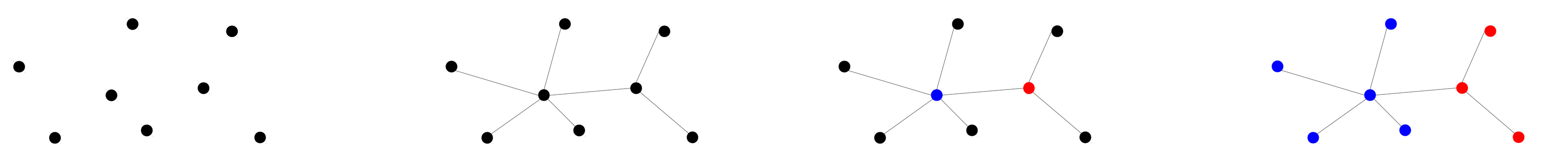
Active Semi-supervised Learning Using Sampling Theory for Graph Signals

Akshay Gadde, Aamir Anis and Antonio Ortega
University of Southern California, Los Angeles
{agadde, aanis}@usc.edu, ortega@sipi.usc.edu

USC University of Southern California

Motivation and Problem Definition

- Unlabeled data is abundant. Labeled data is expensive and scarce.
- **Problem setting**: Pool-based, batch-mode active SSL via graphs.



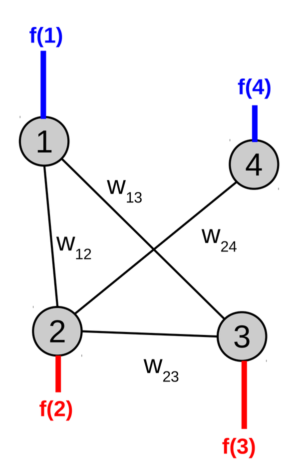
1. *How to predict unknown labels from known labels?*
2. *What is the optimal set of nodes to label, given the learning algorithm?*

Key Ideas

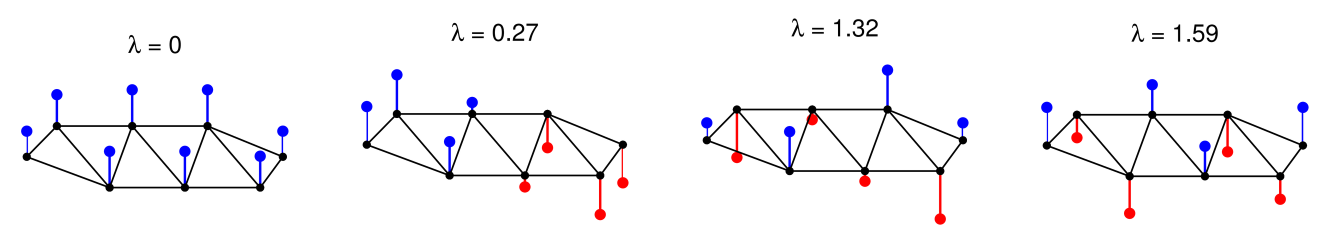
- Class labels $\Longrightarrow$ *Smooth or bandlimited graph signals.*
- Choosing nodes to label $\Longrightarrow$ *Best sampling set selection.*
- Predicting unknown labels $\Longrightarrow$ *Signal reconstruction from samples.*

Background: Graph Signal Processing

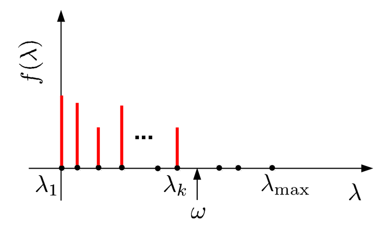
- **Graph** $G = (\mathcal{V}, \mathcal{E})$ with n nodes
  - nodes $\Rightarrow$ data points.
  - w_{ij}: similarity between points i and j.
  - Adjacency matrix $\mathbf{W} = [w_{ij}]_{n\times n}$.
  - Degree matrix $\mathbf{D} = \text{diag}\{\sum_j w_{ij}\}$.
  - Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
  - Normalized Laplacian $\boldsymbol{\mathcal{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.



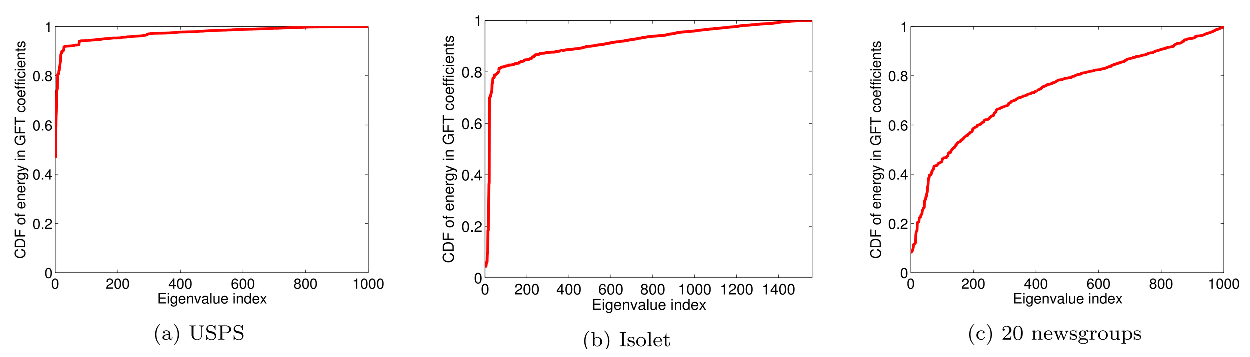
- **Graph signal** $f: \mathcal{V} \to \mathbb{R}$, denoted as $\mathbf{f} \in \mathbb{R}^n$.
  - Membership function $\mathbf{f}_i$: $\mathbf{f}_i(j) = 1 \Rightarrow$ node j belongs to class i.
- Spectrum of $\boldsymbol{\mathcal{L}} \Rightarrow$ spectral representation for graph signals.
  - Frequencies: $\{\lambda_k\} \in [0, 2]$; Fourier basis: $\{\mathbf{u}_k\}$.



- **Graph Fourier Transform** (GFT): $\tilde{\mathbf{f}}(\lambda_k) = \langle \mathbf{f}, \mathbf{u}_k \rangle$ or $\tilde{\mathbf{f}} = \mathbf{U}^\top \mathbf{f}$.
- **PW_ω(G)** : space of ω-bandlimited graph signals
  - Support of GFT $= [0, \omega]$, i.e., $\tilde{\mathbf{f}}(\lambda) = 0 \quad \forall\, \lambda > \omega$



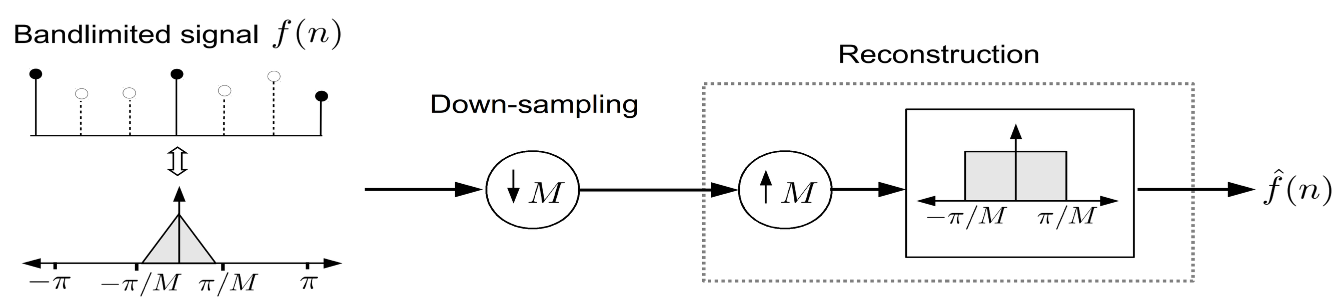
- **Class membership functions are smooth graph signals.**



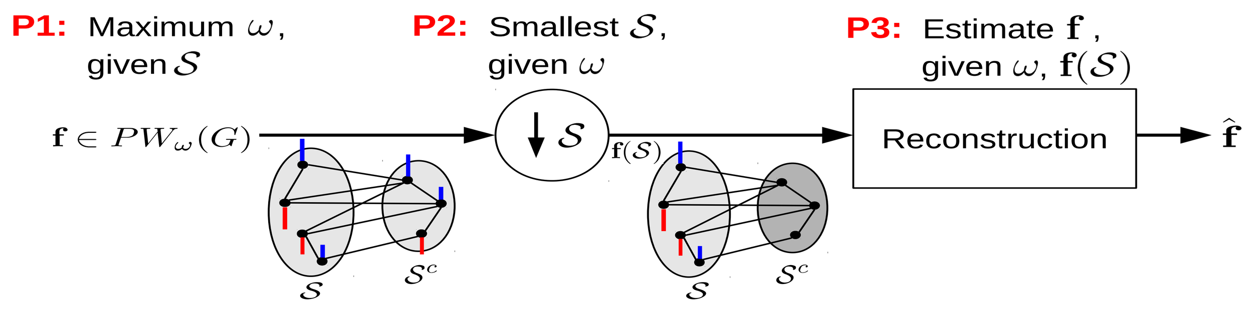
(a) USPS (b) Isolet (c) 20 newsgroups

Sampling Theory for Graph Signals

- Sampling theorem: BW $\omega \Leftrightarrow$ sampling rate for unique representation

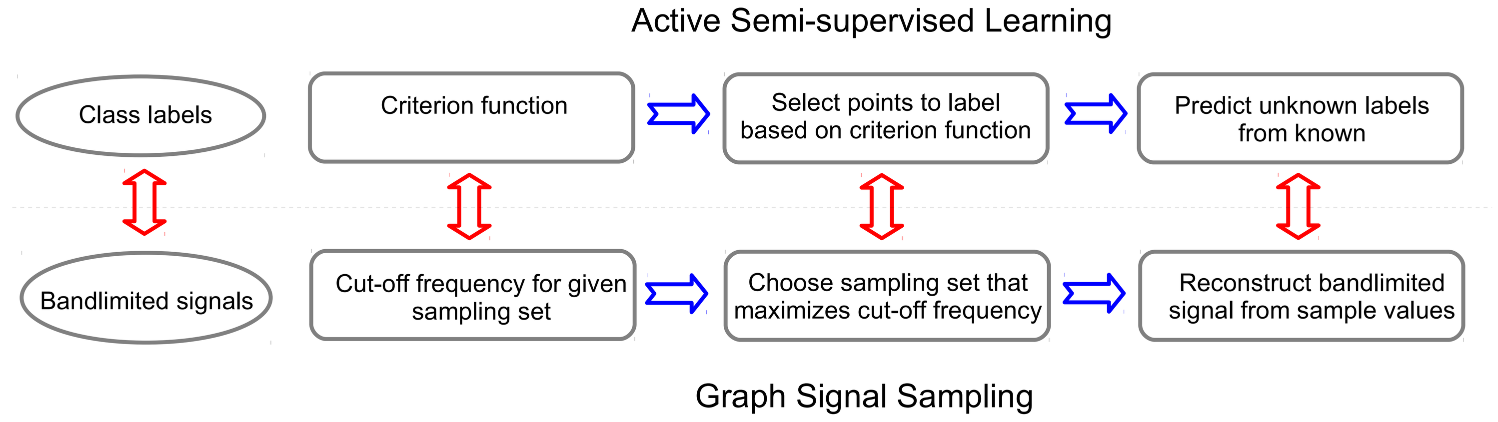


- Sampling theory for graph signals:



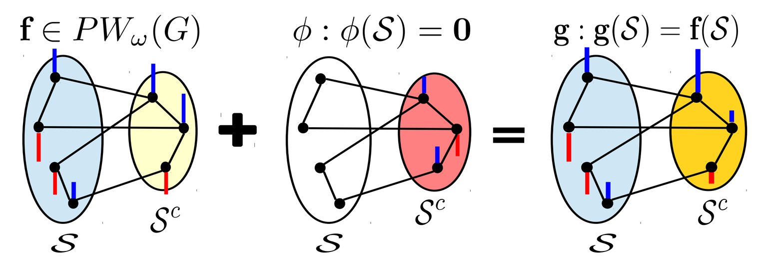
Active Semi-supervised Learning

Approach



P1: Cut-off frequency criterion

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



- Cut-off frequency = smallest BW that a $\phi \in L_2(\mathcal{S}^c)$ can have.
- Estimate by $\min_{\phi(\mathcal{S})=\mathbf{0}} \left(\frac{\phi^\top \boldsymbol{\mathcal{L}}^k \phi}{\phi^\top \phi}\right)^{1/k}$. Higher $k \Rightarrow$ better estimate.
- Let $\{\sigma_{1,k}, \psi_{1,k}\}$ be the smallest eigen-pair of $(\boldsymbol{\mathcal{L}}^k)_{\mathcal{S}^c}$.
- Thus, cut-off estimate $\Omega_k(\mathcal{S}) = (\sigma_{1,k})^{1/k}$ with $\phi_{\text{opt}}(\mathcal{S}^c) = \psi_{1,k}$.

P2: Sampling set selection

- For given size, sampling set must be able to maximally capture signal information; $\mathcal{S}_{\text{opt}} = \arg\max_{|\mathcal{S}|=m} \Omega_k(\mathcal{S})$.

binary relaxation

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

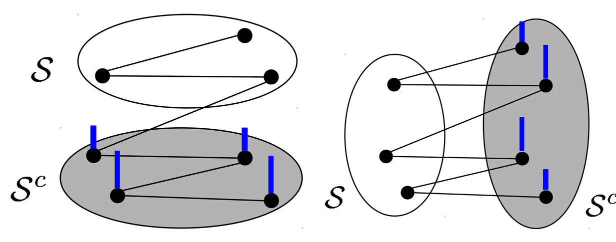
relax the constraint

- $\mathbf{x}_{\text{opt}} \approx \phi_{\text{opt}} \quad \Rightarrow \quad \frac{d\lambda_k^\alpha(\mathbf{t})}{d\mathbf{t}(i)} \approx \alpha(\phi_{\text{opt}}(i))^2$.
- **Greedy algorithm**: $\mathcal{S} \leftarrow \mathcal{S} \cup v$, where $v = \arg\max_j (\phi_{\text{opt}}(j))^2$

Active Semi-supervised Learning (continued)

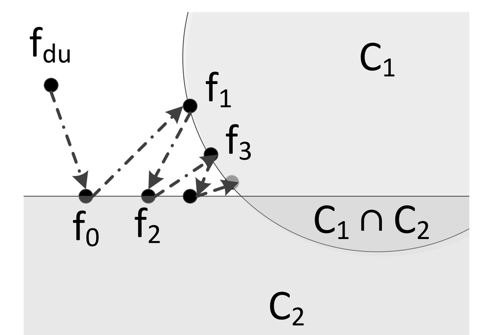
Maximizing Cut-off Frequency $\Rightarrow$ Active Learning

- Cut-off $\Omega_k(\mathcal{S}) \equiv$ variation of smoothest signal ϕ_{opt} in $L_2(\mathcal{S}^c)$.
- Larger cut-off $\Rightarrow$ more variation in $\phi_{\text{opt}} \Rightarrow$ more cross-links.
- Unlabeled nodes are strongly connected to labeled nodes!



P3: Label prediction by reconstruction

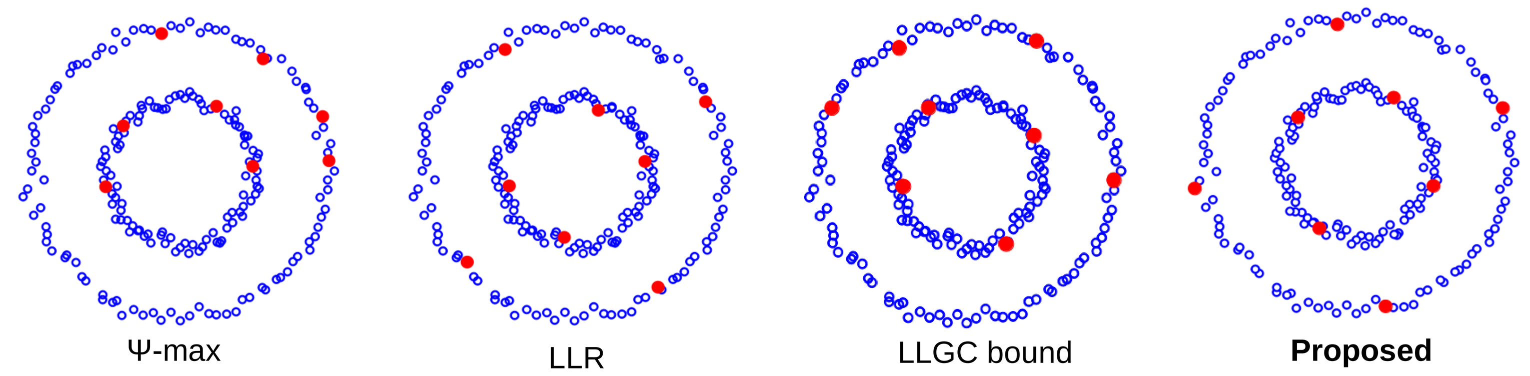
- $\mathcal{C}_1 = \{\mathbf{x} : \mathbf{x}(\mathcal{S}) = \mathbf{f}(\mathcal{S})\}$ and $\mathcal{C}_2 = PW_\omega(G)$.
- Solution: $\mathbf{f} \in \mathcal{C}_1 \cap \mathcal{C}_2$, sampling theory $\Rightarrow$ unique $\mathbf{f}$.
- POCS: $\mathbf{f}_{i+1} = \mathbf{P}_{\mathcal{C}_2}\mathbf{P}_{\mathcal{C}_1}\mathbf{f}_i$, where $\mathbf{f}_0 = [\mathbf{f}(\mathcal{S})^\top, \mathbf{0}]^\top$.
  - $\mathbf{P}_{\mathcal{C}_1}$ resets the samples on $\mathcal{S}$ to $\mathbf{f}(\mathcal{S})$.
  - $\mathbf{P}_{\mathcal{C}_2} = \mathbf{U}h(\boldsymbol{\Lambda})\mathbf{U}^\top$ sets $\tilde{\mathbf{f}}(\lambda) = 0$ if $\lambda > \omega$.



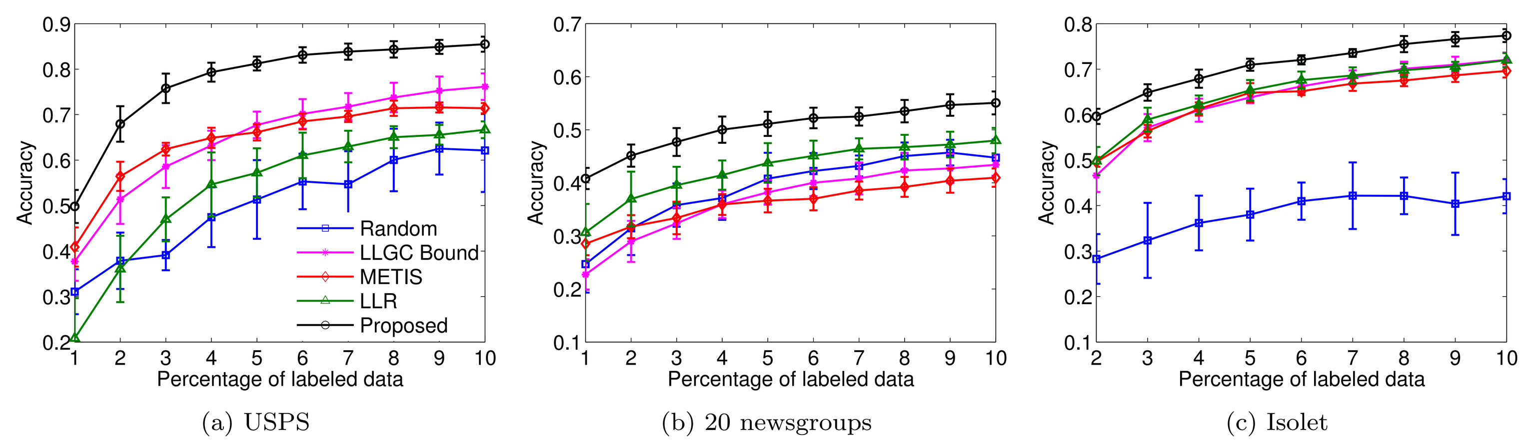
- $\mathbf{P}_{\mathcal{C}_2} \approx \sum_{i=1}^n \left(\sum_{j=0}^p a_j \lambda_i^j\right) \mathbf{u}_i\mathbf{u}_i^\top = \sum_{j=0}^p a_j \boldsymbol{\mathcal{L}}^j \to p$-hop localized

Results

Toy example:

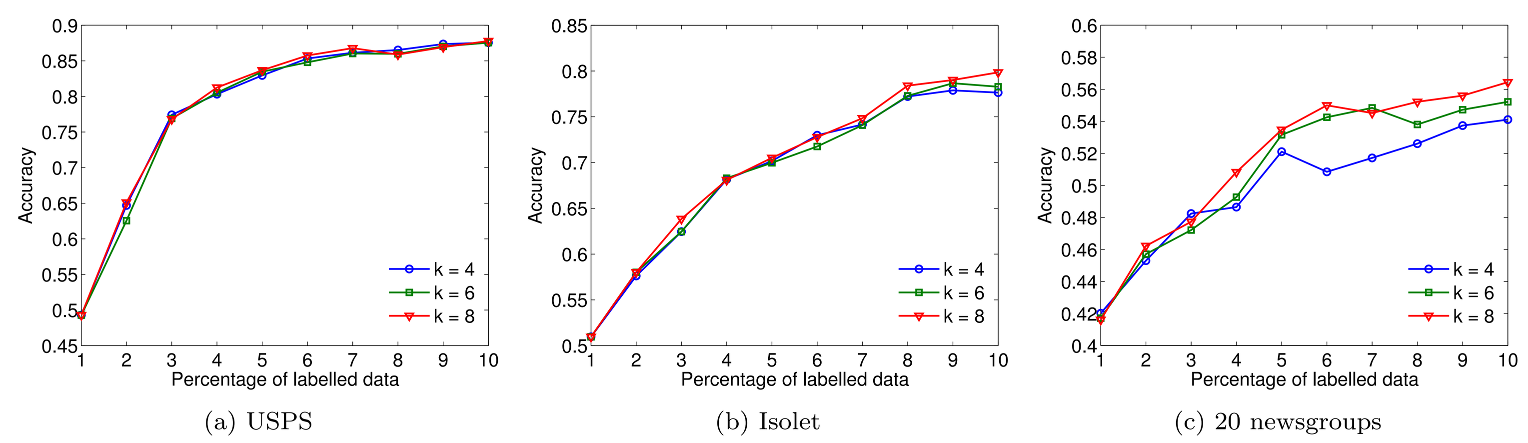


Real datasets:



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

Effect of k:



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

References