Local Spectral Diffusion for Robust Community Detection

Kun He, Pan Shi[*]
Huazhong University of
Science and Technology,
Wuhan 430074, China
brooklet60@hust.edu.cn,
panshi@hust.edu.cn

John E. Hopcroft
Computer Science
Department
Cornell University
Ithaca, NY 14853, USA
jeh@cs.cornell.edu

David Bindel
Computer Science
Department
Cornell University,
Ithaca, NY 14853, USA
bindel@cs.cornell.edu

ABSTRACT

We address a semi-supervised learning problem of identifying all latent members of a local community from very few labeled seed members in large networks. By a simple and efficient sampling method, we conduct a comparatively small subgraph encompassing most of the latent members such that the follow-up membership identification could focus on an accurate local region instead of the whole network. Then we look for a sparse vector, a relaxed indicator vector representing the subordinative probability of the corresponding nodes, that lies in a local spectral subspace defined by an order-d Krylov subspace. The subspace serves as a local proxy for the invariant subspace spanned by leading eigenvectors of the Laplacian matrices. Based on Rayleigh quotients, we relate the local membership identification task as a local RatioCut or local normalized cut optimization problem, and provide some theoretical justifications.

We thoroughly explore different probability diffusion methods for the subspace definition and evaluate our method on four groups with a total of 28 representative LFR benchmark datasets, and eight publicly available real-world networks with labeled ground truth communities across multiple domains. Experimental results exhibit the effectiveness and robustness of the proposed algorithm, and the local spectral communities perform better than those from the celebrated Heat Kernel diffusion [10] and the PageRank diffusion [1].

CCS Concepts

•Mathematics of computing → Graph algorithms; •Information systems → Clustering; •Human-centered computing → Social networks; •Computing methodologies → Spectral methods;

Keywords

Clustering; social networks; local spectral; Krylov subspace; Rayleigh quotient

1. INTRODUCTION

Communities are regarded as densely linked components with sparser external connections [21]. There are many situations in which we are interested to find the social groups or communities for a small set of observed members. Intuitively, the latent members are very close to the exemplary members as evaluated by the shortest paths, especially when the community is small with dozens or a hundred of members, which is the common scale in real-world networks [14]. In other words, they are in the local region around the seeding nodes.

Many seed set expansion methods are proposed for finding the "local community" in the local region around the seeds [2, 11, 10]. A main stream for the seed set expansion is to do probability diffusion from the seeds. PageRank [1, 11], Heat Kernel [10, 5] and local spectral [3, 7, 15] are three main techniques for the probability diffusion. Among which, the local spectral method is a newly proposed technique that exhibits high performance for the local community detection task. Classical spectral clustering methods extract disjoint communities from the leading eigenvectors of the graph Laplacian matrix. Motivated by the classical spectral method, local spectral algorithms [7, 15] seek a sparse indicator vector containing the seeds and lies in the local spectral subspace. By starting from different seeds they detect overlapping communities. Though very successful experimentally, there is still very little understanding of the principles, and unlike PageRank methods that have attracted much attention in the literature, the local spectral methods are not yet fully explored.

We make progress on the LOSP (Local Spectral) algorithm [7], and propose a LOSP++ algorithm. LOSP++ improves the local sampling method of LOSP, defines a simplified local spectral subspace basing on the Krylov subspace, and finds the best probability diffusion way by exploring several random walk diffusion methods with different diffusion degrees. LOSP++ exhibits high accuracy even after removing the pre-processing of strengthening the seeds and the post-processing of reseeding iteration adapted by LOSP. Our main contributions include:

- We focus on the core of the local spectral methods, systematically define and thoroughly investigate variants of the spectral diffusion methods and simplify the

[*] Corresponding Author.

SIGKDD '16 Workshop San Francisco, CA, USA
© 2016 ACM. ISBN 123-4567-24-567/08/06. . . $15.00
DOI: 10.475/123_4

subspace generation. For the four spectral diffusion methods, we show that light lazy random walk and lazy random walk are very stable for different parameters, and outperform the standard and personalized PageRank diffusions.

- Based on the Rayleigh quotients related to the Laplacian matrices, we provide a theoretical analysis for the soundness of the local spectral method.
- LOSP++ shows the robustness on parameter selection for subspace dimension and random walks steps.

In the end, we provide a diverse set of computational evidence on 28 synthetic benchmark graphs and eight real world social and Biological networks that LOSP++ yields much higher accuracy as compared with the well-known personalized PageRank diffusion algorithm `pprpush` [1] and the venerable Heat Kernel algorithm `hk-relax` [10].

2. RELATED WORK

A considerable amount of literature has been published on finding local communities in large networks [2, 11, 10, 16]. It is natural to apply the seed set expansion method initially designed for the global community structure detection to uncover the local community structure from a few observed seed members.

Local seed set expansion. The random walk technique has been extensively adopted as a subroutine for locally expanding the seed set [2]. PageRank [1, 22, 11], Heat Kernel [10, 4, 5] and local spectral [3, 16, 15, 7] are three main techniques for probability diffusion.

Spielman and Teng[18] use degree-normalized, personalized PageRank (DN PageRank) with respect to the start seed and do truncation on small values, leading to the PageRank Nibble method[1]. And the DN PageRank is adopted by several PageRank-based clustering algorithms [2, 22], which are competitive with a sophisticated and popular algorithm METIS [9]. Kloumann and Kleinberg [11] evaluate different variations of PageRank, and find that the standard PageRank yields better performance than the DN PageRank.

The Heat Kernel provides another local graph diffusion [4, 5, 10], and involves the Taylor series expansion of the matrix exponential of the random walk transition matrix. Chung et al. analyze the property of this diffusion theoretically [4], and propose a randomized Monte Carlo method to estimate the diffusion [5]. Kloster et al. propose a deterministic method that uses coordinate relaxation on an implicit linear system that estimates the Heat Kernel diffusion, and show that Heat Kernel outperforms the personalized PageRank by finding smaller sets with substantially higher F1 measures [10].

Spectral methods have been popularly used to extract disjoint communities from a few leading eigenvectors of the graph Laplacian related matrix [19, 8]. Recently, there has been a growing interest in adapting the spectral method to mine the local structure around the seed set [3, 16]. Machael et al. [16] introduce a locally-biased analogue of the second eigenvector for extracting local properties of data graphs near an input seed set, and apply their method for a semi-supervised image segmentation and a local community extraction by finding a sparse-cut around the seed set on a small social network.

He et al. [7] and Li et al. [15] extract the local community by seeking a sparse vector from the local approximate spectral subspaces using ℓ_1 norm optimization, and propose LOSP and LEMON respectively. They apply a power method for the subspace iteration using standard random walk on a modified graph with a self loop on each node, which we call the light lazy random walk. To get relatively small subgraph containing nodes around the seeds of interest, LOSP samples via BFS and LEMON samples via random walk. LOSP [7] strengthens the initial seeds by adding nodes along a shortest path for each pair of seeds if the path length is no greater than a small value like 3. Both LEMON and LOSP apply a reseeding iteration to improve the detection accuracy.

The LOSP++ we proposed is similar in spirit to LOSP. LOSP++ adopts a more effective local sampling method, and extracts the local community structure from a Krylov subspace which is much more efficient for the subspace calculation. We thoroughly evaluate different local spectral diffusion methods, do the parameter study on the diffusion degrees, and provide theoretical analysis for the local spectral method basing on the Rayleigh quotients and quadratic forms. We also remove the pre-processing procedure of strengthening the initial seeds and the post-processing procedure of the reseeding iterations. These slimming strategies makes LOSP++ at least three times quicker, as LOSP usually takes at least several rounds of reseeding. Nevertheless, on the five SNAP real datasets that LOSP reported its results, we obtain considerably higher accuracy.

Metrics for bounding the community. All seed set expansion methods need a stopping criterion for defining the community boundary unless the size of the target community is known as a budget. Conductance is commonly recognized as the best stopping criterion due to its intrinsic local properties [10, 20, 22]. Yang and Leskovec provide widely-used real world datasets with labeled ground truth[22], and find that conductance and triad-partition-ratio (TPR) are the two stopping rules yielding the highest detection accuracy. The Heat Kernel method also adopts conductance as the stopping rule for the local community [10]. He et al. [7] propose two new metrics, TPN and nMod, and compare them with conductance, modularity and TPR. They find conductance and TPN consistently outperforms other metrics including modularity, TPR and nMod.

Seeding strategies. The seeding strategy is a key component for seed set expansion algorithms. Local community detection tasks provide some known members as the prior for the semi-supervised learning. Kloumann and Kleinberg[11] compare random seeds with high degree seeds, and discover that random seeds are superior to high degree seeds, and they suggest domain experts provide seeds with a diverse degree distribution. He et al. [7] compare low degree, random, high triangle participation (number of triangles the seed involved inside the community) and low escape seeds (judged by probability reserved after short random walks), and find all four types of seeds yield almost the same accuracy. They observe that low degree seeds spread out the probabilities slowly and better preserve the local information, and random seeds are similar to low degree seeds due to the power law distribution of the node degrees. High triangle participation seeds and low escape seeds follow another philosophy in that they choose seeds more cohesive to the target community.

3. PRELIMINARIES

3.1 Problem formulation

Consider a connected, undirected graph $G = (V, E)$ with $|V| = n$ nodes and $|E| = m$ edges. Let $\mathbf{A} \in \{0,1\}^{n \times n}$ be the associated adjacency matrix whose $ij-$entry is a_{ij}, $\mathbf{I}$ the identity matrix, and $\mathbf{e}$ the vector of all ones. Let $\mathbf{d} = \mathbf{Ae}$ be the vector of node degrees, and $\mathbf{D} = diag(\mathbf{d})$ the diagonal matrix of node degrees. Let $\mathbf{s} \in \{0,1\}^n$ be a binary indicator vector representing the exemplary members $S \subset V$.

We formalize the local community finding task as a semi-supervised learning problem. Let $T = (V_t, E_t)$ $(S \subset V_t \subset V$, $|V_t| \ll |V|)$ be the labeled ground truth community. We want to identify the remaining latent members in the target community T. The identification accuracy is evaluated by the F_1 score for the detected community $C = (V_c, E_c)$.

$$F_1(C,T) = \frac{2|V_c \cap V_t|}{|V_c| + |V_t|}.$$

which is the harmonic mean of precision and recall. See [7] for details of the definition. Then we could formalize this semi-supervised learning problem as an optimization problem:

$$\begin{aligned} max \quad & F_1(C,T) \\ s.t. \quad & (1) \quad S \subset V_c \subset V \\ & (2) \quad C = (V_c, E_c) \text{ is connected} \end{aligned}$$

Whether C is connected could be judged by its algebraic connectivity, the second smallest eigenvalue of its Laplacian matrix $\mathbf{L_c} = \mathbf{D_c} - \mathbf{A_c}$.

Further, when the size of the target community is also known as a budget, which is reasonable as we may want to extract the target community of a given scale for the seed members, we could formalize the problem as follows.

$$\begin{aligned} max \quad & F_1(C,T) \\ s.t. \quad & (1) \quad S \subset V_c \subset V \\ & (2) \quad |V_c| = |V_t| \\ & (3) \quad C = (V_c, E_c) \text{ is connected} \end{aligned}$$

3.2 Datasets

To thoroughly evaluate the performance of the proposed algorithm, we consider four groups with a total of 28 synthetic datasets, 5 SNAP datasets in social, product, or collaboration domains, and 3 biology networks for a comprehensive evaluation on the proposed local spectral algorithm.

3.2.1 LFR Benchmark Graphs

Lancichinetti, Fortunato, and Radicchi [13, 12] proposed an algorithm for generating LFR[1] benchmark graphs with a built-in community structure, and simulate properties of real networks accounting for heterogeneity of node degree and community size distributions. In a recent survey paper, Xie et al. [21] performed a thorough performance comparison of different global overlapping community detection algorithms on LFR benchmark datasets.

We adopt the same set of parameter settings used in the survey paper for evaluating different community detection algorithms [21] and generate four groups with a total of 28 LFR benchmark graphs. Table 1 summarizes the parameter settings we used for the LFR datasets. Among which the mixing parameter μ has a big impact on the network topology, and it controls the average fraction of neighboring nodes that do not belong to any community for each node. μ is usually set to be 0.1 or 0.3 and the detection accuracy usually decays for a larger μ. b and s provide two ranges of typical community sizes, big and small. Each node belongs to either one community or *om* overlapping communities, and the number of nodes in overlapping communities is specified by *on*. A larger *om* or *on* indicates more overlaps that are harder for the community detection tasks.

Parameter	**Description**
$n = 5000$	number of nodes in the graph
$\mu = 0.3$	mixing parameter
$\bar{d} = 10$	average degree of the nodes
$d_{max} = 50$	maximum degree of the nodes
$s : [10, 50], b : [20, 100]$	range of the community size
$\tau_1 = 2$	node degree distribution exponent
$\tau_2 = 1$	community size distribution exponent
$om \in \{2, 3..., 8\}$	overlapping membership
$on \in \{500, 2500\}$	number of overlapping nodes

Table 1: Parameters for the LFR benchmarks.

For four groups of configuration basing on the community size and *on*, we vary *om* from 2 to 8 to get seven networks in each group. Denote them as:
1) LFR_s_0.1 for $\{s : [10, 50], on = 500\}$;
2) LFR_s_0.5 for $\{s : [10, 50], on = 2500\}$;
3) LFR_b_0.1 for $\{b : [20, 100], on = 500\}$;
4) LFR_b_0.5 for $\{b : [20, 100], on = 2500\}$.

3.2.2 Real-world Networks

We consider five real-world network datasets with labeled ground truth from the Stanford Network Analysis Project (SNAP)[2] and three genetic networks with labeled ground truth from the Isobase website[3].

- **SNAP**: The five SNAP networks, **Amazon, DBLP, LiveJ, YouTube, Orkut**, are in the domains of social, product, and collaboration [22]. We preprocess the ground truth to remove identical copies of ground truth communities.
- **Isobase**: The three genetic networks from the Isobase website describe the interactions between proteins. **HS** describes these interactions in humans, **SC** in *S. cerevisiae*, a type of yeast, and **DM** in *D. melanogaster*, a type of fruit fly. Such networks are interesting as communities may correspond to different genetic functions, such as metabolism regulation.

Table 2 summarizes the networks and their ground truth communities. We calculate the average and standard deviation of the community sizes, and the average conductance.

Network			**Ground truth communities**	
Name	**#Nodes**	**#Edges**	**Avg. ± Std. Size**	**Avg. Cond.**
Amazon	334,863	925,872	14 ± 20	0.07
DBLP	317,080	1,049,866	37 ± 356	0.40
LiveJ	3,997,962	34,681,189	29 ± 65	0.36
YouTube	1,134,890	2,987,624	21 ± 58	0.83
Orkut	3,072,441	117,185,083	242 ± 418	0.73
DM	15,326	486,970	215 ± 741	0.88
HS	10,296	54,654	166 ± 258	0.88
SC	5,523	82,656	159 ± 208	0.90

Table 2: Statistics for real-world networks and their ground truth communities.

[1] http://santo.fortunato.googlepages.com/inthepress2

[2] http://snap.stanford.edu

[3] http://groups.csail.mit.edu/cb/mna/isobase/

4. SPECTRAL DIFFUSION

4.1 Local sampling

In large networks with millions or billions of nodes, we first apply a heuristic method to sample a subgraph with thousands of nodes around the seeds and then do membership identification on the small subgraph rather than on the whole network. We need the subgraph to be large enough to contain most of the latent members, but not too large to contain many irrelevant nodes. How to sample plays a key role in the follow-up step of identifying the latent members effectively and efficiently.

According to the small world phenomenon and "six degrees of separation", most members should be at most two or three steps far away from the seed members if we want to identify a small community of size hundreds. If we apply a short random walk to expand the subgraph, the subgraph will be much larger than we expected due to some very popular nodes with thousands of neighbors in the large network. Thus, we mainly use BFS and filter some very popular nodes during the BFS expansion, then consider a short random walk as the post processing if the sampled subgraph surpasses our upper bound threshold.

Starting from each seed s_k, at each round we do a one-step BFS and use a Filter procedure on the frontiers to choose high inward ratio nodes (evaluated by the fraction of inward edges to the BFS subgraph) until the total out-degree is greater than 3000. We do two rounds of such BFS and Filter, and add one more round BFS and Filter if we have fewer than N_1 nodes. N_1 is set to 300 for Orkut and 30 for other real world datasets. Then we conduct one more step of BFS on the selected nodes, some very popular nodes filtered by previous expansion may still be included by this step. In the end, we union all BFS subgraphs obtained from each seed, and use $k = 3$ steps of random walk to filter some low probability nodes if the amalgamated subgraph contains more than $N_2 = 5000$ nodes. All parameters are determined experimentally by considering the statistics of the networks and the ground truth shown in Table 2.

For simplicity of notation, we denote the sampled subgraph as $G = (V, E)$ in the following discussion, and extract the community from this comparatively small subgraph instead of the original large network. Experiments in Section 5 will show that the above sampling method yields a subgraph which is only a 0.09% fraction of the original network but covers 96% of the nodes in the ground truth communities on average. This pre-processing procedure largely reduces the computation load for the follow-up community detection and guarantees the detection accuracy.

4.2 Spectra and quadratic forms

Let $\mathbf{L} = \mathbf{D} - \mathbf{A}$ be the Laplacian matrix of G, and the two normalized graph Laplacian matrices are

$$\mathbf{L_{rw}} = \mathbf{I} - \mathbf{N_{rw}} = \mathbf{D^{-1}L},$$

$$\mathbf{L_{sym}} = \mathbf{I} - \mathbf{N_{sym}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}},$$

where $\mathbf{N_{rw}} = \mathbf{D}^{-1}\mathbf{A}$ is the transition matrix, and $\mathbf{N_{sym}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix. The eigenvalue decompositions of the Laplacian matrices are closely related to minimum cuts related to conductance optimization, and give rise to the success of spectral clustering.

Let $\mathbf{y} \in \{0, 1\}^n$ be a binary indicator vector representing a community C. We can express some properties of C via quadratic forms [3]:

- number of nodes in C: $\mathbf{y^T y}$
- number of edges in C: $\frac{1}{2}\mathbf{y^T A y}$
- total degree of nodes in C: $\mathrm{Vol}(\mathbf{A_c}) = \mathbf{y^T D y}$
- number of cutting edges from C to the remainder of the graph: $\mathbf{y^T L y} = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(y_i - y_j)^2$

If we are looking for a minimum cut subgraph containing the seeds in the scale of the ground truth community $T = (V_t, E_t)$, the problem could be written in a quadratic optimization form:

$$\min \quad \mathbf{y^T L y}$$

$$s.t.\ (1) \quad \mathbf{y^T y} = |V_t|,\ (2)\ y_i \in \{0, 1\},\ (3)\ y_i = 1, i \in S.$$

The Rayleigh quotients related to the Laplacian matrices and the indicator vector $\mathbf{y}$ are intimately tied to the community finding task and the spectra of the corresponding matrix [3].

1) **Local RatioCut**. The Rayleigh quotient $\mathbf{R(L, y)} = \frac{\mathbf{y^T L y}}{\mathbf{y^T y}}$ indicates the fraction of cutting edges to the community size, which is related to the RatioCut for spectral clustering [6]. Let $\mathbf{L} = \mathbf{Q \Lambda Q^T}$ be the eigen decomposition, where $\mathbf{Q}$ is an orthonormal matrix and $\mathbf{\Lambda} = diag(\lambda_1, ...\lambda_n)$, $\lambda_1 \leq ... \leq \lambda_n$. $\mathbf{R(L, y)}$ is the weighted average of the eigenvalues

$$\mathbf{R(L, y)} = \frac{\mathbf{y^T L y}}{\mathbf{y^T y}} = \frac{\sum_{i=1}^{n} \lambda_i x_i^2}{\sum_{i=1}^{n} x_i^2}, \tag{1}$$

where $x_i = \mathbf{v_i^T y}$ is the projection coordinate of $\mathbf{y}$ on the ith eigenvector $\mathbf{v_i}$. Thus $\mathbf{R(L, y)} \in [\lambda_{min}, \lambda_{max}]$. If we want to minimize the cut as compared with the internal nodes in community C, i.e. to minimize $\mathbf{R(L, y)}$, the indicator vector $\mathbf{y}$ should be very close to the dominant eigenvectors having smaller eigenvalues. This leads to our idea of finding a sparse indicator vector containing the seeds in the span of the dominant eigenvectors with smaller eigenvalues of $\mathbf{L}$.

2) **Local NCut**. The conductance Φ_c of community $C = (V_c, E_c)$ could be written as a generalized Rayleigh quotient, which is related to the normalized cut, Ncut, for spectral clustering [17].

$$\mathbf{R(L, D, y)} = \frac{\mathbf{y^T L y}}{\mathbf{y^T D y}}. \tag{2}$$

As $\mathbf{R(L, D, y)} = \mathbf{R}(\mathbf{L_{sym}}, \mathbf{D}^{\frac{1}{2}}\mathbf{y})$, conductance Φ_c is in the range of the eigenvalues of $\mathbf{L_{sym}}$. Similarly, if we want to minimize the conductance of community C, the scaled indicator vector $\mathbf{D}^{\frac{1}{2}}\mathbf{y}$ should be very close to the dominant eigenvectors with smaller eigenvalues of $\mathbf{L_{sym}}$.

According to the spectral clustering theory [19], λ is an eigenvalue of $\mathbf{L_{rw}}$ with eigenvector $\mathbf{v}$ if and only if λ is an eigenvalue of $\mathbf{L_{sym}}$ with eigenvector $\mathbf{D^{1/2}v}$:

$$\mathbf{L_{rw}v} = \lambda\mathbf{v} \quad \text{iff} \quad \mathbf{L_{sym}}(\mathbf{D}^{\frac{1}{2}}\mathbf{v}) = \lambda(\mathbf{D}^{\frac{1}{2}}\mathbf{v}).$$

So the indicator vector $\mathbf{y}$ should be very close to the dominant eigenvectors of $\mathbf{L_{rw}}$. This leads to our idea of finding a sparse indicator vector containing the seeds in the span of the dominant eigenvectors with smaller eigenvalues of $\mathbf{L_{rw}}$.

$$\min \quad |\mathbf{y}|_0 = |\mathbf{y}|_1 = e^T y$$

$$s.t.\ (1)\ \exists \mathbf{x}, \mathbf{y} = \mathbf{V_d x},\ (2)\ y_i \in \{0, 1\},\ (3)\ y_i = 1, i \in S.$$

$\mathbf{V_d}$ is formed by the dominant eigenvectors of $\mathbf{L_{rw}}$ and $\mathbf{e}$ the vector of all ones. The objective is in ℓ_0 norm for minimizing the number of nonzero elements, which is equivalent to the ℓ_1 norm when elements of $\mathbf{y}$ are restricted to 0 or 1.

4.3 Spectral diffusion

As $\mathbf{L_{rw}} = \mathbf{I} - \mathbf{N_{rw}}$, the eigenvalue decompositions of the

Laplacian matrices are also closely related to expansion of rapid mixing of random walks. As

$$\mathbf{L_{rw}v} = (\mathbf{I} - \mathbf{N_{rw}})\mathbf{v} = \lambda\mathbf{v} \quad \Leftrightarrow \quad \mathbf{N_{rw}v} = (1-\lambda)\mathbf{v},$$

$\mathbf{L_{rw}}$ and $\mathbf{N_{rw}}$ share the same set of eigenvectors and the corresponding eigenvalue of $\mathbf{N_{rw}}$ is $1-\lambda$ where λ is the eigenvalue of $\mathbf{L_{rw}}$. Equivalently, we could find a sparse indicator vector containing the seeds in the span of the dominant eigenvectors with larger eigenvalues of $\mathbf{N_{rw}}$.

Further, instead of using the eigenvalue decomposition, we consider short random walks for the probability diffusion starting from the seed set to get the "local spectra". We define several variants of the spectral diffusion based on different transition matrices for the random walks.

1) **Standard Random Walk**. The standard random walk uses the transition matrix $\mathbf{N_{rw}}$ for the probability diffusion.

$$\mathbf{N_{rw}} = \mathbf{D}^{-1}\mathbf{A} \tag{3}$$

2) **Light Lazy Random Walk**. Light lazy random walk keeps some probability at the current node for the random walks.

$$\mathbf{N_{rw}} = (\mathbf{D} + \alpha\mathbf{I})^{-1}(\alpha\mathbf{I} + \mathbf{A}) \tag{4}$$

where $\alpha \in N^{0+}$. $\alpha = 0$ degenerates to the standard random walk and $\alpha = 1, 2, 3, ...$ corresponds to a random walk in the modified graph with $1, 2, 3, ...$ loops at each node.

3) **Lazy Random Walk**.

$$\mathbf{N_{rw}} = (\mathbf{D} + \alpha\mathbf{D})^{-1}(\alpha\mathbf{D}+\mathbf{A}) = \frac{\alpha}{1+\alpha}\mathbf{I} + \frac{1}{1+\alpha}\mathbf{D}^{-1}\mathbf{A} \tag{5}$$

where $\alpha \in [0, 1]$. E.g. $\alpha = 0.1$ corresponds to a random walk that always retains $\frac{0.1}{1+0.1}$ probability on the current nodes during the diffusion process. $\alpha = 0$ degenerates to the standard random walk.

4) **Personalized PageRank**.

$$\mathbf{N_{rw}} = \alpha\mathbf{S} + (1-\alpha)\mathbf{D}^{-1}\mathbf{A} \tag{6}$$

where $\alpha \in [0, 1]$ and $\mathbf{S}$ the diagonal matrix with binary indicators for the seed set S. E.g. $\alpha = 0.1$ corresponds to a random walk that always retains 10 percent of the probability on the seed set. $\alpha = 0$ degenerates to the standard random walk.

One step of random walk is defined as $\mathbf{N_{rw}^T p}$ for a probability vector $\mathbf{p}$, and the probability density for a random walk of length k is given by a Markov chain

$$\mathbf{p_k} = \mathbf{N_{rw}^T p_{k-1}} = (\mathbf{N_{rw}^T})^{\mathbf{k}}\mathbf{p_0}$$

where $\mathbf{p_0}$ is the initial probability density evenly assigned on the seeds. Instead of using the eigenvalues and eigenvectors as the "global spectra", we conduct short random walks for the "local spectra" representing the local structure around the seeds. Different random walks yield different local spectral diffusions.

4.4 Local spectral clustering

Local Spectral Subspace. Instead of using the invariant subspace spanned by the leading eigenvectors of $\mathbf{N_{rw}}$, we define the invariant subspace approximation on an order-d Krylov subspace.

$$\mathbf{V_d^{(k)}} = [\mathbf{p_k}, \mathbf{p_{k+1}}, ..., \mathbf{p_{k+d-1}}] \tag{7}$$

Here k and d are some modest numbers. This local spectral subspace is essentially the same as the definition of LOSP [7], which is defined based on subspace iterations. But our definition is simplified and also more efficient for the subspace calculation.

We define variants of the local spectral subspaces based on different local spectral diffusions as defined in Section 4.3. A thorough investigation on different local spectral diffusions is provided in Section 5.

Local Community Detection. We relax the indicator vector $\mathbf{y}$ to be in $[0, 1]$ and look for a vector in the local spectral subspace by solving a linear programming problem:

$$\min \quad |\mathbf{y}|_1 = e^T y$$

$$s.t. \ (1)\ \exists \mathbf{x}, \mathbf{y} = \mathbf{V_d^{(k)}x},\ (2)\ \mathbf{y} \geq 0,\ (3)\ y_i \geq \frac{1}{|S|},\ i \in S.$$

where $\mathbf{e}$ is the vector of all ones. This is an ℓ_1 norm approximation for finding a sparse vector indicating a small community containing the seeds with $\mathbf{y}$ in the span of $\mathbf{V_d^{(k)}}$. $\mathbf{y} \leq 1$ is not required for the minimization on $\mathbf{y}$. $\mathbf{y_i}$ indicates the belonging likelihood of node i in the target community. We then sort the values in $\mathbf{y}$ in the non-descending order and select the corresponding $|V_t|$ nodes with the higher belonging likelihood as the output community.

A relaxed version of the ℓ_1 norm [7] requires the indicator vector to contain at least some seeds, and relaxes constraint (3) to be $\mathbf{s}^T\mathbf{y} \geq 1$ where $\mathbf{s}$ is the binary indicator vector representing the seeds. Another alternative is to solve a quadratic programming problem and replace the objective by $\mathbf{y^T L y}$ or $(\mathbf{D^{\frac{1}{2}}y})^\mathbf{T}\mathbf{L_{sym}}(\mathbf{D^{\frac{1}{2}}y})$ but this involves computations with the Laplacian matrices.

5. EXPERIMENTAL RESULTS

We implement LOSP++ in Matlab and thoroughly compare LOSP++ with state-of-the-art localized community finding algorithms on the 28 LFR datasets as well as the eight real world networks across multiple domains. For the five SNAP datasets, we randomly locate 500 labeled ground truth communities on each dataset, and randomly pick three exemplary seeds from each target community. For the 28 LFR datasets and the three Biology datasets, we deal with every ground truth community and randomly pick three exemplary seeds from each ground truth community. We do sampling as a preprocessing for the real data, and directly apply the local spectral method without sampling on the 28 LFR benchmarks as there are only $n = 5000$ nodes for each network.

5.1 Parameter setup

Some modest values for the number of random walk steps k and the subspace dimension d are needed such that the probability spreads out to all members in the local community but doesn't reach the global stationary probability. We did parameter study on all datasets, and found that $d = 2$ and $k = 2$ perform the best in general. Figure 1 illustrates the F1 scores for different (k, d) combinations on Amazon. By fixing the random walk steps to 2, Figure 2 shows that $d = 2$ yields the best for all real networks.

5.2 Compare with local spectral methods

For related works in finding localized community from the local spectral subspace, we show the progress of LOSP++ as compared with LEMON [15] and LOSP [7].

Sampling. LEMON and LOSP also perform sampling as a preprocessing on large networks. LEMON samples based on random walks while LOSP samples based on BFS search. Table 3 provides statistics after applying sampling method on the SNAP networks and compares the sampling quality with LEMON and LOSP. Coverage ratio indicates the average fraction of ground truth covered by the sampled subgraph, and sampling rate indicates the average fraction of subgraph size as compared with the original network scale. Results show that LOSP++ has a higher coverage ratio with reasonable sample size, covering about 96% ground truth with a small sampling rate (the fraction of sampled subgraph as compared with the original network) of less than one in thousand on average. As compared with LOSP, LOSP++ covers considerably more nodes and increase the coverage ratio significantly by 10 percent.

Local community detection. Figure 3 shows the detection accuracy on SNAP datasets. In order to remove the impact of different methods in finding a local minimum, we use the ground truth size as a budget for three algorithms. LOSP++ improves the detection accuracy as compared with the labeled ground truth, especially on LiveJ and Orkut. Note that LOSP++ hasn't added the reseeding process, while LEMON and LOSP both use reseeding to further improve the detection quality.

5.3 Evaluation on spectral diffusions

We thoroughly evaluate different spectral diffusion methods: light lazy, lazy and personalized random walks (RW) with different α parameters on all datasets. Note that the three variants all degenerates to the standard RW when $\alpha = 0$. Figure 4 shows the average F1 scores on the sampled 500 ground truth for each SNAP network. Results show that:

- LOSP++ is robust for different spectral diffusion methods using different α parameters.
- Light lazy RW and lazy RW perform better than the standard RW, and the standard RW performs better than the personalized RW.
- Parameter α has very little impact for light lazy, being stable at the F1 score for each of the five SNAP datasets.
- For the light lazy RW and lazy RW, the average F1 score increases slightly with a higher α. The rising trend is more apparent for YouTube.
- The detection accuracy decays with a higher α for the personalized RW. $\alpha = 0.1$ or 0.15 performs the best, which achieves the same score as that of standard RW. One exception is on YouTube.

Experiments on the four groups of LFR datasets show similar results. LOSP++ is very stable for light lazy and lazy, and the quality decays for personalized RW when α increases. To save space, we only show in Figure 5 the trends for the seven networks in group LFR_s_0.5, which embed smaller communities with more overlaps.

Experiments on the three Biology datasets, as shown in Figure 6, also witness the stabilization of LOSP++ on light lazy RW and lazy RW, and there is some fluctuation on the personalized RW.

5.4 Clusters extracted vs. ground-truth

For the final comparison, we thoroughly evaluate LOSP++ on all datasets with two venerable local diffusion algorithms, HK for `hk-relax` [10] and PR for `pprpush` [1]. PR is based on the PageRank diffusion while HK is based on Heat Kernel graph diffusion. To make a fair comparison, we run the three algorithms on the same three seeds randomly chosen from the ground truth communities. In the following discussion, LS stands for LOSP++.

Evaluation on LFR benchmarks. Figure 7 illustrates

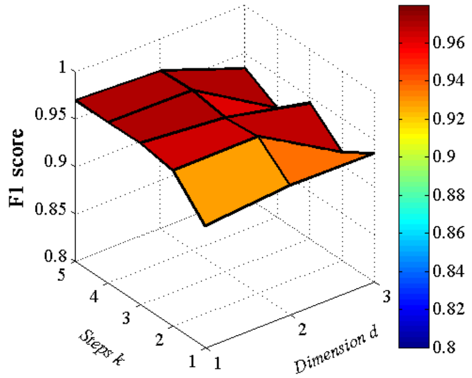


Figure 1: The k, d combinations for Amazon. $k = 2, d = 2$ yields the highest F1 score, and the scores are robust for different k, d combinations.

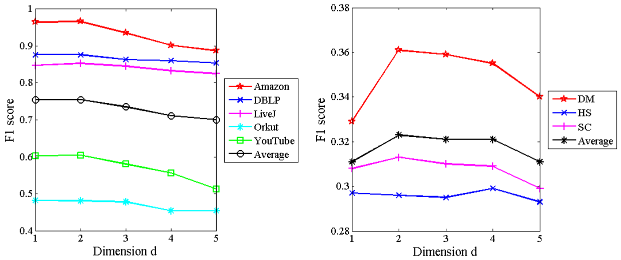


Figure 2: Parameter study on the subspace dimension d for steps $k = 2$ on all real networks. $d = 2$ yields the highest F1 score, and the scores are robust for different parameters.

SNAP	LEMON		LOSP		LOSP++		
Dataset	C_{ratio}	S_{size}	C_{ratio}	S_{size}	C_{ratio}	S_{size}	S_{rate}
Amazon	1.00	2913	0.99	19	0.99	34	0.0001
DBLP	0.98	2409	0.98	171	0.98	198	0.0002
LiveJ	0.63	4398	0.99	293	1.00	629	0.0002
YouTube	0.66	3745	0.90	906	0.95	3237	0.0028
Orkut	0.64	3379	0.45	313	0.87	4035	0.0013
Average	0.78	3369	0.86	340	0.96	1627	0.0009

Table 3: Statistics of the mean values for the sampling method on SNAP datasets. C_{ratio}, S_{size} and S_{rate} correspond to the mean coverage ratio, sampling size and sampling rate.

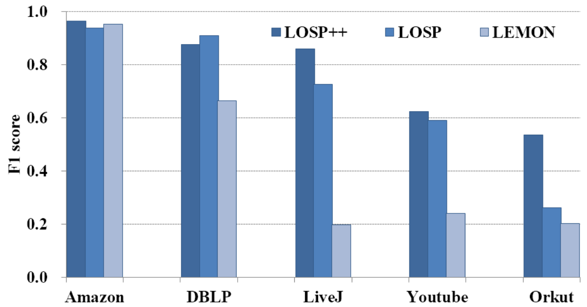


Figure 3: Comparison with local spectral methods.

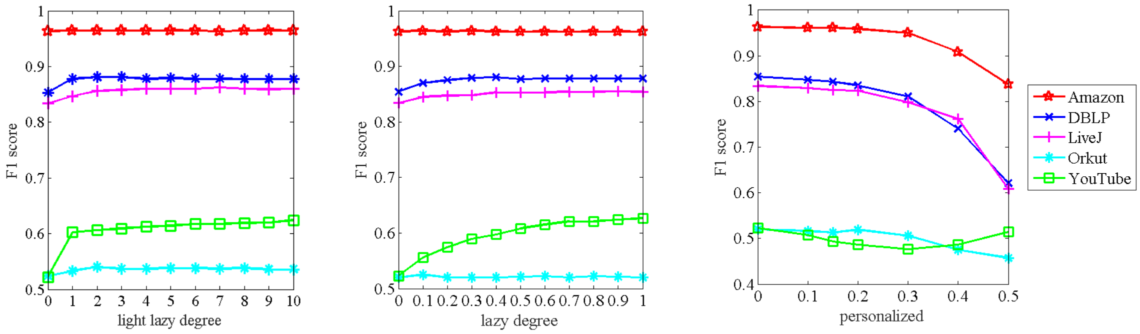


Figure 4: Evaluation of different spectral diffusions on SNAP.

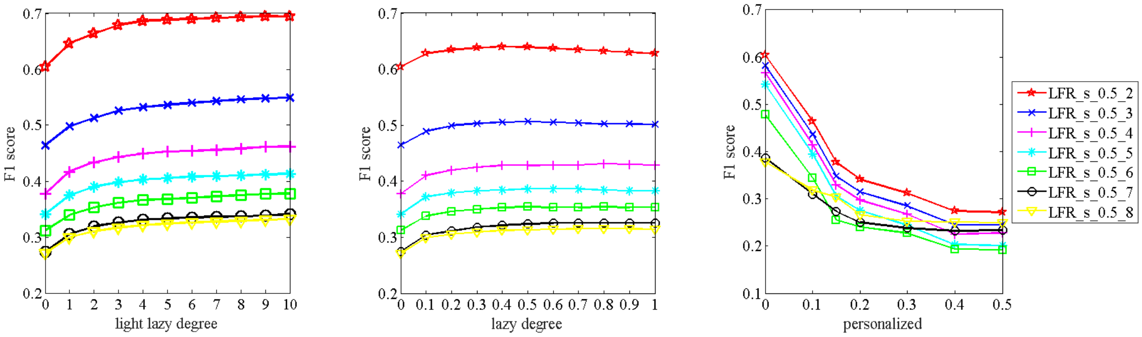


Figure 5: Evaluation of different spectral diffusions on LFR_s_0.5 group of benchmarks.

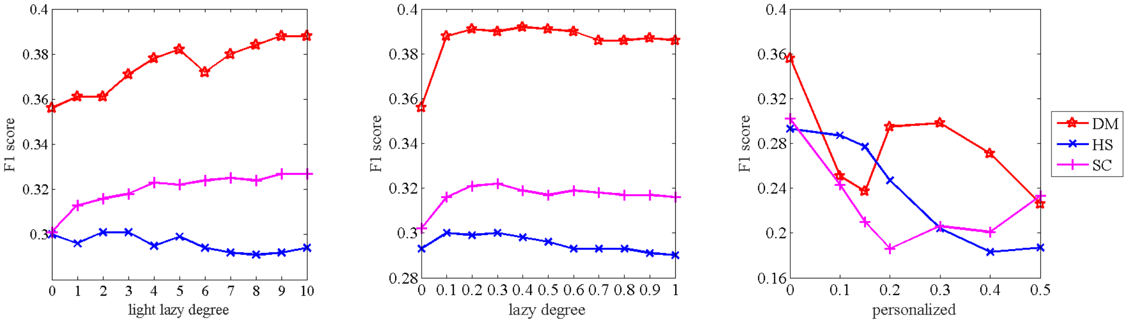
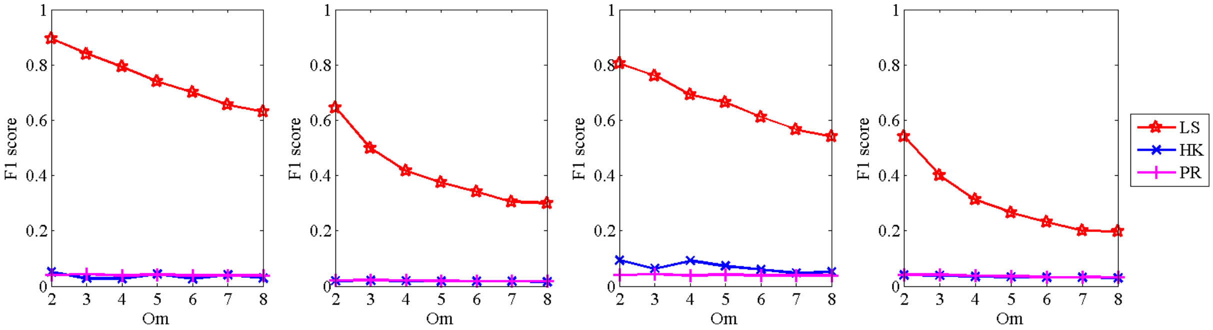


Figure 6: Evaluation of different spectral diffusions on Biology networks.



(a) LFR_s_0.1 (b) LFR_s_0.5 (c) LFR_b_0.1 (d) LFR_b_0.5

Figure 7: Evaluation on the LFR Benchmark Graphs.

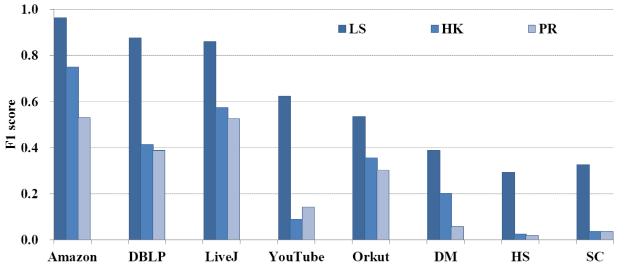


Figure 8: Comparison with heat kernel(HK) and PageRank(PR).

the comparison on four groups of LFR benchmark graphs. LOSP++ substantially outperforms HK and PR. It is reasonable that the detection accuracy decays on graphs with more overlappings indicated by the higher *om* and *on*.

Evaluation on real data. Figure 8 illustrates the detection accuracy of LS (for LOSP++), HK and PR on the five SNAP networks and the three Biology networks. LOSP++ apparently outperforms HK and PR on accuracy. Especially on the Biology networks, HK and PR rarely find the protein-protein-interaction (PPI) communities but LOSP++ detects a considerable fraction.

6. CONCLUSIONS

Based on Rayleigh quotients related to the Laplacian matrices, we provide theoretical justifications for finding community structure from the local spectral subspace, a local approximation for the invariant subspace spanned by dominant eigenvectors of the Laplacian matrices. Experimental results suggest that LOSP++ is a worthy competitor for the semi-supervised learning task of extracting the target community from very few seed members in large networks, and considerably improves the detection accuracy on LOSP.

There are a number of interesting issues for further investigation. We evaluate several variants for the spectral diffusions. Light lazy and lazy random walks outperform standard and personalized random walks among all synthetic as well as real networks we considered. A theoretical analysis on different diffusion methods will be very valuable for the local spectral methods. We also wish to give analysis on the impact of parameter α, considering the structural properties of the networks and the communities.

Acknowledgments

This research work was supported by US Army Research Office(W911NF-14-1-0477), NSF of China(61472147) and NSF of Hubei Province(2015CFB566).

7. REFERENCES

[1] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *FOCS*, pages 475–486, 2006.

[2] R. Andersen and K. J. Lang. Communities from seed sets. In *WWW*, pages 223–232. ACM, 2006.

[3] D. Bindel. Communities, spectral clustering, and random walks. In *Workshop on Algorithms for Modern Massive Data Sets (MMDS)*, 2012.

[4] F. Chung. The heat kernel as the pagerank of a graph. *PNAS*, 104(50):19735–19740, 2007.

[5] F. Chung and O. Simpson. Solving linear systems with boundary conditions using heat kernel pagerank. In *Algorithms and Models for the Web Graph(WAW)*, pages 203–219, 2013.

[6] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11(9):1074–1085, 1992.

[7] K. He, Y. Sun, D. Bindel, J. Hopcroft, and Y. Li. Detecting overlapping communities from local spectral subspaces. In *ICDM*, pages 769–774, 2015.

[8] R. Kannan, S. Vempala, and A. Yetta. On clusterings - good, bad and spectral. In *FOCS*, pages 367–377, 2012.

[9] G. Karypis and V. Kumar. Metis-unstructured graph partitioning and sparse matrix ordering system. version 2.0, 1995.

[10] K. Kloster and D. F. Gleich. Heat kernel based community detection. In *KDD*, pages 24–27. ACM, August 2014.

[11] I. M. Kloumann and J. M. Kleinberg. Community membership identification from small seed sets. In *KDD*, pages 1366–1375. ACM, August 2014.

[12] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):016118, 2009.

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

[14] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW*, pages 695–704, 2008.

[15] Y. Li, K. He, D. Bindel, and J. Hopcroft. Uncovering the small community structure in large networks. In *WWW*, pages 658–668, 2015.

[16] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi. A local spectral method for graphs: with applications to improving graph partitions and exploring data graphs locally. *The Journal of Machine Learning Research*, 13(1):2339–2365, 2012.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[18] D. A. Spielman and S. Teng. Nearly-linear time algorithms for graph partitiongn, graph sparsification, and solving linear systems. In *STOC*, pages 81–90, 2004.

[19] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[20] J. J. Whang, D. F. Gleich, and I. S. Dhillon. Overlapping community detection using seed set expansion. In *CIKM*, pages 2099–2108, October 2013.

[21] J. Xie, S. Kelley, and B. K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):43, 2013.

[22] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *ICDM*, pages 745–754, December 2012.