Uncovering the Small Community Structure in Large Networks: A Local Spectral Approach

Yixuan Li [1], Kun He [2], David Bindel [1] and John E. Hopcroft [1]

[1]Cornell University, USA

[2]Huazhong University of Science and Technology, China

May 20th, 2015

Uncover Small Community Structure

How can we find a community of hundreds in a network of billions?

Figure : Visualization of Facebook friendship data[1].

[1] www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

Motivation

Global Structure → Local Structure

- Reduce Complexity

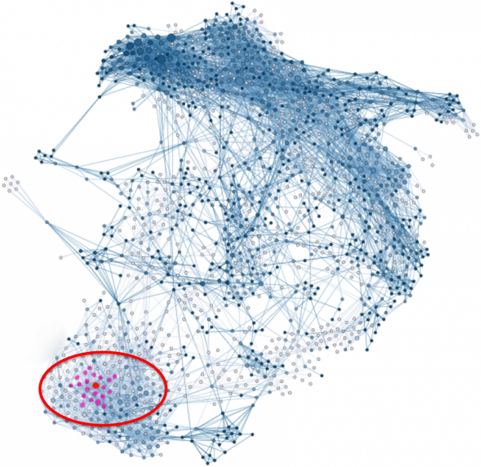
Enables finding communities in time functional to the size of the community ($\sim$100) rather than the size of the entire graph ($\sim$ billions)

- Improve Accuracy

Enables finding out how many communities is an individual in, and what are these communities.

Local expansion

- What is a **seed set** [2][3][4]?
- What is local expansion useful for?
- How can we conduct local expansion? (short-step random walks)



[2] K. Kloster and D. F. Gleich. Heat kernel based community detection. In KDD'14.

[3] I. M. Kloumann and J. M. Kleinberg. Community membership identification from small seed sets. In KDD'14.

[4] J. J. Whang, D. F. Gleich, and I. S. Dhillon. Overlapping community detection using seed set expansion. In CIKM'13.

Datasets

Synthetic datasets: LFR benchmark graphs[5]

- Built-in community structure, power-law distribution.
- *om*: overlapping memership. ($2 \sim 8$)
- μ (mixing parameter): controls the fraction of links for each vertex to connect outside. ($\mu = 0.1$, $\mu = 0.3$)

Real datasets[6]

amazon	dblp computer science bibliography	orkut	YouTube
Product Networks	**Collaboration Networks**	**Social Networks**	**Social Networks**

[5] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. Physical Review E, 78(4):046110, 2008.

[6] Stanford Network Analysis Project: `http://snap.stanford.edu`

Existing Spectral Methods

- Consider an undirected graph $G = (V, E)$ with n nodes and m edges.
- $\mathbf{A}$ is the adjacency matrix; $\mathbf{N} = \mathbf{D}^{-1}\mathbf{A}$ is the transition matrix where $\mathbf{D}$ is the diagonal matrix of node degrees.
- Find out the dominant eigenvectors of $\mathbf{N}$.

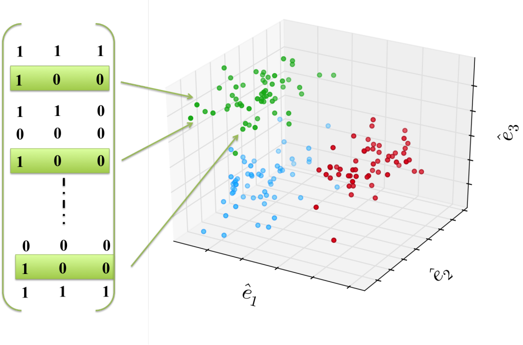


Figure : Spectral clustering methods

Existing Spectral Methods

Drawbacks:

- Inefficient computation of eigenvectors
- Unable to partition overlapping communities

Local Spectral Clustering

Find *local spectra*:

- Random walks starting from a seed set $\mathcal{S}$.
- Consider the span of ℓ-dimensional probability vectors $\mathbf{P}_{0,\ell} = [\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_\ell]$.
- Initial invariant subspace: $\mathbf{V}_{0,\ell}$.
- Use the following recurrence to calculate the local spectra $\mathbf{V}_{k,\ell}$ after k steps of random walk

$$\mathbf{V}_{k,\ell}\mathbf{R}_{k,\ell} = \mathbf{V}_{k-1,\ell}\bar{\mathbf{A}}, \tag{1}$$

where $\bar{\mathbf{A}} = \mathbf{D}^{-1/2}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-1/2}$ is the normalized adjacency matrix of the graph.

Local Spectral Clustering

To find out the members in the same community as the seed set $\mathcal{S}$ belong to, is equivalent to find rows in $\mathbf{V}_{k,\ell}$ that point in nearly the same direction as nodes in $\mathcal{S}$.

To seek a sparse vector $\mathbf{y}$ in the span of $\mathbf{V}_{k,\ell}$ such that seed nodes are in its support.

$$\begin{aligned} \min \quad & \mathbf{e}^T\mathbf{y} = ||\mathbf{y}||_1 \\ s.t. \quad & \mathbf{y} = \mathbf{V}_{k,\ell}\mathbf{x}, \\ & \mathbf{y} \geq \mathbf{0}, \\ & \mathbf{y}(\mathcal{S}) \geq 1 \end{aligned}$$

Community Size Determination

We can obtain a “candidate community” by truncating sorted sparse vector $\hat{\mathbf{y}}$. But we still need to pin down to answer the following question:

Q.1

What defines “good” communities and when do they emerge as we expand the seed set?

Community Size Determination

The random walk techniques produce communities with conductance guarantees[7].

$$\psi(V) = \frac{|\partial(V)|}{\min(Vol(V), Vol(\bar{V}))}, \tag{2}$$

where $|\partial(V)|$ denote the cut size, and Vol(V) is the sum of node degree in set V.

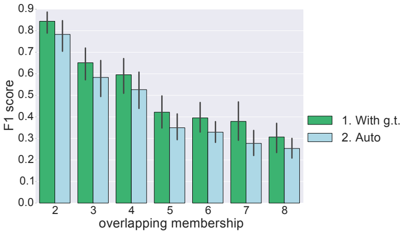
A.1

The expansion of seed set can stop and form a natural community when it encounters a low-conductance cut.

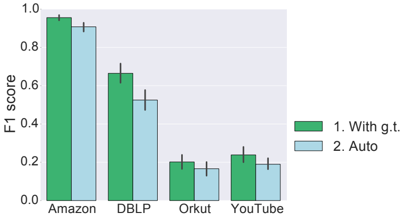
[1] Andersen, Reid, and Kevin J. Lang. "Communities from seed sets." Proceedings of the 15th international conference on World Wide Web. ACM, 2006.

Community Size Determination

Comparison of the average F1 score with ground truth and automatic size determination:



LFR benchmark graphs ($\mu = 0.3$)



Real datasets

Complexity Reduction by Sampling

If one wants to uncover a small community within a network with billions of vertices, it would be very costly to take the whole graph into account.

Q.2

How to find a small community in time functional to the size of the community rather than that of the entire graph?

Complexity Reduction by Sampling

In practice, the unknown members in the target community are more likely to be around the seed members, and are usually a few steps away from the seeds.

A.2

Sample the graph by cutting off the redundant nodes with low probability being reached after short random walk.

Dataset	Coverage ratio	Sample rate	$\lvert\mathcal{C}\rvert_{\text{avg}}$	Subgraph size
Amazon	1.00	0.0087	39	2913
DBLP	0.98	0.0076	251	2409
YouTube	0.66	0.0033	79	3745
Orkut	0.64	0.0011	83	3379

Table : Statistics of the mean values for the sampling method on real datasets.

Seeding Method

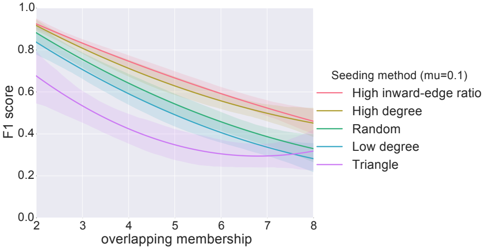
The quality of seed set is crucial to the detection accuracy. The alternative seeding methods can be strategically applied by domain experts in different scenarios based on the availability of candidate seeds.

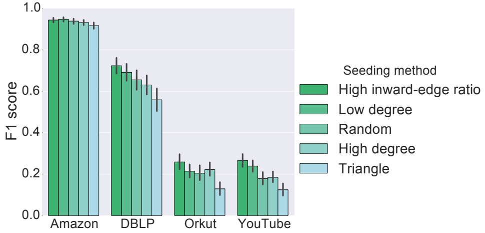
Q.3: What defines a “good” seed set and how many seeds are needed in order to define a community?

Seeding Method

Adopt $|\mathcal{S}| = 3$ seeds for each of the seeding method:

- High degree seeding
- Low degree seeding
- Random seeding
- Triangle seeding
- High inward-edge ratio seeding





Seed Set Size

For LFR benchmark graphs, we test with five different seeding ratios: 2%, 4%, 6%, 8% and 10%. For real networks, varying seed set size has little effect on the performance.

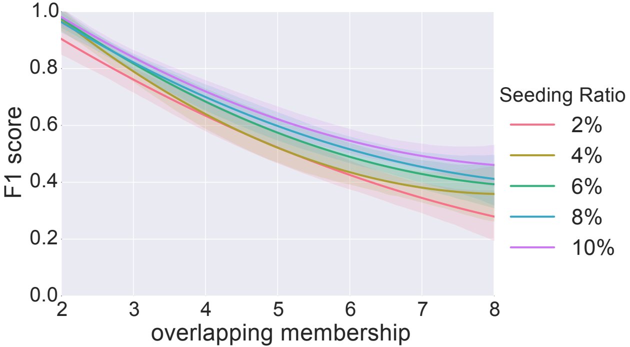


Figure : LFR benchmark graphs with μ = 0.1.

Comparison with localized algorithms

- Heat Kernel (Kloster et. al, KDD'14)
- PageRank (Kloumann et. al, KDD'14)
- Seed Set Expansion (Whang et. al, CIKM'13)
- Local Expansion via Minimum One Norm

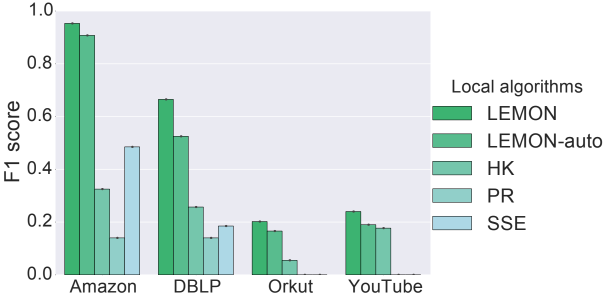


Figure : Comparison of the average F1 score with state-of-the-art local detection algorithms.

Discussion

Networks are not all similar and we cannot assume one algorithm works for finding communities in a network will behave the same on the other networks.

Q.4

How the local expansion approach is suited for uncovering communities in different types of networks?

Discussion

Empirical comparison between synthetic and real networks:

A.4

- *LEMON* is less sensitive to the random walk step and subspace dimension on real networks than that on LFR benchmark graphs.
- *LEMON* is less sensitive to the seed set size on real networks than that on LFR benchmark graphs.
- *LEMON* is more sensitive to the high-degree seeds on real networks than that on LFR benchmark graphs.

Future Work

Hierarchical Structure

Look further into some larger low-conductance communities and see if a hierarchical structure exists. In this case, some large social group consisting of several small cliques is likely to be discovered.

Membership Detection

The local spectral clustering method could be potentially applied to the membership detection problem, i.e., finding all the communities that an arbitrary vertex belongs to.

Q & A

Evaluation Metric

Use F1 score to quantify the similarity between the algorithmic community $\mathcal{C}$ and the ground truth community $\mathcal{C}^*$:

$$F_1(\mathcal{C}, \mathcal{C}^*) = \frac{2 \cdot \mathit{Precision}(\mathcal{C}, \mathcal{C}^*) \cdot \mathit{Recall}(\mathcal{C}, \mathcal{C}^*)}{\mathit{Precision}(\mathcal{C}, \mathcal{C}^*) + \mathit{Recall}(\mathcal{C}, \mathcal{C}^*)}, \tag{3}$$

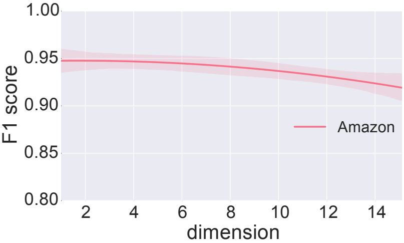
where the precision and recall are defined as:

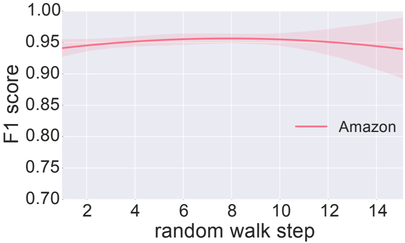
$$\mathit{Precision}(\mathcal{C}, \mathcal{C}^*) = \frac{|\mathcal{C} \cap \mathcal{C}^*|}{|\mathcal{C}|}, \tag{4}$$

$$\mathit{Recall}(\mathcal{C}, \mathcal{C}^*) = \frac{|\mathcal{C} \cap \mathcal{C}^*|}{|\mathcal{C}^*|}. \tag{5}$$

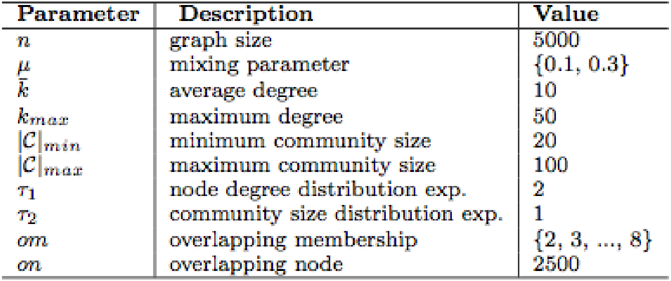
Parameter Selection

- Fix $k = 3$, vary ℓ from 1 to 15
- Fix $\ell = 3$, vary k from 1 to 15
- The observation holds for the remaining datasets as well.





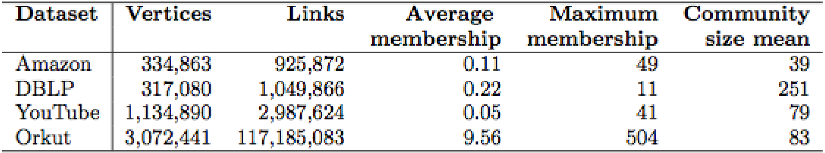
Parameters of LFR graphs



Parameter	Description	Value
n	graph size	5000
μ	mixing parameter	$\{0.1, 0.3\}$
$\bar{k}$	average degree	10
k_{max}	maximum degree	50
$\|C\|_{min}$	minimum community size	20
$\|C\|_{max}$	maximum community size	100
τ_1	node degree distribution exp.	2
τ_2	community size distribution exp.	1
om	overlapping membership	$\{2, 3, ..., 8\}$
on	overlapping node	2500

Figure : Parameters for the LFR datasets.

Stats of Real Datasets



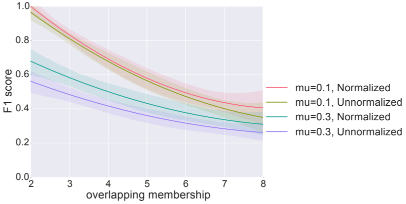
Dataset	Vertices	Links	Average membership	Maximum membership	Community size mean
Amazon	334,863	925,872	0.11	49	39
DBLP	317,080	1,049,866	0.22	11	251
YouTube	1,134,890	2,987,624	0.05	41	79
Orkut	3,072,441	117,185,083	9.56	504	83

Figure : Statistics for the real networks.

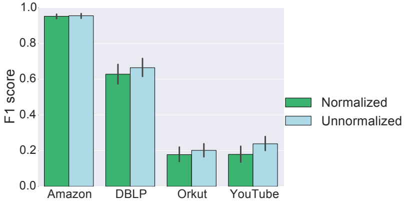
Further Extension

We enforce a bias towards the high-degree vertices at the beginning of the random walk by normalizing the initial probability vector:

$$p_0(v_i) = \begin{cases} d(v_i)/\mathrm{Vol}(\mathcal{S}) & \text{if } \; v_i \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$



LFR benchmark graphs



Real datasets