Hidden Community Detection in Social Networks

Kun He, ***Yingru Li,*** Sucheta Soundarajan, John E. Hopcroft

szrlee@hust.edu.cn

Huazhong University of S&T, Cornell University

Outline

- Clustering and Community Detection
- Hidden Structure and Hidden Community Detection

Clustering

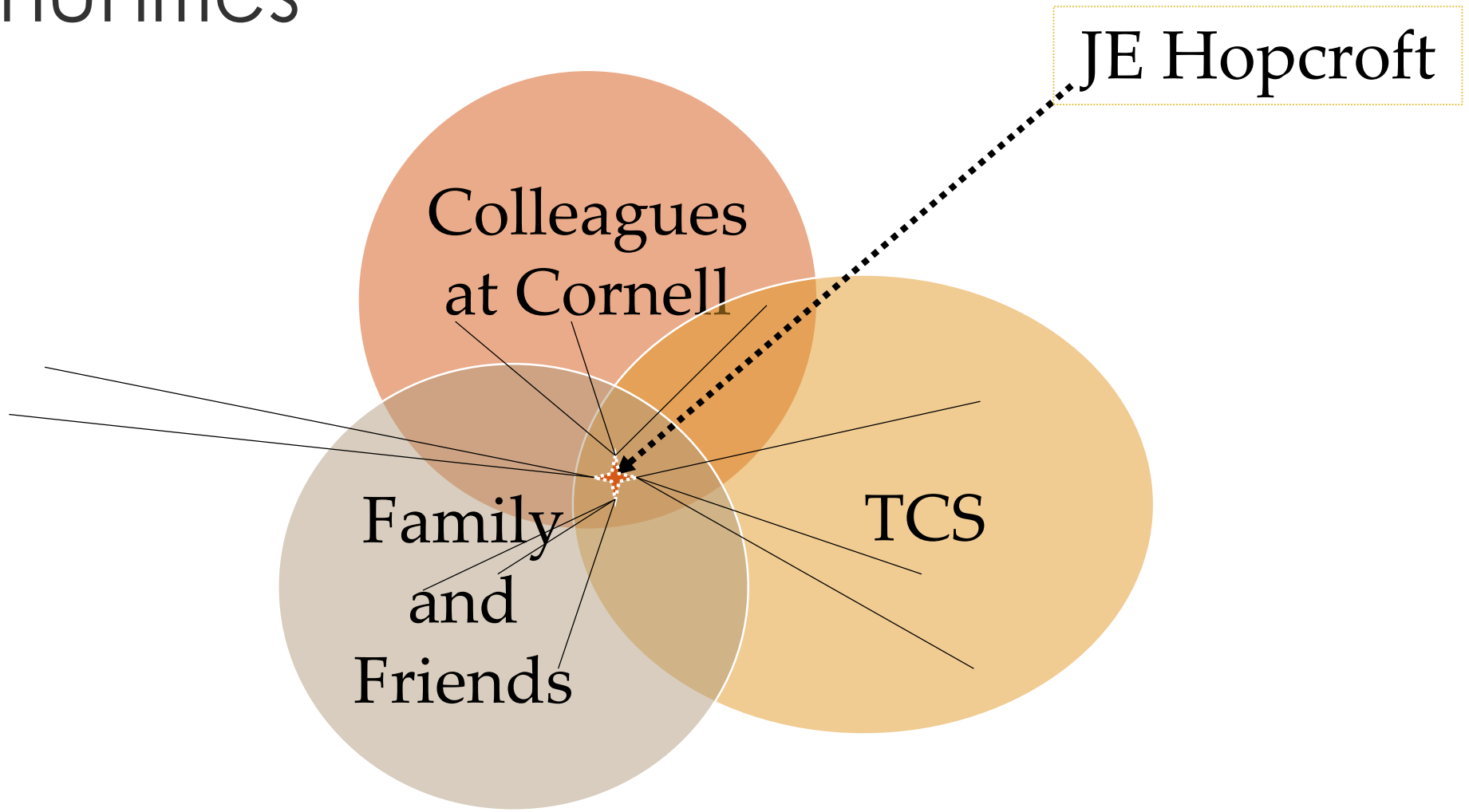
- 
- 
- 

- 
- 
- 

- 
- 
- 

- Multi View Clustering
  - D. Niu, J. G. Dy, and M. I. Jordan. Multiple non-redundant spectral clustering views. In ICML, pages 831–838, 2010.
  - Etc.

Communities

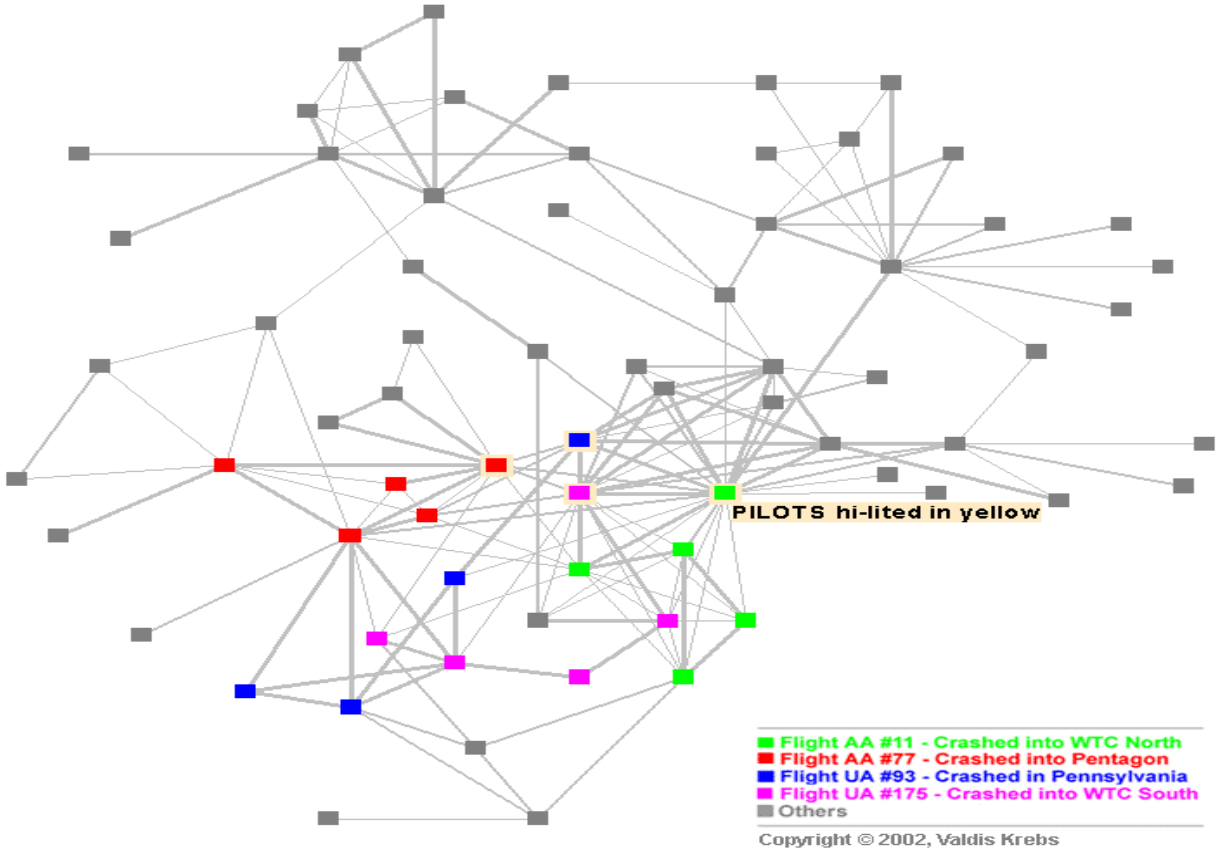


Community Detection

- Disjoint Community Detection
  - **Louvain Method (MOD)**
    - V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10:P10008, 2008.
  - **Infomap (IM)**
    - M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. PNAS, 105(4):1118–1123, 2008.
- Overlapping Community Detection
  - **Link Communities (LC)**
    - Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. Nature, 466(7307):761–764, 2010.
  - **OSLOM (OS)**
    - A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. Finding statistically significant communities in networks. PLoS ONE, 6(4):e18961, 2011.

Hidden Structure

- Secret organizations
- College Students from same high schools
- Brotherhood
- Terrorist Networks



Hidden Structure

- Sparse in Network
- Comparatively weakly concentrated in Network view
- With weaker homophily
- Most of its members veiled by stronger communities
- Hard to be discovered.
- Interfere with accurate detection of the dominant structure.
- Still meaningful!

Preliminaries

- Modularity [Newman et al.]
  - Measure the quality of a partition (or a set of overlapping comms [S. Zhang et al.])

$$Q = \sum_{k=1}^{c} Q_k = \sum_{k=1}^{c} \left[\frac{e_{kk}}{M} - \left(\frac{d_k}{2M} \right)^2 \right]$$

- Normalized Mutual Information (NMI) [L. Danon et al.]
  - Measure the similarity of two partitions (overlapping sets [A. F. McDaid et al.])

$$NMI(X,Y) = \frac{2I(X,Y)}{H(X) + H(Y)}$$

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad p(x) = |x|$$

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \qquad p(x,y) = |x \cap y|$$

Formal Definitions

- Given a network $G = (V, E)$
- Given a set of overlapping communities $\mathcal{C} = \{C_1, C_2, ..., C_k\}$
- Given a metric $\mathcal{M}$ that assigns a quality score to a community
  - E.g. modularity
- $\mathcal{M}_i$ denotes the quality score of community C_i in graph G
  - Higher modularity score indicates Stronger community
  - Metric can be trivially modified. (e.g. conductance)

Formal Definitions

- Definition 1. Hiddenness Value.
  - The hiddenness value H_{C_i} of community C_i is the fraction of nodes belonging to various communities with a higher $\mathcal{M}$ score.
- Definition 2. Layer.
  - A layer is a set of communities that partitions or covers the network.
- Definition 3. Hiddenness Value of a Layer.
  - The hiddenness value of a layer is the weighted average hiddenness values of the communities in this layer.

Formal Definitions

- Dominant Layer
  - the layer with the lowest hiddenness value

- Hidden Layer
  - the layer with a comparatively high hiddenness value

Example

- Caltech Facebook network dataset
- Annotations:
  - the year of matriculation (‘Year’), the residence address (‘Dorm’).
- Suppose there are only communities from the ‘Year’ and ‘Dorm’.
- 79% of the nodes in the ‘Year’ Layer belong to a stronger community in ‘Dorm’ Layer
- 8% on the contrary
  - since year or dorm itself is not overlapping

Hidden Community Detection

- Not only uncover all communities on the ‘surface’ layer
- But also reveal the hidden communities veiled by other stronger communities in deep layer
- Simultaneously and Iteratively

HICODE – A Meta Approach

- Base Detection Algorithms
  - Feed a Graph $G = (V, E)$
  - Output a set of Communities $\mathcal{C} = \{C_1, C_2, ..., C_k\}$

HICODE – A Meta Approach

- Stage 1. Identification
  - Identify a layer of communities via the base algorithm
  - Weaken the structure of the detected layer
  - Repeat until the appropriate number of layers are found

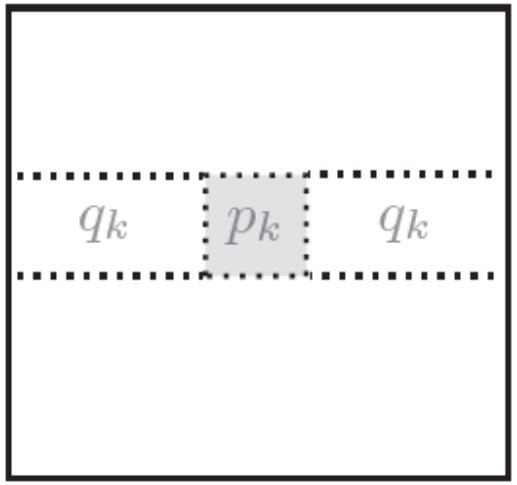
- Stage 2. Refinement
  - Weaken the structures of all other layers from the original network to obtain a reduced network
  - Apply the base algorithm to the resulting network

HICODE – A Meta Approach

- Reducing (Weaken) Methods
  - RemoveEdge
    - removing all intra-community edges
  - ReduceEdge
    - This method approximates each layer as a stochastic blockmodel, where other edges are regarded as background noise.
    - Observe:

$$p_k = \frac{e_{kk}}{0.5n_k(n_k - 1)}$$

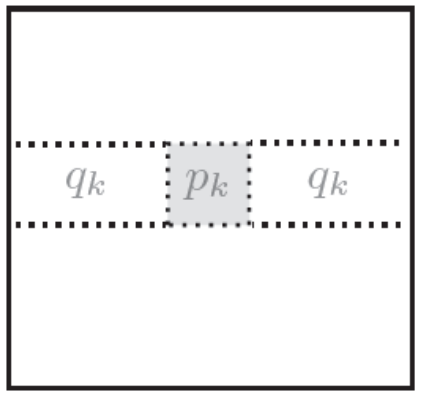
$$q_k = \frac{d_k - 2e_{kk}}{n_k(n - n_k)}$$



HICODE – A Meta Approach

- Reducing (Weaken) Methods
  - ReduceEdge
    - Observed edge probability p_k in community C_k as the superposition of the underlying edge probability p'_k and background block probability q_k
    - Thus, $p_k = 1 - (1 - p'_k)(1 - q_k)$
    - Edge in C_k generated by the background noise is

$$q'_k = 1 - p'_k = \frac{1 - p_k}{1 - q_k} = \frac{q_k}{p_k}$$



    - ReduceEdge removes each edge within community C_k with probability $1 - q'_k$
    - it keeps each internal edge with probability q'_k (keep background noise)

HICODE – A Meta Approach

- ReduceWeight.
  - This method reduces the weight of each edge within community C_k by a factor of q'_k
  - Deterministic

HICODE – A Meta Approach

- Selecting the Number of Layers
  - Q_t the average modularity of all the detected layers at step t:
  - Simple Greedy Method
    - 1. Start with Number of Layers = 2
    - 2. Perform HICODE with T iterations and evaluate Q_t, $t \in \{1, \ldots, T\}$. (set T = 10)
    - 3. Increase by one and goto 2 until $R_T = \frac{\sum_{t=1}^{T} Q_t}{T \cdot Q_0}$ decrease for the first time.
  - Probably a theorem here.

HICODE – A Meta Approach

- HICODE with Base Algorithms
  - MOD + HICODE ➔ HC:MOD
  - IM + HICODE ➔ HC:IM
  - LC + HICODE ➔ HC:LC
  - OS + HICODE ➔ HC:OS

Empirical Results

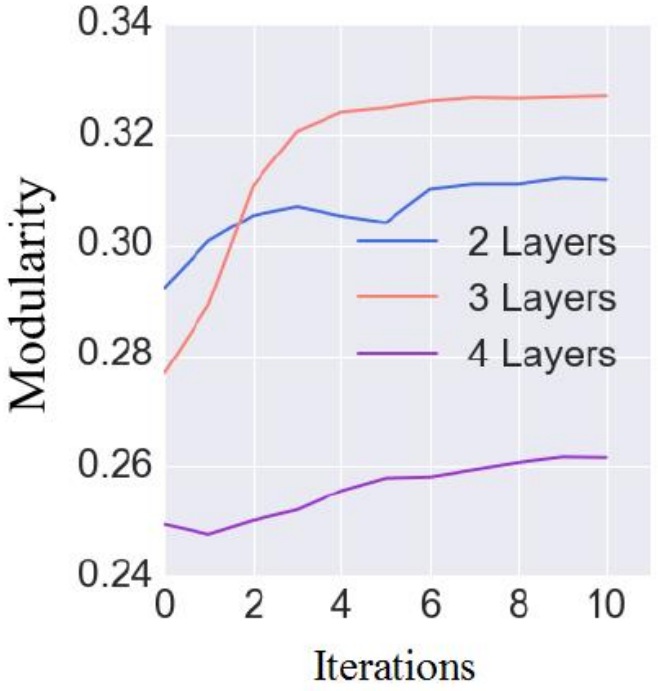
- Synthetic Data

Dataset	$\|V\|$	$\|E\|$	#Layers	$\#C$	$\|C\|_{avg}$	block probability	modularity	NMI_{max}	F_{max}
SynL2_200	200	960	2	5, 4	40, 50	0.12, 0.10	0.398, 0.391	0.05	0.02
SynL2	3,000	14,446	2	100,50	30, 60	0.16, 0.08	0.491, 0.492	0.20	0.03
SynL3	3,000	21,510	3	100,50,30	30, 60, 100	0.16, 0.08, 0.048	0.332, 0.321, 0.323	0.20	0.04

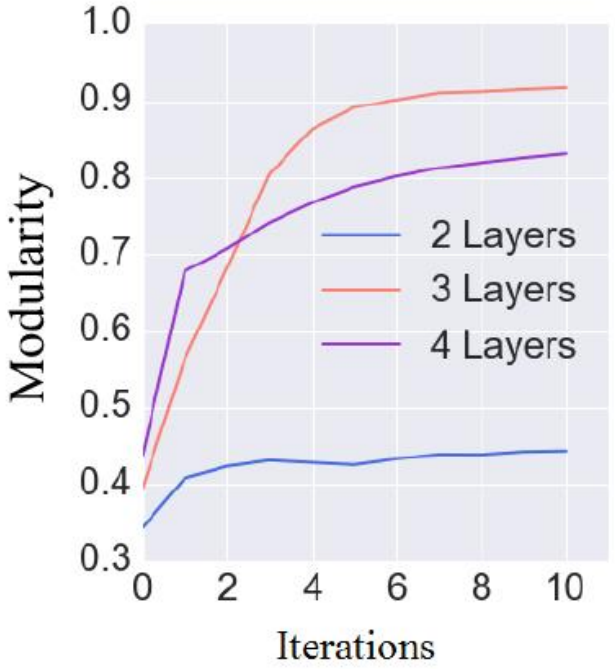
- Real Data

Source	Domain	Dataset	$\|V\|$	$\|E\|$
Facebook	**Social**	Caltech	769	16,656
		Smith	2,970	97,133
		Rice	4,087	184,828
		Vassar	3,068	119,161
		Wellesley	2,970	94,899
		Bucknell	3,826	158,864
		Carnegie	6,637	249,967
		UIllinois	30,809	1,264,428
SNAP	**Social**	YouTube	31,150	202,130
	Products	Amazon	13,288	41,730
	Collaboration	DBLP	49,097	170,284

Empirical Results (Number of Layers)



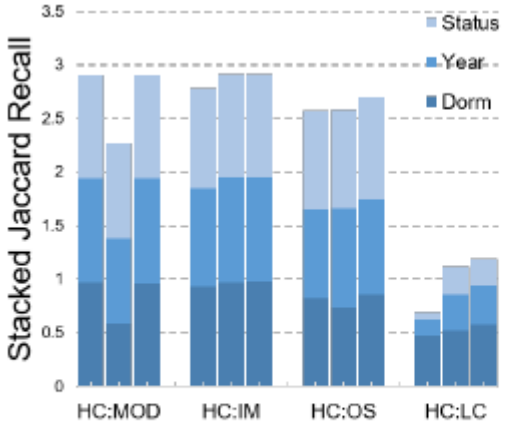
(a) in original network



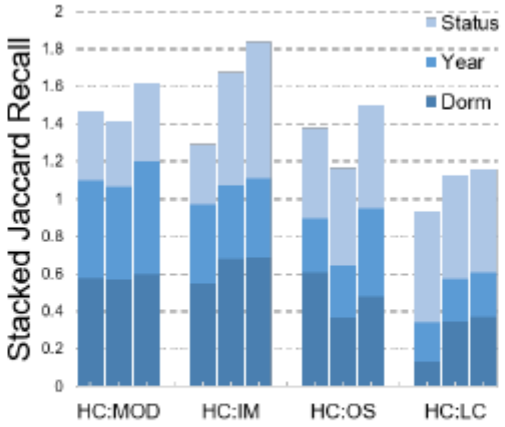
(b) in reduced network

Fig. 2. Change in average modularity when detecting 2, 3, or 4 fixed number of layers on SynL3.

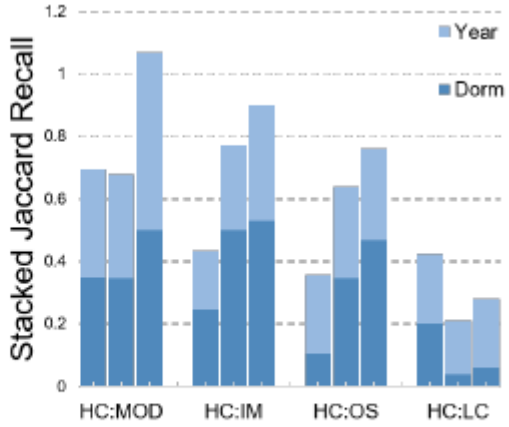
Empirical Results (Reducing Methods)



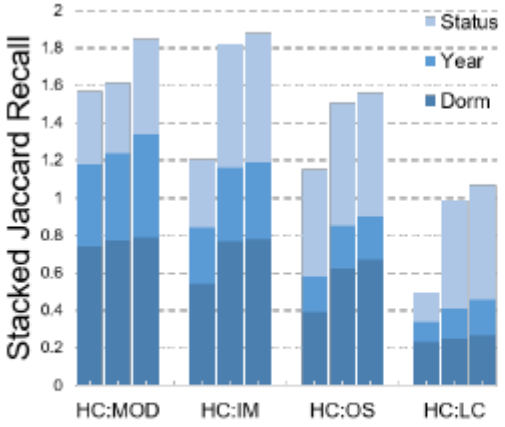
(a) SynL3



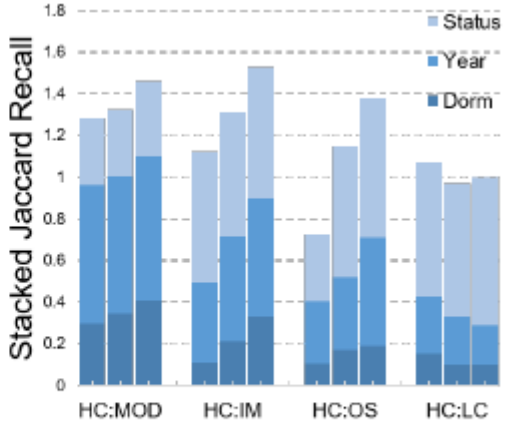
(b) Caltech



(c) Smith

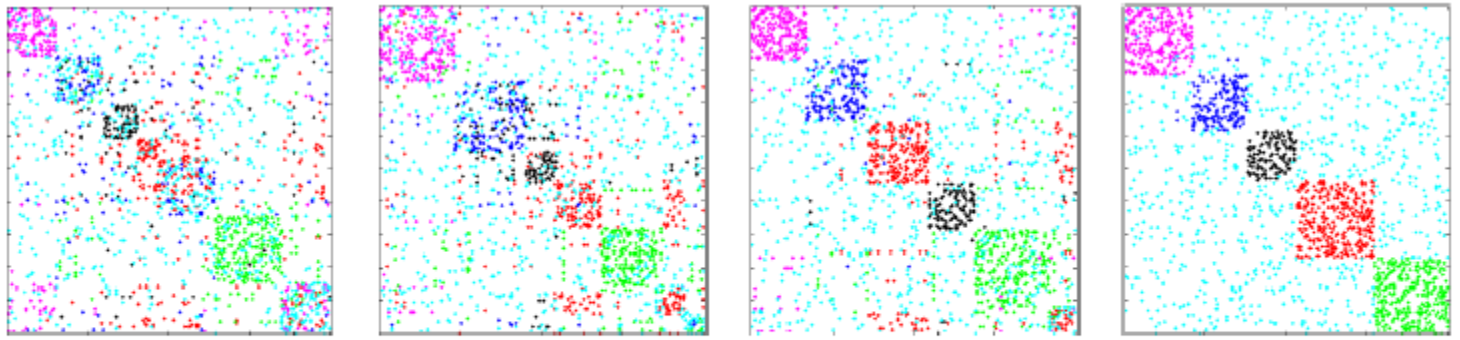


(d) Rice



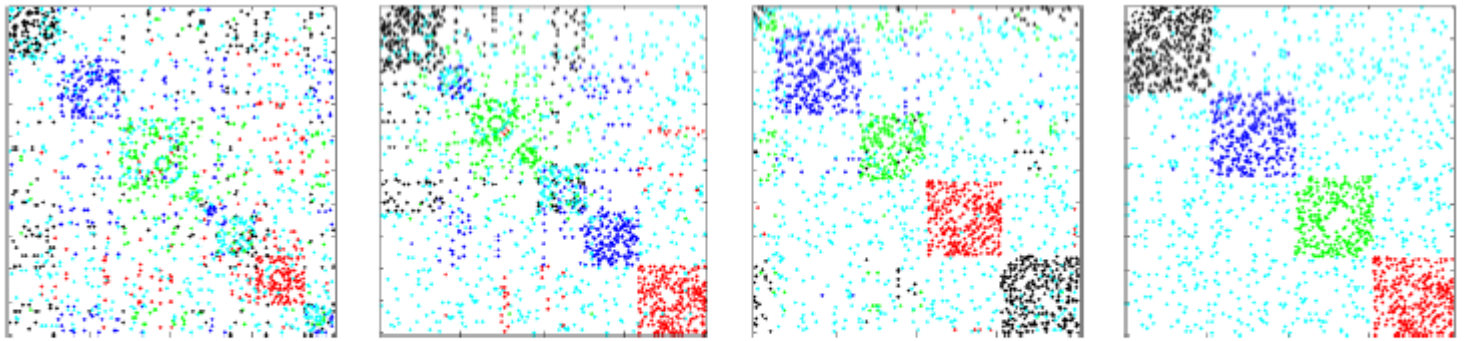
(e) Vassar

Result Visualization



(a) $t = 0$ (b) $t = 5$ (c) $t = 10$ (d) $t = 20$

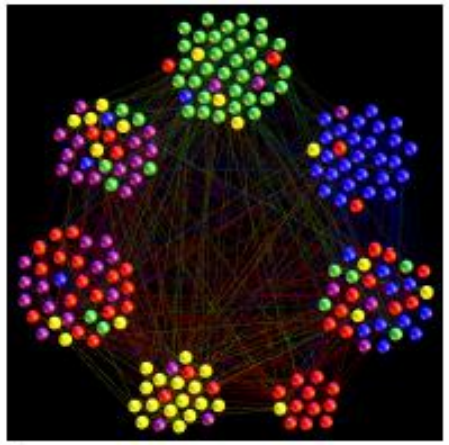
Fig. 3. Refinement of layer 1 on SynL2_200.



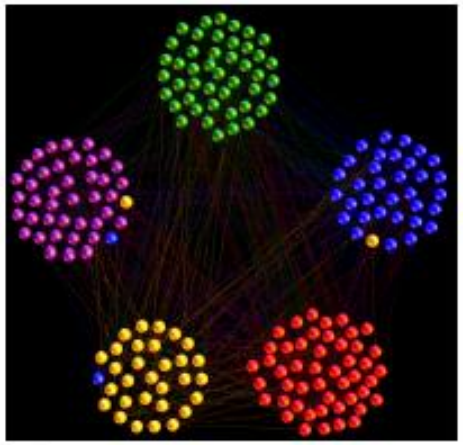
(a) $t = 0$ (b) $t = 5$ (c) $t = 10$ (d) $t = 20$

Fig. 4. Refinement of layer 2 on SynL2_200.

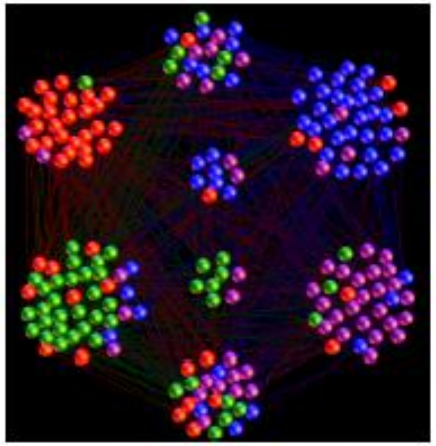
Result Visualization



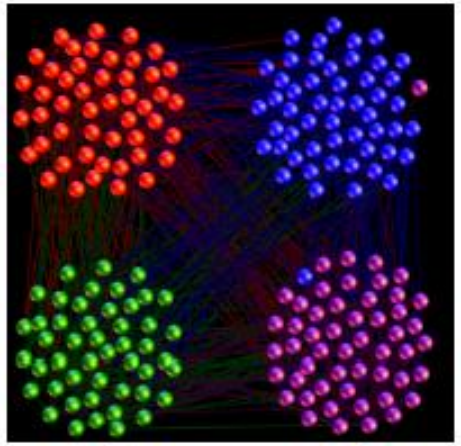
(a) Initial output of layer 1



(b) Final output of layer 1

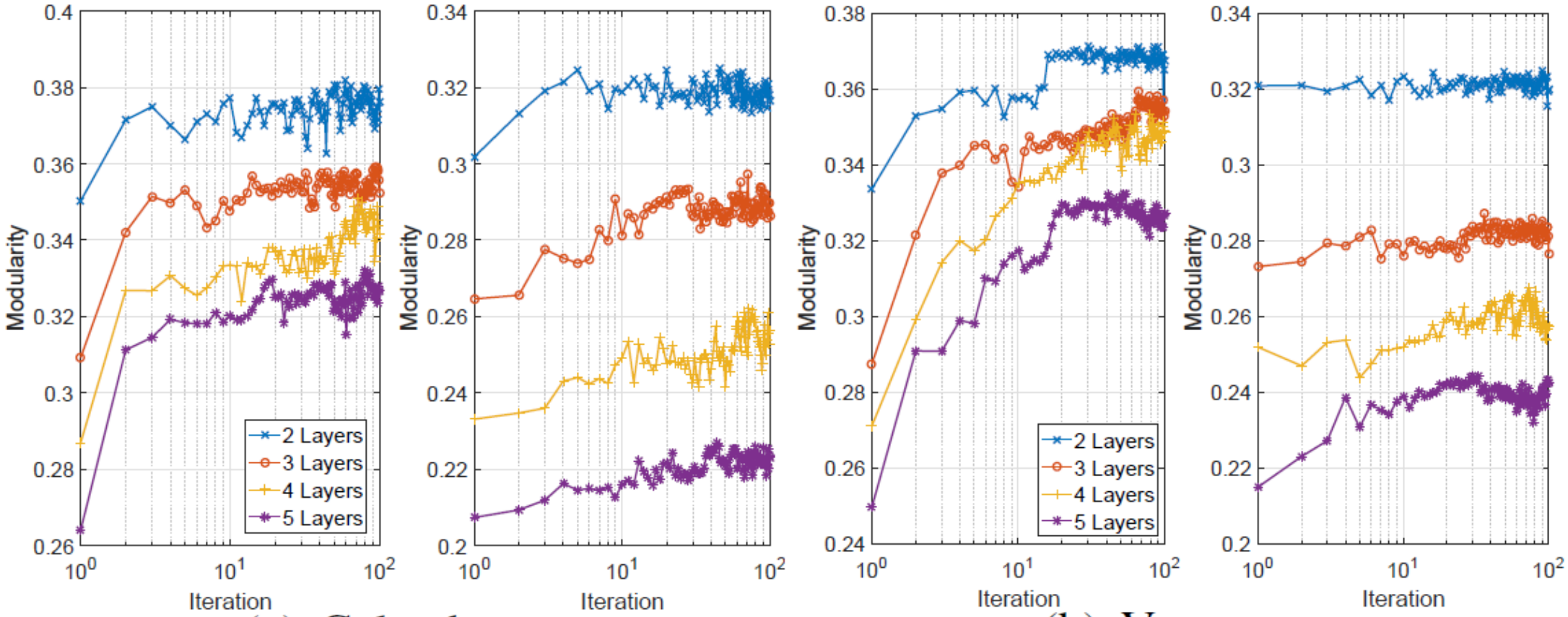


(c) Initial output of layer 2

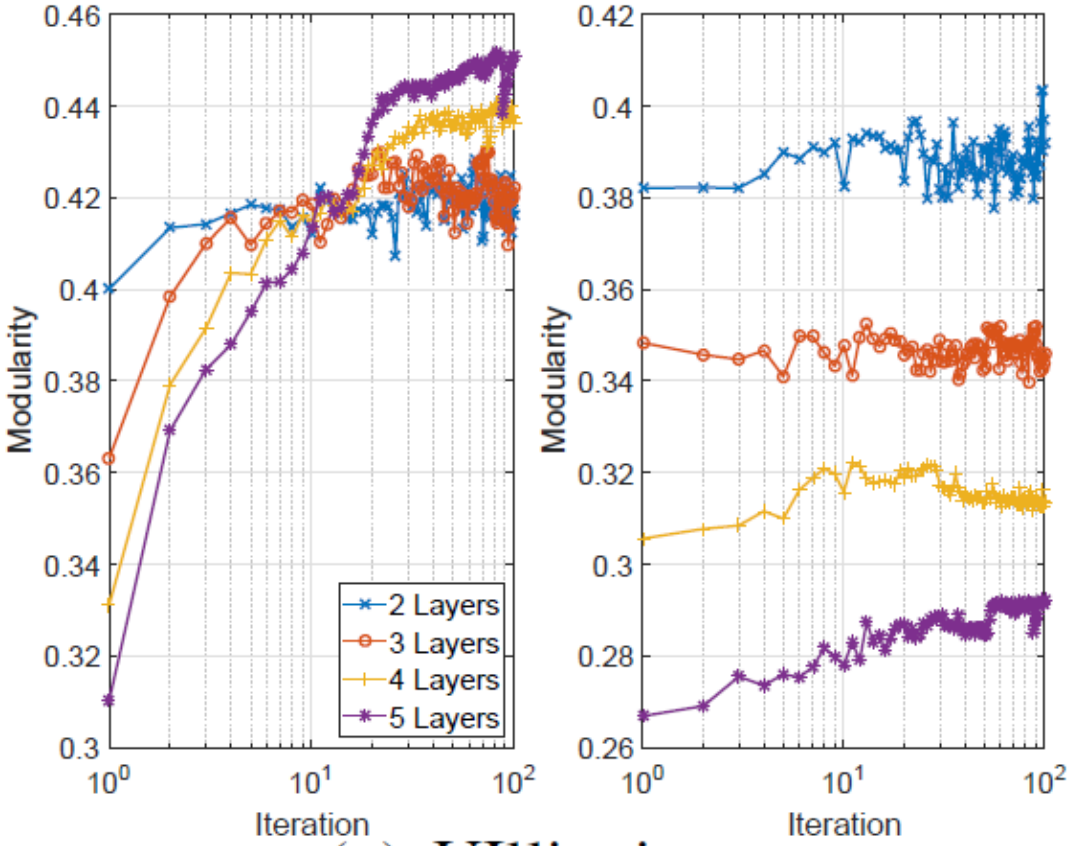


(d) Final output of layer 2

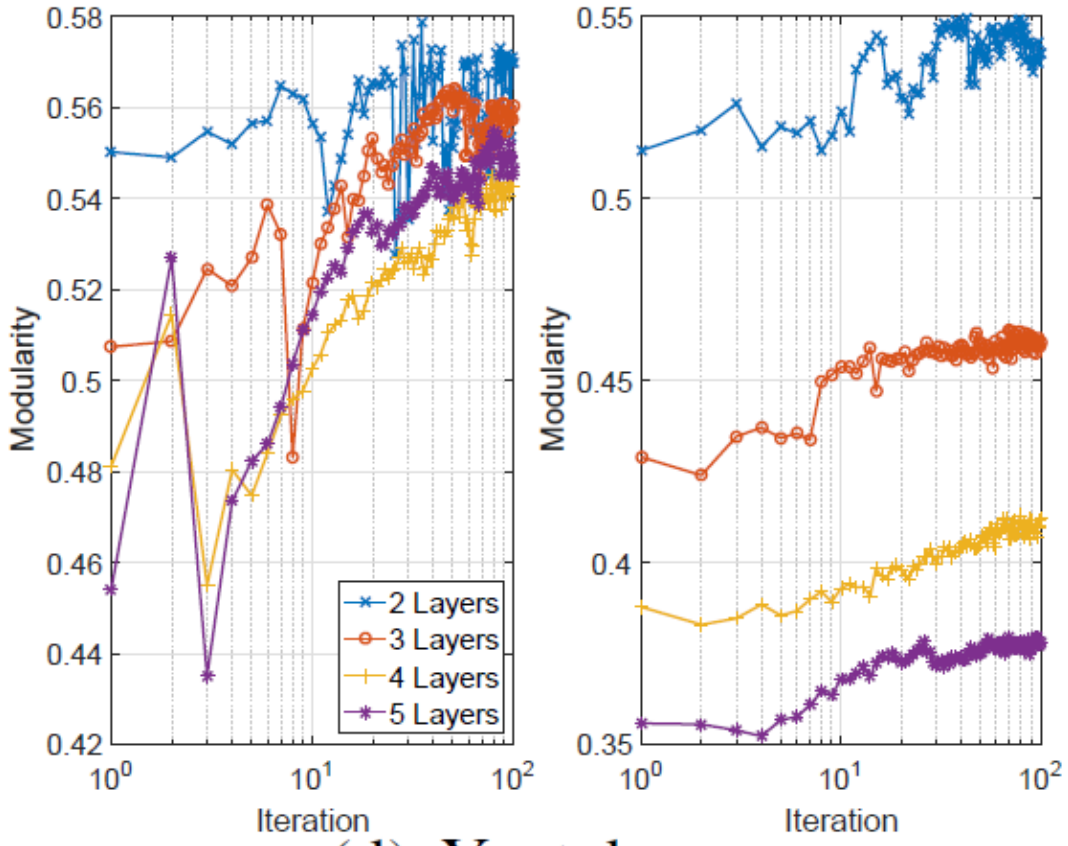
Improvement on Modularity



Improvement on Modularity

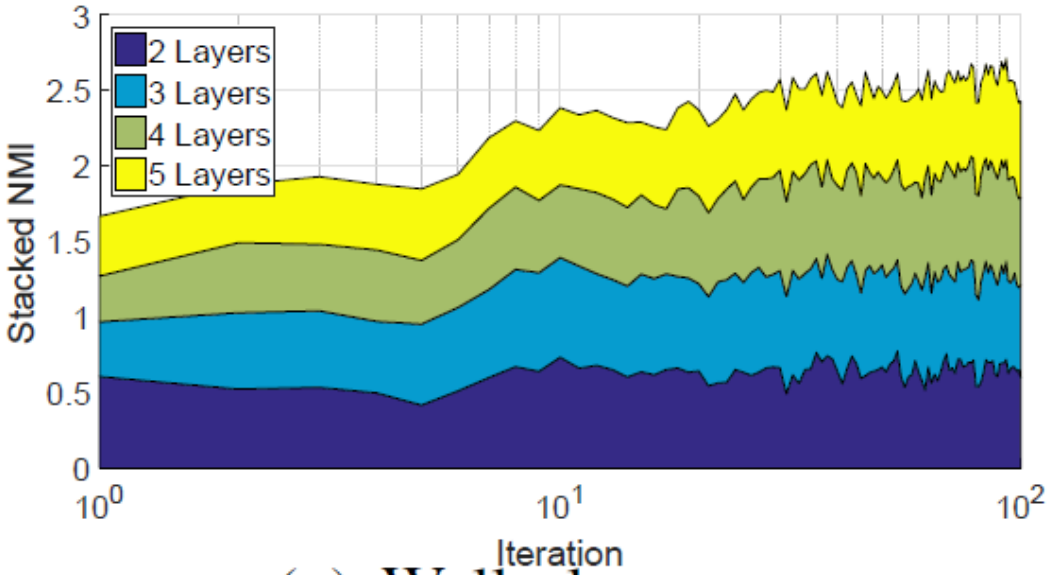


(c) UIllinois

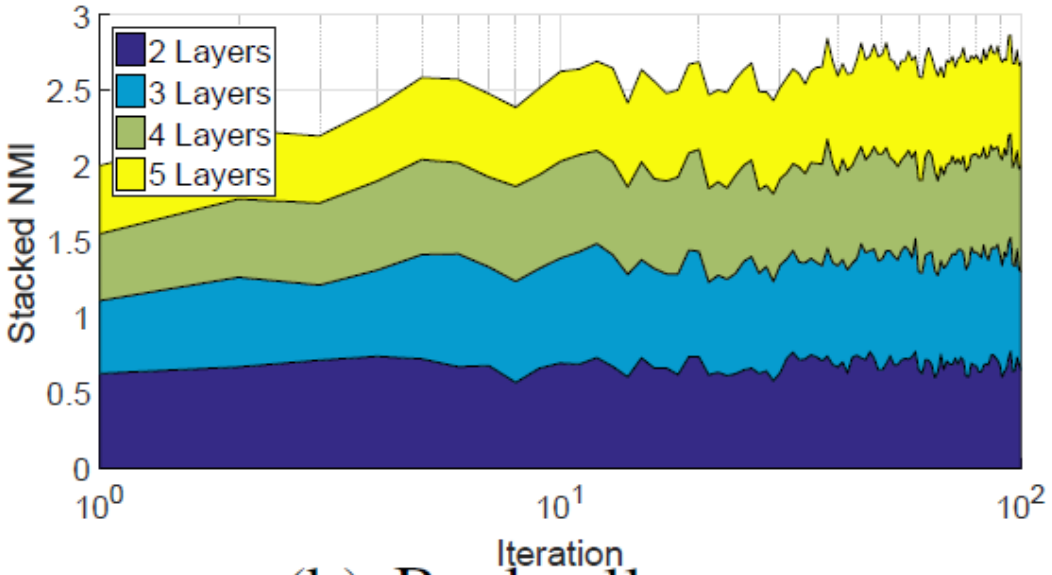


(d) Youtube

Convergence

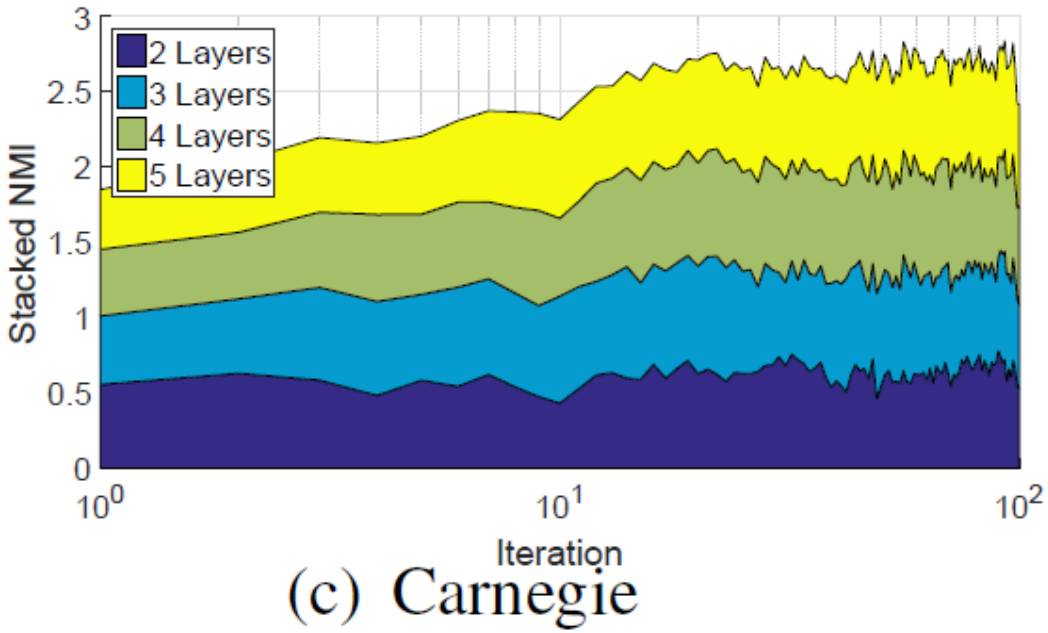


(a) Wellesley

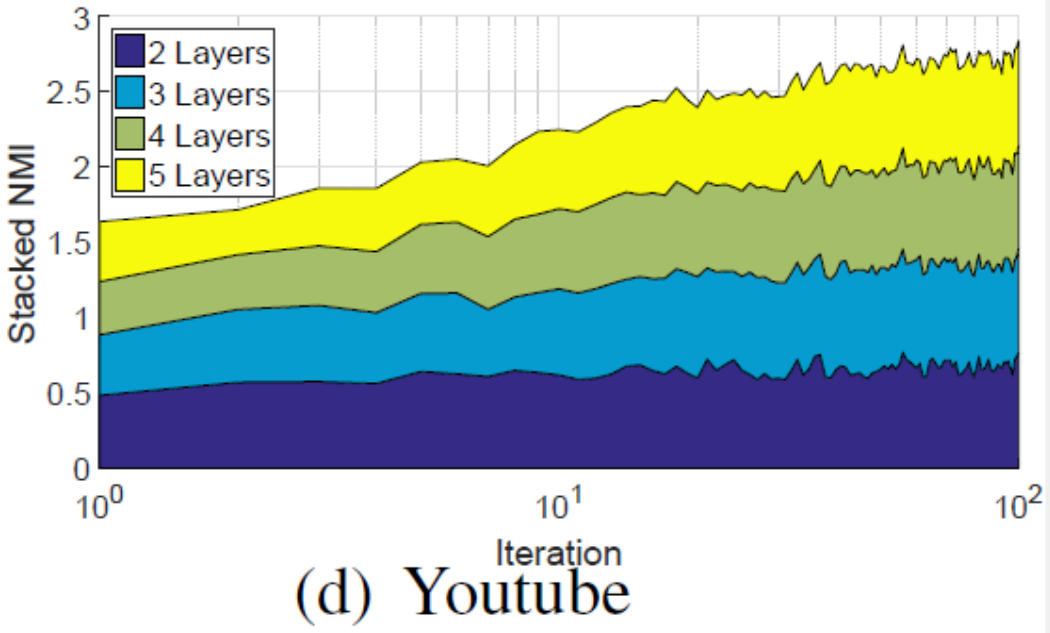


(b) Bucknell

Convergence



(c) Carnegie



(d) Youtube

Evaluation methods

- Individual Jaccard Precision
- Jaccard Precision P
- Individual Jaccard Recall
- Jaccard Recall R

$$P(D_i) = \max_{G_j \in \mathcal{G}} \frac{|G_j \cap D_i|}{|G_j \cup D_i|}$$

$$R(G_j) = \max_{D_i \in \mathcal{D}} \frac{|G_j \cap D_i|}{|G_j \cup D_i|}$$

- Jaccard F_1 Score F_J as the harmonic mean of P and R: $2PR/(P+R)$

Dataset Evaluation

Datasets	Annotated communities				Communities found by HC:MOD			
	Annotations (modularity, hiddenness value)	NMI_{max}	F_{max}	F_{avg}	#Layers	Modularity	NMI_{max}	F_{max}	F_{avg}
Synthetic									
SynL2	PL_1(0.49, 0.50) PL_2(0.49, 0.50)	0.00	0.04	0.04	2	0.49, 0.49	0.00	0.04	0.04
SynL3	PL_1(0.33, 0.59) PL_2(0.32, 0.70) PL_3(0.32, 0.71)	0.00	0.04	0.03	3	0.33, 0.32, 0.32	0.00	0.04	0.03
Facebook									
Caltech	Dorm(0.30, 0.08) Year(0.19, 0.84) Status(0.08, 0.85)	0.13	0.32	0.18	2	0.40, 0.25	0.00	0.18	0.18
Smith	Dorm(0.23, 0.42) Year(0.23, 0.49)	0.00	0.04	0.04	2	0.51, 0.34	0.03	0.11	0.11
Rice	Dorm(0.37, 0.02) Year(0.23, 0.94) Status(0.13, 0.91)	0.11	0.26	0.14	3	0.50, 0.36, 0.34	0.02	0.20	0.13
Vassar	Year(0.34, 0.06) Dorm(0.15, 0.98) Status(0.13, 0.89)	0.18	0.30	0.18	3	0.45, 0.32, 0.31	0.04	0.21	0.15
Bucknell	Year(0.40, 0.05) Status(0.12, 0.91) Dorm(0.11, 0.98)	0.15	0.31	0.17	3	0.51, 0.37, 0.31	0.05	0.30	0.22
Carnegie	Year(0.28, 0.29) Major(0.12, 0.84) Status(0.11, 0.78) Dorm(0.08, 0.92)	0.07	0.24	0.09	4	0.40, 0.42, 0.34, 0.32	0.06	0.23	0.17
Wellesley	Year(0.30, 0.10) Status(0.15, 0.79)	0.15	0.28	0.28	2	0.37, 0.26	0.00	0.15	0.15
UIllinois	Year(0.27, 0.24) High_school(0.15, 0.82) Dorm(0.14, 0.76)	0.00	0.07	0.03	2	0.45, 0.34	0.04	0.16	0.14

Results Evaluation

SynL3		HC:MOD			HC:IM			HC:OS			HC:LC			Partitioning		Overlapping	
		L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3	Mod	IM	OS	LC
PL_1	F_J	**0.92**	0.04	0.03	**0.97**	0.03	0.03	0.05	0.04	**0.80**	**0.58**	0.06	0.07	0.26	0.83	0.51	0.36
HV = 0.59	NMI	**0.93**	0.00	0.00	**0.94**	0.00	0.00	0.01	0.00	**0.82**	**0.43**	0.00	0.00	0.14	0.72	0.50	0.19
PL_2	F_J	0.04	**0.97**	0.04	0.03	**0.97**	0.03	**0.89**	0.03	0.04	0.05	**0.30**	0.10	0.05	0.05	0.32	0.09
HV = 0.70	NMI	0.00	**0.96**	0.00	0.00	**0.93**	0.00	**0.87**	0.00	0.00	0.00	**0.11**	0.00	0.00	0.00	0.29	0.00
PL_3	F_J	0.03	0.04	**0.97**	0.03	0.03	**0.97**	0.04	**0.90**	0.04	0.05	0.07	**0.17**	0.05	0.03	0.14	0.04
HV = 0.71	NMI	0.00	0.00	**0.95**	0.00	0.00	**0.85**	0.00	**0.87**	0.00	0.00	0.00	**0.02**	0.00	0.00	0.06	0.00

TABLE IV

R, P, F AND NMI SCORES WHEN EVALUATED ON THE SYNL3 COMMUNITY LAYERS.

Caltech		HC:MOD		HC:IM			HC:OS		HC:LC		Partitioning		Overlapping	
		L_1	L_2	L_1	L_2	L_3	L_1	L_2	L_1	L_2	Mod	IM	OS	LC
Dorm	F_J	**0.58**	0.11	**0.65**	0.11	0.11	**0.48**	0.11	**0.21**	0.10	0.51	0.51	0.49	0.18
HV = 0.08	NMI	**0.39**	0.00	**0.48**	0.01	0.01	**0.32**	0.01	**0.16**	0.02	0.36	0.42	0.28	0.14
Year	F_J	0.11	**0.60**	0.12	**0.40**	0.20	0.14	**0.45**	0.07	**0.15**	0.13	0.14	0.12	0.07
HV = 0.84	NMI	0.00	**0.38**	0.03	**0.19**	0.09	0.00	**0.29**	0.02	**0.10**	0.05	0.13	0.00	0.03
Status	F_J	0.17	**0.37**	0.12	0.38	**0.64**	0.02	**0.36**	0.23	**0.51**	0.16	0.16	0.12	0.03
HV = 0.85	NMI	**0.16**	**0.11**	0.15	0.14	**0.32**	0.00	**0.22**	0.12	**0.25**	0.06	0.30	0.19	0.13

TABLE V

JACCARD F_1 AND NMI SCORES OF ALL ALGORITHMS ON CALTECH COMMUNITY CATEGORIES.

Smith		HC:MOD		HC:IM		HC:OS		HC:LC		Partitioning		Overlapping	
		L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2	Mod	IM	OS	LC
Dorm	F_J	**0.45**	0.04	**0.50**	0.04	**0.40**	0.07	0.05	0.01	0.25	0.43	0.38	0.04
HV = 0.42	NMI	**0.26**	0.00	**0.36**	0.00	**0.25**	0.01	0.01	0.00	0.14	0.31	0.23	0.00
Year	F_J	0.12	**0.56**	0.10	**0.35**	0.15	**0.24**	0.01	**0.20**	0.21	0.18	0.16	0.18
HV = 0.49	NMI	0.00	**0.37**	0.03	**0.16**	0.05	**0.11**	0.00	**0.03**	0.06	0.06	0.06	0.00

TABLE VI

JACCARD F_1 AND NMI SCORES OF ALL ALGORITHMS ON SMITH COMMUNITY CATEGORIES.

Results Evaluation

		HC:MOD			HC:IM			HC:OS			HC:LC			Partitioning		Overlapping	
Rice		L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3	Mod	IM	OS	LC
Dorm	F_J	**0.79**	0.11	0.07	**0.74**	0.08	0.08	**0.61**	0.22	0.22	**0.20**	0.10	0.07	0.70	0.55	0.54	0.10
HV = 0.02	NMI	**0.71**	0.00	0.00	**0.45**	0.00	0.00	**0.50**	0.10	0.11	**0.10**	0.01	0.00	0.59	0.32	0.29	0.04
Year	F_J	0.08	0.22	**0.55**	0.08	0.20	**0.34**	0.09	**0.22**	0.20	0.07	0.14	**0.15**	0.07	0.08	0.14	0.05
HV = 0.94	NMI	0.00	0.07	**0.29**	0.00	0.06	**0.16**	0.00	**0.13**	0.10	0.01	0.01	**0.07**	0.05	0.05	0.00	0.00
Status	F_J	0.11	**0.42**	0.23	0.10	**0.61**	0.23	0.13	0.32	**0.58**	0.05	**0.61**	0.12	0.10	0.08	0.14	0.03
HV = 0.91	NMI	0.00	**0.20**	0.11	0.01	**0.25**	0.09	0.01	0.16	**0.42**	0.00	**0.05**	0.02	0.04	0.01	0.00	0.01

TABLE VII

JACCARD F_1 AND NMI SCORES OF ALL ALGORITHMS ON RICE COMMUNITY CATEGORIES.

		HC:MOD			HC:IM			HC:OS			HC:LC		Partitioning		Overlapping	
Vassar		L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	Mod	IM	OS	LC
Year	F_J	**0.67**	0.16	0.10	**0.54**	0.17	0.15	0.14	**0.44**	0.22	**0.17**	0.08	0.68	0.47	0.37	0.16
HV = 0.06	NMI	**0.47**	0.01	0.00	**0.31**	0.07	0.02	0.06	**0.19**	0.06	**0.04**	0.02	0.38	0.27	0.14	0.04
Dorm	F_J	0.13	**0.41**	0.08	0.11	0.09	**0.25**	0.11	**0.15**	0.11	0.07	**0.08**	0.12	0.12	0.16	0.08
HV = 0.98	NMI	0.03	**0.24**	0.02	0.02	0.01	**0.08**	0.01	**0.07**	0.02	0.02	**0.03**	0.00	0.02	0.00	0.00
Status	F_J	**0.34**	0.23	0.12	0.34	**0.57**	0.19	0.52	0.23	**0.61**	**0.63**	0.27	0.33	0.33	0.23	0.61
HV = 0.89	NMI	**0.15**	0.06	0.01	0.10	**0.21**	0.06	0.20	0.09	**0.27**	**0.04**	**0.08**	0.15	0.10	0.07	0.04

TABLE VIII

R, P, F AND NMI SCORES OF ALL ALGORITHMS ON VASSAR COMMUNITY CATEGORIES.

Open questions

- Theoretical justification on why weakening the dominant layer can help uncovering the hidden layer?
  - How many layers does it guaranties to recover.
- Does weakening process hurt the current structure?
  - If does, to what extent? If not, why? Need Theory.
  - Is there a upper or lower bound of the number of layers that would hurt structure in a comfortable extent?
- How hard is the hidden layers to be detected?
- To what extent does the hidden structure interfere with the accurate detection of the dominant structure?

Open questions

- Convergence Guarantees of the Algorithm?
- Another methods on Selecting Numbers of Layers?
- Base Algorithm Ensembling?
- Hidden Community Generating Model

- We hope this work creates a new line of research in the area of community detection!

Thank you!

Questions?