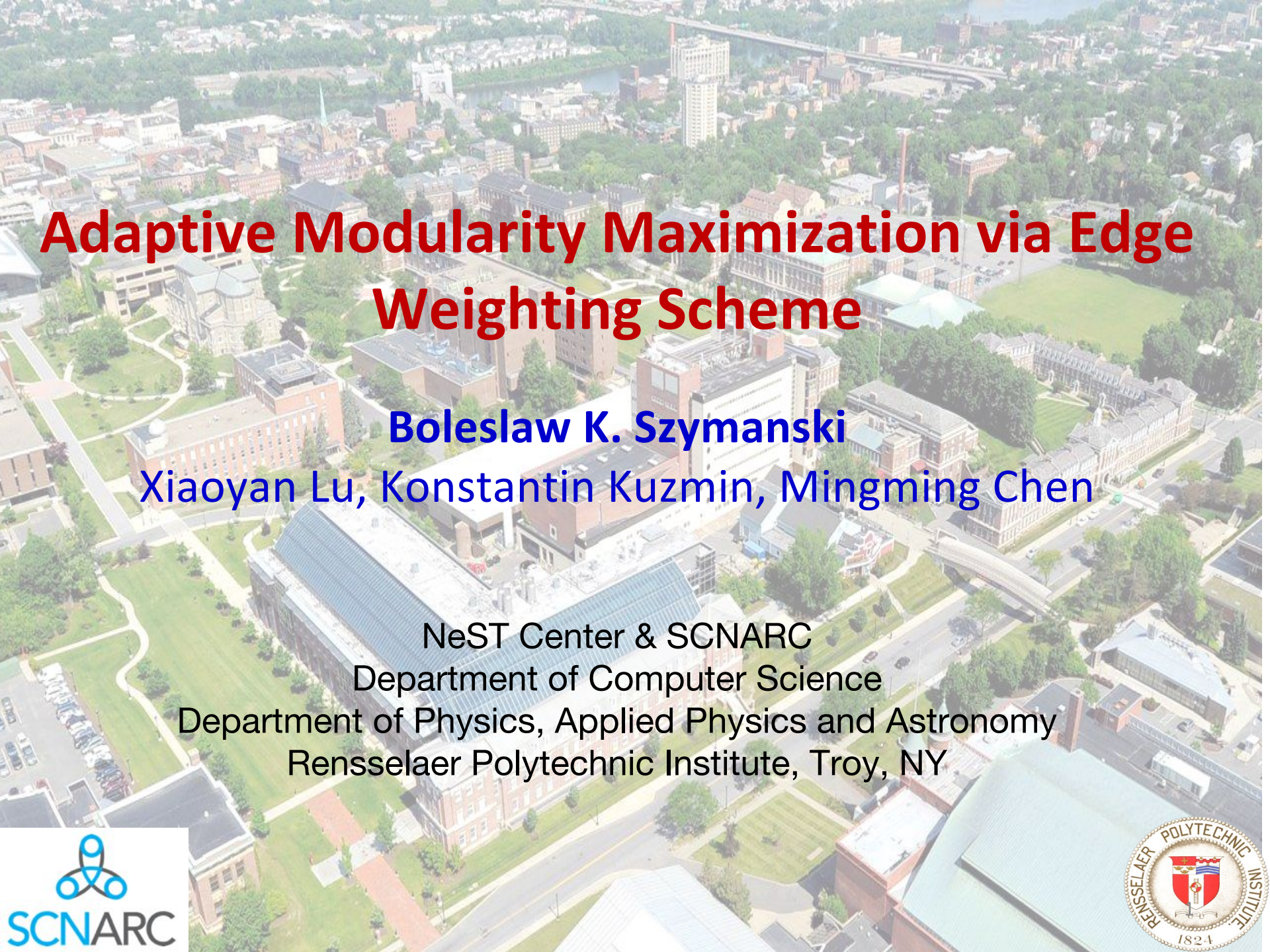


Adaptive Modularity Maximization via Edge Weighting Scheme

Boleslaw K. Szymanski

Xiaoyan Lu, Konstantin Kuzmin, Mingming Chen

NeST Center & SCNARC
Department of Computer Science
Department of Physics, Applied Physics and Astronomy
Rensselaer Polytechnic Institute, Troy, NY

SCNARC

RENSSELAER POLYTECHNIC INSTITUTE 1824

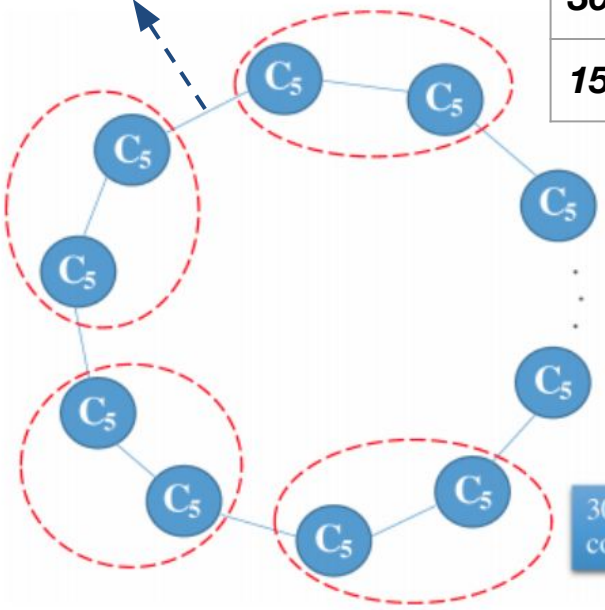
Motivation

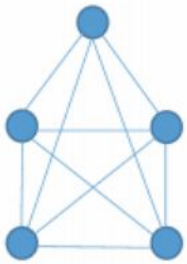
The modularity optimization approaches:

- ➢ Resolution limit problem arises from the definition of modularity.
- ➢ The optimization process traps at some **local** optimums.

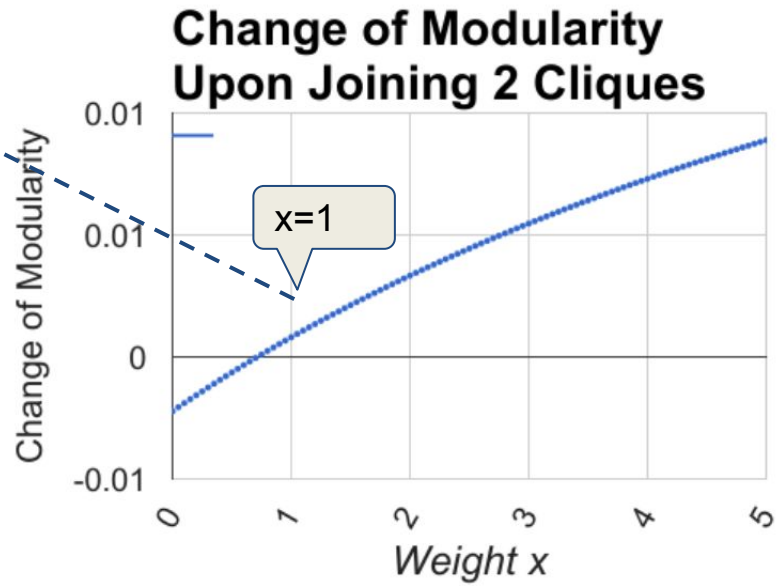
Denote the weight of the edge between two cliques by x

#Comm	*Modularity*
30	*0.8758*
15	*0.8879*





30 cliques, each has 5 nodes, connected with a single edge.



Edge weighting scheme

- The definition of the weighted modularity:

$$Q^w = \sum_{c \in C} \left[\frac{W_c^{in}}{W} - \left(\frac{W_c}{2W} \right)^2 \right]$$

W is the summation of all edges' weights.
W_c is the "weight" of the community c.

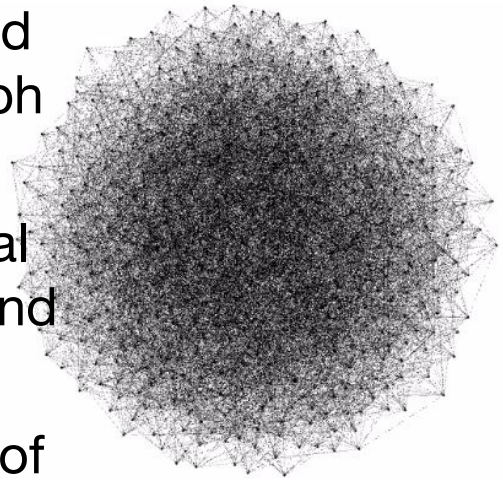
- The change in weighted modularity upon joining *any pair of small ground truth communities* should be negative.

$$\Delta Q^w_{c_i^1, c_i^2} \leq 0 \qquad \text{for } i = 1, 2, \ldots, I$$

- **Idea**: Assign proper weights to edges so that the constraints can be satisfied. A successful edge weighting scheme leads to a decent community detection performance.

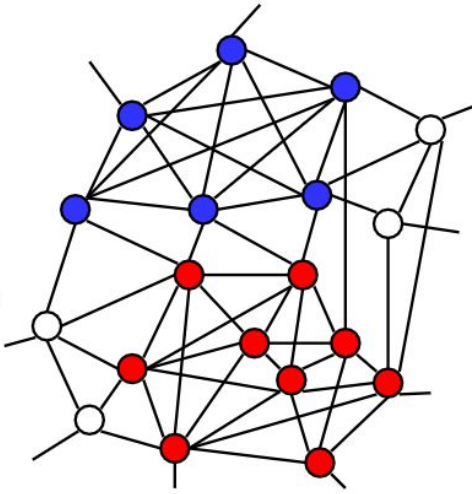
Adaptive Modularity Maximization

- Sample a large unweighted network to obtain the graph parameters.
- Construct a similar artificial network with **known** ground truth communities.
- Extract topology features of the edges $x_e = [f_1, f_2, ...]$.
  - f_1: Common Neighbors.
  - f_2: Jaccard-coefficient.
  - f_3: Adamic-Adar index.
  - etc.
- Train a feature-based **linear** regression model on the edge weight w_e.



Large unweighted networks

Sampled graph parameters



Artificial network with ground truth communities

Unsupervised Learning

Training data for regression

f1	f2	f3	f4	⋯	edge weight
				⋯	
				⋯	
				⋯	
⋯	⋯	⋯	⋯	⋯	⋯
				⋯	
				⋯	

$$w_e = p^T x_e$$

Feature-based regression model

Optimization Task

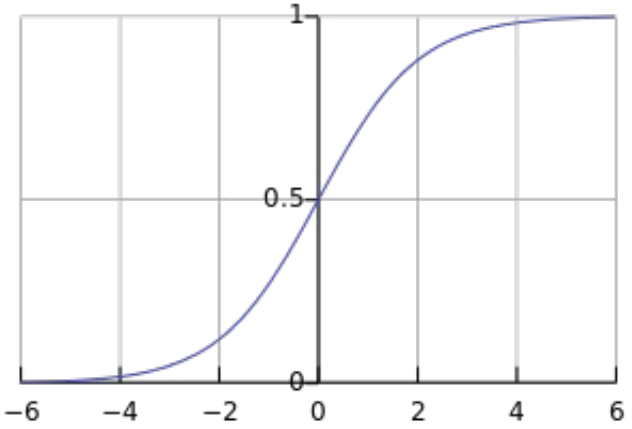
- **Penalization** on the change of modularity upon joining pairs of small ground truth communities:

$$\min_{p} \quad F(p) = (\bar{w} - 1)^2 + \lambda_1 \sigma_w^2 + \lambda_2 \sum_{1 \le i \le l} h(\Delta Q_{c_i^1, c_i^2})$$

- **Regularization** on the variance of the edge weights σ_w^2 to control the portion of negative weights and the average edge weight $\bar{w}$ which is expected to be close to 1.

- **Loss function** such as the sigmoid function is used to improve the robustness against outliers.

$$h(x) = \frac{1}{1 + e^{-x}}$$



Inferring Algorithm

- We can apply the quasi-Newton method, such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which requires the first order derivative of the objective function only.
- The summation of weights of all the edges in the graph is

$$W = \sum_{e \in E} w_e = \sum_{e \in E} p^T x_e = p^T \sum_{e \in E} x_e$$

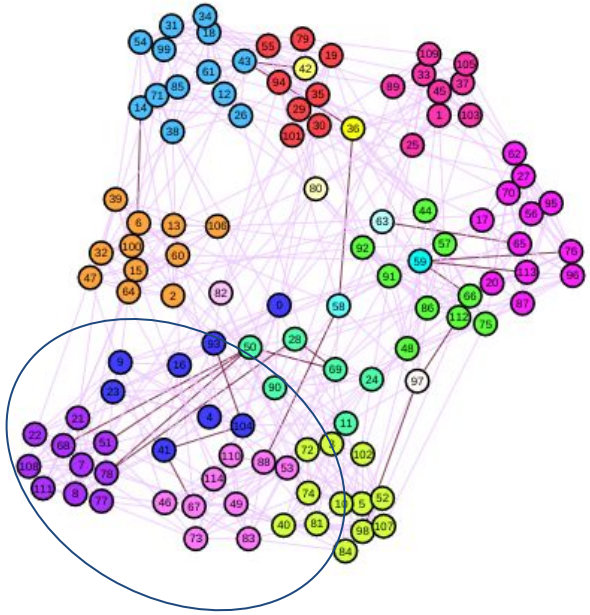
Compute $\sum_{e \in E} x_e$ once at the beginning and update W as the vector product in every BFGS iteration.

- The other terms can also be computed in a similar manner. The time complexity is linear in the number of the edges in **sampled** ground truth communities.

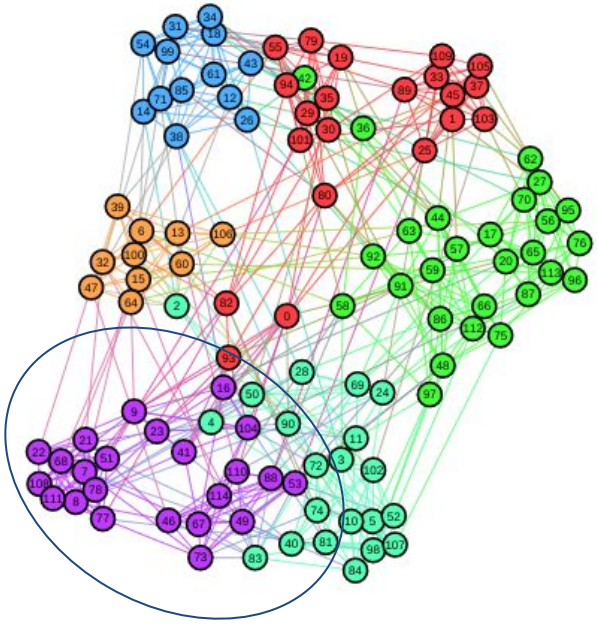
$$O(\#\text{iterations} \times \#\text{Edges})$$

American College Football Network

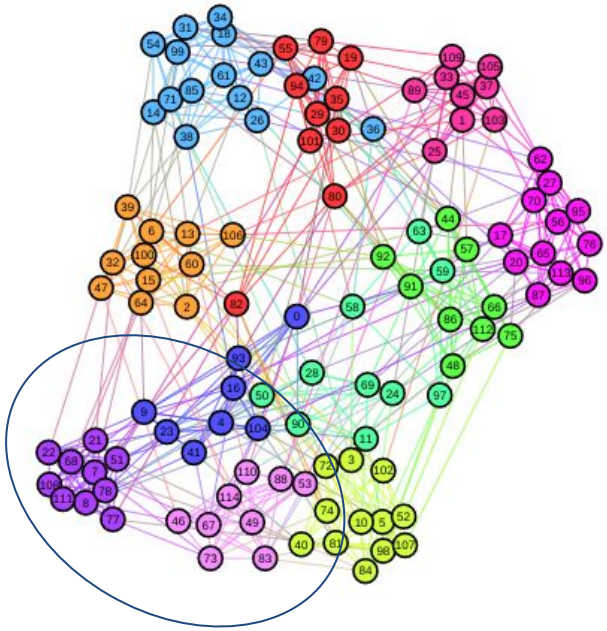
- The American college football network consists of 115 college football teams. The teams in the same conference play with each other more frequently than those in different conferences..



(a) 19 ground truth communities defined as 11 conferences and 8 independent teams. Edges in black are assigned **negative** weights by our model.



(b) 6 communities detected on the **unweighted** graph by the Fast Greedy modularity maximization method.



(c) 11 communities detected on the **weighted** graph by the Fast Greedy modularity maximization method.

American College Football Network

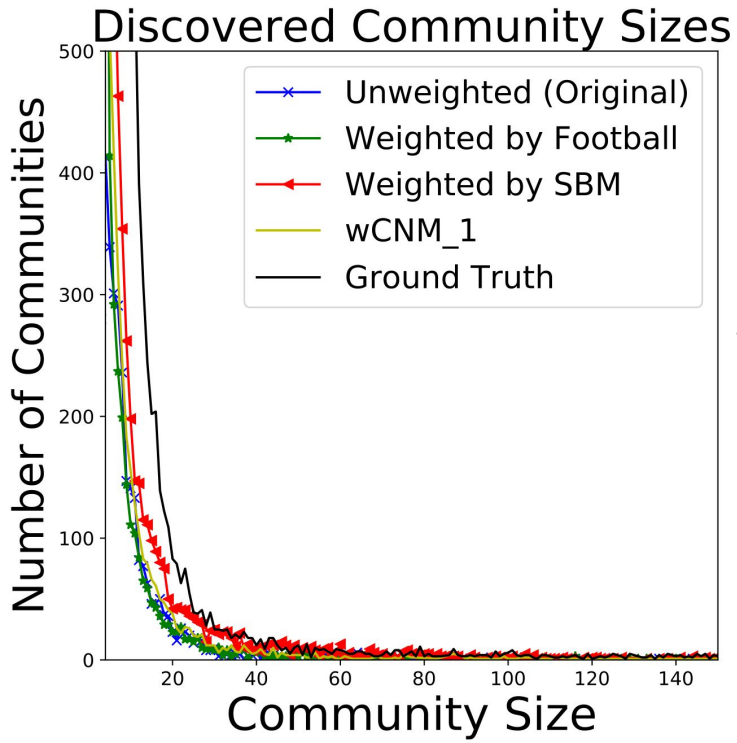
- The community structures are discovered in the **original unweighted** graph or the corresponding **weighted graph** produced by our model.
- Normalized mutual information (NMI), adjusted rand index (ARI), modularity and modularity density are computed over the **original unweighted** graph.
- Modularity maximization algorithm can avoid local optimums in the weighted graph produced by our model.

FG: Fast Greedy algorithm by Clauset et al.
LE: Leading eigenvector method by Newman et al.
LP: Label propagation algorithm by Raghavan et al.
RW: Community detection based on random walks by Pons et al.
ML: Multilevel algorithm by Blondel et al.

Metric	Graph	FG	LE	LP	RW	ML
NMI	Original	0.58528	0.58140	0.76962	0.83833	0.83391
	Weighted	0.91117	0.85903	0.92635	0.91117	0.87272
ARI	Original	0.49333	0.49441	0.71749	0.86938	0.85815
	Weighted	0.94723	0.88982	0.91539	0.94723	0.90085
Q	Original	0.56860	0.49326	0.57668	0.60337	0.60503
	Weighted	0.60140	0.59338	0.57315	0.60140	0.60356
Q_{ds}	Original	0.15877	0.13661	0.21106	0.23650	0.23626
	Weighted	0.25696	0.23893	0.24025	0.25696	0.24889

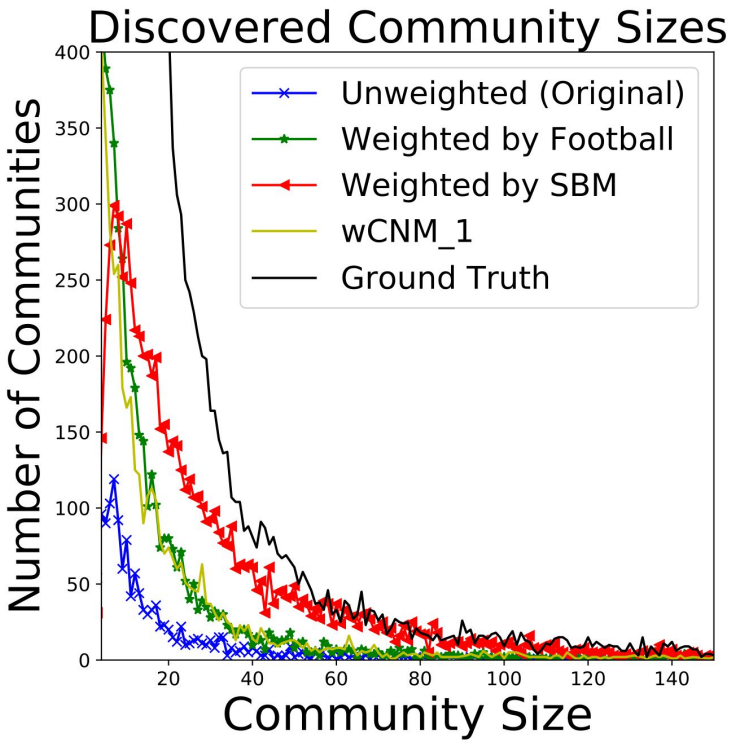
Amazon and DBLP Networks

- Amazon co-purchasing network: 334,863 Amazon products with two frequently co-purchased products linked.
- DBLP co-authorship network: 317,080 researchers in computer science with every pair of co-authors linked.
- The number of communities detected in weighted and unweighted Amazon network in relation to the community sizes:



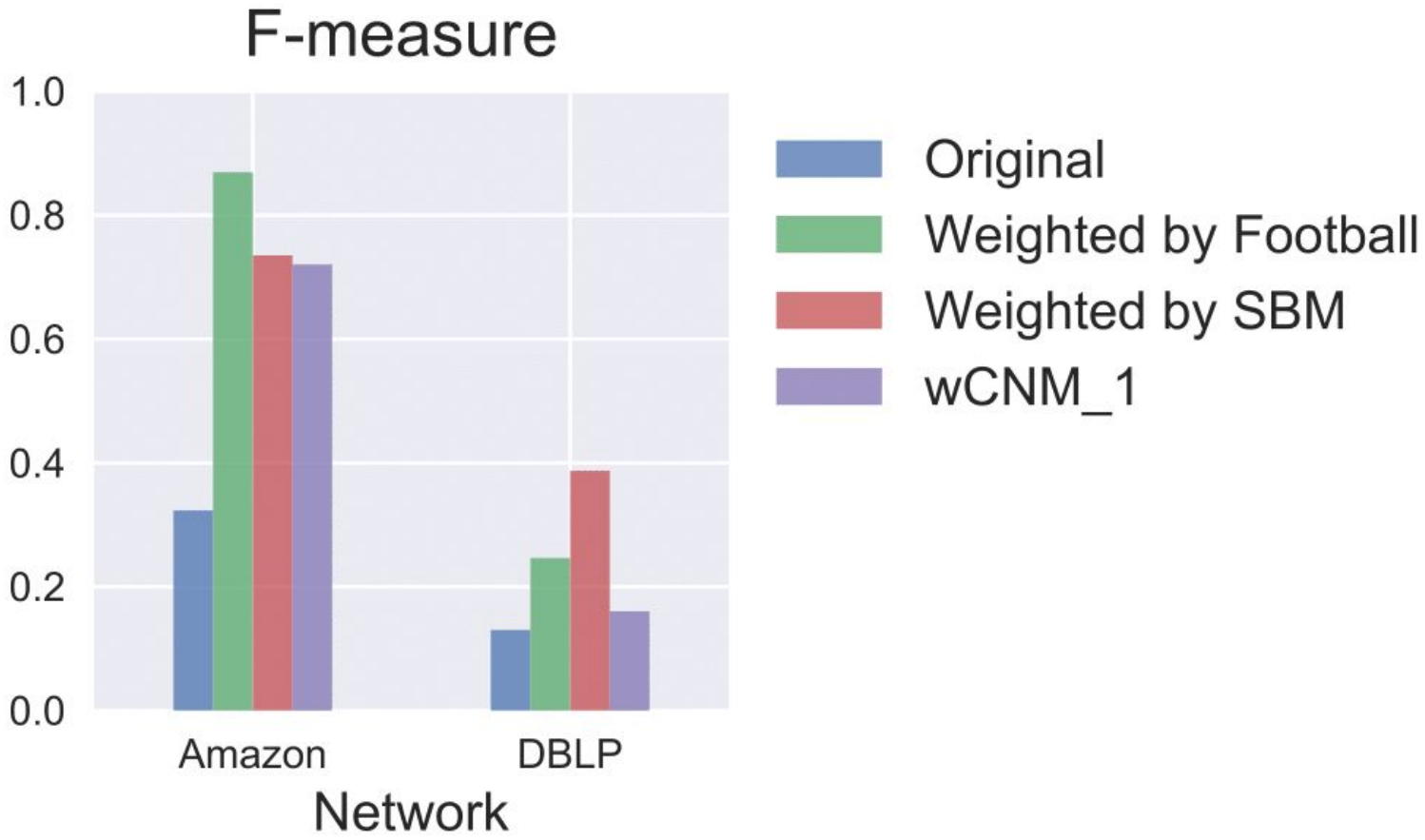
Note: the regression model is trained by the ground truth in

- *Football network*
- *Stochastic blockmodel with sampled graph parameters.*



Amazon and DBLP Networks

- F-measure of the detected communities using different ground truth networks as the training data.



Amazon and DBLP Networks

- The execution time including training time and the edge weighting time (+disk I/O time)