Analysis of Spectral Space Properties of Directed Graphs using Matrix Perturbation Theory with Application in Graph Partition

Yuemeng Li*, Xintao Wu† and Aidong Lu*
* *University of North Carolina at Charlotte, Email: {yli60,alu1}@uncc.edu*
† *University of Arkansas, Email: xintaowu@uark.edu*

***Abstract*—The eigenspace of the adjacency matrix of a graph possesses important information about the network structure. However, analyzing the spectral space properties for directed graphs is challenging due to complex valued decompositions. In this paper, we explore the adjacency eigenspaces of directed graphs. With the aid of the graph perturbation theory, we emphasize on deriving rigorous mathematical results to explain several phenomena related to the eigenspace projection patterns that are unique for directed graphs. Furthermore, we relax the community structure assumption and generalize the theories to the perturbed Perron-Frobenius simple invariant subspace so that the theories can adapt to a much broader range of network structural types. We also develop a graph partitioning algorithm and conduct evaluations to demonstrate its potential.**

***Keywords*-Directed graphs; Asymmetric adjacency matrices; Matrix perturbation; Spectral projection; Graph partition**

I. Introduction

Researchers have developed approaches and algorithms to deal with the clustering in directed graphs because relationships in many networks are asymmetric. Refer to [1] for a recent survey. Roughly speaking, they can be classified into two categories. In the first category, the directed graph is converted into an undirected one, either unipartite or bipartite, where edge direction is preserved, e.g., via edge weights of the produced unipartite graph [2] or edges in the produced bipartite graph [3]. Clustering algorithms for undirected weighted graphs are then applied. Methods in the second category are mainly based on the idea of extending clustering objective functions and methodologies to directed networks. In those approaches, the graph clustering is expressed as an optimization problem and the desired clustering properties are captured in the modified objective criterion. For example, researchers developed the directed versions of modularity [4]–[6], the objective function of weighed cuts in directed graphs [7], and the spectral graph clustering based on the Laplacian matrix of the directed graphs [8], [9]. However, it is unclear to what extent the information about the directionality of the edges is retained by these approaches.

In this paper we study whether we can directly analyze the spectral properties of the adjacency matrix of the underlying directed network instead of transforming the directed network to undirected or developing the directed versions of the objective criterion used in graph clustering. When the concern is with directed graphs, one main difficulty for spectral clustering is to deal with the complex values for eigenpairs associated with the asymmetric adjacency matrix. The problem of how to select a set of eigenvectors to produce a meaningful partition result becomes very complicated. Another major difficulty associated with analyzing the spectral spaces of asymmetric adjacency matrices is that the eigenvectors do not form an orthonormal basis naturally. This complicates the process of analyzing the behaviors of nodes in the spectral space.

We conduct theoretical analysis to address the above difficulties by leveraging the spectral graph perturbation theory. The spectral graph perturbation focuses on analyzing the changes in the spectral space of a graph after new edges are added or deleted. We provide a theoretical analysis of the properties of the eigenspace for directed graphs and develop a method to circumvent the issue of complex eigenpairs. Our analysis utilizes the connectedness property of the components of a network to screen out irrelevant eigenpairs and thus eliminating the need for dealing with complex eigenpairs. We demonstrate how to derive the approximations of the eigenvectors by leveraging the constructed orthonormal basis when treating the graph as a perturbation from a block matrix. Furthermore, the derived theories are generalized to the perturbed Perron-Frobenius simple invariant subspace. The significance of such a spectral subspace is that it is a real subspace with some unique properties that contains all the spectral clustering information of a graph. We develop an algorithm to partition directed graphs without transforming the adjacency matrices or modifying the objective functions. Empirical evaluations show effectiveness of our algorithm.

II. Preliminaries

In this study, we focus on directed graphs without self-loops having nonnegative edge weights. A directed graph can be represented as its adjacency matrix $A_{n\times n}$ with $a_{ij} = 1$ if there exists an edge pointing from node V_i to node V_j and $a_{ij} = 0$ otherwise. Let λ_i be an eigenvalue of A with its eigenvector $\boldsymbol{x}_i$ $(i = 1, \cdots, n)$. The eigenvector $\boldsymbol{x}_i$ is represented as a column vector. For undirected graphs, the eigenvectors $\boldsymbol{x}_i$ $(i = 1, \cdots, K)$ corresponding to the K largest real eigenvalues contain the most topological information of the corresponding K communities of the graph in the spectral space. The K-dimensional spectral space is

spanned by $(\boldsymbol{x}_1, \cdots \boldsymbol{x}_K)$. When a node u is projected in the K-dimensional subspace with $\boldsymbol{x}_i$ as the basis, the row vector $\boldsymbol{\alpha}_u = (x_{1u}, x_{2u}, \cdots, x_{Ku})$ is its coordinate in this spectral subspace. For directed graphs in this study, the chosen K real eigenvectors will be used to perform the projections.

Table I
SYMBOLS AND DEFINITIONS

A	Adjacency matrix of a graph
$\|A\|_2$	The spectral norm of A, i.e., the largest singular value of A
P	Permutation matrix
$\widetilde{A}$	Perturbed matrix of A
$\mathcal{L}(L)$	The set of eigenvalues of L
$\Re(X)$	An invariant subspace of A spanned by a basis X
$\Re(X)^{\perp}$	The orthogonal complement of $\Re(X)$
$\rho(A)$	The spectral radius of A
$(q_1, \cdots, q_n)$	An orthonormal basis of A
$(\lambda_1, \cdots, \lambda_n)$	Eigenvalues of A
$(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$	Eigenvectors of A
A^H	Conjugate transpose of A
Unitary	A is unitary if $A^{-1} = A^H$

For a square matrix A with a perturbation E, the matrix after perturbation can be written as $\widetilde{A} = A + E$. For the perturbed matrix, $\tilde{\lambda}_i$ and $\tilde{\boldsymbol{x}}_i$ denote the perturbed eigenpairs. When the matrix perturbation theory is applied to analyze the spectral properties of directed graphs, the most difficult problem is that the perturbed eigenvectors cannot be estimated using simple linear combinations of other eigenvectors, since the eigenvectors do not form orthonormal basis naturally. This problem was solved by working with spectral resolutions and using orthogonal reduction to block triangular. Therefore, the estimations for perturbed eigenvectors can be expressed by the spectral resolution of A with respect to its simple invariant subspaces. We reference the relevant definitions and theorems from [10] as follows:

Lemma 1: Let the columns of X be linearly independent and let columns of Y span $\Re(X)^{\perp}$. Then $\Re(X)$ is an invariant subspace of A if and only if $Y^H AX = 0$. In this case $\Re(Y)$ is an invariant subspace of A^H.

Lemma 2: Let $\Re(X)$ be an invariant subspace of A, columns of X form an orthonormal basis for $\Re(X)$, and (X, Y) be unitary. Then the decomposition of A has the reduced form:

$$(X,Y)^H A(X,Y) = \begin{pmatrix} L_1 & H \\ \mathbf{0} & L_2 \end{pmatrix}, \tag{1}$$

where $L_1 = X^H AX$, $L_2 = Y^H AY$, $AX = XL_1$ and $H = X^H AY$. Furthermore, eigenvalues of L_1 are the eigenvalues of A associated with $\Re(X)$. The rest eigenvalues of A are those of L_2.

Definition 1: Let $\Re(X)$ be an invariant subspace of A, and let (1) be its reduced form with respect to the unitary matrix (X, Y). Denote $\mathcal{L}(L)$ as the set of the eigenvalues of L. Then $\Re(X)$ is a **simple invariant subspace** if $\mathcal{L}(L_1) \cap \mathcal{L}(L_2) = \emptyset$.

Lemma 3: Theorem 2.7 in chapter V of [10]. Let $\widetilde{A} = A + E$, $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ be a basis of A and denote $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ and $Y = (\boldsymbol{x}_{K+1}, \cdots, \boldsymbol{x}_n)$. Suppose that (X, Y) is unitary, and suppose that $\Re(X)$ is a **simple invariant subspace** of A so that it has the reduced form as Equation (1). For $i \in (1, \cdots, K)$, the perturbed eigenvectors $\tilde{\boldsymbol{x}}_i$ can be approximated as:

$$\tilde{\boldsymbol{x}}_i \approx \boldsymbol{x}_i + Y(\lambda_i I - L_2)^{-1} Y^H E \boldsymbol{x}_i, \tag{2}$$

when the following conditions hold:

1) $\delta = \inf_{\|T\|=1} \|THT\|_2 - \|X^H EX\|_2 - \|Y^H EY\|_2 > 0$, where $H = X^H AY$ and $t_i \approx (\lambda_i I - L_2)^{-1} Y^H E \boldsymbol{x}_i$ for column vectors in T.
2) $\gamma = \|Y^H EX\|_2 < \frac{1}{2}\delta$.

III. MODELING OBSERVED GRAPHS AS PERTURBATIONS

We assume that the observed graph $\widetilde{A}$ of a network has K communities namely $C_1, \cdots, C_K$. According to the pattern based criterion, we make the assumption that each community C_i in a directed graph should be a strongly connected component. This assumption makes an easy starting point to study directed graphs and will be relieved in the next section. Formally, for any observed directed graph containing multiple communities with the above defined community structure, its adjacency matrix $\widetilde{A}$ has the reducible form:

$$P\widetilde{A}P^{-1} = \begin{pmatrix} A_1 & & E \\ & \ddots & \\ \mathbf{0} & & A_K \end{pmatrix}, \tag{3}$$

where A_is are strongly connected components corresponding to communities and E contains the edges connecting A_is. The permutation matrix P has the same effect of switching the nodes. We can also consider $\widetilde{A}$ as the perturbation of the matrix A with disconnected communities by links connecting each other as E. Hence, the observed graph $\widetilde{A}$ in the form of Equation (3) can be regarded as the perturbation from the diagonal block form as:

$$\widetilde{A} = A + E = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ \mathbf{0} & & A_K \end{pmatrix} + E. \tag{4}$$

A. Perron-Frobenius Eigenpair

For the strongly connected component C_i, the following lemma shows the relationship between the connectedness and reducibility of a graph.

Lemma 4: [11] For all i, let A_i be the adjacency matrix representation of a component C_i, then C_i is strongly connected iff A_i is irreducible (i.e., it cannot be reduced into the form of Equation 3).

We introduce the Perron-Frobenius theorem for non-negative irreducible components.

Lemma 5: Chapter 8 of [12]. Let C be an irreducible and non-negative $c \times c$ matrix corresponding to a strongly connected component. Let $\lambda_1, \cdots, \lambda_c$ be its (real or complex) eigenvalues. Then its spectral radius $\rho(C)$ is defined as:

$$\rho(C) \stackrel{\text{def}}{=} \max_p(|\lambda_p|). \tag{5}$$

It is called the Perron-Frobenius eigenvalue of C and the corresponding eigenvector is called the Perron-Frobenious eigenvector. The following properties hold:

1) The spectral radius $\rho(C)$ is a positive real number and it is a simple eigenvalue of C.
2) The only eigenvector that has all positive entries is the one associated with $\rho(C)$. All the other eigenvectors have mixed signed entries.

Lemmas 4 and 5 simply suggest that there exists a bijective mapping from the set of communities to the set of spectral radii of all the communities. Therefore, if the network has a clear community structure, we can identify the underlying community structure by analyzing its spectral projection in the subspace spanned by Perron-Frobenius eigenvectors. This selection could essentially avoid the complex valued eigenpairs in asymmetric adjacency matrices.

B. Networks with Disconnected Communities

Lemma 6: For an adjacency matrix A of a directed graph with K disconnected communities in the form of Equation (4). For $i = 1, \cdots, K$, the following results hold:

1) The K Perron-Frobenius eigenvalues λ_{Ci}s corresponding to communities C_is are real, positive, simple eigenvalues, and are also the eigenvalues of A.
2) Let $\boldsymbol{x}_{C_i}$ be the Perron-Frobenius eigenvectors of communities. The eigenvectors $\mathbf{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ of A corresponding to λ_{Ci}s are the only eigenvectors whose non-zero components are all positive, all the entries of $\mathbf{x}$ are real valued and have the following form:

$$(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_K) = \begin{pmatrix} \boldsymbol{x}_{C_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_{C_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_{C_K} \end{pmatrix}.$$

3) There is only one location of the row vector $\boldsymbol{\alpha}_{\boldsymbol{u}}$ that has a non-zero value with the form:

$$\boldsymbol{\alpha}_{\boldsymbol{u}} = (0, \cdots, 0, x_{iu}, 0, \cdots, 0). \tag{6}$$

The location of x_{iu} indicates the community which node u belongs to and the value of x_{iu} denotes the influence of node u to that community.

Since the matrix A is of diagonal block form, the eigenvectors of A will be of the the same form corresponding to each block and the eigenvalues of A will be the union of those of A_is. If we perform the spectral projection using this set of eigenvectors, nodes from different communities will form orthogonal lines.

IV. Perturbed Eigenspace

As discussed in Section II, the set of eigenvectors from directed graphs do not form orthonormal basis naturally. The perturbation theory, introduced in Lemma 2, requires the simple invariant subspace to produce a similarity reduction of the asymmetric adjacency matrix. Hence, in order to give explicit approximations explaining the spectral projection patterns observed, we need to find an unitary orthonormal basis that satisfies the conditions in Lemma 1 and Definition 1 to achieve the orthonormal reduction for a given asymmetric matrix. It is important to emphasize that such process is not needed for undirected graphs because the eigenvectors form orthonormal basis naturally.

A. Orhtonormal Basis Construction

The following proposition sets up such a basis by using the Gram-Schmidt process.

Proposition 1: Let $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$ be the eigenvalues for $\widetilde{A}$, and $\mathbf{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ be the eigenvectors. Assume that all the eigenvectors are linearly independent, and, without loss of generality, let $\sigma = (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_n)$ be an orthonormal basis formed by Gram-Schmidt process. Suppose that there exist K eigenvectors $\boldsymbol{x}_i = \boldsymbol{q}_i$ for $i \in (1, \cdots, n)$ that are part of this orthonormal basis and relabel their indices as $(\boldsymbol{q}_1, \cdots, \boldsymbol{q}_K)$ along with the corresponding eigenvalues. Denote $X = (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_K)$ and Q as the rest of the orthonrmal basis. If $\lambda_1, \cdots, \lambda_K$ are simple, then the following results hold:

1) $(X, Q)^H = (X, Q)^{-1}$ is unitary. $Q^H A X = 0$, thus $\Re(X)$ is a simple invariant subspace of A.
2) A can be reduced to a block triangular form:

$$(X, Q)^H A (X, Q) = \begin{pmatrix} L_1 & H \\ \mathbf{0} & L_2 \end{pmatrix}, \tag{7}$$

where $L_1 = X^H A X$, $L_2 = Q^H A Q$ is upper triangular, $AX = XL_1$ and $H = X^H A Q$. The eigenvalues of L_1 are the eigenvalues of A associated with $\Re(X)$. The rest eigenvalues of A are those of L_2.

Due to the space limit, we skip all proofs in this paper. Refer to [13] for details. By item 2 of Proposition 1, the orthornormal reduction results in an upper triangular matrix. In the symmetric case, the result is a diagonal matrix containing only eigenvectors, since the eigenvectors diagonalize the matrix. In the next section, we will give the approximations for the perturbed Perron-Frobenius eigenvectors corresponding to the K communities that span the K dimensional subspace. The approximations will be used to explain several phenomena in this particular subspace.

B. Approximation

When we treat the observed graph as the perturbed graph from Equation (4), we are able to 1) use Lemma 5, Lemma 6 and Proposition 1 to show the Perron-Frobenius eigenvectors $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ form a **simple invariant subspace**; and 2)

use the perturbation theory shown in Lemma 3 to derive the approximation of the perturbed Perron-Frobenius subspace.

Theorem 1: Let the observed graph be $\widetilde{A} = A + E$ with K communities and the perturbation E denotes the edges connecting communities $C_1, \cdots, C_K$. Assume that E satisfies the conditions in Lemma 3. Let $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ be the relabeled Perron-Frobenius eigenvectors of A for all communities, and Q be the rest of the orthronormal basis constructed using Proposition 1. Then $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ is a simple invariant subspace of A, and the perturbed Perron-Frobenius spectral space for $\widetilde{A}$ can be approximated as:

$$(\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_K) \approx (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K) + \nabla E(\frac{\boldsymbol{x}_1}{\lambda_1}, \cdots, \frac{\boldsymbol{x}_K}{\lambda_K}), \quad (8)$$

where $\nabla = Q(I - \frac{L_2}{\lambda_i})^{-1}Q^H$.

When spectral projection is performed on the subspace spanned by the eigenvectors corresponding to the K Perron-Frobenius eigenvalues, we can use Theorem 1 to derive the approximation of spectral coordinate of $\boldsymbol{\alpha}_u$ using the following simplified result that only takes into account the influences of neighboring nodes from other communities. Since the edge direction indicates the flow of information, we define the outer community neighbours of a node $u \in C_i$ to be any node $v \notin C_i$ that has an edge pointing to u.

Theorem 2: For node $u \in C_i$, let Γ_u^j denote its set of neighbors in C_j for $j \in (1, \cdots, K)$. The simplified spectral coordinates $\boldsymbol{\alpha}_u$ can be approximated as:

$$\boldsymbol{\alpha}_u \approx x_{iu} I_i + \left(\sum_{j=1}^n \nabla_{uj} \sum_{v \in \Gamma_u^1} \frac{e_{jv} x_{1v}}{\lambda_1}, \cdots, \sum_{j=1}^n \nabla_{uj} \sum_{v \in \Gamma_u^K} \frac{e_{jv} x_{Kv}}{\lambda_K}\right), \quad (9)$$

where I_i is the i-th row of a K-by-K identity matrix, e_{jv} is the (j, v) entry of E and ∇ is defined in Theorem 1.

The entry $\sum_{j=1}^n \nabla_{uj} \sum_{v \in \Gamma_u^i} \frac{e_{jv} x_{iv}}{\lambda_i}$ in the i-th column position of the spectral coordinate in Equation (9) is responsible for determining the influence of the perturbation to the current community members. For general perturbation, the perturbation could occur inside the community or even onto the node itself. Therefore, for the spectral coordinate of node u, this term will be 0 only when the perturbation does not appear on the column positions of E corresponding to the community which the node u belongs to. On the other hand, if perturbations occur inside the column positions of E corresponding to the community where the node u is, the values of $\boldsymbol{\alpha}_{iu} (\forall u \in C_i)$ will be altered. This phenomenon is reasonable, since all members in a community are strongly connected. Hence, the perturbation influence affects the entire community.

C. Inference

Before the perturbation, when the adjacency matrix A is of the diagonal block form, the second part of right hand side of Equation (9) will be 0, so nodes from the community C_i will lie on line I_i. Since $I_i \cdot I_m = 0$ for $i \neq m$, the nodes from different communities lie on different orthogonal lines. After the matrix is perturbed, suppose that the perturbation happens on the C_m region of the v-th column of E, then $E\boldsymbol{x}_i = 0$, since the C_m region of $\boldsymbol{x}_i = 0$ by Equation (6). Then the coordinates of all the nodes in C_i with respect to the two-dimensional subspace become $(\boldsymbol{x}_{iu}, \sum_{j=1}^n \nabla_{uj} \sum_{v \in \Gamma_u^m} \frac{e_{jv} x_{mv}}{\lambda_m})$ for $u \in C_i$. Likewise, the coordinates of the nodes in C_m are: $(0, \boldsymbol{x}_{mw} + \sum_{j=1}^n \nabla_{wj} \sum_{v \in \Gamma_u^m} \frac{e_{jv} x_{mv}}{\lambda_m})$ for $w \in C_m$. The dot products of any two rows are not 0, so the projections of nodes do not form strict orthogonal lines. Due to the sum of the product of scalar ∇_{ij} and the remaining terms, the spectral projections of all the nodes u of the same community C_i will deviate from the original line at different rates depending on the values in ∇.

We illustrate the above proposition with an example of a graph of two communities. Suppose nodes u and v are from community C_1 and C_2 respectively, the perturbation matrix E adds an edge from u to v as $u \to v$. Then the spectral coordinates for nodes u and v in the two dimensional space would be:

$$\begin{pmatrix} \boldsymbol{x}_{1u} & \nabla_{u1} \frac{e_{uv}}{\lambda_2} x_{2v} \\ \mathbf{0} & \boldsymbol{x}_{2v} + \nabla_{v1} \frac{e_{uv}}{\lambda_2} x_{2v} \end{pmatrix}. \quad (10)$$

D. Discussion

At the beginning of this paper, we made the assumption that communities are strongly connected components. However, this requirement is unrealistic in real-world applications because nodes within one community may not be strongly connected. Our theoretical result based on the matrix perturbation can be extended to a general case.

Theorem 3: Suppose that a community has a large strongly connected core and a small portion of leaf nodes which all have edges pointing to the members in the core. Let the leaf edges be a perturbation matrix E and treat the core as A. If the norm of E satisfies the conditions in Lemma 3, all the leaf nodes can be clustered according to their correlations with communities in the perturbed Perron-Frobenius spectral subspace.

Based on this theorem, we can replace the original assumption that all communities are strongly connected components with a weaker one: if all communities have strongly connected cores, all nodes can be assigned to communities based on the geometric relationships in the spectral space spanned by the perturbed Perron-Frobenius eigenvectors.

V. Algorithm

According to Theorem 1, we have $\boldsymbol{q}_i = \boldsymbol{x}_i (i = 1, \cdots, K)$, so $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ are part of the orthornormal basis. Then by Lemma 5 and Definition 1, $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K)$ is a simple invariant subspace. Hence, all the results derived for observed graphs can be generalized into the perturbed spectral space from the simple invariant subspace spanned by the Perron-Frobenius eigenvectors. By combining the results with the Perron-Frobenius Theorem, this particular simple invariant subspace has many unique properties: it is

real valued, values in each column vector are same signed with small or no incoming perturbations, and its dimension equals to the number of the communities.

With all the observations and theoretical results, a spectral clustering algorithm follows immediately, as shown in Algorithm 1. Our algorithm includes the following major steps: diagonalization of the adjacency matrix; normalization of the eigenvectors; selecting the initial set of eigenvectors with same signed components whose corresponding eigenvalues are real valued, positive and amongst the largest real positive eigenvalues of the adjacency matrix; projection of the nodes onto a unit sphere; clustering the nodes according to their location on the unit sphere using the classic k-means clustering algorithm; screen all the potential eigenpairs based on the modularity to find meaningful partitions.

As discussed previously, there are several factors that can affect the signs of the the components of the perturbed eigenvectors. Therefore, the initial set of eigenpairs may not include all the perturbed Perron-Frobenius eigenvectors. As a result, we need to search through all the real eigenvectors to select the ones that could increase the modularity. This process will cross validate all the newly added eigenvectors with the selected set of perturbed Perron-Frobenius ones. As a result, it would reduce the workload while avoiding producing a partition that deviates from the true structure by selecting non-Perron-Frobenius eigenvectors in the beginning. Since the communities in directed graphs are not defined by the density based criterion, maximizing modularity could no longer suffice as the objective. Ideally, we could use a combination of objective functions to determine the communities, but due to limited space, we will only test modularity with a tuning factor α. The rationale behind this approach is: although adding some eigenvector to partition the graph reduces the overall modularity by an insignificant amount, this partition could still be meaningful. As it turns out in our empirical evaluation, this approach can detect overlaps of communities if used properly.

Algorithm 1 *Aug_Adj*: Simple Invariant Subspace based Spectral Clustering for Directed Graphs

Input: A, τ, α
Output: clustering result CL

Compute Z, the set of eigenvectors of A corresponding to the largest τ real eigenvalues;
Normalize the eigenvectors $\bar{\boldsymbol{\alpha}}_u = \frac{\boldsymbol{\alpha}_u}{\|\boldsymbol{\alpha}_u\|}$;
$C \leftarrow$ eigenvectors from Z with same signed components;
$S \leftarrow Cardinality(C)$, $M \leftarrow 0$;
for each $c \in \emptyset \cup E \setminus C$ **do**
  Apply k-means algorithm on $C \cup c$ to get clustering result R;
  Compute the modularity score M_{temp};
  if $M_{temp} \geq \alpha M$ **then**
    $S \leftarrow S + 1$,$C \leftarrow C \cup c$, $CL \leftarrow R$, $M \leftarrow M_{temp}$;
  end if
end for
Return S and clustering result CL;

VI. Empirical Evaluation

In this evaluation, we mainly compare our algorithm *Aug_Adj* with several representative spectral clustering methods on synthetic data in terms of accuracy. We leave the evaluation of scalability and robustness on a large Twitter data set in [13]. All the data and code are available [1]. We compare with the random walk based Normalized cut (Ncut) [9], the random walk based Laplacian method (Lap) [7], the adjacency matrix based method using symmetrization (AdjCl) [14], and the SVD based method which works on the eigenspace associated with $A^H A$ and AA^H (SpokeEn) [15] on synthetic graphs under various conditions. Note that both Ncut and Lap are spectral clustering methods for directed graphs and the transition matrices used are based on the classic PageRank method. In our evaluation, we set the default damping factor 0.85 in the PageRank when calculating the transition matrices used in Ncut and Lap.

Table II
Synthetic Data Results

	Synth-1			Synth-2			Synth-3			Synth-4			Synth-5		
Method	Det	Acc	M	Det	Acc	M	Det	Acc	M	Det	Acc	M	Det	Acc	M
Lap	7	1.000	0.761	7	1.000	0.762	6	0.801	0.362	6	0.806	**0.390**	6	**0.806**	**0.356**
Ncut	7	1.000	0.761	7	1.000	0.762	7	0.995	0.419	6	0.802	0.384	6	**0.806**	**0.356**
SpokeEn	6	0.985	0.760	6	0.997	0.761	6	0.977	0.469	4	0.724	0.380	5	0.784	0.355
AdjCl	6	0.985	0.760	7	1.000	0.762	7	**1.000**	**0.472**	4	0.724	0.380	5	0.784	0.355
Aug_Adj	7	1.000	0.761	7	1.000	0.762	7	**1.000**	**0.472**	**7**	**0.997**	0.364	6	**0.806**	**0.356**

The synthetic graphs are generated based on 8 strongly connected components, $C_0, \cdots, C_7$, each with 18, 28, 74, 120, 194, 268, 240 and 314 nodes respectively. The densities (defined as the number of edges divided by the square of the number of nodes) of these components are: 0.4722, 0.4235, 0.3629, 0.3435, 0.3280, 0.3202, 0.3218, and 0.3178, respectively. We set $\alpha = 0.9$ in our *Aug_Adj* algorithm and set τ (the number of eigen-pairs to search for) as 10.

We generate five synthetic graphs, each of which is composed of 7 components. Synth-1 contains 7 isolated components $C_0, C_2, \cdots, C_7$ with no inter-cluster edges. Synth-2 contains 7 isolated components $C_1, \cdots, C_7$ with no inter-cluster edges. The difference between Synth-1 and Synth-2 is that we purposely include a tiny component with 18 nodes (C_0) in Synth-1, which is used to demonstrate the existence of eigen-gaps would not be a reliable criterion alone to determine the eigenvectors used for clustering. Synth-3 is generated by adding inter-community edges with probabilities 0.1 between all pairs of components for both incoming and outgoing directions to Synth-2. Synth-4 is generated by increasing the probability between C_6 and C_7 to 0.27 whereas Synth-5 is generated by increasing the

[1] https://github.com/gnemeuyil/Aug_ADJ

probability between C_6 and C_7 to 0.5. The clustering results are shown in Table II, where Det indicates the number of clusters detected, M is the modularity, and Acc is the accuracy.

For Synth-1, we find that by using the naive symmetrization with AdjCl algorithm, only 6 clusters are detected. This is because the spectral radius corresponding to the perturbed Perron-Frobenius eigenvector of C_0 does not fall in the range of the largest 10 eigenvalues. If τ is increased to 15, the algorithm detects this eigen-pair and then can correctly identify 7 clusters. SpokeEn can only detect 6 clusters even if $\tau = 50$, which indicates that it is more susceptible to noises in the spectral space. To test our hypothesis, we increase the size of the smallest component in Synth-2. We can see AdjCl successfully identifies the correct number of components and assigns the corresponding nodes correctly to their components for Synth-2 with $\tau = 10$. However, SpokeEn fails again to detect 7 communities in this setting even if τ is increased to 50.

For Synth-3, the results are very stable for all the methods with different symmetrization techniques. In this setup, the densities of inter-community edges are smaller than those of inner-community edges, so most algorithms can find 7 components. The Lap and SpokeEn detect 6 components. It is possible that the weighted symmetrization assigned some boundary nodes to incorrect clusters. Adjacency based methods outperform the other methods due to correct selection of eigenvectors for a well conditioned adjacency matrix.

For Synth-4, components C_6 and C_7 are on the verge of merging together. Our *Aug_Adj* algorithm identifies 7 clusters and assigns most nodes correctly. For Synth-5, components C_6 and C_7 are merged together due to dense inter-cluster edges. Lap, Ncut and *Aug_Adj* correctly report 6 components while the other methods report 5.

From the above results, we have some consistent observations: methods based on matrix transformations tend to introduce both redundant information and noises that will cause graph partitions to be inaccurate; when the community sizes are not well balanced, naive symmetrization could fail to detect small communities in the adjacency eigenspace. These observations coincide with our discussions in previous sections. The results for Synth-4 lead to the speculation that the down tuned significance requirement for objective functions could lead our method to detect certain hidden structures of components. This could potentially be useful for studying micro-structures of components or overlapping problems of communities.

VII. Conclusion and Future Work

In this research, the properties of the adjacency eigenspaces of directed graphs were studied. We started our work by learning from the observations in the Perron-Frobenius eigenspaces. Then we began the theoretical work from networks with disconnected communities by making the assumption that each community should be a strongly connected component. By using the matrix perturbation theory, we constructed an orthonormal basis containing the Perron-Frobenius eigenvectors corresponding to all communities to achieve the orthornormal reduction of the adjacency matrices of directed graphs and described mathematically how the projections of nodes would behave in the perturbed Perron-Frobenius simple invariant subspace of an observed graph. Then, we extended our theories by replacing the original assumption of community structures with a weaker one that only requires a community to have a strongly connected core component, so that they can be used to study the networks without clear community structures. A spectral clustering algorithm was developed and compared with other representative methods within the same domain on various synthetic data sets.

For future work, the theories and algorithms will be extended to analyze various graph related problems including but not limited to: studying the microstructure of a community, analyzing the changes of the macrostructure of a network, and detecting network anomalies.

References

[1] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks:a survey," *Physics Reports*, 553, pp. 95–142, 2013.

[2] V. Satuluri and S. Parthasarathy, "Symmetrizations for clustering directed graphs," in *EDBT* 2011, pp. 343–354.

[3] D. Zhou, B. Schölkopf, and T. Hofmann, "Semi-supervised learning on directed graphs," in *NIPS*, 2005.

[4] E. A. Leicht and M. E. Newman, "Community structure in directed networks," *Physical review letters*, 100(11):118703, 2008.

[5] Y. Kim, S.-W. Son, and H. Jeong, "Finding communities in directed networks," *Physical Review E*, 81(1):016103, 2010.

[6] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics*, vol. 03, p. 3024, 2009.

[7] M. Meila and W. Pentney, "Clustering by weighted cuts in directed graphs," in *SDM*, 2007.

[8] F. Chung, "Laplacians and the cheeger inequality for directed graphs," *Annals of Combinatorics*, 9(1):1-19, 2005.

[9] D. Zhou, J. Huang, and B. Schölkopf, "Learning from labeled and unlabeled data on a directed graph," in *ICML*, 2005.

[10] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*. Academic Press, 1990.

[11] R. S. Varga, *Matrix Iterative Analysis*.Springer, 2009.

[12] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[13] Y. Li, X. Wu, and A. Lu, "Analysis of Spectral Space Properties of Directed Graphs using Matrix Perturbation Theory with Application in Graph Partition," *Technical Report, DPL-2014-002*, University of Arkansas, 2014.

[14] L. Wu, X. Ying, X. Wu, and Z.-H. Zhou, "Line orthogonality in adjacency eigenspace with application to community partition," in *IJCAI*, 2011.

[15] B. A. Prakash, A. Sridharan, M. Seshadri, S. Machiraju, and C. Faloutsos, "Eigenspokes: Surprising patterns and scalable community chipping in large graphs," in *PAKDD*, 2010.