Sketch-based Community Detection via Representative Node Sampling

Andre Beckus*†, Mahlagha Sedghi*, and George K. Atia

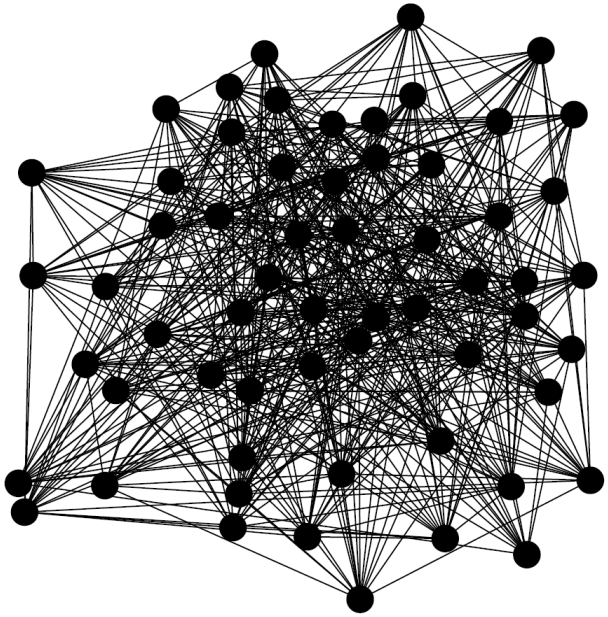
Dept. of Electrical and Computer Engineering
University of Central Florida

ICPR 2020
Paper #2726
Poster Session PS T1.9:
Thursday, January 14th, 2021
12:30 PM CET

* Authors contributed equally to this work
† Currently at the Air Force Research Laboratory (Information Directorate)

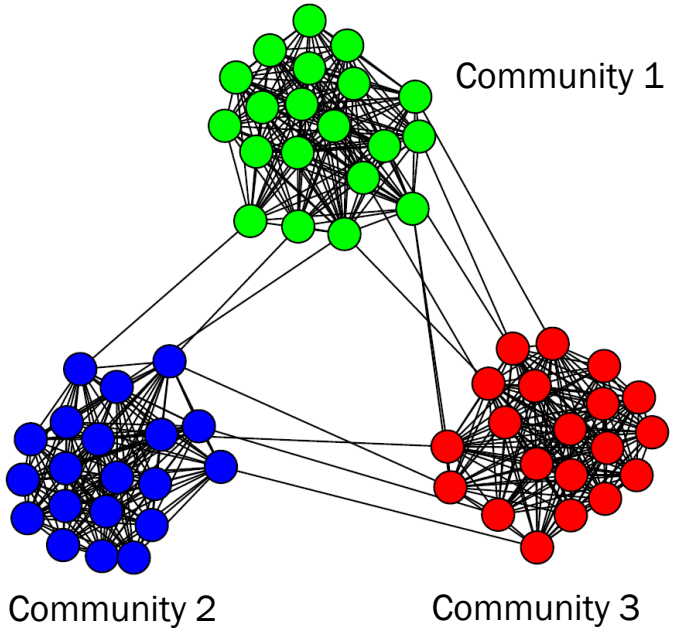
Community Detection

Graph



Community Detection Algorithm

Clustered Graph



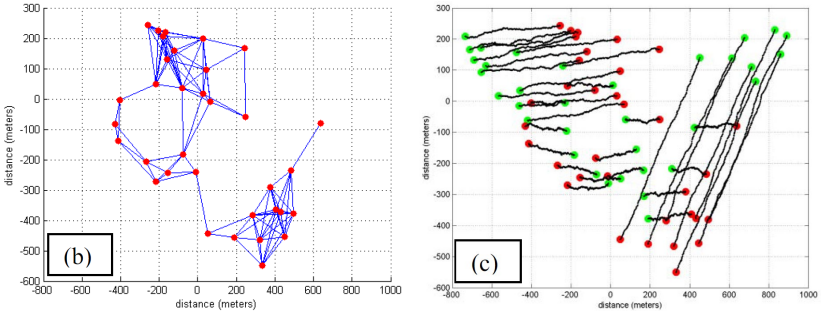
Community: Set of nodes with higher edge density

Community Detection

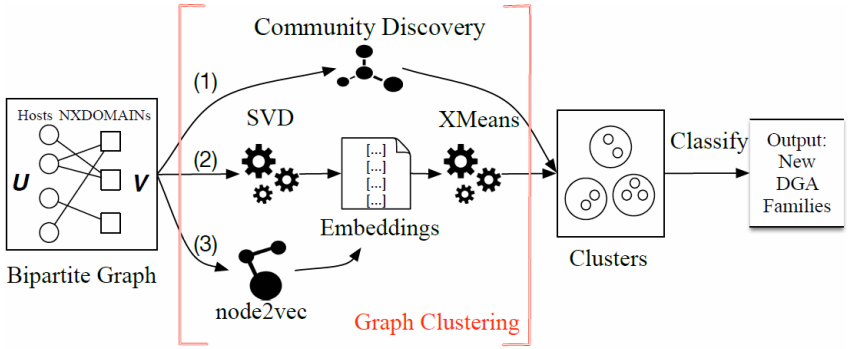
- Applications

Image Segmentation [Shi & Malik, '00]

Organizing Mobile Ad-hoc Wireless Networks
[Parker & McEachen, '16]



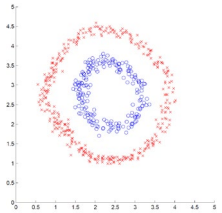
Detecting Computer Network Attacks
[Chen et al.,2017], [Antonakakis et al, '12]



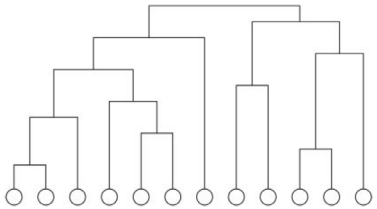
Community Detection

Algorithms

Spectral Clustering
[Shi & Malik, '00], [Ng et al. '01]



Hierarchical [Girvan, '02]



Convex Optimization
e.g. matrix decomposition [Chen, '14] [Vinayak, '14]

$$\min_{\mathbf{B},\mathbf{K}} \quad \lambda \|\mathbf{B}\|_1 + \|\mathbf{K}\|_*$$
$$\text{s.t.} \quad \mathcal{P}_{\Omega_{obs}}(\mathbf{B}+\mathbf{K}) = \mathcal{P}_{\Omega_{obs}}(\mathbf{A}+\mathbf{I})$$

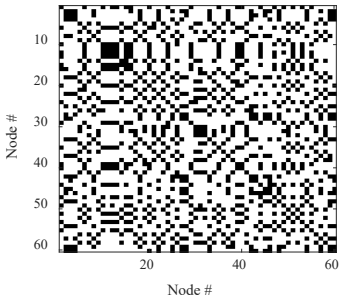
Issues

- Clustering can be slow
- Difficulties in handling small clusters

Sketch-based Clustering

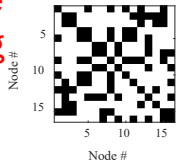
Main Bottleneck

Input
$(N \times N)$



Step 1
Representative Node Sampling

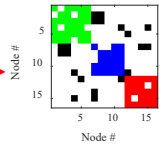
Sketch
$(N' \times N')$



Step 2
Community Detection

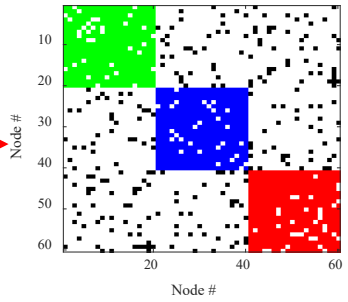
Using any community detection algorithm

Node Labels (Sketch)



Step 3
Infer

Node Labels (Full)



Representative Node Sampling

Given: Graph with N nodes

Positive Semi-definite similarity matrix: $\mathbf{S} \in \mathbb{R}^{N \times N}$

Goal: Find representation matrix $\mathbf{R} \in \mathbb{R}^{N \times N}$

Encodes representation power of each node for describing others

$$\min_{\mathbf{R}} \operatorname{tr}\left\{\mathbf{R}^T\mathbf{S}\mathbf{R} - 2\mathbf{S}\mathbf{R}\right\} + \lambda\|\mathbf{R}^T\|_{1,2}$$

Penalize similarity among chosen representatives

Structured Regularization

1. Enforces subset size constraints
2. Avoids trivial solution

Reward representing other samples

For more details see:

M. Sedghi, M. Georgiopoulos, and G. K. Atia, "A multi-criteria approach for fast and robust representative selection from manifolds," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

Similarity via Average Commute Time

Average commute time $C(i,j)$:
Time for a random walk from node i to node j and then back again.

Why is it useful?
1) It reflects community structure
2) We can form embedding where distance between nodes is $\sqrt{C(i,j)}$
3) We have the relation $C(i,j) = V_G(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+)$

$\mathbf{L}^+$ is the pseudoinverse of the Laplacian

$\Rightarrow \mathbf{L}^+$ is the gram matrix for the embedded nodes.

Our similarity matrix $\mathbf{S}$ is the cosine similarity derived from $\mathbf{L}^+$:

$$s_{ij} = \frac{l_{ij}^+}{\sqrt{l_{ii}^+ l_{jj}^+}}$$

Free benefit: We can use same similarity matrix for the inference step as well

Algorithm

Problem is convex

ADMM-based algorithm:

Algorithm 2 Proposed Representative Node Sampling Method

Require: Similarity matrix $\mathbf{S}$, Regularization parameter λ, initialize all optimization variables to zero.

Ensure: Sampled Representative Node Indices $\mathbf{index} = [\,]$

1: Obtain the optimal representation matrix
  while not converged
    Update $\mathbf{\Delta}, \mathbf{R}, \mathbf{Q}$ according to (7) - (10)
  end while
  $\mathbf{R}^* = \mathbf{R}^t$
2: Locate Representative Samples
  for $i = 1 \ldots n$, **do**
    $rn(i) = \left\| \mathbf{R}^*_{i,\cdot} \right\|_2$
    if $rn(i) \neq 0$
      $\mathbf{index} = [\mathbf{index}, i]$
    end if
  end for
3: **return** $\mathbf{index}$

Both ADMM algorithm and calculation $\mathbf{L}^+$ are amenable to parallelization.

$$\mathcal{O}(N^{1.373}\lceil N/P \rceil) \qquad P = \text{number of processors}$$

Experiments

Clustering step: Convexified Modularity Maximization (CMM)

Compare against other sampling based techniques:

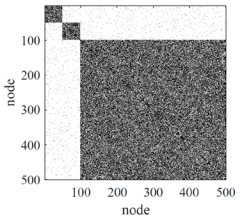
- Uniform Random Sampling (URS) [1]
- Spatial Random Sampling (SRS) [1]
- SamPling Inversely proportional to Node Degree (SPIN) [2]

[1] M. Rahmani, A. Beckus, A. Karimian, and G. K. Atia, "Scalable and robust community detection with randomized sketching," *IEEE Trans. Signal Process.*, vol. 68, pp. 962–977, 2020

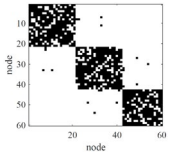
[2] A. Beckus and G. K. Atia, "Scalable community detection in the heterogeneous stochastic block model," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process*, Oct 2019, pp. 1–6

Experiments - SBM

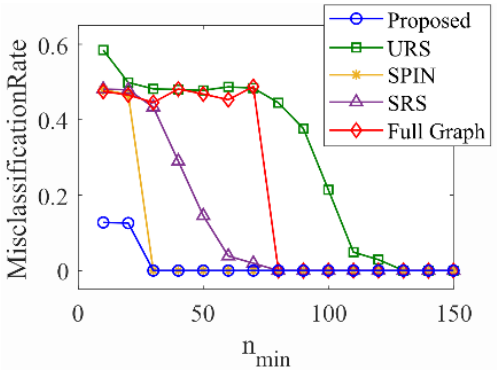
Example Full Graph



Sketch (proposed method)



Homogeneous SBM

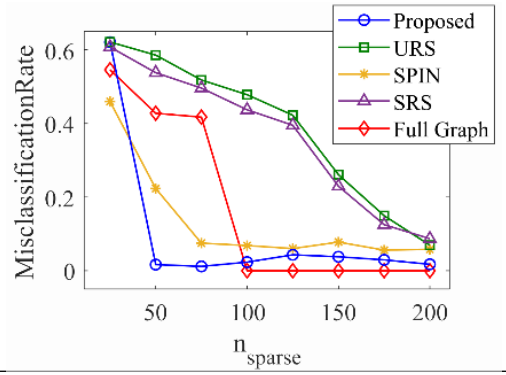


- Uniform intra-cluster edge density for *graph*
- Three clusters - n_{min} is size of smallest cluster
- Smaller n_{min} indicates more imbalance

Clustering time in seconds

N	Sketch-based	Full Graph
400	4×10^{-2}	3.0×10^{0}
800	1.8×10^{-1}	1.4×10^{1}
1600	7.2×10^{-1}	9.3×10^{1}
3200	1.9×10^{0}	7.8×10^{2}
6400	9.7×10^{0}	6.3×10^{3}
12800	6.7×10^{1}	4.6×10^{4}

Heterogeneous SBM



- Intra-cluster edge density varies for each *cluster*
- Smaller n_{sparse} indicates more imbalance.

Experiments – Real World

Misclassification rate for the discrete DCSBM example

p	Full Graph	Sketch-based Algorithm			
		Proposed	URS	SPIN	SRS
0.05	0.095	**0.102**	0.513	0.578	0.525
0.10	0	**0**	0.131	0.338	0.159
0.15	0	**0**	0.024	0.085	0.037

Misclassification rate for power law DCSBM example

p	Full Graph	Sketch-based Algorithm			
		Proposed	URS	SPIN	SRS
0.05	0.064	**0.052**	0.326	0.432	0.362
0.10	0.004	**0.004**	0.122	0.333	0.150
0.15	0.002	**0.002**	0.034	0.214	0.058
0.20	0	**0**	0.011	0.090	0.019

Misclassification rate for the Political Blogs dataset

Dataset	Full Graph	Sketch-based Algorithm			
		Proposed	URS	SPIN	SRS
Full	0.050	**0.052**	0.178	0.438	0.218
Unbalanced	0.437	**0.142**	0.224	0.334	0.289

For Degree Corrected SBM:
Average degree varies for each *node*

Take Away

- Way to sample nodes from graph
  - Samples nodes with good representation power
- Sketch-based community detection
  - Speed-up clustering step
  - Handle unbalanced graphs

ICPR 2020
Paper #2726
Poster Session PS T1.9:
Thursday, January 14th, 2021
12:30 PM CET

Acknowledgement

This work was supported in part by NSF CAREER Award CCF-1552497.