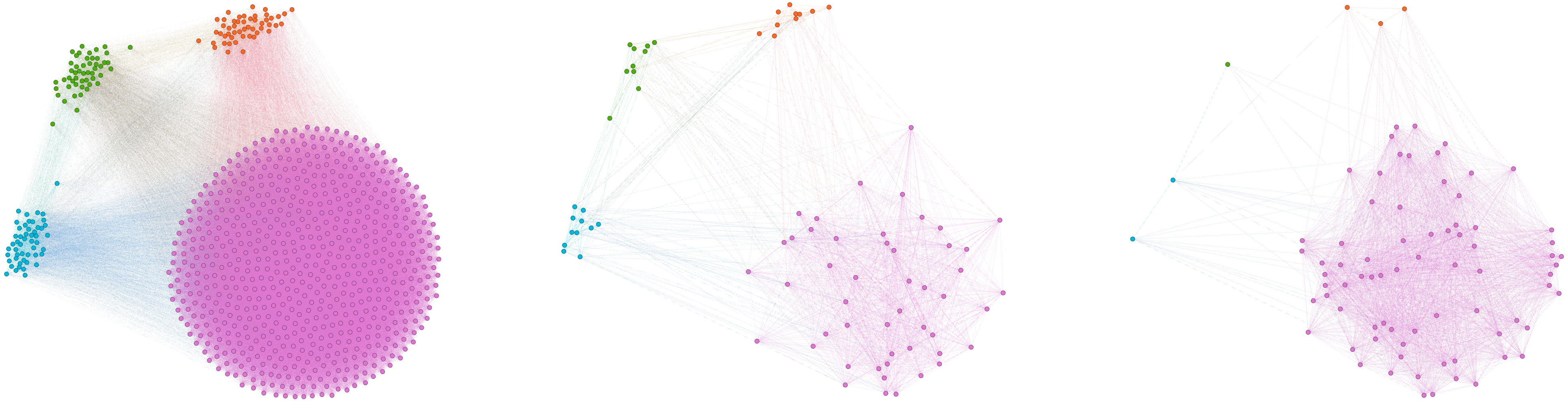
Ollivier Ricci (OR) Curvature





OR Curvature Gradient-based (ORG) Graph Subsampling

$$\left(x^{(i+1)}, y^{(i+1)}\right) = \text{argmax}_{(x,y)\in\Delta\left(\left(x^{(i)},y^{(i)}\right)\right)}\left|\kappa(x,y), \kappa\left(x^{(i+1)}, y^{(i+1)}\right)\right|$$



Given arbitrary community B_i, we have:

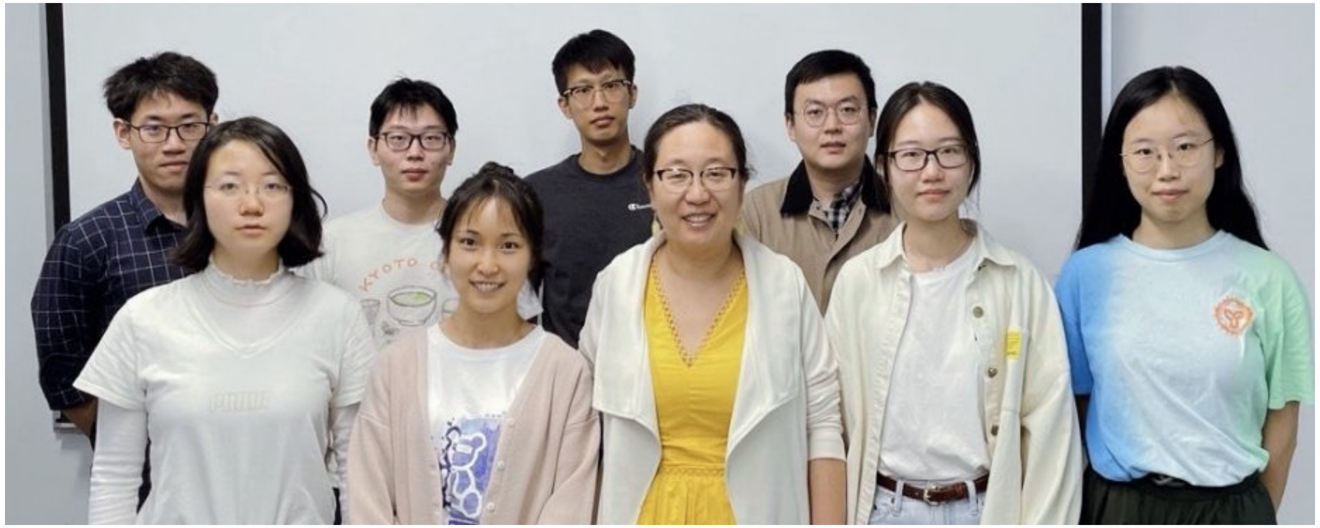
$$P(\exists v_{RW} \in G_{RW}[S]: v_{RW} \in B_i) < P(\exists v_{ORG} \in G_{ORG}[S]: v_{ORG} \in B_i)$$

Experiment Results

Dataset	Prop.	ORG-sub	MHRW	CSE	FFS	Snowball	RW	MDRW
Polbooks (T: 1.88 s)	10%	**0.00** **(T: 0.10 s)**	1.20	0.62	2.68	0.48	0.33	0.00
Polblogs (T: 48.6 s)	5%	**0.00** **(T: 0.23 s)**	1.87	0.90	2.00	0.43	1.03	0.30
PubMed (T: NA)	2%	**0.00** **(T: 4.42 s)**	0.30	0.80	0.40	0.20	1.20	1.80

Time of estimation of M is the much lower than full sample!
Error of estimation of M is the lowest!

Acknowledgement



https://github.com/SavannahWu99/Graph-Subsampling

The University of Georgia
Department of Statistics

big data analytics lab

NSF

NIH