Graph Transduction via Alternating Minimization

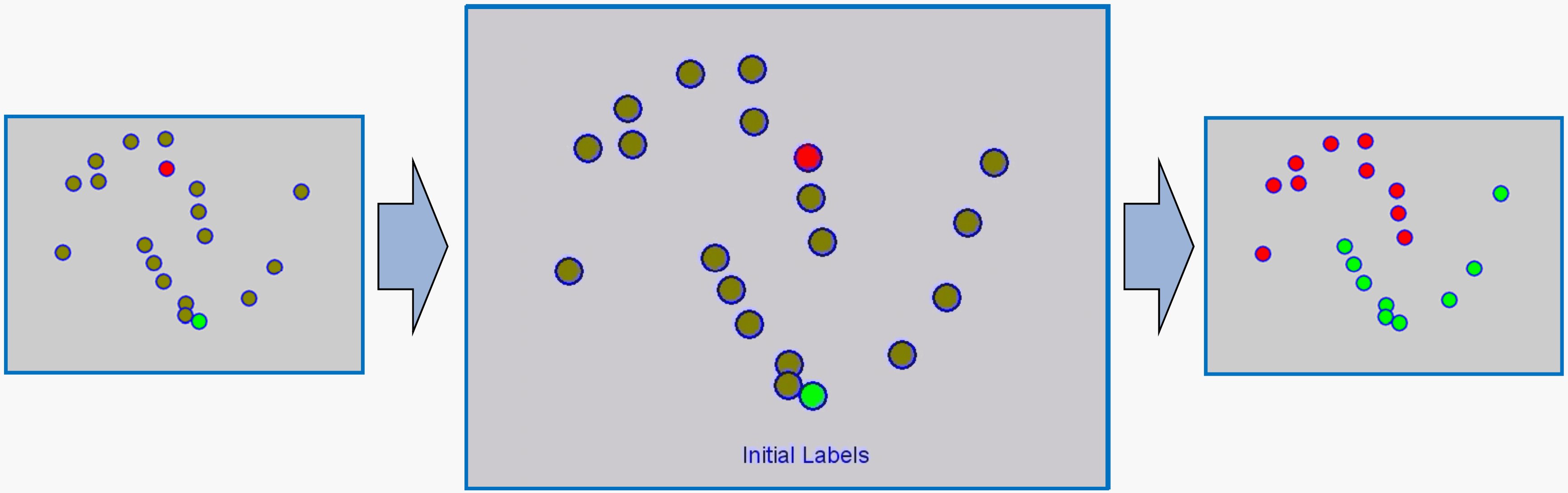
Jun Wang, Tony Jebara, and Shih-fu Chang

Outline of the presentation

- Brief introduction and related work
- Problems with Graph Labeling
  - Imbalanced labels
  - Weak or uninformative labels
  - Noisy and non-separable data
- Proposed Method
  - Graph transduction via Alternating minimization (GTAM)
  - A *bivariate* optimization over graph function and graph labels
  - Label regularizer terms for handling imbalances
- Experiments

Graph Transduction –Review

- Label propagation on graphs



Input samples with sparse labels | Label propagation on graph | Label inference results

Unlabeled | Positive | Negative

Positive ▬ Negative

J. Wang, T. Jebara, and S. F. Chang

Graph Transduction - Review

- Given a dataset $\mathcal{X} = (\mathcal{X}_l, \mathcal{X}_u)$ of labeled samples $\mathcal{X}_l$, and unlabeled samples $\mathcal{X}_u$
- Graph transduction here uses an *undirected* graph $\mathcal{G} = \{\mathcal{X}, \mathcal{E}\}$ of samples $\mathcal{X}$ as nodes and edges $\mathcal{E}$ weighted by sample similarity $w_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$;
- Define weight matrix $\mathbf{W} = \{w_{ij}\}$,
Node degree $\mathbf{D} = diag([d_1, \cdots, d_n])$,
graph Laplacian $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$,
and normalized Laplacian $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{\Delta}\mathbf{D}^{-1/2}$
label matrix $\mathbf{Y}$,

$$\begin{array}{cc} & \begin{array}{cc} \text{positive} & \text{negative} \\ \Downarrow & \Downarrow \end{array} \\ \mathbf{Y} = & \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \end{array}$$

Graph Transduction – Review

- Function estimation through optimization
  - A continuous valued classification function is estimated by minimizing a cost $\mathcal{Q}$

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \mathcal{Q}(\mathbf{F}) = \arg\min_{\mathbf{F}} \left\{ Q_{smooth}(\mathbf{F}) + Q_{fit}(\mathbf{F}) \right\}$$

  - Trades off smoothness over graph with fitness on given labels

Graph Transduction – Review

- Previous choices for cost: $\mathcal{Q}$

➢ Gaussian fields and Harmonic functions *GFHF*

(Zhu, Ghahramani, and Lafferty ICML03)

$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|F_{i\cdot} - F_{j\cdot}\|^2$$

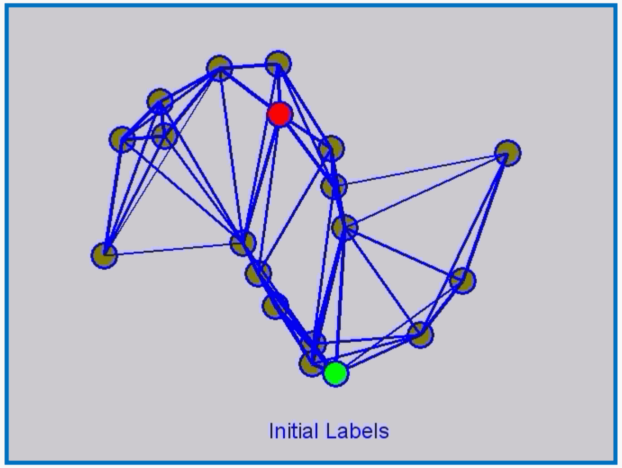
➢ Local and global consistency *LGC*

(Zhou, Bousquet, Lal, Weston, and Scholkopf NIPS04)

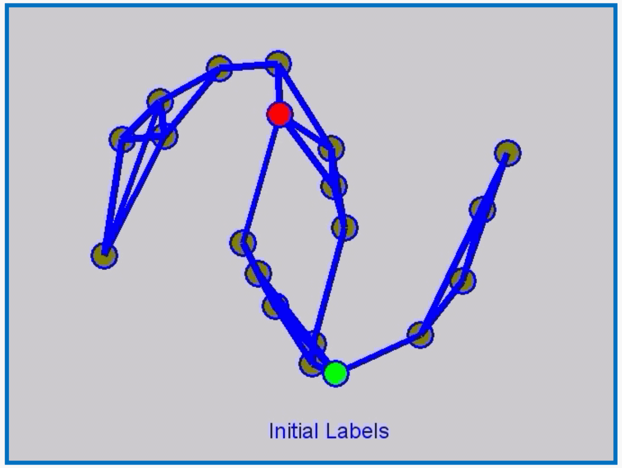
$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left\|\frac{F_{i\cdot}}{\sqrt{D_{ii}}} - \frac{F_{j\cdot}}{\sqrt{D_{ii}}}\right\|^2 + \mu\sum_{i=1}^{n}\|F_{i\cdot} - Y_{i\cdot}\|^2$$

J. Wang, T. Jebara, and S. F. Chang

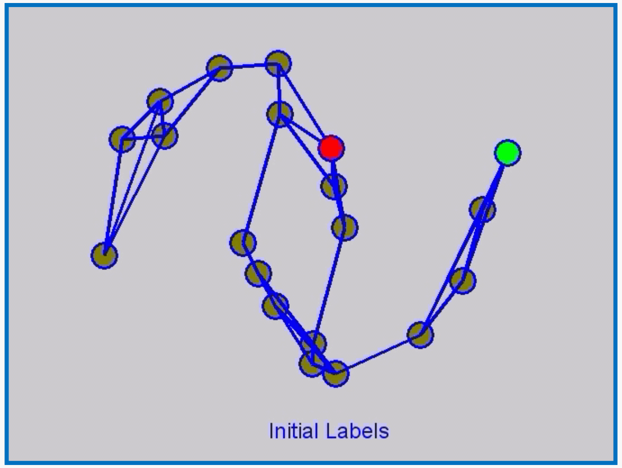
Graph Transduction: Problemistic Cases



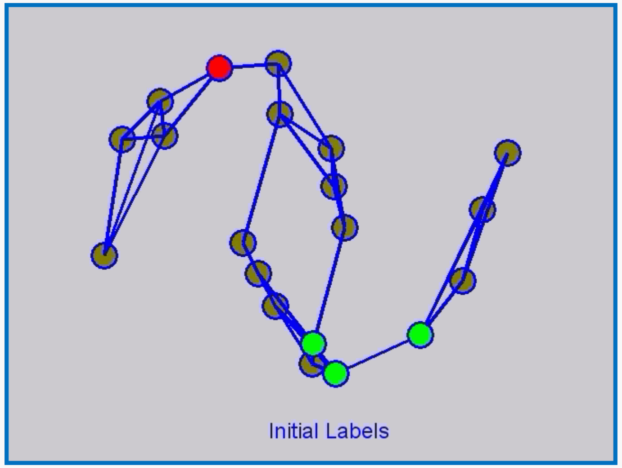
Over connected graph



Improper edge weighting



Difficult label location



Imbalance labels

J. Wang, T. Jebara, and S. F. Chang

Methodology – Our Choice for $\mathcal{Q}$

1) Start with *LGC*'s Cost

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2}\mathrm{tr}\left\{\mathbf{F}^T\mathbf{L}\mathbf{F} + \mu(\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y})\right\}$$

2) Make into a bivariate optimization over $\mathbf{F}$ and $\mathbf{Y}$

$$\mathcal{Q}(\mathbf{F},\mathbf{Y}) = \frac{1}{2}\mathrm{tr}\left\{\mathbf{F}^T\mathbf{L}\mathbf{F} + \mu(\mathbf{F}-\mathbf{V}\mathbf{Y})^T(\mathbf{F}-\mathbf{V}\mathbf{Y})\right\}$$

3) Introduce label regularizer terms

$$\mathbf{v} = \sum_{j=1}^{c} \frac{\mathbf{Y}_{\cdot j} \odot \mathbf{D}\vec{\mathbf{1}}}{\mathbf{Y}_{\cdot j}^T\mathbf{D}\vec{\mathbf{1}}} \qquad \mathbf{V} = diag\{\mathbf{v}\}$$

Methodology– Label Regularizer

- Normalize labels among classes to handle imbalance
- Weight labels based on the degrees;

Example:

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$\begin{array}{cc} \text{positive} & \text{negative} \\ \Downarrow & \Downarrow \end{array}$$

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\Longrightarrow$$

$$\mathbf{V} = \begin{bmatrix} \frac{1}{1+3} & 0 & 0 & 0 \\ 0 & \frac{2}{2} & 0 & 0 \\ 0 & 0 & \frac{3}{1+3} & 0 \\ 0 & 0 & 0 & \frac{0}{4} \end{bmatrix}$$

J. Wang, T. Jebara, and S. F. Chang

Methodology – Optimize $\mathbf{F}$

- Minimizing $\mathcal{Q}(\mathbf{F},\mathbf{Y})$ is mixed integer program

$$\mathcal{Q}(\mathbf{F},\mathbf{Y}) = \frac{1}{2}\mathrm{tr}\left\{\mathbf{F}^T\mathbf{L}\mathbf{F} + \mu(\mathbf{F}-\mathbf{V}\mathbf{Y})^T(\mathbf{F}-\mathbf{V}\mathbf{Y})\right\}$$

- Try greedy solution via alternating minimization

- Solve for continuous valued $\mathbf{F}$: $\mathbf{P} = (\mathbf{L}/\mu + \mathbf{I})^{-1}$

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{F}^*} = 0 \Rightarrow \mathbf{F}^* = (\mathbf{L}/\mu + \mathbf{I})^{-1}\mathbf{V}\mathbf{Y} = \mathbf{P}\mathbf{V}\mathbf{Y}$$

- Insert the solution gives NP hard problem for $\mathbf{Y}$ (slightly nonlinear MAXCUT)

$$\mathcal{Q}(\mathbf{Y}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Y}^T\mathbf{V}^T\left[\mathbf{P}^T\mathbf{L}\mathbf{P} + \mu(\mathbf{P}^T-\mathbf{I})(\mathbf{P}-\mathbf{I})\right]\mathbf{V}\mathbf{Y}\right)$$

J. Wang, T. Jebara, and S. F. Chang

Methodology – Gradient Greedy (1)

- Optimization on binary valued $\mathbf{Y}$ with constraint
- Almost Max Cut problem where Greedy is 0.5 optimal

$$\mathcal{Q}(\mathbf{Y}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Y}^T\mathbf{V}^T\left[\mathbf{P}^T\mathbf{L}\mathbf{P} + \mu(\mathbf{P}^T - \mathbf{I})(\mathbf{P} - \mathbf{I})\right]\mathbf{V}\mathbf{Y}\right)$$

$$\mathbf{Z} = \mathbf{V}\mathbf{Y} \qquad \mathbf{A} = \mathbf{P}^T\mathbf{L}\mathbf{P} + (\mathbf{P}^T - \mathbf{I})(\mathbf{P} - \mathbf{I})$$

$$\mathcal{Q}(\mathbf{Z}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^T\mathbf{A}\mathbf{Z}\right)$$

- We do 'gradient greedy' on our problem. Find which entry of $\mathbf{Y}$ will most reduce cost and select it for labeling

$$\frac{\partial\mathcal{Q}}{\partial\mathbf{Y}} = \frac{\partial\mathcal{Q}}{\partial\mathbf{Z}}\frac{\partial\mathbf{Z}}{\partial\mathbf{Y}} \qquad \frac{\partial\mathcal{Q}}{\partial\mathbf{Z}} = \mathbf{A}\mathbf{Z} = \mathbf{A}\mathbf{V}\mathbf{Y}$$

J. Wang, T. Jebara, and S. F. Chang

Methodology – Gradient Greedy (2)

- Find location with the steepest descent of the value of the loss function

$$(i^*, j^*) = \arg\min_{\mathbf{x}_i \in \mathcal{X}_u, 1 \le j \le c} \nabla_{\mathbf{Z}_{ij}} \mathcal{Q}$$

- Label the corresponding node: $\mathbf{Y}_{i^* j^*} = 1$

$$\mathbf{Y}^t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \xrightarrow[i^* = 3, j^* = 1]{\nabla_{\mathbf{Z}} \mathcal{Q}^t = \begin{bmatrix} * & * \\ * & * \\ -0.31 & 0.07 \\ -0.17 & -0.04 \end{bmatrix}} \mathbf{Y}^{t+1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

- Iterative repeat the above procedure until all the nodes are labeled

J. Wang, T. Jebara, and S. F. Chang

Final Algorithm

1) Calculate gradient matrix

$$\nabla_{\mathbf{Z}} \mathcal{Q}^t = \mathbf{A} diag(\mathbf{v}^t) \mathbf{Y}^t$$

2) Label the most beneficial node with largest cost reduction

$$\begin{aligned}(i^*, j^*) &= \arg\min_{\mathbf{x}_i \in \mathcal{X}_u, 1 \le j \le c} \nabla_{\mathbf{Z}_{ij}} \mathcal{Q}^t \\ \mathbf{Y}^{t+1}_{i^* j^*} &= 1\end{aligned}$$

3) Update the label regularizer

$$\mathbf{v}^{t+1} = \sum_{j=1}^{c} \frac{\mathbf{Y}^{t+1}_{\cdot j} \odot \mathbf{D}\vec{1}}{{\mathbf{Y}^{t+1}_{\cdot j}}^T \mathbf{D}\vec{1}}$$

4) Update labeled and unlabeled sets

$$\mathcal{X}_l^{t+1} \leftarrow \mathcal{X}_l^t + \mathbf{x}_{i^*}; \quad \mathcal{X}_u^{t+1} \leftarrow \mathcal{X}_u^t - \mathbf{x}_{i^*}; \quad t \leftarrow t+1$$

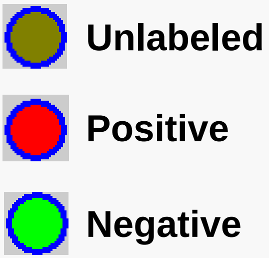
5) If $\mathcal{X}_u^{t+1} = \emptyset$, output the labels; else go to 1)

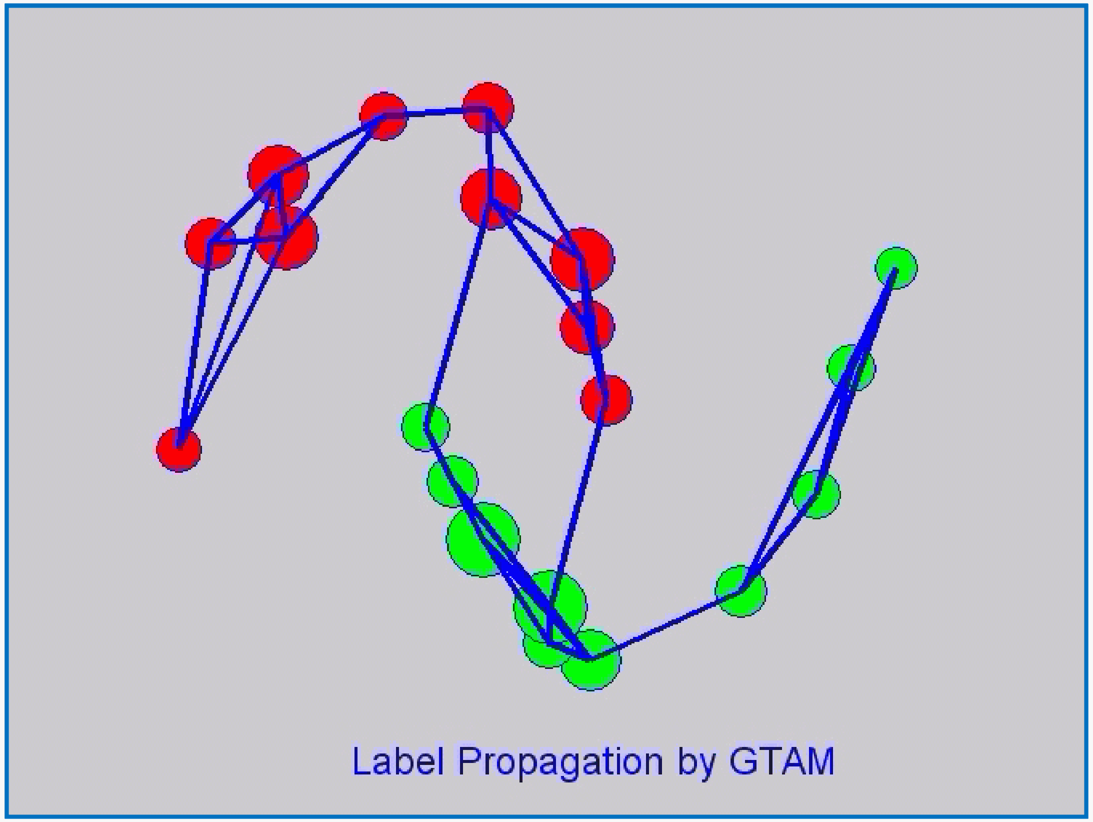
Some Intuition

- Previous methods (e.g. *LGC* and *GFHF*) prematurely commit to an erroneous labeling;
- Our method iteratively infers labels with the current given labels and each step only assign label to the most beneficial node with highest cost reduction;
- Greedy MaxCut is not bad (0.5), best solution is SDP 0.878 (but too slow).

J. Wang, T. Jebara, and S. F. Chang

Intuition





Label propagation by GTAM

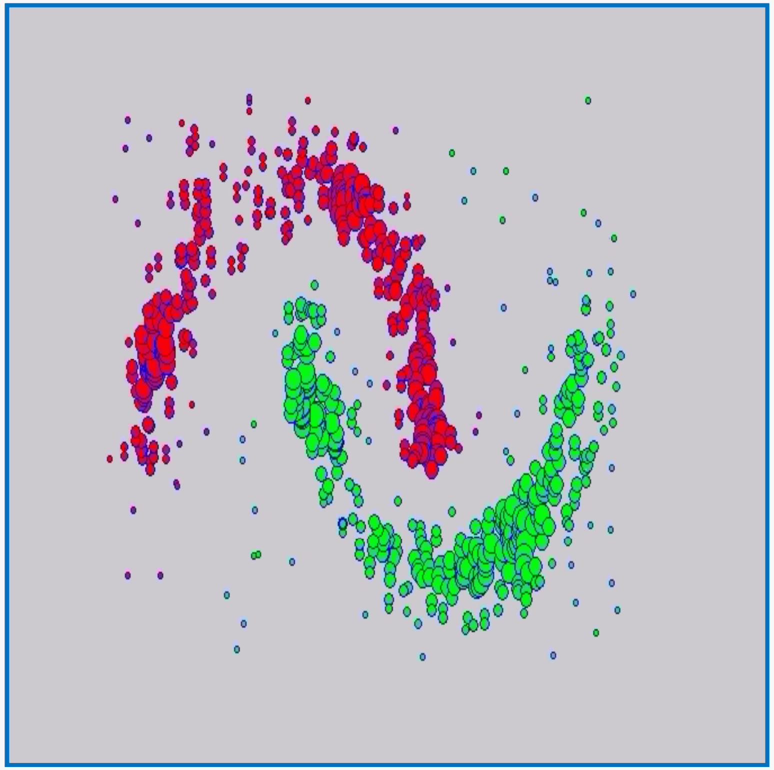
The scale of labeled nodes denotes the value of label regularizer

Computation Efficiency

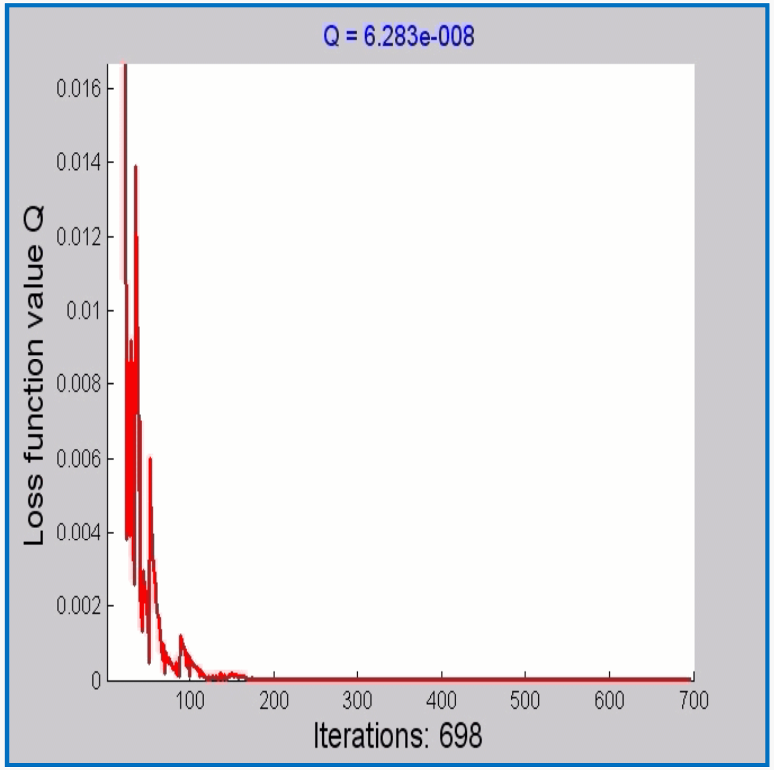
- Complexity is $\mathcal{O}(n^3)$
- Can run more efficiently
  - Applying superposition approach to achieve incremental updating (*Wang, Chang et al CVPR08*)
  - Can early-stop greedy algorithm after enough labeling
  - Can do multiple nodes labeling in each iteration

J. Wang, T. Jebara, and S. F. Chang

Experiments – Toy Data



Label propagation by GTAM



Convergence procedure

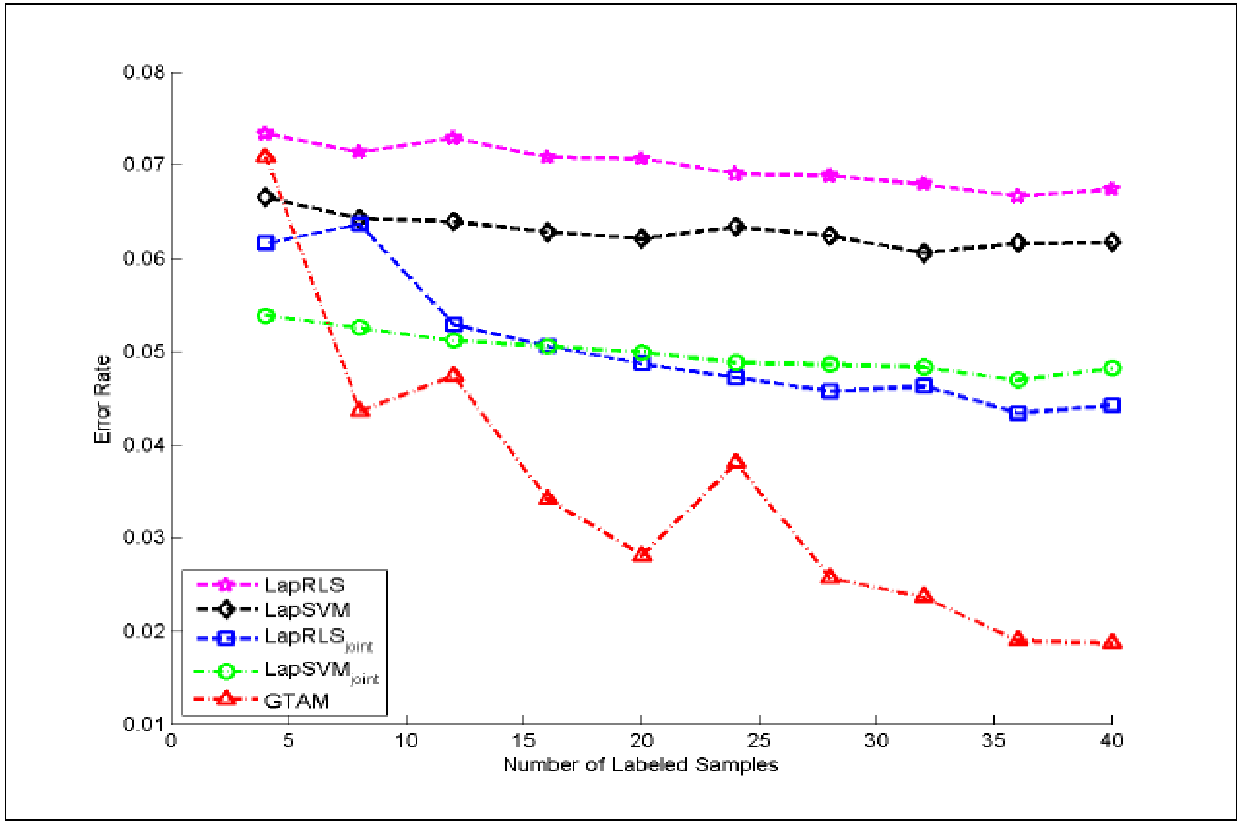
(non-monotonic due to gradient greedy discrete step size)

Unlabeled   Positive   Negative

J. Wang, T. Jebara, and S. F. Chang

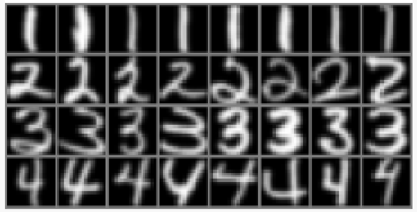
Experiments – WebKB Data

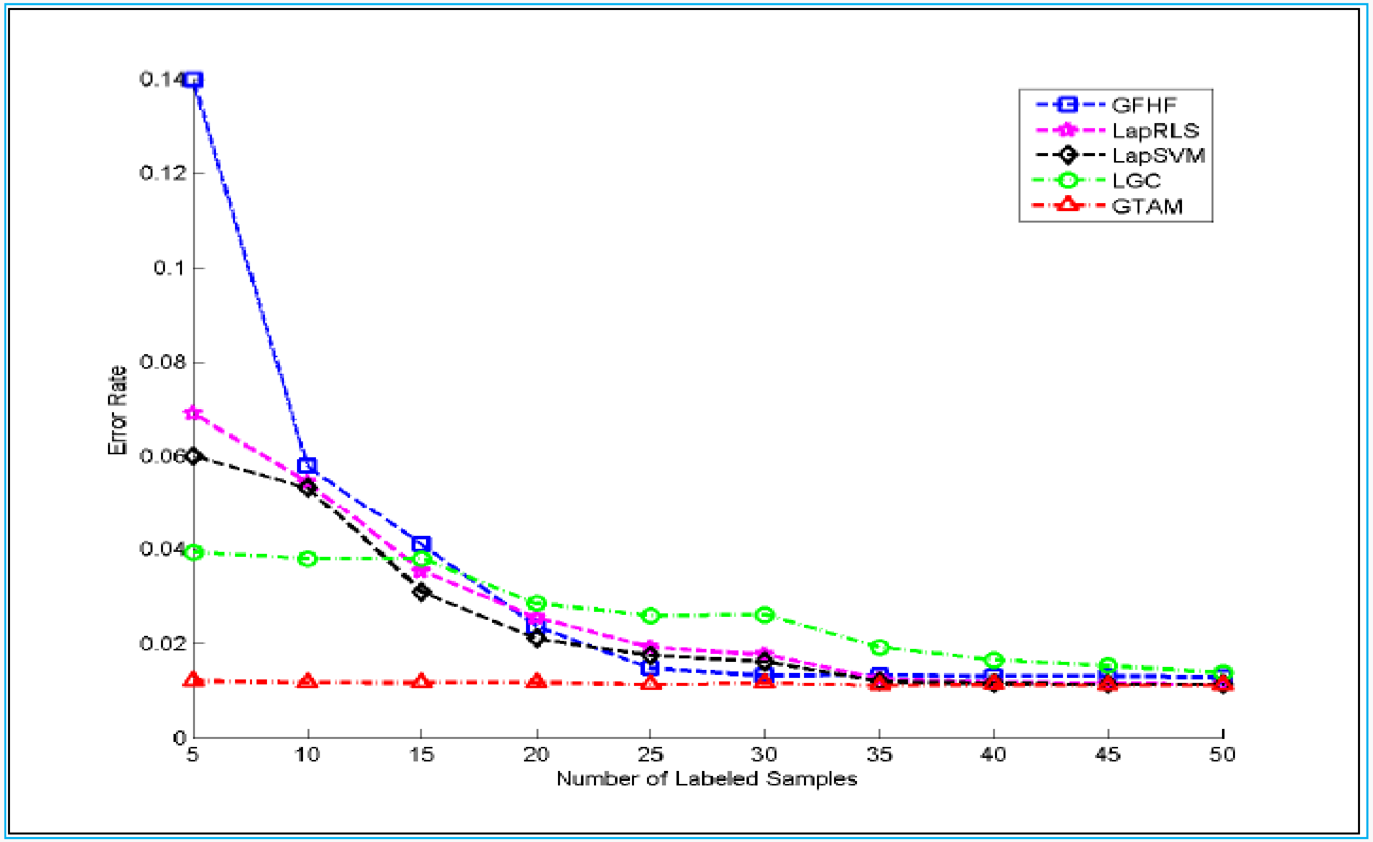
1051 documents (course & non-course) containing **1840** page + **3000** link attributes;
Comparing with approaches reported in *Sindhwani, Niyogi, and Belkin ICML 2005;*
100 random test , evaluation based average error rate;



Experiments – USPS Digits Data

3874 digit samples (16*16 image) containing four digits 1, 2, 3, 4;
Comparing with LGC, GFHF and LapSVM et al;
20 random test , evaluation based average error rate;





Summary

- Cast graph transduction as cost over labels $\mathbf{Y}$ and graph functions $\mathbf{F}$
- Add label normalization terms
- Greedy alternating optimization of $\mathbf{F}$ and $\mathbf{Y}$ (reminiscent of MaxCut)
- This produces gradual propagation-style algorithm
- Fast and robust to labeling degeneracies
- Reduces error rate of existing approaches on WebKB and USPS digits by more than half
- Open questions ….