Semi-supervised Graph Clustering: A Kernel Approach

Brian Kulis, Sugato Basu, Inderjit Dhillon, Raymond Mooney

[kulis, sugato, inderjit, mooney]@cs.utexas.edu

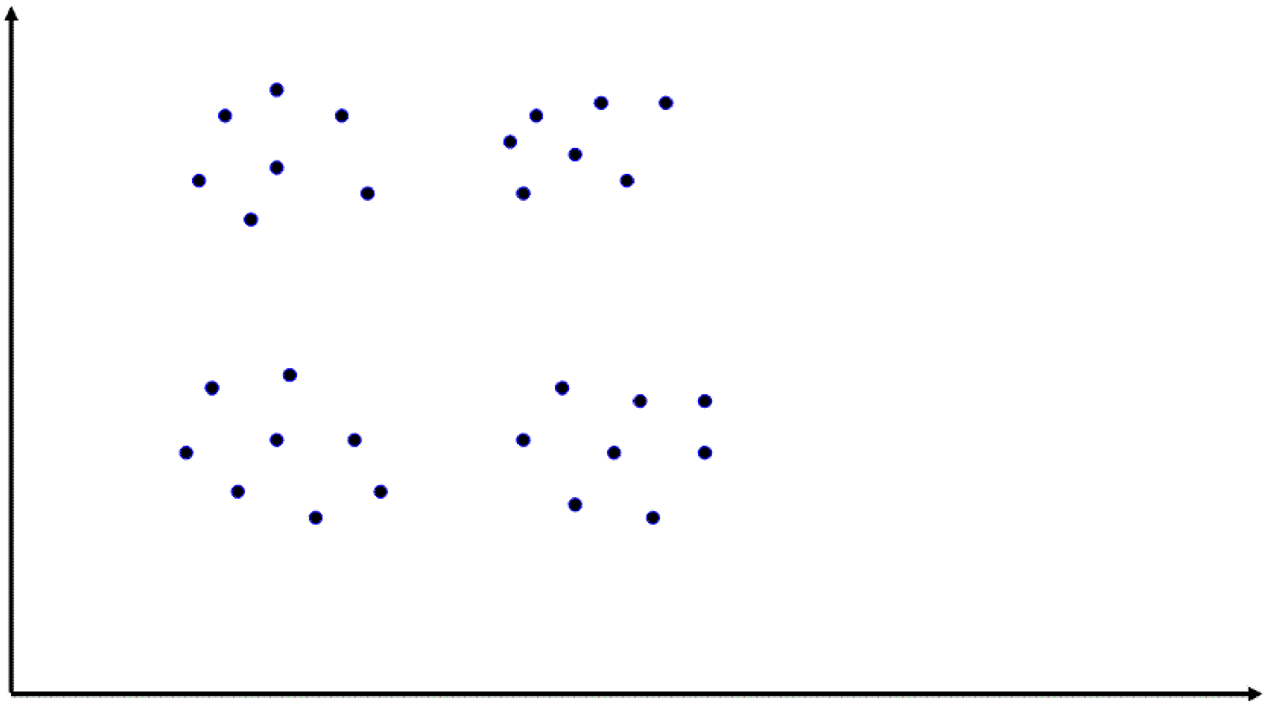
Department of Computer Sciences

The University of Texas at Austin

Austin, TX 78712

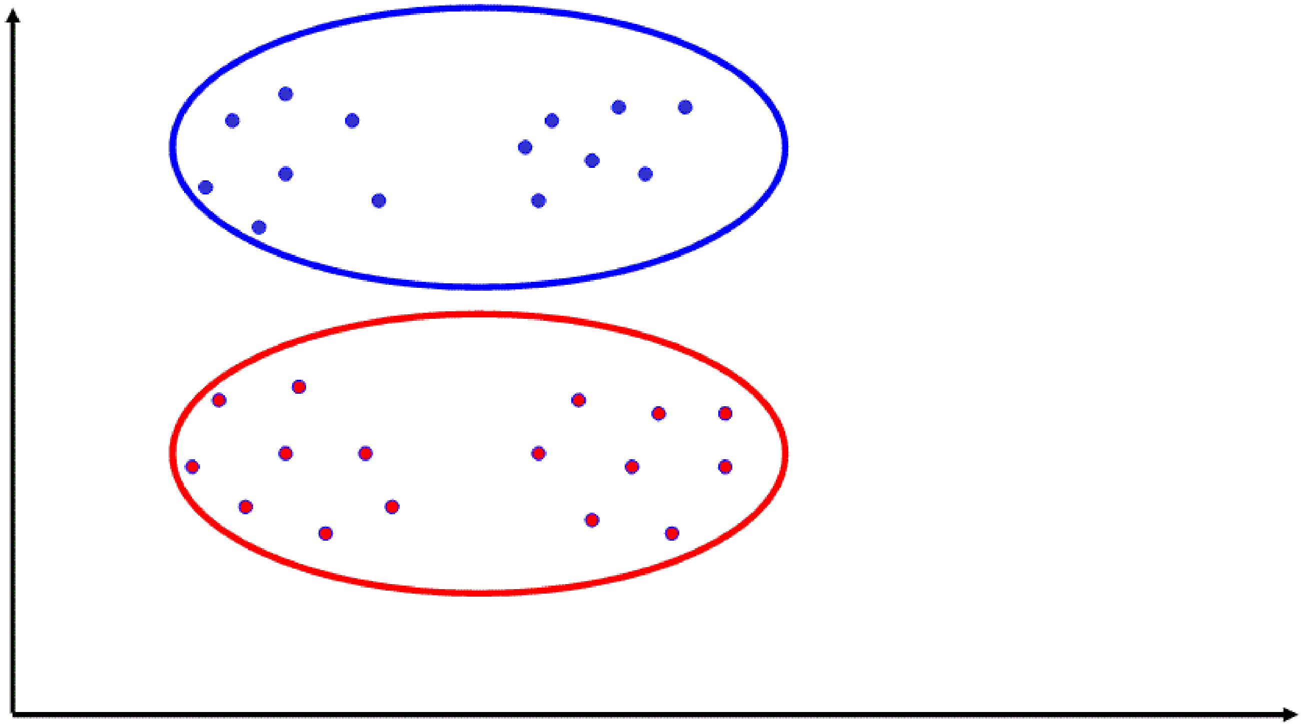
Semi-supervised Clustering

- Grouping together of similar objects given some knowledge about the cluster structure



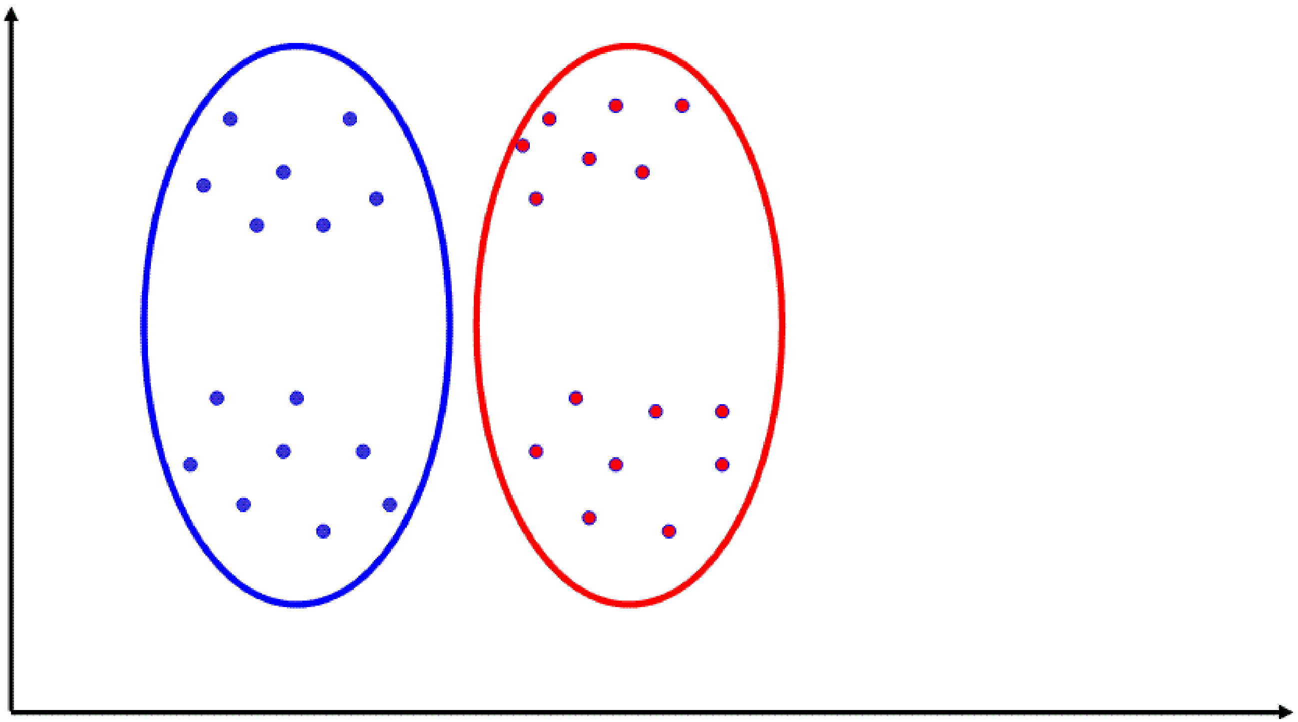
Semi-supervised Clustering

- Grouping together of similar objects given some knowledge about the cluster structure



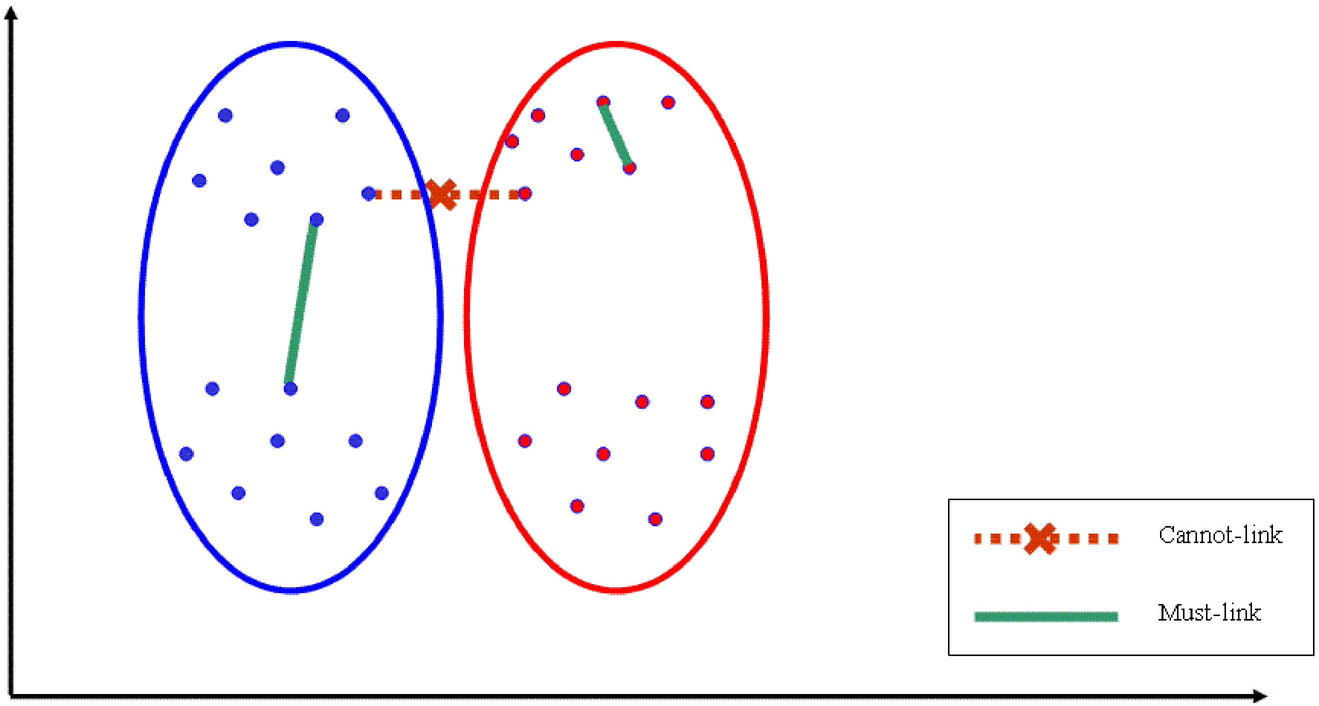
Semi-supervised Clustering

- Grouping together of similar objects given some knowledge about the cluster structure



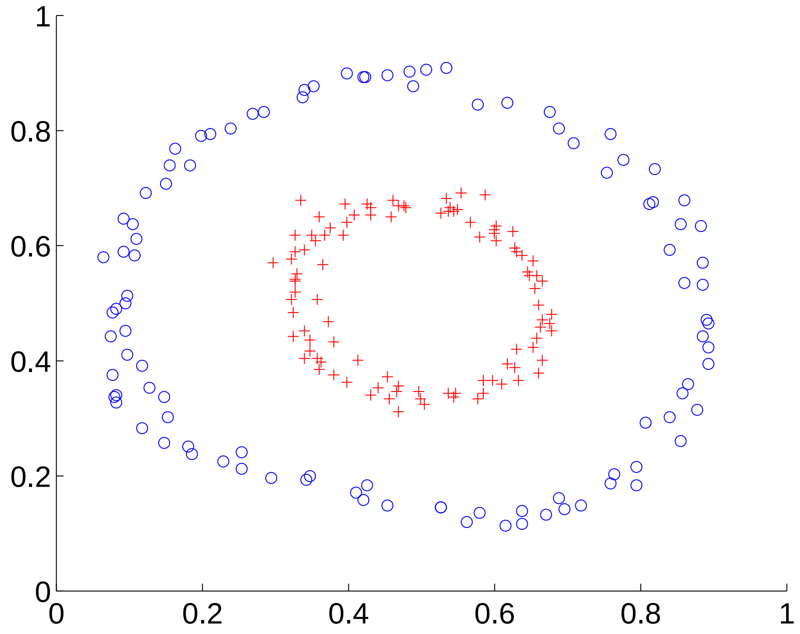
Semi-supervised Clustering

- HMRF_KMeans: framework for semi-supervised clustering based on Hidden Markov Random Fields [Basu04]



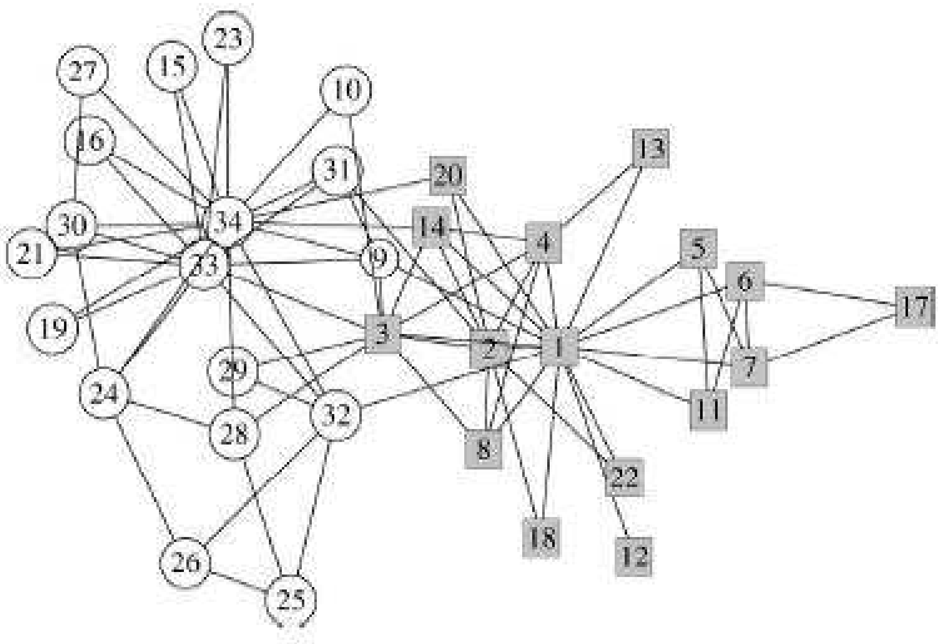
Two Circles

- Some data is not linearly separable
- Algorithms such as HMRF_KMeans cannot recover true clusters



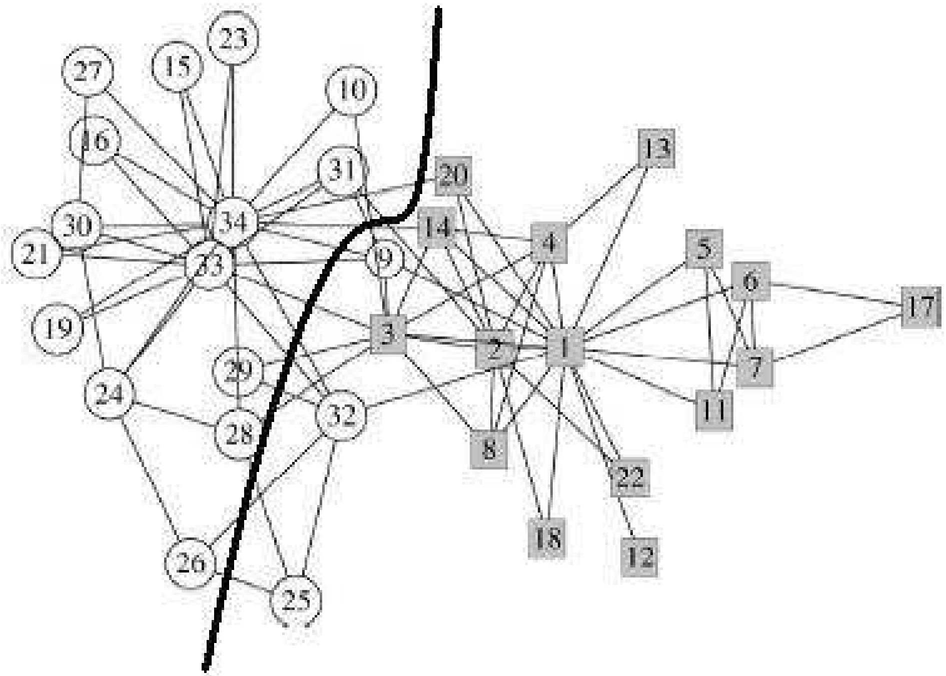
Graph-Based Data

- Data may also be in form of graph



Graph-Based Data

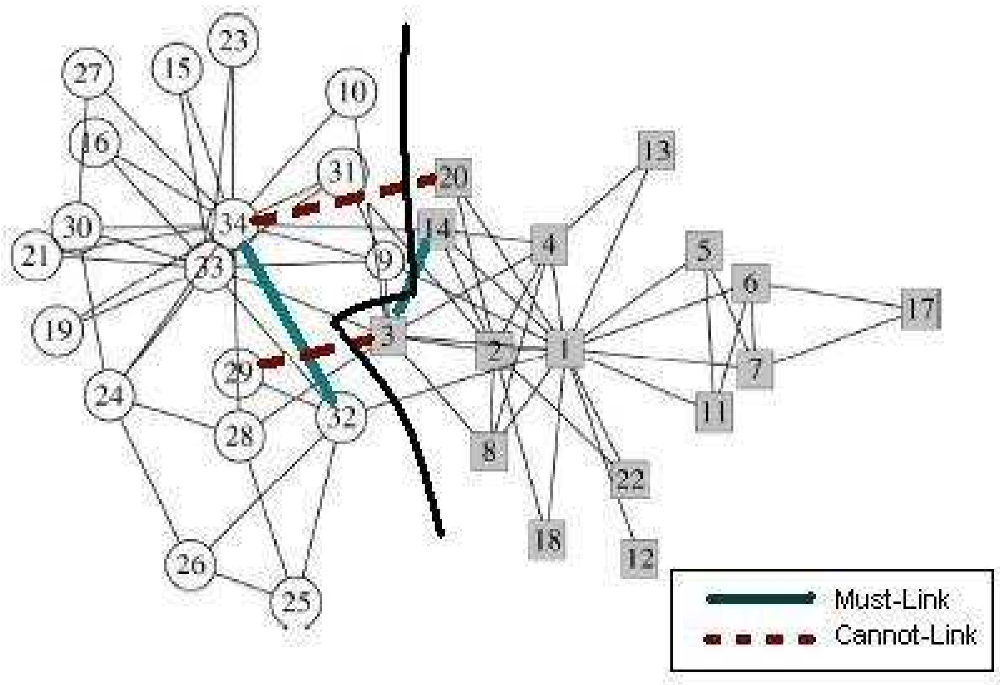
- Vector-based algorithms are inappropriate



- Example goal: minimizing the normalized cut [Shi00]

Graph-Based Data

- Spectral Learning algorithm [Kam03] for this kind of data



- Yeast gene interaction network

Main Contributions

- Theoretical equivalence between weighted kernel k-means and graph clustering

Main Contributions

- Theoretical equivalence between weighted kernel k-means and graph clustering
- Unifies vector-based and graph-based semi-supervised clustering using a kernel approach

Main Contributions

- Theoretical equivalence between weighted kernel k-means and graph clustering
- Unifies vector-based and graph-based semi-supervised clustering using a kernel approach
  - Implication: One algorithm for semi-supervised clustering of graph-based and vector-based data

Main Contributions

- Theoretical equivalence between weighted kernel k-means and graph clustering
- Unifies vector-based and graph-based semi-supervised clustering using a kernel approach
  - Implication: One algorithm for semi-supervised clustering of graph-based and vector-based data
  - HMRF_KMeans, Spectral Learning and several other graph clustering objectives are special cases

Main Contributions

- Theoretical equivalence between weighted kernel k-means and graph clustering
- Unifies vector-based and graph-based semi-supervised clustering using a kernel approach
  - Implication: One algorithm for semi-supervised clustering of graph-based and vector-based data
  - HMRF_KMeans, Spectral Learning and several other graph clustering objectives are special cases
- Empirical results validate superior performance on real-life data sets

Weighted Kernel k-means [Dhi04]

- Seek k-way partitioning $\{\pi_c\}_{c=1}^k$ that minimizes:

$$\mathcal{D}(\{\pi_c\}_{c=1}^k) = \sum_{c=1}^{k} \sum_{\mathbf{x}_i \in \pi_c} \gamma_i \|\phi(\mathbf{x}_i) - \mathbf{m}_c\|^2,$$

$$\text{where } \mathbf{m}_c = \frac{\sum_{\mathbf{x}_i \in \pi_c} \gamma_i \phi(\mathbf{x}_i)}{\sum_{\mathbf{x}_i \in \pi_c} \gamma_i}$$

- If all weights γ_i are set to 1 and ϕ is the identity, reduces to standard k-means
- Algorithm is kernelizable and is analogous to standard k-means
- Any PSD matrix K can be interpreted as a kernel matrix

Graph Clustering

- Many graph clustering objectives are special cases of the weighted kernel k-means objective function [Dhi04]

Objective	Node Weights	Kernel Matrix
Ratio Cut	$1 \;\; \forall$ nodes	$K = \sigma I - L$
Normalized Cut	Deg. of node	$K = \sigma D^{-1} + D^{-1} A D^{-1}$

- A: graph affinity matrix
  D: diagonal degree matrix
  L: Laplacian matrix

HMRF_KMeans Clustering

- Framework for semi-supervised clustering based on Hidden Markov Random Fields [Basu04]
- HMRF_KMeans Objective:

$$\sum_{c=1}^{k} \sum_{\mathbf{x}_i \in \pi_c} \|\mathbf{x}_i - \mathbf{m}_c\|^2 - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{|\pi_{l_i}|} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{|\pi_{l_i}|}.$$

- $\mathcal{M}$ set of must-link pairs, $\mathcal{C}$ set of cannot-link pairs, l_i is the cluster label for $\mathbf{x}_i$

HMRF_KMeans Clustering

- Framework for semi-supervised clustering based on Hidden Markov Random Fields [Basu04]
- HMRF_KMeans Objective:

$$\sum_{c=1}^{k} \sum_{\mathbf{x}_i \in \pi_c} \|\mathbf{x}_i - \mathbf{m}_c\|^2 - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{|\pi_{l_i}|} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{|\pi_{l_i}|}.$$

- $\mathcal{M}$ set of must-link pairs, $\mathcal{C}$ set of cannot-link pairs, l_i is the cluster label for $\mathbf{x}_i$
- $K = S + W$, $S_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$

HMRF_KMeans Clustering

- Framework for semi-supervised clustering based on Hidden Markov Random Fields [Basu04]
- HMRF_KMeans Objective:

$$\sum_{c=1}^{k} \sum_{\mathbf{x}_i \in \pi_c} \|\mathbf{x}_i - \mathbf{m}_c\|^2 - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{|\pi_{l_i}|} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{|\pi_{l_i}|}.$$

- $\mathcal{M}$ set of must-link pairs, $\mathcal{C}$ set of cannot-link pairs, l_i is the cluster label for $\mathbf{x}_i$
- $K = S + W$, $W_{ij} = w_{ij}$ if $(i, j) \in \mathcal{M}$, $-w_{ij}$ if $(i, j) \in \mathcal{C}$

Semi-supervised Graph Clustering

- Semi-Supervised Normalized Cut
- 3 terms, as in HMRF objective

$$\sum_{c=1}^{k} \frac{\text{links}(\mathcal{V}_c, \mathcal{V} \setminus \mathcal{V}_c)}{\text{degree}(\mathcal{V}_c)} - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})}.$$

Semi-supervised Graph Clustering

- Semi-Supervised Normalized Cut
- 3 terms, as in HMRF objective

$$\sum_{c=1}^{k} \frac{\text{links}(\mathcal{V}_c, \mathcal{V} \setminus \mathcal{V}_c)}{\text{degree}(\mathcal{V}_c)} - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})}.$$

- $K = D^{-1}AD^{-1} + D^{-1}WD^{-1}$
- Note: kernel k-means node weights should be set to the degrees of the nodes

Semi-supervised Graph Clustering

- Semi-Supervised Normalized Cut
- 3 terms, as in HMRF objective

$$\sum_{c=1}^{k} \frac{\text{links}(\mathcal{V}_c, \mathcal{V} \setminus \mathcal{V}_c)}{\text{degree}(\mathcal{V}_c)} - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})}.$$

- $K = D^{-1}AD^{-1} + D^{-1}WD^{-1}$
- Note: kernel k-means node weights should be set to the degrees of the nodes

Semi-supervised Graph Clustering

- Semi-Supervised Normalized Cut
- 3 terms, as in HMRF objective

$$\sum_{c=1}^{k} \frac{\text{links}(\mathcal{V}_c, \mathcal{V} \setminus \mathcal{V}_c)}{\text{degree}(\mathcal{V}_c)} - \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})} + \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}, l_i = l_j} \frac{w_{ij}}{\text{deg}(\mathcal{V}_{l_i})}.$$

- Can generalize to ratio cut and other graph clustering objectives
- Spectral Learning [Kam03] can be viewed as spectral relaxation to semi-supervised ratio cut

Algorithm

- Input: Similarity matrix S, constraint matrix W, diagonal node weight matrix Γ
  - For SS-NormCut, $S = A$; for SS-RatioCut, $S = A - D$
  - For SS-NormCut, $\Gamma = D$; for SS-RatioCut and HMRF_KMeans, $\Gamma = I$
- Form $K = \Gamma^{-1}(S + W)\Gamma^{-1} + \sigma\Gamma^{-1}$
- Get initial clusters using the constraints
- Run weighted kernel k-means (with weights from Γ) on K using the initial clustering
- Return the resulting clusters

Experimental Methodology

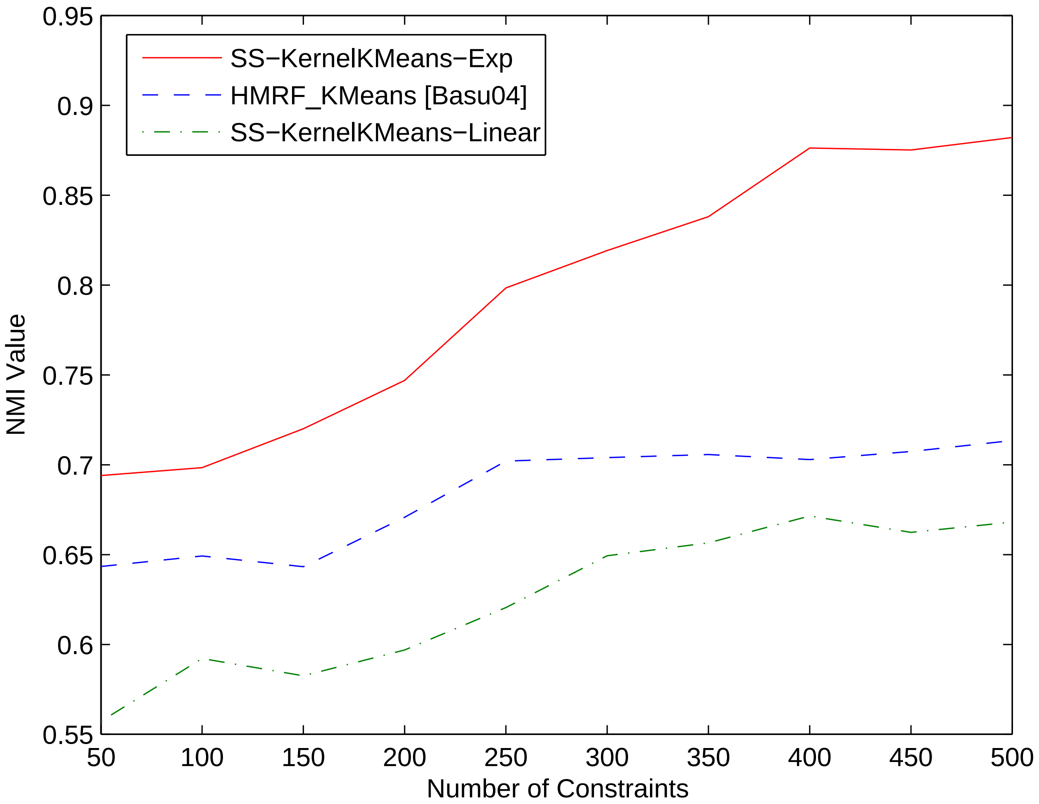
- Choose data sets with pre-existing labels
- Clustering done on entire data set
- Constraints chosen randomly from points in the training set
- Clustering accuracy computed using only the test set

Experimental Methodology

- Plotted learning curves with averages of 10 runs of 2-fold cross validation
- x-axis corresponds to increasing number of constraints
- y-axis corresponds to Normalized Mutual Information

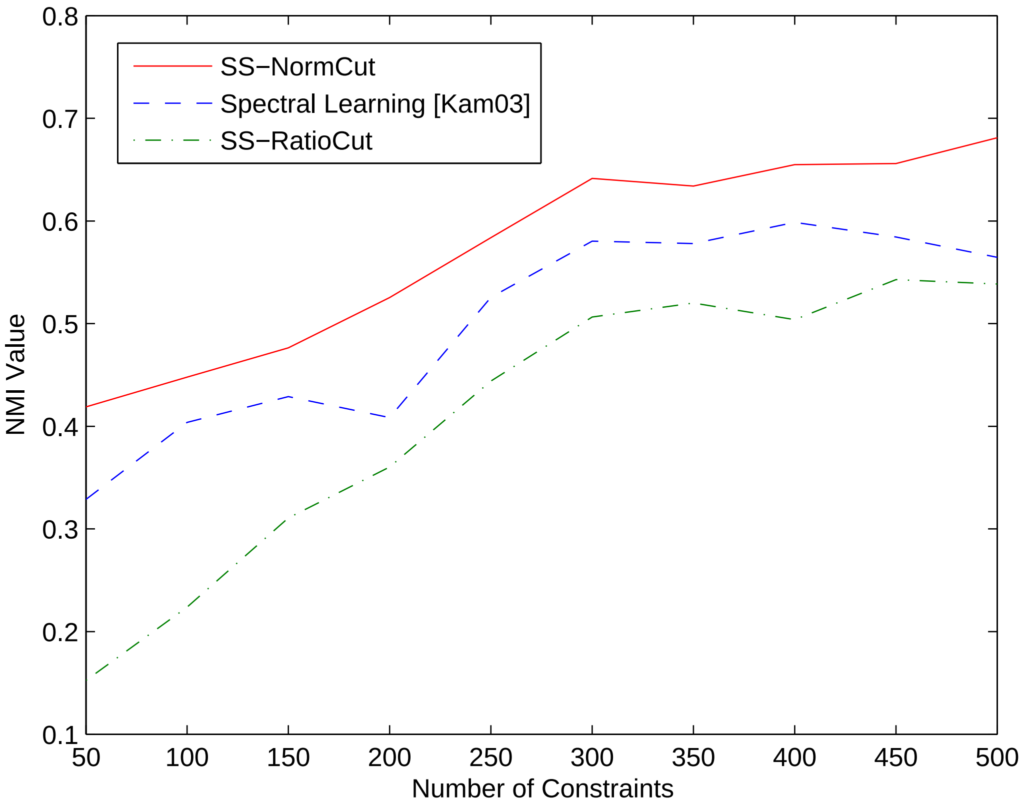
Handwritten Digits Data

- Subset of digits 3,8,9 from Pendigits (317 points in 16-D space)



Yeast Gene Network Data

- Interaction network for 216 yeast genes, labeled by KEGG functional pathway labels



Conclusion

- Introduced a framework that unifies graph-based and vector-based semi-supervised clustering
- Captures a number of semi-supervised clustering objectives, including HMRF_KMeans, Spectral Learning and new semi-supervised graph clustering objectives
- Kernel-based approach able to obtain better results on real-life data sets

References

Basu03 Basu, S., Bilenko, M., and Mooney, R. *A Probabilistic Framework for Semi-supervised Clustering*. KDD, 2004.

Dhi04 Dhillon, I., Guan, Y., and Kulis, B. *A Unified View of Kernel k-means, Spectral Clustering and Graph Partitioning*. TR-04-25, UT Austin, 2004.

Kam03 Kamvar, S., Klein, D., and Manning, C. *Spectral Learning*. IJCAI, 2003.

Shi00 Shi, J., and Malik, J. *Normalized Cuts and Image Segmentation*. IEEE PAMI, 2000.