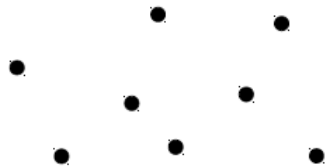
Active Semi-Supervised Learning Using Sampling Theory for Graph Signals (KDD2014)

2019.4.26

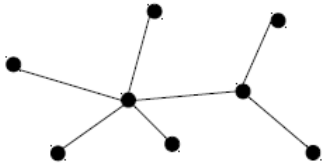
Motivation and Problem Definition

- Unlabeled data is abundant. Labeled data is expensive and scarce.
- Solution: **Active Semi-supervised Learning** (SSL).

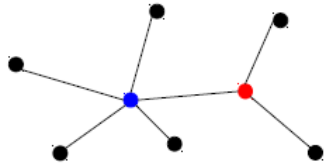
- **Problem setting**: Offline, pool-based, batch-mode active SSL via graphs



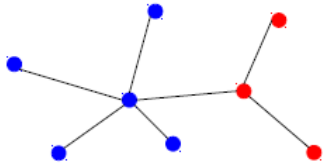
Data points in feature space



Construct similarity graph



Choose points to label



Predict labels for the rest

1. *How to predict unknown labels from the known labels?*
2. *What is the optimal set of nodes to label given the learning algorithm?*

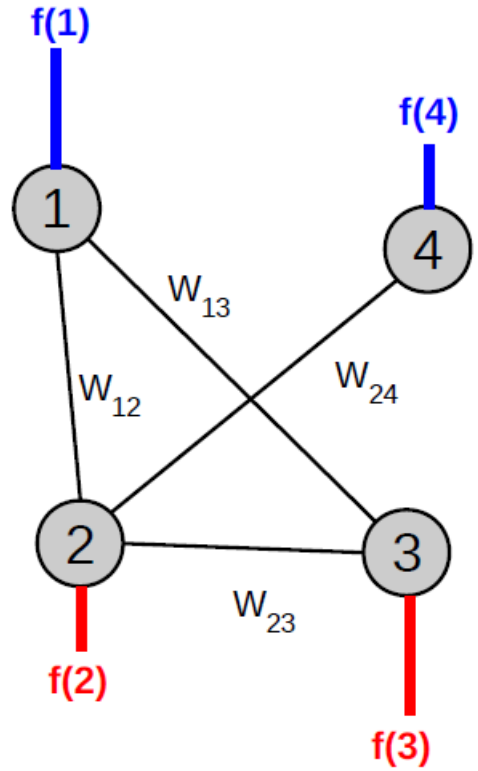
Graph Signal Processing

- **Graph** $G = (\mathcal{V}, \mathcal{E})$ with N nodes
- nodes $\equiv$ data points; w_{ij}: similarity between i and j.

- Adjacency matrix $\mathbf{W} = [w_{ij}]_{n \times n}$.
- Degree matrix $\mathbf{D} = \text{diag}\{\sum_j w_{ij}\}$.
- Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
- Normalized Laplacian $\boldsymbol{\mathcal{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.

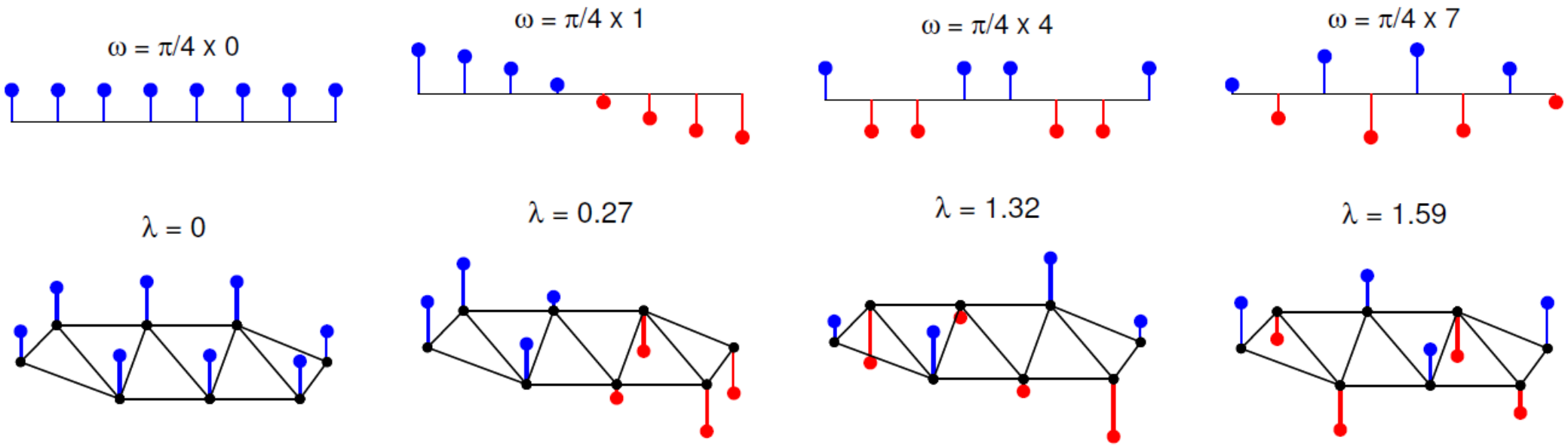
- **Graph signal** $f : \mathcal{V} \to \mathbb{R}$, denoted as $\mathbf{f} \in \mathbb{R}^N$.
- Class membership functions are graph signals.

$$\mathbf{f}^c(j) = \begin{cases} 1, & \text{if node } j \text{ is in class } c \\ 0, & \text{otherwise} \end{cases}$$



Spectrum of $\mathcal{L}$ provides frequency interpretation:

- $\lambda_k \in [0, 2]$: *graph frequencies*.
- $\mathbf{u}_k$: *graph Fourier basis*.



- *Fourier coefficients* of $\mathbf{f}$: $\tilde{\mathbf{f}}(\lambda_i) = \langle \mathbf{f}, \mathbf{u}_i \rangle$.
- *Graph Fourier Transform* (GFT):

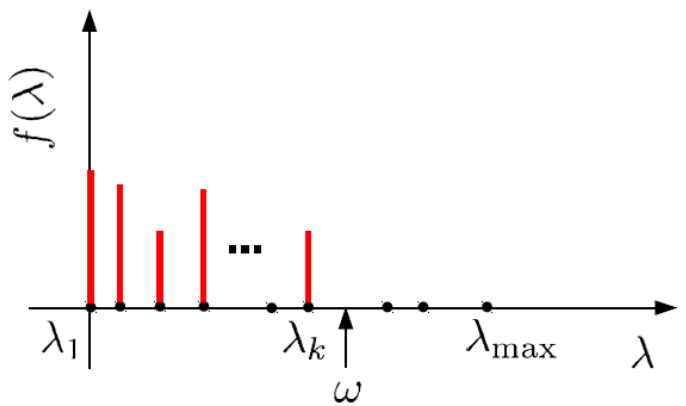
$$\tilde{\mathbf{f}} = \mathbf{U}^T \mathbf{f}.$$

A smooth or low-pass graph signal can be obtained by forcing high frequency GFT coefficients to vanish, *w-bandlimited signal* on a graph have zero GFT coefficients for frequencies above its bandwidth *w*.

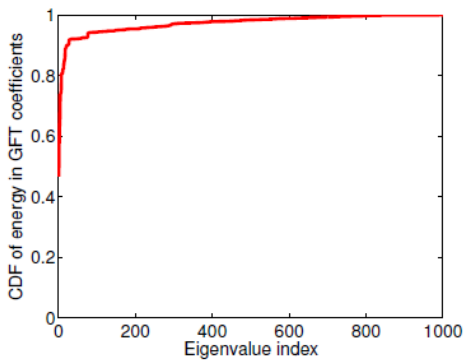
- ω-**bandlimited signal**: GFT has support $[0, \omega]$.
- **Paley-Wiener space** $PW_\omega(G)$: Space of all ω-bandlimited signals.
  - $PW_\omega(G)$ is a subspace of $\mathbb{R}^N$.
  - $\omega_1 \leq \omega_2 \Rightarrow PW_{\omega_1}(G) \subseteq PW_{\omega_2}(G)$.

- **Bandwidth of a signal:**

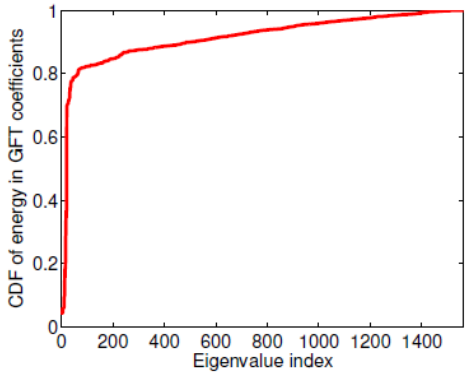
$$\omega(\mathbf{f}) = \arg\max_{\lambda} \tilde{\mathbf{f}}(\lambda) \text{ s.t. } |\tilde{\mathbf{f}}(\lambda)| \geq 0$$



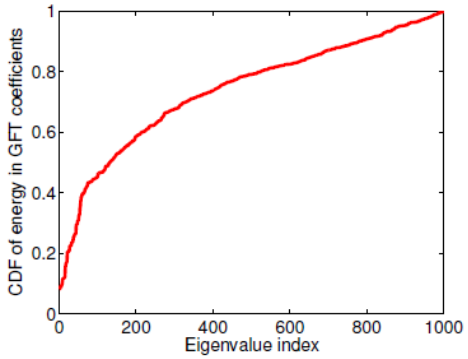
- *Class membership functions can be approximated by bandlimited graph signals.*



(a) USPS



(b) Isolet



(c) 20 newsgroups

P1:Cut-off frequency

$L_2(S^c)$ denote the space of all graph signals that are zero everywhere except possibly on the nodes in S^c,i.e., $\forall \phi \in L_2(S^c), \phi(S) = 0$. Also, let $\omega(\phi)$ denote the bandwidth of a graph signal ϕ, i.e., the value of the maximum non-zero frequency of that signal.

Sampling theorem. For a graph G, with normalized Laplacian L, any signal $\mathbf{f} \in PW_\omega(G)$ can be perfectly recovered from its values on a subset of nodes $S \subset V$ if and only if

$$\omega < \omega_c(S) \triangleq \inf_{\phi \in L_2(S^c)} \omega(\phi)$$

Intuitively, a signal $\phi \in L_2(S^c)$ can be added to any input signal $\mathbf{f}$ without affecting its sampled version (since ϕ is identically zero for all vertices that are sampled, i.e., those in S).

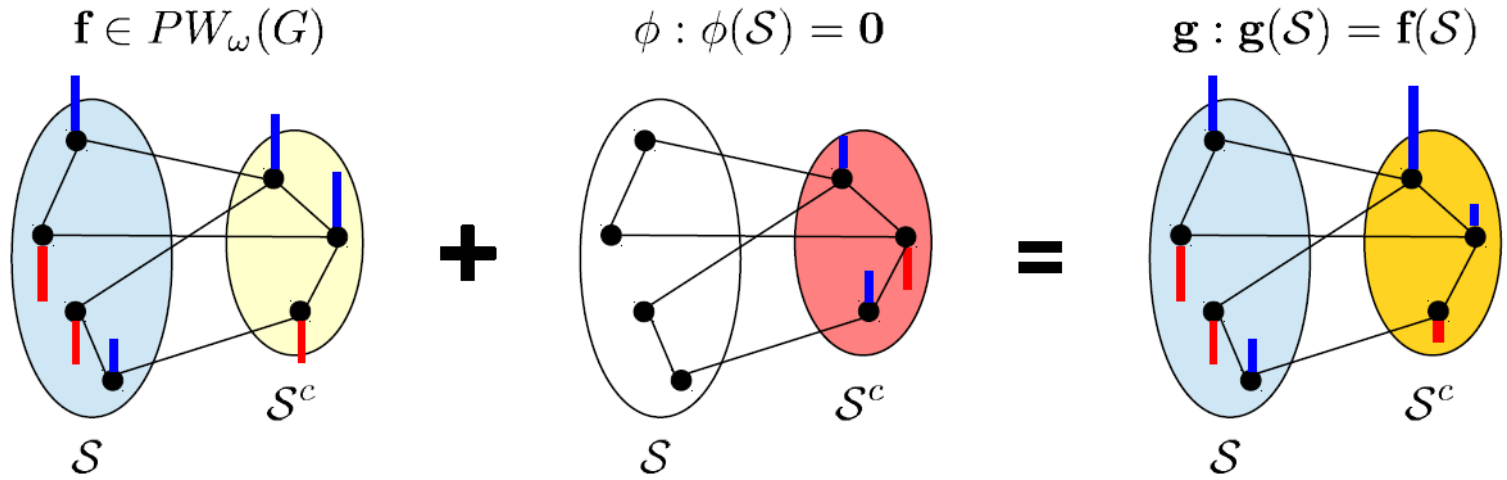
Thus, if there existed a $\phi \in L_2(S^c)$ such that $\phi \in PW_\omega(G)$ we would have that both $\mathbf{f}$ and $\phi + \mathbf{f}$ belong to $PW_\omega(G)$ and lead to the same set of samples on S. So clearly it would not be possible to recover them both, and thus sampling of such signals in $PW_\omega(G)$ would not be possible. The conditions in **sampling theorem** ensures that $PW_\omega(G) \cap L_2(S^c) = \{0\}$ and no such ϕ exists.

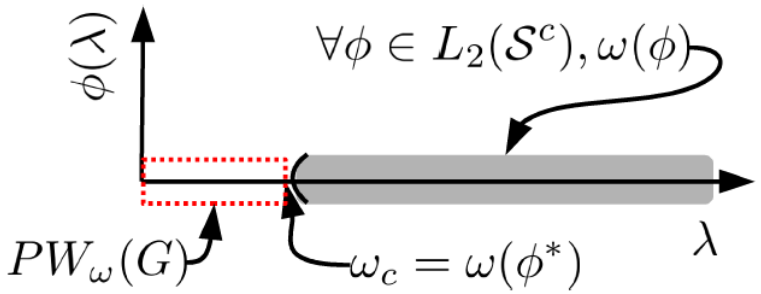
Approximate the bandwidth of any signal ϕ for a given integer parameter $k > 0$ as follows

$$\omega_k(\phi) = \left(\frac{\phi^t L^k \phi}{\phi^t \phi}\right)^{1/k} \qquad \Omega_k(S) = \inf_{\phi \in L_2(S^c)} \omega_k(\phi) = \inf_{\phi \in L_2(S^c)} \left(\frac{\phi^t L^k \phi}{\phi^t \phi}\right)^{1/k}$$

Numerically, $\Omega_k(S)$ and ϕ_k^* can be determined from the smallest eigenpair $(\sigma_{1,k}, \psi_{1,k})$ of the reduced matrix $(L^k)_{S^c}$:

$$\Omega_k(S) = \sigma_{1,k}, \phi_k^*(S^c) = \psi_{1,k}\ \phi_k^*(S) = 0.$$





Sampling Theorem

$\mathbf{f}$ can be perfectly recovered from $\mathbf{f}(\mathcal{S})$ iff

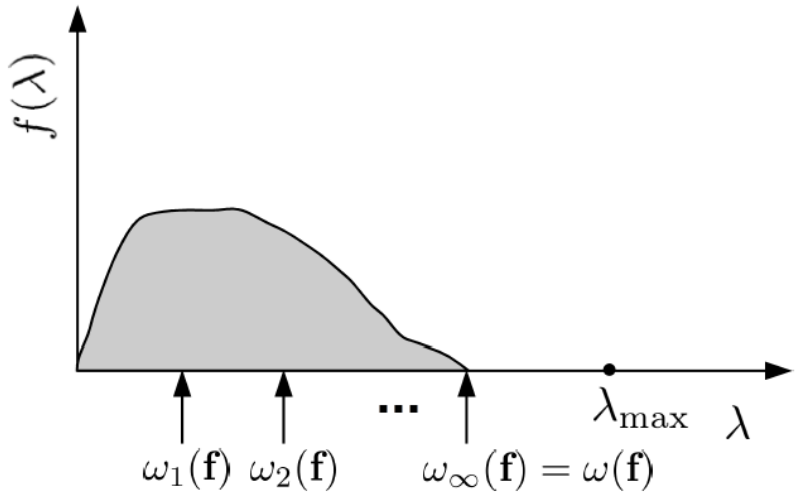
$$\omega(\mathbf{f}) \leq \omega_c(\mathcal{S}) \triangleq \inf_{\phi_{L_2(\mathcal{S}^c)}} \omega(\phi)$$

Condition for unique sampling of $PW_\omega(G)$ on $\mathcal{S}$

Let $L_2(\mathcal{S}^c) = \{\phi : \phi(\mathcal{S}) = \mathbf{0}\}$. Then, we need $PW_\omega(G) \cap L_2(\mathcal{S}^c) = \{\mathbf{0}\}$.

Approximate bandwidth of a signal

$$\omega_k(\mathbf{f}) \triangleq \left(\frac{\mathbf{f}^\top \boldsymbol{\mathcal{L}}^k \mathbf{f}}{\mathbf{f}^\top \mathbf{f}}\right)^{1/k}, \text{ where } k \in \mathbb{Z}^+$$



- *Monotonicity:* $\forall \mathbf{f}, k_1 < k_2 \Rightarrow \omega_{k_1}(\mathbf{f}) \leq \omega_{k_2}(\mathbf{f})$.
- *Convergence:* $\lim_{k\to\infty} \omega_k(\mathbf{f}) = \omega(\mathbf{f})$.

Minimize approximate bandwidth over $L_2(\mathcal{S}^c)$ to estimate cut-off frequency

$$\Omega_k(\mathcal{S}) \triangleq \min_{\phi \in L_2(\mathcal{S}^c)} \omega_k(\phi) = \min_{\phi:\phi(\mathcal{S})=\mathbf{0}} \left(\frac{\phi^T \boldsymbol{\mathcal{L}}^k \phi}{\phi^T \phi}\right)^{1/k} = \left(\min_{\psi} \underbrace{\frac{\psi^T (\boldsymbol{\mathcal{L}}^k)_{\mathcal{S}^c} \psi}{\psi^T \psi}}_{\text{Rayleigh quotient}}\right)^{1/k}$$

Let $\{\sigma_{1,k}, \psi_{1,k}\} \rightarrow$ smallest eigen-pair of $(\boldsymbol{\mathcal{L}}^k)_{\mathcal{S}^c}$.

$$\text{Estimated cutoff frequency } \Omega_k(\mathcal{S}) = (\sigma_{1,k})^{1/k},$$

$$\text{Corresponding smoothest signal } \phi_k^{\text{opt}}(\mathcal{S}^c) = \psi_{1,k},\ \phi_k^{\text{opt}}(\mathcal{S}) = \mathbf{0}.$$

P2:Sampling set

- Optimal sampling set should maximally capture signal information.
- $\mathcal{S}_{\text{opt}} = \arg\max_{|\mathcal{S}|=m} \Omega_k(\mathcal{S}) \rightarrow$ combinatorial!
- Greedy gradient-based approach.
  - Start with $\mathcal{S} = \{\emptyset\}$.
  - Add nodes one by one while ensuring maximum increase in $\Omega_k(\mathcal{S})$.

The hope is that $\Omega_k(S)$ reaches the target cut-off ω_c with minimum number of node additions to S.
To understand which nodes should be included in S, we introduce a binary relaxation of out cut-off formulation by defining the following matrix

$$\mathbf{M}_k^{\alpha}(\mathbf{t}) \stackrel{\triangle}{=} \boldsymbol{\mathcal{L}}^k + \alpha\, \boldsymbol{\mathcal{D}}(\mathbf{t}), \quad k \in \mathbb{Z}^+, \alpha > 0, \mathbf{t} \in \mathbb{R}^N$$

$$\mathbf{M}_k^{\alpha}(\mathbf{t}) \triangleq \boldsymbol{\mathcal{L}}^k + \alpha\, \boldsymbol{\mathcal{D}}(\mathbf{t}), \quad k \in \mathbb{Z}^+, \alpha > 0, \mathbf{t} \in \mathbb{R}^N$$

$\boldsymbol{\mathcal{D}}(\mathbf{t})$ is a diagonal matrix with $\mathbf{t}$ on its diagonal. $(\lambda_k^{\alpha}(\mathbf{t}), \mathbf{x}_k^{\alpha}(\mathbf{t}))$ denote the smallest eigen-pair of $\mathbf{M}_k^{\alpha}(\mathbf{t})$

$$(\Omega_k(\mathcal{S}))^k = \min_{\phi(\mathcal{S})=\mathbf{0}} \frac{\phi^\top \boldsymbol{\mathcal{L}}^k \phi}{\phi^\top \phi} \approx \min_{\mathbf{x}} \left(\frac{\mathbf{x}^\top \boldsymbol{\mathcal{L}}^k \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} + \alpha \frac{\mathbf{x}^\top \mathrm{diag}(\mathbf{t}) \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \right) \Bigg|_{\mathbf{t}=\mathbf{1}_{\mathcal{S}}} = \lambda_k^{\alpha}(\mathbf{t})|_{\mathbf{t}=\mathbf{1}_{\mathcal{S}}}$$

binary relaxation

relax the constraint

If $\mathbf{1}_s: V \to \{0,1\}$ denotes the indicator function for the subset $S(i.e.\ \mathbf{1}(S) = \mathbf{1}\ and\ \mathbf{1}(S^c) = \mathbf{0})$

$$\lambda_k^{\alpha}(\mathbf{1}_{\mathcal{S}}) = \inf_{\mathbf{x}} \left(\frac{\mathbf{x}^t \boldsymbol{\mathcal{L}}^k \mathbf{x}}{\mathbf{x}^t \mathbf{x}} + \alpha \frac{\mathbf{x}(\mathcal{S})^t \mathbf{x}(\mathcal{S})}{\mathbf{x}^t \mathbf{x}} \right)$$

When $\alpha \gg 1$, the compoents $\mathbf{x}(S)$ are highly penalized during minimization. If $x_k^{\alpha}(\mathbf{1}_S)$ is the minimizer, then $[x_k^{\alpha}(\mathbf{1}_S)](S) \to 0$, i.e. the values on nodes S tend to be very small.

$$\phi_k^* \approx \mathbf{x}_k^{\alpha}(\mathbf{1}_{\mathcal{S}}), \quad (\Omega_k(\mathcal{S}))^k \approx \lambda_k^{\alpha}(\mathbf{1}_{\mathcal{S}}) \qquad \left.\frac{d\lambda_{\alpha}^k(\mathbf{t})}{d\mathbf{t}(i)}\right|_{\mathbf{t}=\mathbf{1}_{\mathcal{S}}} = \alpha\left([\mathbf{x}_{\alpha}^k(\mathbf{1}_{\mathcal{S}})](i)\right)^2 \approx \alpha(\boldsymbol{\phi}_k^*(i))^2$$

Start with an empty $S(\mathbf{1}_s = \mathbf{0})$, if at each step, include the node on which the smoothest signal $\phi_k^* \in L_2(S^c)$ has maximum energy(i.e., $\mathbf{1}_s(i) \leftarrow 1, \mathrm{i} = \max_j[(\phi_k^*(j))^2]$), then the cut-off estimate $\Omega_k(S)$ tends to increase maximally.

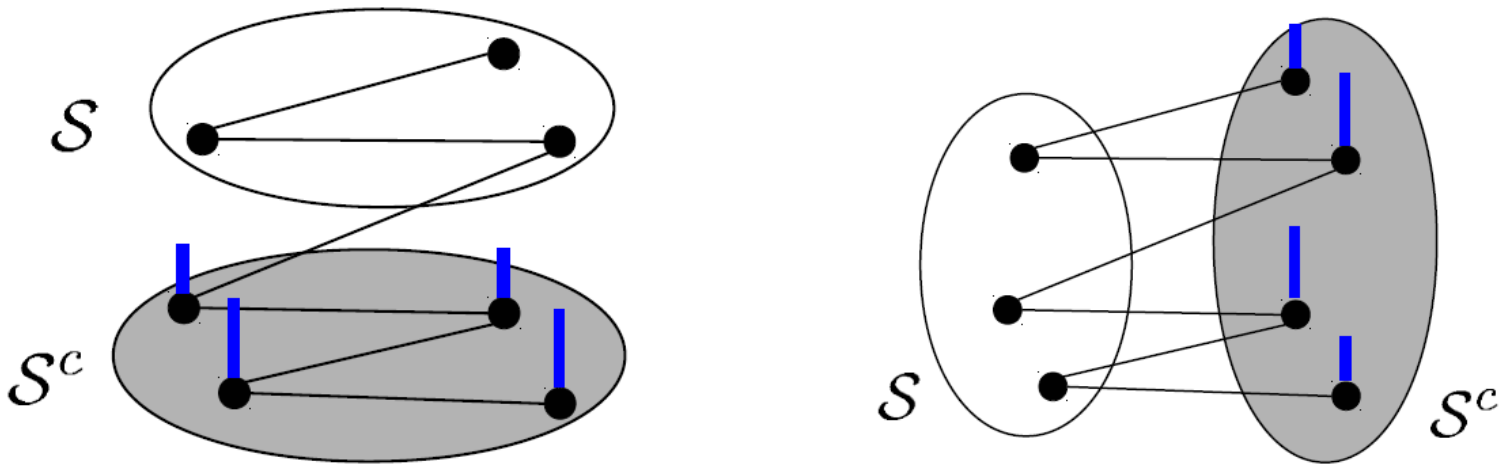
Algorithm 1 Greedy heuristic for finding $\mathcal{S}_L^*$

Input: $G = \{\mathcal{V}, E\}$, $\boldsymbol{\mathcal{L}}$, target size m, parameter $k \in \mathbb{Z}^+$.
Initialize: $\mathcal{S} = \{\emptyset\}$.

1: **while** $|\mathcal{S}| \leq m$ **do**
2: For $\mathcal{S}$, compute the smoothest signal $\phi_k^* \in L_2(\mathcal{S}^c)$ using (4) and (5).
3: $v \leftarrow \arg\max_i \left[(\phi_k^*(i))^2\right]$.
4: $\mathcal{S} \leftarrow \mathcal{S} \cup v$.
5: **end while**
6: $\mathcal{S}_L^* \leftarrow \mathcal{S}$.

$$\Omega_k(\mathcal{S}) = \sigma_{1,k},$$
$$\phi_k^*(\mathcal{S}^c) = \psi_{1,k}, \ \phi_k^*(\mathcal{S}) = \mathbf{0}.$$

- Cut-off function $\Omega_k(\mathcal{S}) \equiv$ variation of smoothest signal in $L_2(\mathcal{S}^c)$.
- Larger cut-off function $\Rightarrow$ more variation in $\phi_{\text{opt}} \Rightarrow$ more cross-links.



Intuition

Unlabeled nodes are strongly connected to labeled nodes!

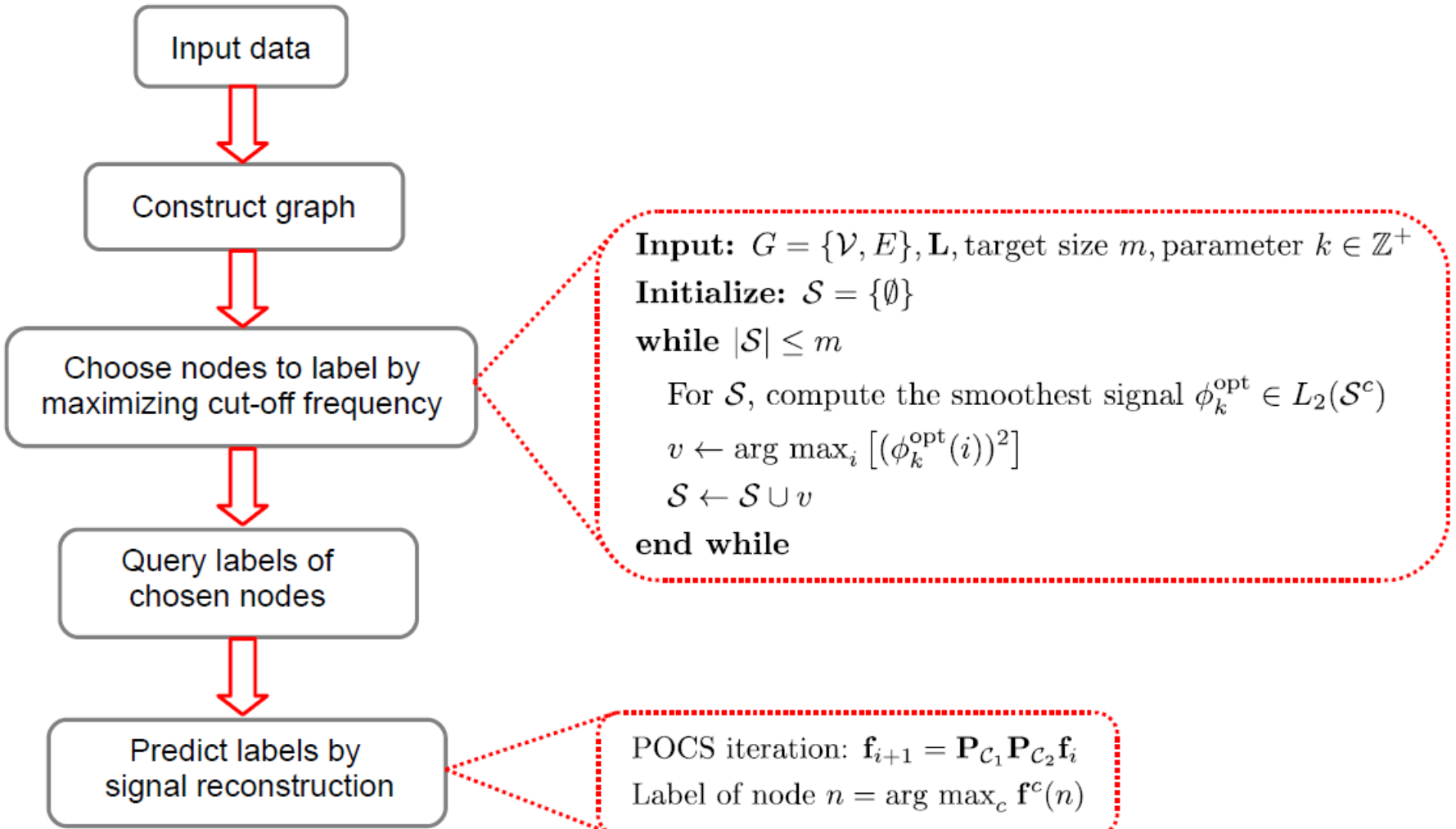
P3:Reconstruction

A graph signal $\mathbf{f} \in PW_{\omega}(G)$ can be written as a linear combination eigenvecctors of L with eigenvalues less than ω,

$$\hat{\mathbf{f}}(\mathcal{S}^c) = \mathbf{U}_{\mathcal{S}^c,\mathcal{K}}\boldsymbol{\alpha}^* \text{ where, } \boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{U}_{\mathcal{S},\mathcal{K}}\boldsymbol{\alpha} - \mathbf{f}(\mathcal{S})\|$$

$\mathcal{K}$ is the index of eigenvectors with eigenvalues less than the cut-off $\omega_c(S)$

If the true signal $\mathbf{f} \in PW_{\omega}(G)$, then the prediction is perfect. However, this is not the case in most problems, the prediction error $\|\mathbf{f} - \hat{\mathbf{f}}\|$ roughly equals the portion of energy of the true signal in $[\omega_c(S), \lambda_N]$ frequency band.


Input data
Construct graph
Choose nodes to label by maximizing cut-off frequency
Query labels of chosen nodes
Predict labels by signal reconstruction
Input: $G = \{\mathcal{V}, E\}, \mathbf{L}$, target size m, parameter $k \in \mathbb{Z}^+$
Initialize: $\mathcal{S} = \{\emptyset\}$
while $|\mathcal{S}| \leq m$
For $\mathcal{S}$, compute the smoothest signal $\phi_k^{\text{opt}} \in L_2(\mathcal{S}^c)$
$v \leftarrow \arg\max_i \left[(\phi_k^{\text{opt}}(i))^2\right]$
$\mathcal{S} \leftarrow \mathcal{S} \cup v$
end while
POCS iteration: $\mathbf{f}_{i+1} = \mathbf{P}_{\mathcal{C}_1}\mathbf{P}_{\mathcal{C}_2}\mathbf{f}_i$
Label of node $n = \arg\max_c \mathbf{f}^c(n)$

Graph Theoretic Interpretation

Minimal Dirichlet energy partitions for graphs(2014)

A new measure of optimality for graph partitions, based on the sum of the Dirichlet eigenvalues of the partition components

$$\min_{V=\amalg_{i=1}^{k} V_i} \sum_{i=1}^{k} \lambda(V_i)$$

Define the Dirichlet energy of a subset $S \subset V$

$$\lambda(S) := \inf_{\substack{\|\psi\|_V=1 \\ \psi|_{S^c}=0}} \|\nabla\psi\|_{w,E}^2$$

$$\|\nabla\psi\|_{w,E}^2 := \sum_{(i,j)\in E} w_{ij}(\psi_i - \psi_j)^2 \qquad \|\psi\|_S^2 := \sum_{i\in S} d_i^r \psi_i^2 \qquad d_i := \sum_j w_{ij} \qquad r \in [0,1]$$

If $\psi = \chi_S$, then $\|\nabla_\psi\|_{w,E}^2$ simply reduces to $\sum_{i\in S, j\in S^c} w_{i,j} \equiv |\partial S|$, implying that **measures variations across the boundary of** S. $\lambda(S)$ is a measure of the connectedness of S that takes into both interior similarity as well as similarity to the rest of the graph.

$\lambda(S)$ satisfies the following Dirichlet eigenvalue problem in S, for some corresponding eigenvector,$\psi = \psi(S)$.

$$\begin{aligned} \Delta_r \psi &= \lambda \psi \qquad \text{on } S \subset V \\ \psi &= 0 \qquad\quad \text{on } S^c. \end{aligned}$$

$$\Delta_r := D^{-r}(D - W)$$

This expression appears more commonly as part of discrete Dirichlet eigenvalue problems on graphs. Specially, it is equal to the Dirichlet energy of the subset S^c.

$$\Omega_1(\mathcal{S}) = \inf_{\substack{\mathbf{x}(\mathcal{S})=\mathbf{0} \\ ||\mathbf{x}||=1}} \mathbf{x}^t \boldsymbol{\mathcal{L}} \mathbf{x}$$

Expand the objective function for any $\mathbf{x}$ with constraint $\mathbf{x}(S) = \mathbf{0}$

$$\begin{aligned} \mathbf{x}^t \boldsymbol{\mathcal{L}} \mathbf{x} &= \sum_{i \sim j} w_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 \\ &= \sum_{\substack{i \sim j \\ i \in \mathcal{S}, j \in \mathcal{S}^c}} w_{ij} \left(\frac{x_j^2}{d_j} \right) + \sum_{\substack{i \sim j \\ i,j \in \mathcal{S}^c}} w_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 \end{aligned}$$

$$\mathbf{x}^t \boldsymbol{\mathcal{L}} \mathbf{x} = \sum_{i \sim j} w_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2$$

$$= \sum_{\substack{i \sim j \\ i \in \mathcal{S}, j \in \mathcal{S}^c}} w_{ij} \left(\frac{x_j^2}{d_j} \right) + \sum_{\substack{i \sim j \\ i, j \in \mathcal{S}^c}} w_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2$$

$$\mathbf{x}^t \boldsymbol{\mathcal{L}} \mathbf{x} \approx \sum_{j \in \mathcal{S}^c} \left(\frac{p_j}{d_j} \right) x_j^2$$

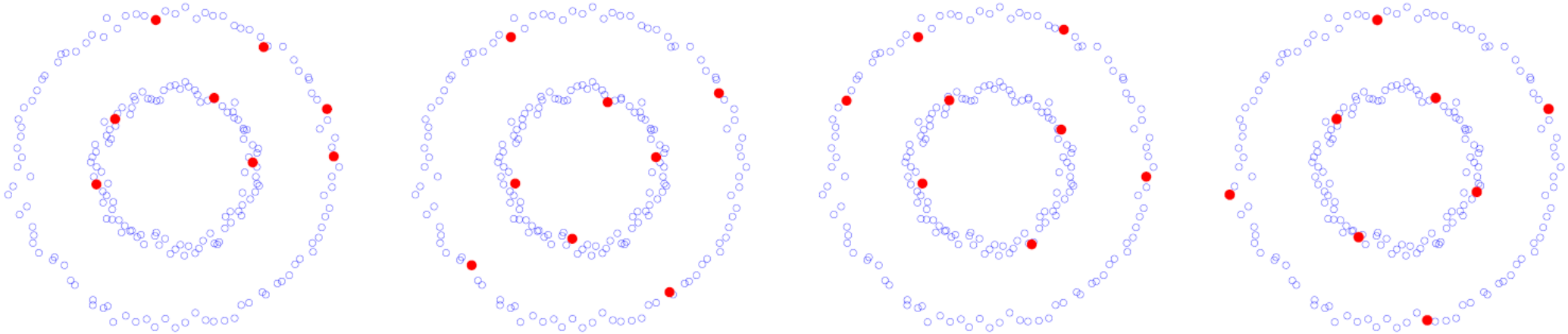
$$p_j = \sum_{i \in \mathcal{S}} w_{ij}$$

Therefore, give a current selected S, the greedy algorithm selects the next node that maximizes the increase in

$$\Omega_1(\mathcal{S}) \approx \inf_{||\mathbf{x}||=1} \sum_{j \in \mathcal{S}^c} \left(\frac{p_j}{d_j} \right) x_j^2$$

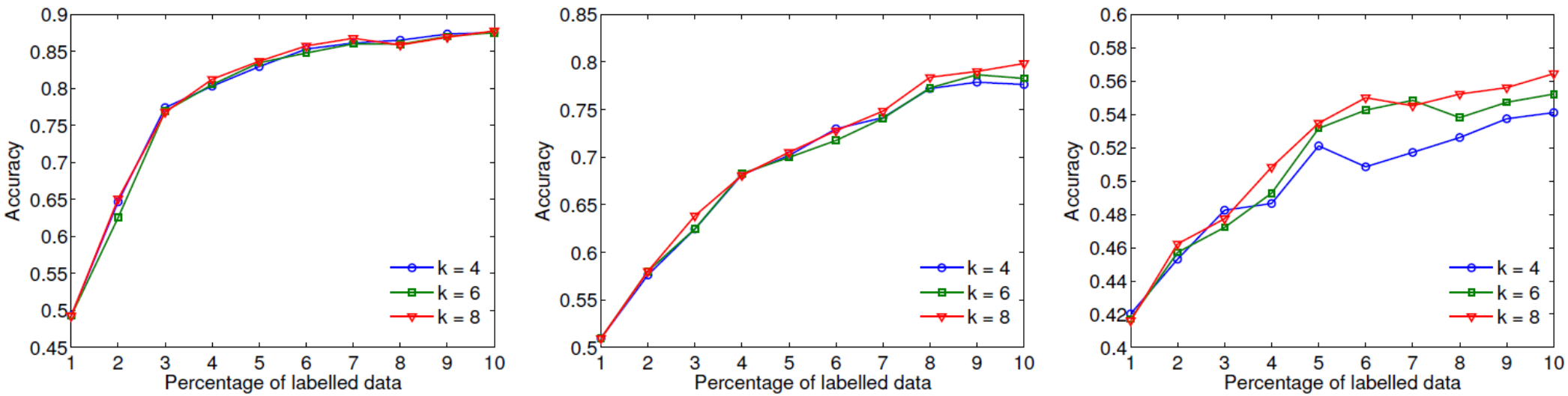
Due to the constraint $\|\mathbf{x}\| = 1$, the expression being minimized is essentially an infimum over a convex combination of the fractional out-degrees and its value is largely determined by nodes $j \in S^c$ for which p_j/d_j is small.

Thus, in the simplest case, out selection algorithm tries to remove those nodes from the unlabeled set that are weakly connected to nodes in the labeled set.



(a) Ψ-max (b) LLR (c) LLGC Bound (d) Proposed

Figure 4: Toy example comparing the nodes selected using different active learning methods



(a) USPS (b) Isolet (c) 20 newsgroups