A Tutorial on

Graph-based Semi-Supervised Learning Algorithms for NLP

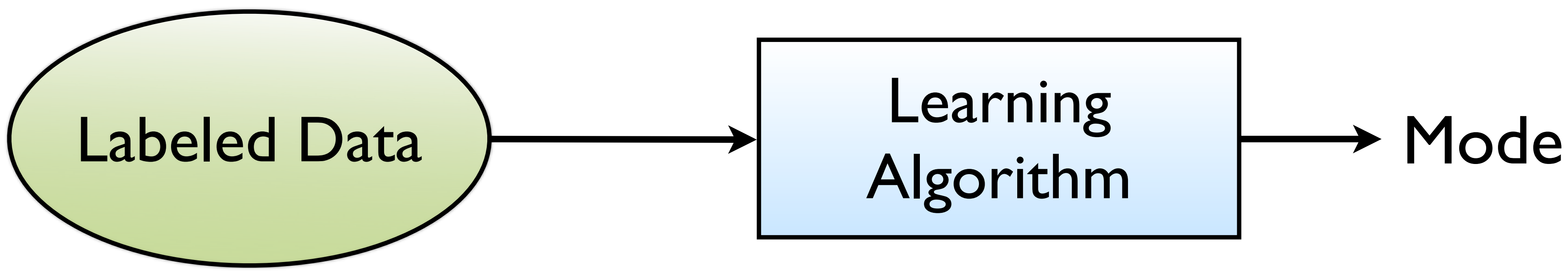
Amarnag Subramanya
(Google Research)

Partha Pratim Talukdar
(Carnegie Mellon University)

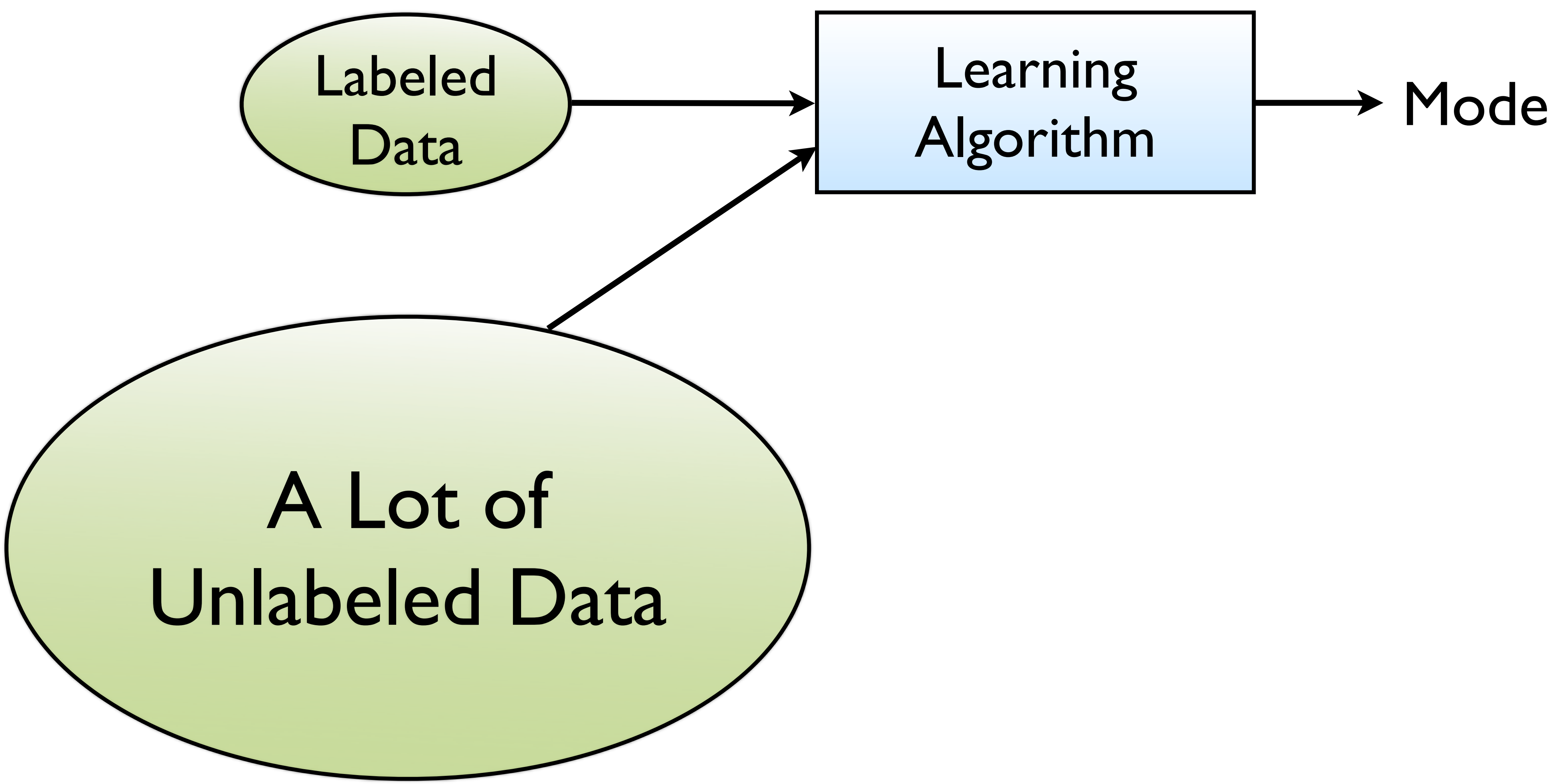
ACL 2012, South Korea

http://graph-ssl.wikidot.com/

Supervised Learning

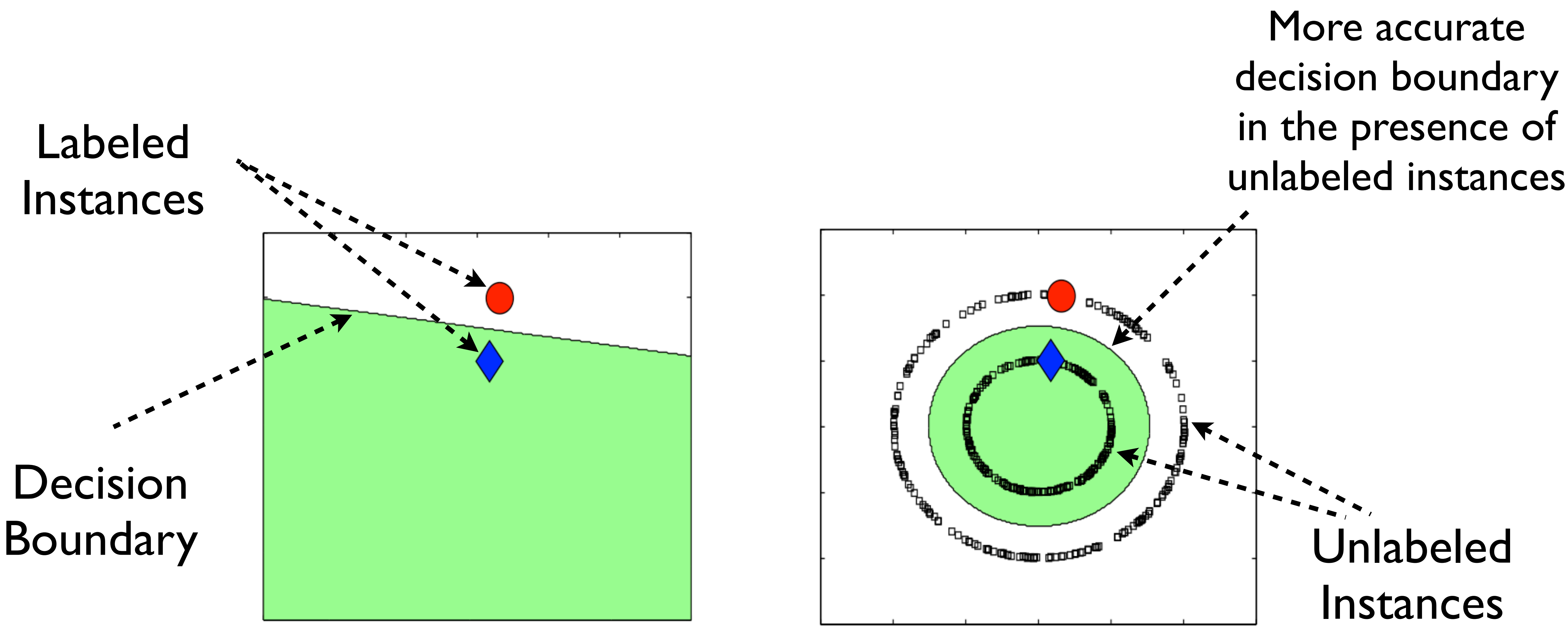


Semi-Supervised Learning (SSL)



Why SSL?

How can unlabeled data be helpful?



Without Unlabeled Data

With Unlabeled Data

Example from [Belkin et al., JMLR 2006]

Inductive vs Transductive

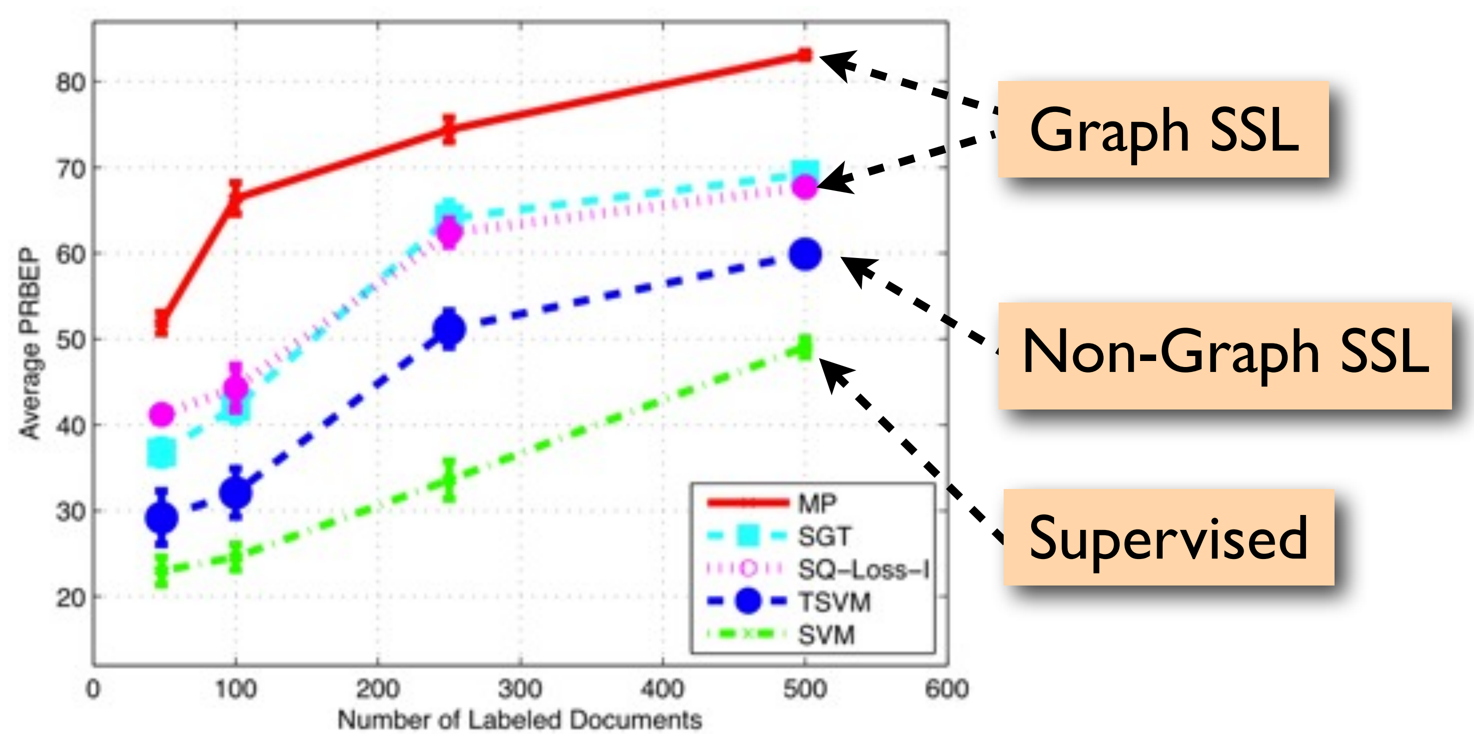
	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)	Manifold Regularization	Graph SSL algorithms

Most Graph SSL algorithms are non-parametric

See Chapter 25 of SSL Book: http://olivier.chapelle.cc/ssl-book/discussion.pdf

Why Graph-based SSL?

- Some datasets are naturally represented by a graph
  - web, citation network, social network, ...
- Uniform representation for heterogeneous data
- Easily parallelizable, scalable to large data
- Effective in practice

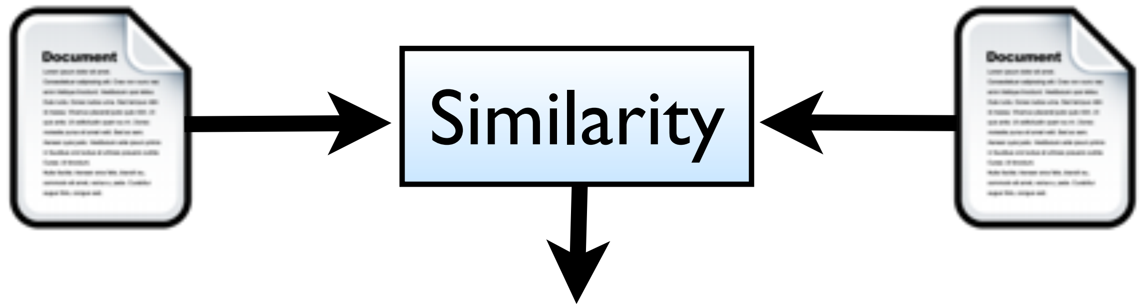


Text Classification

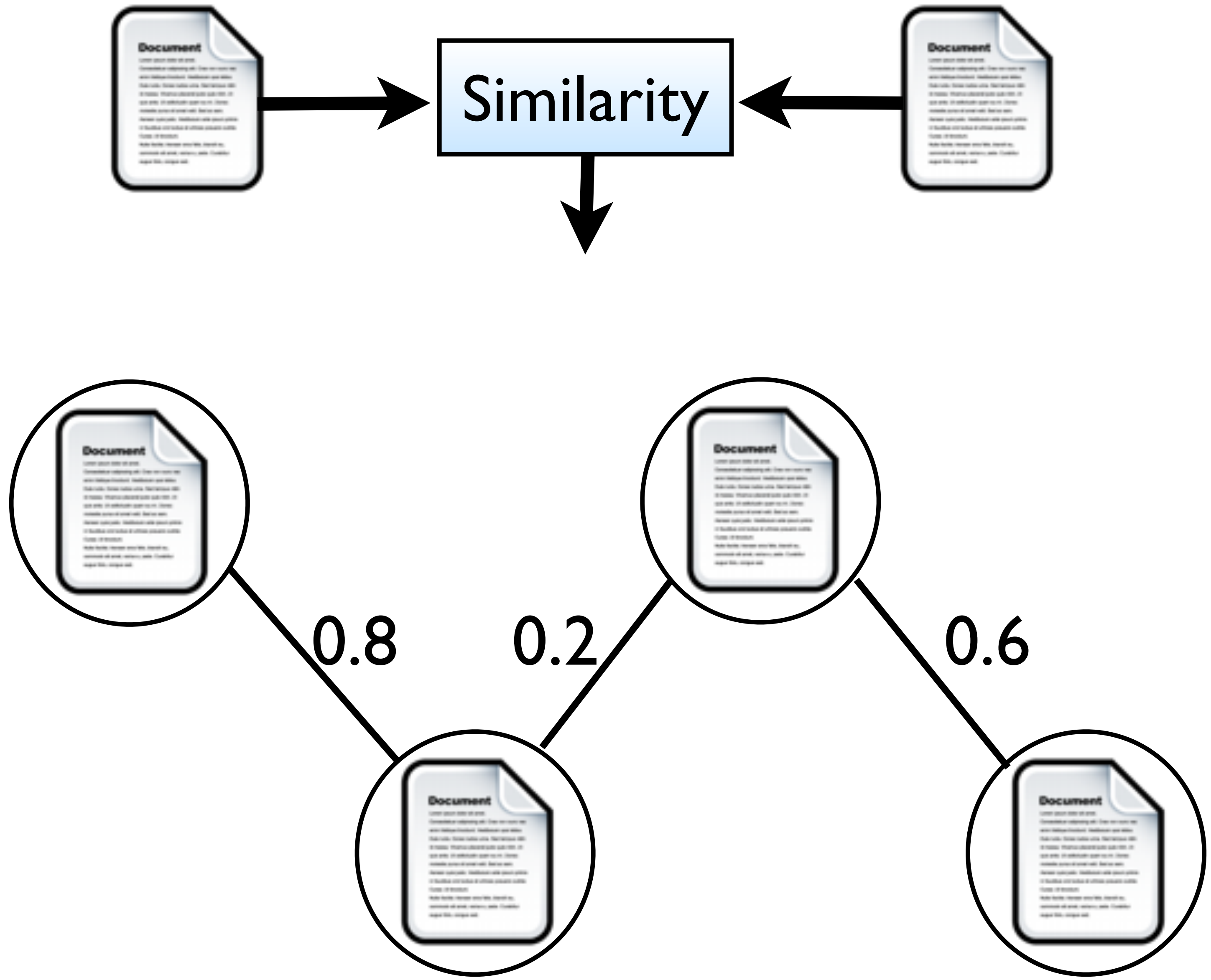
Graph-based SSL

Graph-based SSL

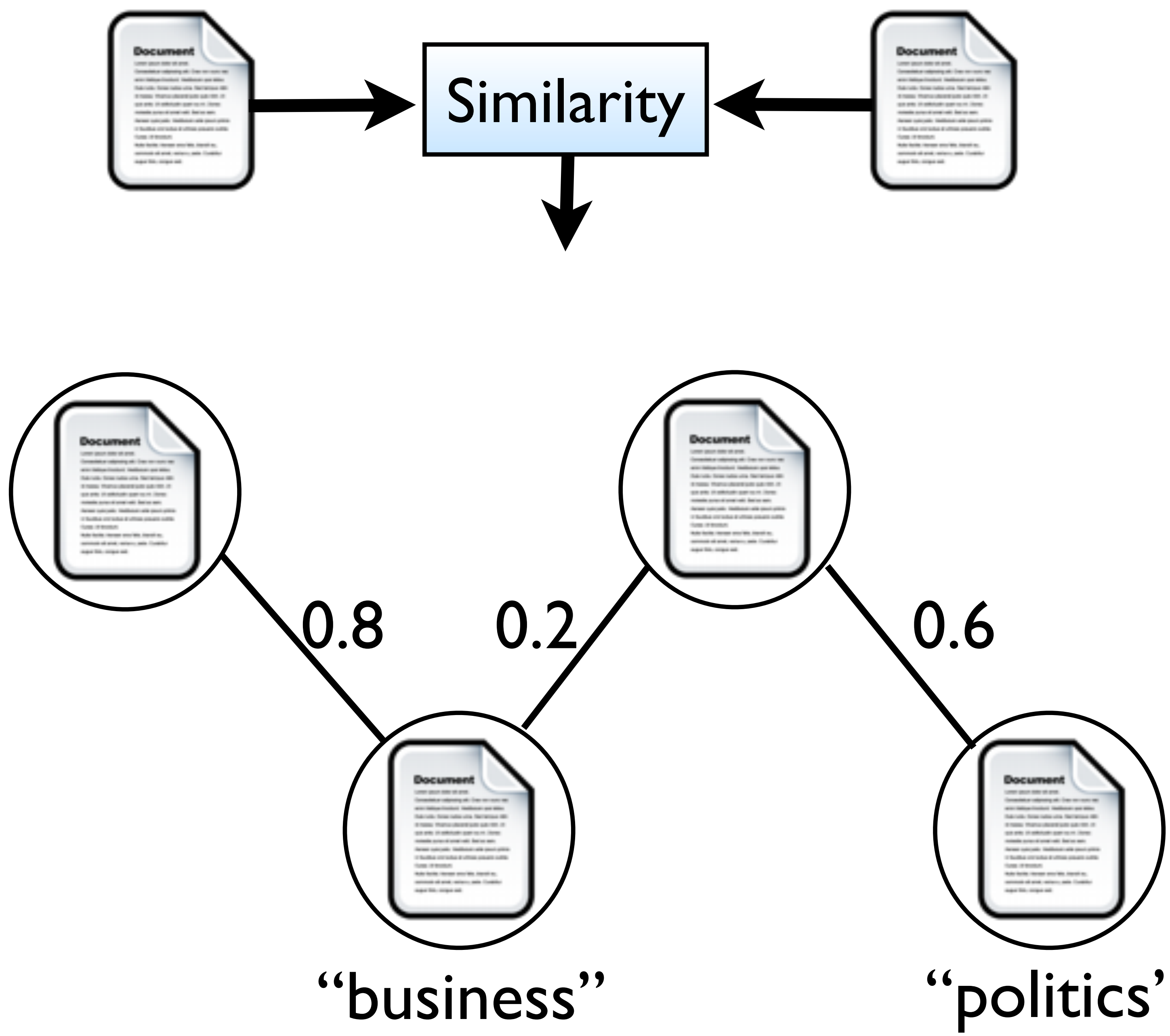
Graph-based SSL



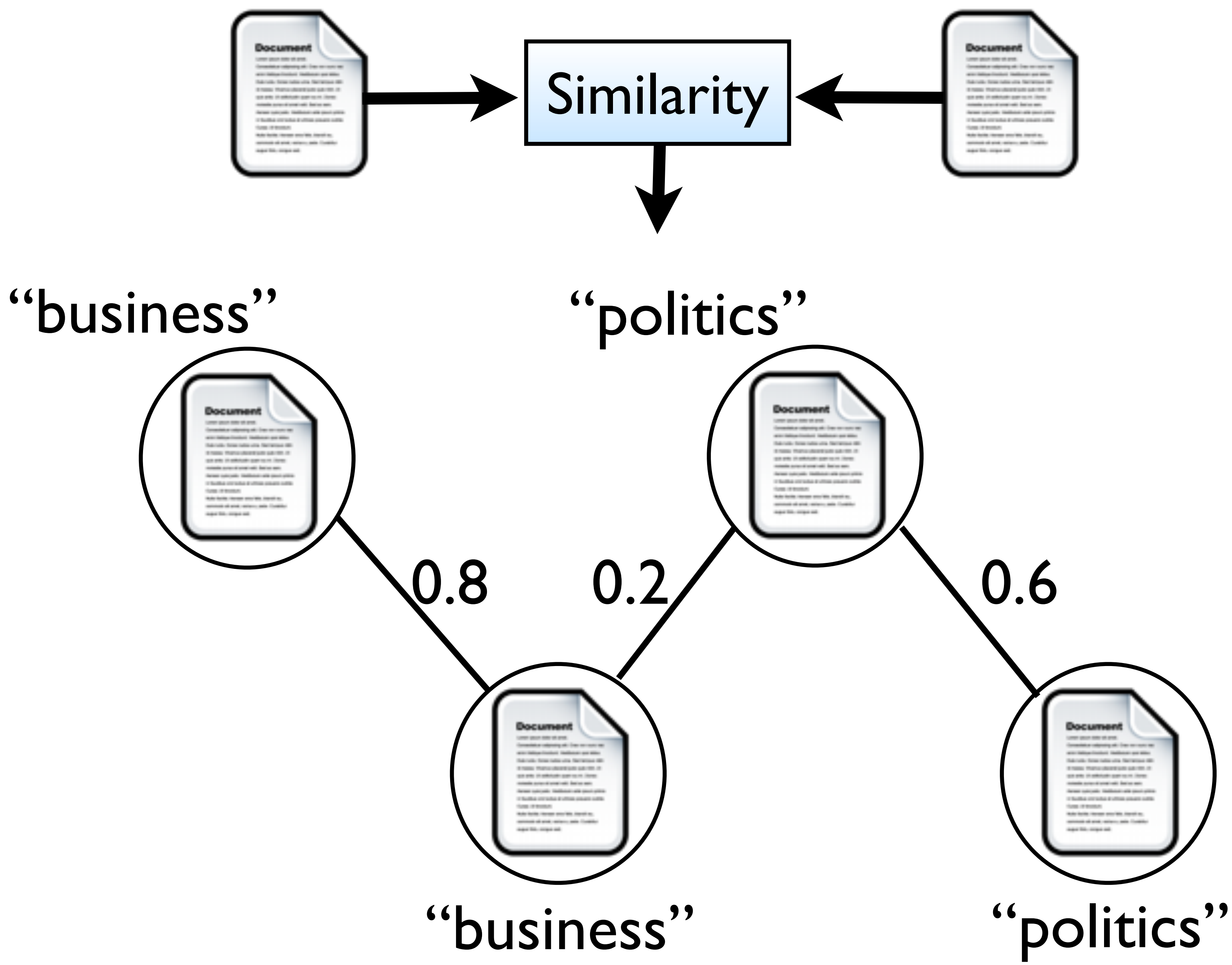
Graph-based SSL



Graph-based SSL



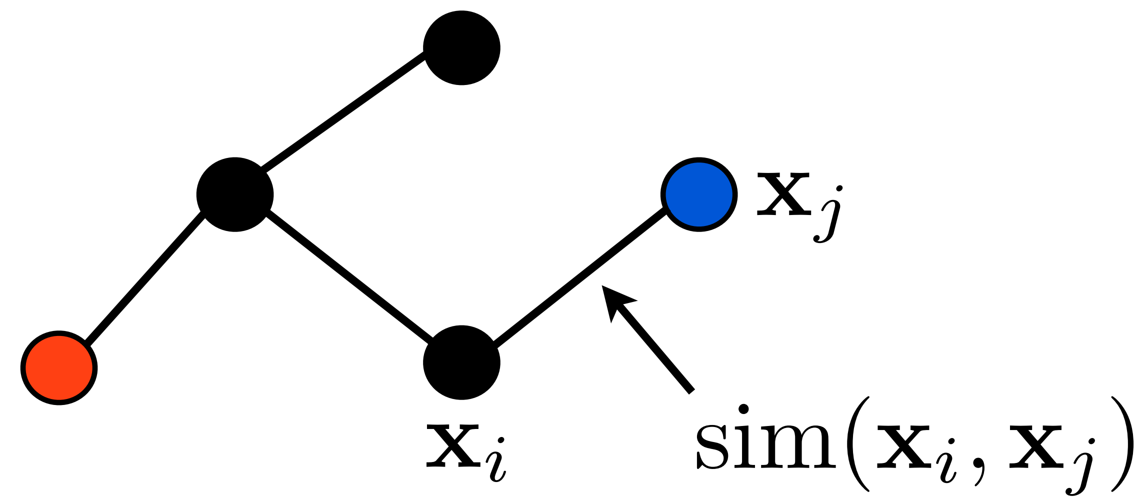
Graph-based SSL



Graph-based SSL

Smoothness Assumption

If two instances are <u>similar</u> according to the graph, then <u>output labels</u> should be <u>similar</u>



- Effective for both relational and IID data
- Two stages
  - Graph construction (if not already present)
  - Label Inference

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

Outline

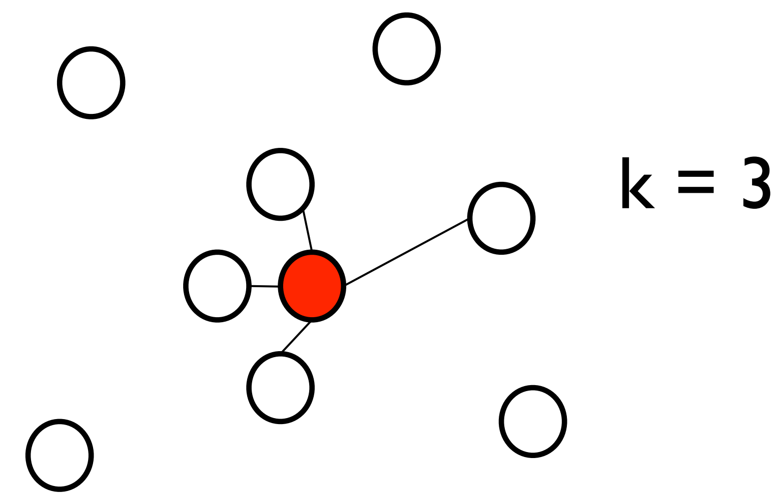
- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

Graph Construction

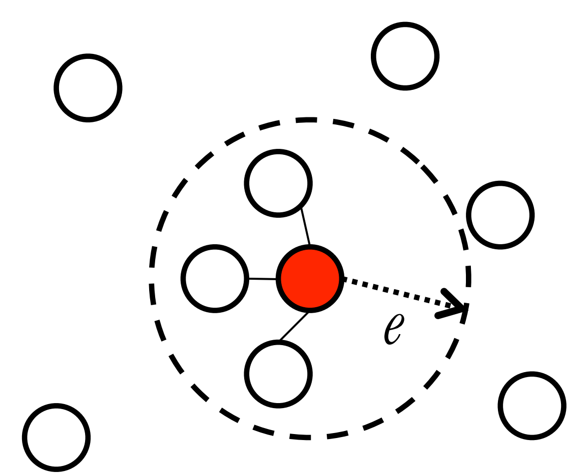
- Neighborhood Methods
  - k-NN Graph Construction
  - e-Neighborhood Method
- Metric Learning
- Other approaches

Neighborhood Methods

- k-Nearest Neighbor (k-NN)
  - add edges between an instance and its k-nearest neighbors

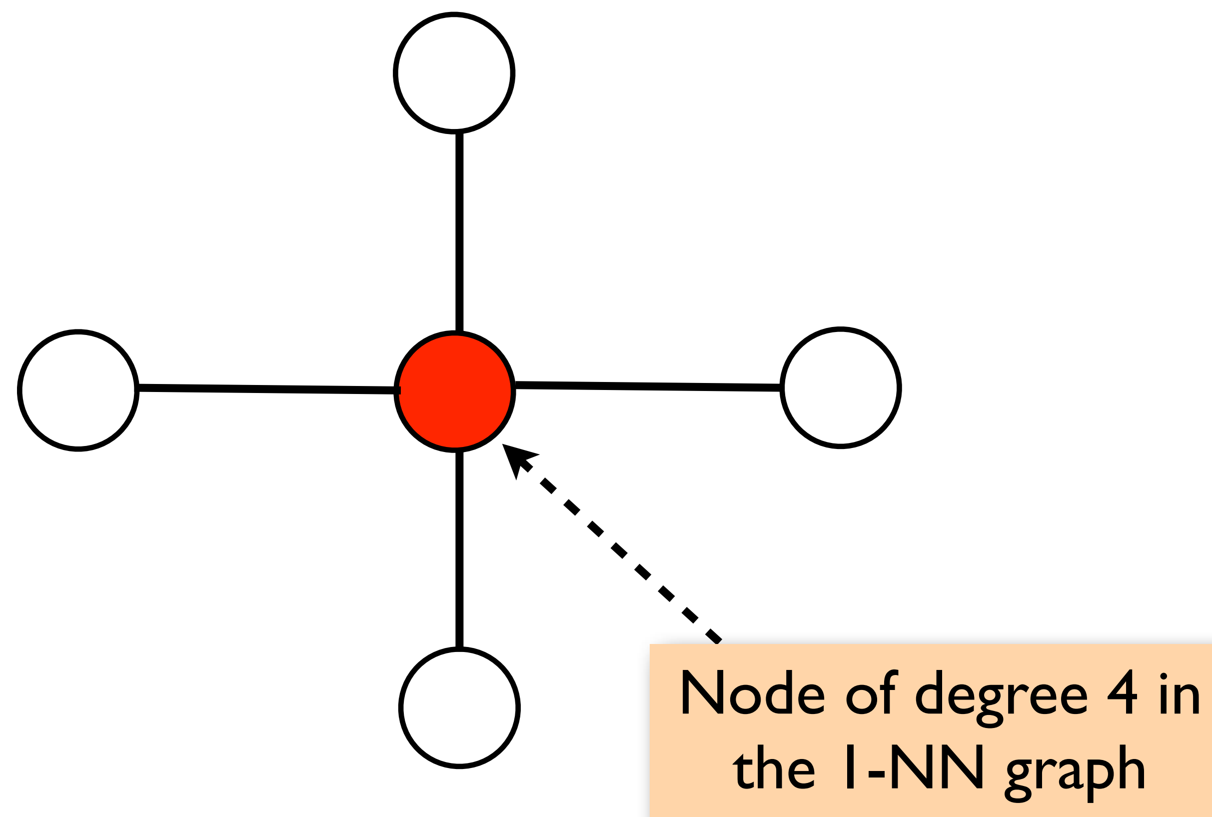


- e-Neighborhood
  - add edges to all instances inside a ball of radius e



Issues with k-NN graphs

- Not scalable (quadratic)
- Results in an asymmetric graph
- Results in irregular graphs
  - some nodes may end up with higher degree than other nodes



Issues with *e*-Neighborhood

- Fragmented Graph: disconnected components
- Sensitive to value of e : not invariant to scaling
- Not scalable

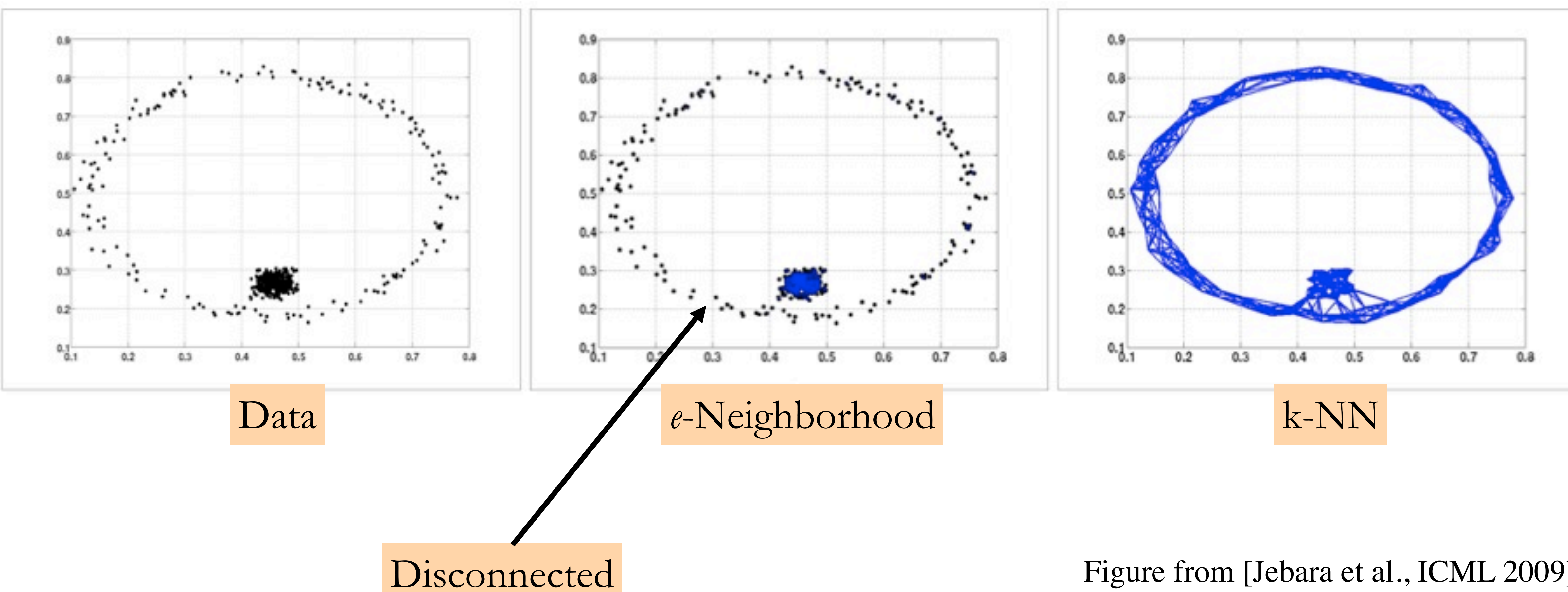
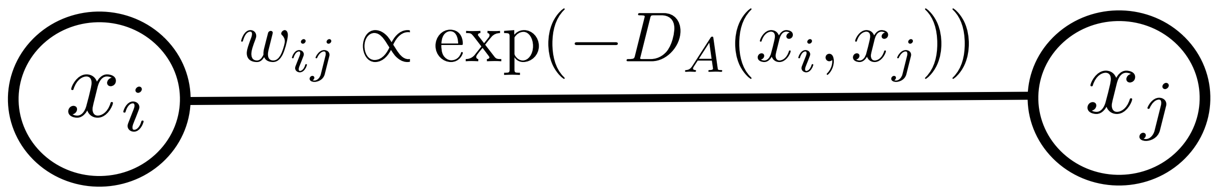


Figure from [Jebara et al., ICML 2009]

Graph Construction using Metric Learning



$$D_A(x_i, x_j) = (x_i - x_j)^T A(x_i - x_j)$$

Estimated using Mahalanobis metric learning algorithms

- Supervised Metric Learning
  - ITML [Kulis et al., ICML 2007]
  - LMNN [Weinberger and Saul, JMLR 2009]
- Semi-supervised Metric Learning
  - IDML [Dhillon et al., UPenn TR 2010]

Benefits of Metric Learning for Graph Construction

Graph constructed using **supervised** metric learning

Graph constructed using **semi-supervised** metric learning [Dhillon et al, 2010]

Datasets	Original	RP	PCA	ITML	LMNN	IDML-LM	IDML-IT
Amazon	0.4046	0.3964	0.1554	0.1418	0.2405	0.2004	**0.1265**
Newsgroups	0.3407	0.3871	0.3098	**0.1664**	0.2172	0.2136	**0.1664**
Reuters	0.2928	0.3529	0.2236	0.1088	0.3093	0.2731	**0.0999**
EnronA	0.3246	0.3493	0.2691	0.2307	0.1852	**0.1707**	0.2179
Text	0.4523	0.4920	0.4820	0.3072	0.3125	0.3125	**0.2893**
USPS	**0.0639**	0.0829	–	0.1096	0.1336	0.1225	0.0834
BCI	0.4508	0.4692	–	0.4217	0.3058	**0.2967**	0.4081
Digit	**0.0218**	0.0250	–	0.0281	0.1186	0.0877	0.0281

Table 3. Comparison of transductive classification performance over graphs constructed using different methods (see Section 6.1), with $n_l = 100$ and $n_u = 1400$.

Careful graph construction is critical!

[Dhillon et al., UPenn TR 2010]

Other Graph Construction Approaches

- Local Reconstruction
  - Linear Neighborhood [Wang and Zhang, ICML 2005]
  - Regular Graph: b-matching [Jebara et al., ICML 2008]
  - Fitting Graph to Vector Data [Daitch et al., ICML 2009]
- Graph Kernels
  - [Zhu et al., NIPS 2005]

Outline

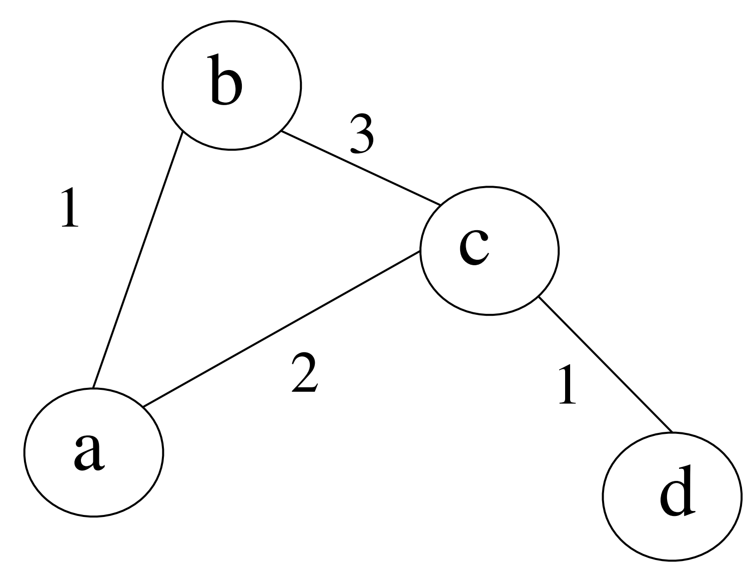
- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Graph Laplacian

- Laplacian (un-normalized) of a graph:

$$L = D - W, \text{where } D_{ii} = \sum_j W_{ij},\ \ D_{ij(\neq i)} = 0$$

$$\begin{array}{c c} & \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{array}{c} a \\ b \\ c \\ d \end{array} & \begin{pmatrix} \mathbf{3} & \mathbf{-1} & \mathbf{-2} & \mathbf{0} \\ \mathbf{-1} & \mathbf{4} & \mathbf{-3} & \mathbf{0} \\ \mathbf{-2} & \mathbf{-3} & \mathbf{6} & \mathbf{-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{1} \end{pmatrix} \end{array}$$



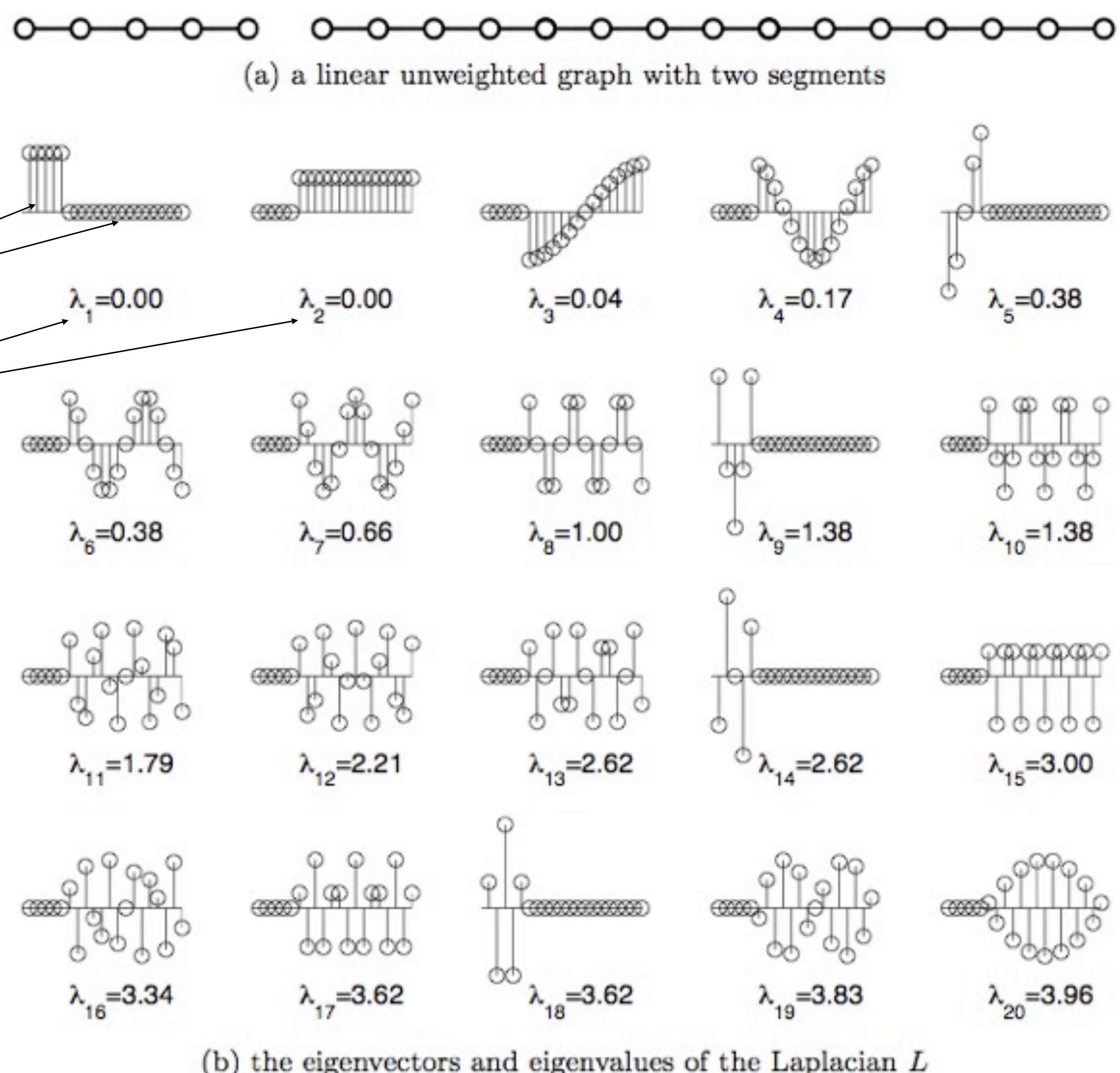
Graph Laplacian (contd.)

- *L* is positive semi-definite (with non-negative weights)
- Smoothness of function *f* over the graph in terms of the Laplacian:

$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$

Measure of Smoothness

Spectrum of the Laplacian



Constant within component

Number of connected components = Number of 0 eigenvalues

Higher Eigenvalue, Irregular Eigenvector, Less smoothness

Figure from [Zhu et al., 2005]

Notations

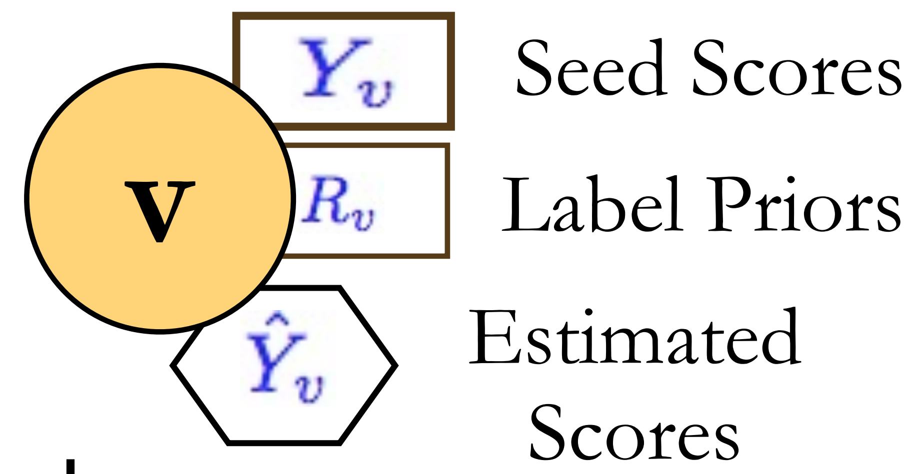
$\hat{Y}_{v,l}$: score of estimated label l on node v

$Y_{v,l}$: score of seed label l on node v

$R_{v,l}$: regularization target for label l on node v

S : seed node indicator (diagonal matrix)

W_{uv} : weight of edge (u, v) in the graph



Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

LP-ZGL [Zhu et al., ICML 2003]

Smooth

$$\arg\min_{\hat{Y}} \boxed{\sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2} = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l$$

such that $\boxed{Y_{ul} = \hat{Y}_{ul},\ \forall S_{uu} = 1}$

Match Seeds (hard)

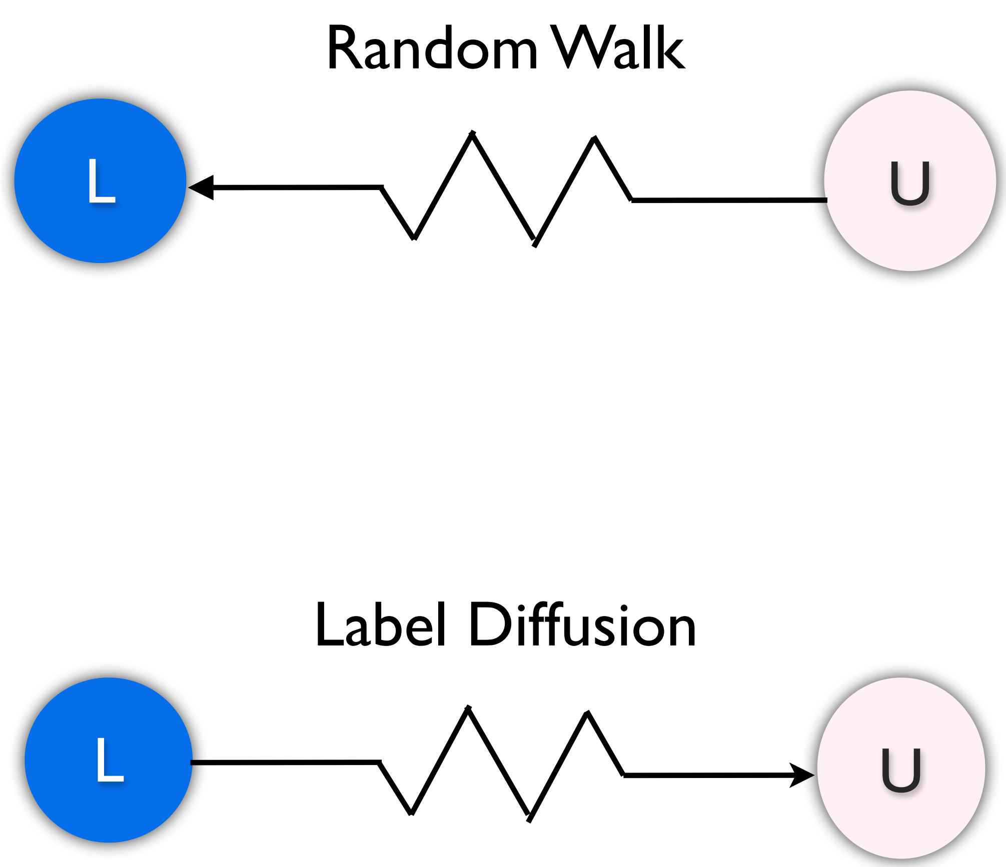
Graph Laplacian

- **Smoothness**
  - two nodes connected by an edge with high weight should be assigned similar labels
- Solution satisfies harmonic property

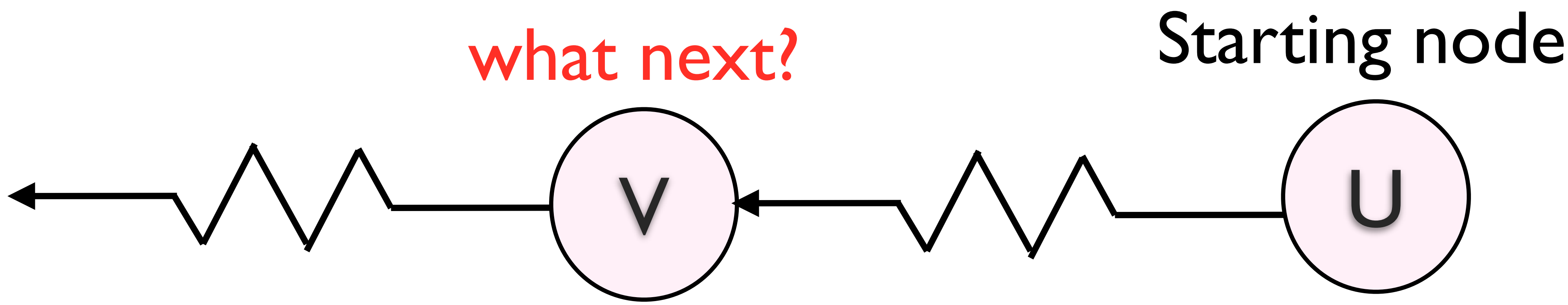
Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Two Related Views



Random Walk View

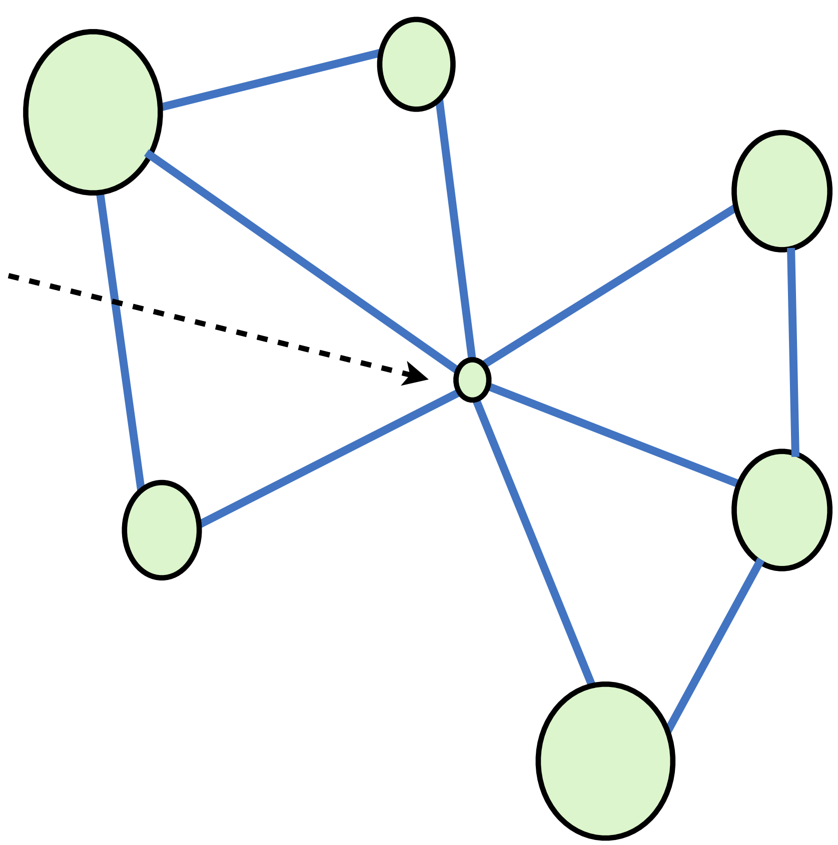


- Continue walk with probability $\mathbf{p_v^{cont}}$
- Assign V's seed label to U with probability $\mathbf{p_v^{inj}}$
- Abandon random walk with probability $\mathbf{p_v^{abnd}}$
  - assign U a dummy label

Discounting Nodes

- Certain nodes can be unreliable (e.g., high degree nodes)
  - do not allow propagation/walk through them
- Solution: increase abandon probability on such nodes:

$$\mathbf{p_v^{abnd}} \propto \mathrm{degree(v)}$$



Redefining Matrices

$$W'_{uv} = p_u^{cont} \times W_{uv}$$

New Edge Weight

$$S_{uu} = \sqrt{p_u^{inj}}$$

$$R_{u\top} = p_u^{abnd}$$, and 0 for non-dummy labels

Dummy Label

Modified Adsorption (MAD)

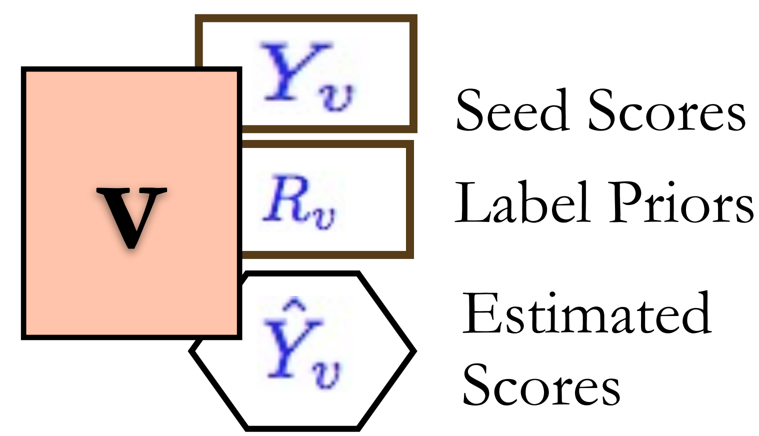
[Talukdar and Crammer, ECML 2009]

Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\mathbf{Y}}} \sum_{l=1}^{m+1} \left[\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2 + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, +1 dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v

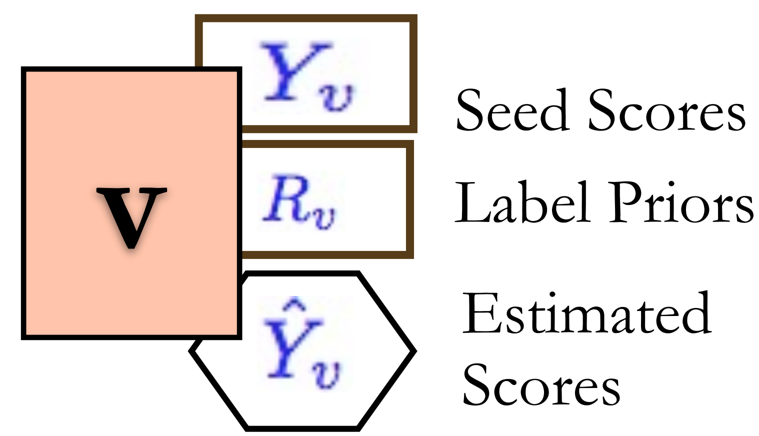


Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\overset{\text{Match Seeds (soft)}}{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}} + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv} (\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v

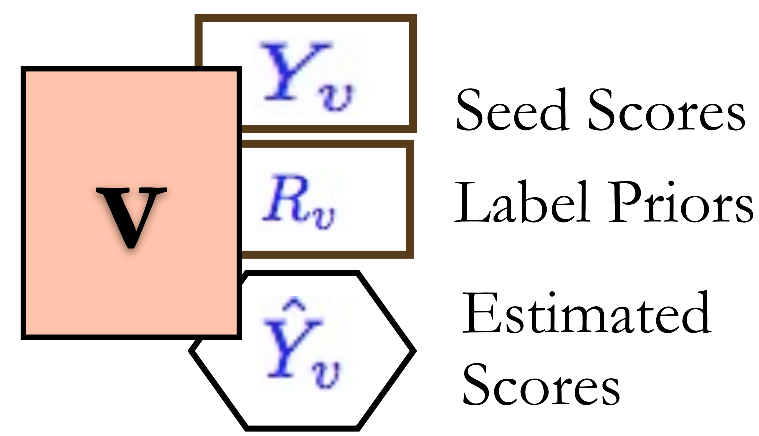


Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\mathbf{Y}}} \sum_{l=1}^{m+1} \left[\overset{\text{Match Seeds (soft)}}{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}} + \mu_1 \overset{\text{Smooth}}{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2}} + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v

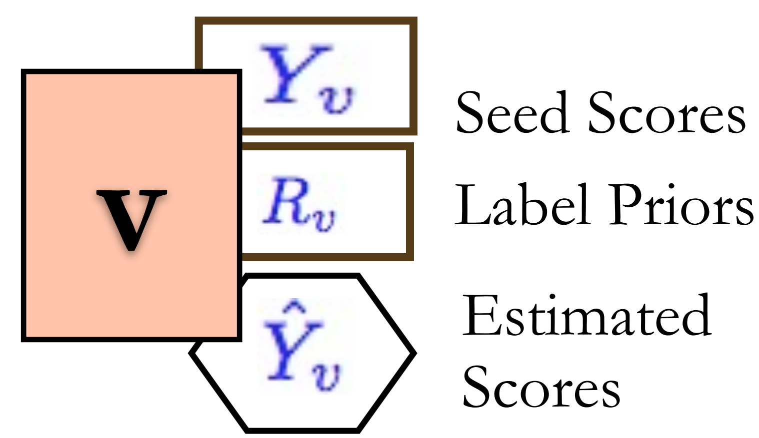


Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\mathbf{Y}}} \sum_{l=1}^{m+1} \left[\overbrace{\boxed{\|\boldsymbol{S}\hat{\mathbf{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}}^{\text{Match Seeds (soft)}} + \mu_1 \overbrace{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\mathbf{Y}}_{ul} - \hat{\mathbf{Y}}_{vl})^2}}^{\text{Smooth}} + \overbrace{\boxed{\mu_2 \|\hat{\mathbf{Y}}_l - \boldsymbol{R}_l\|^2}}^{\text{Match Priors (Regularizer)}} \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



Modified Adsorption (MAD)

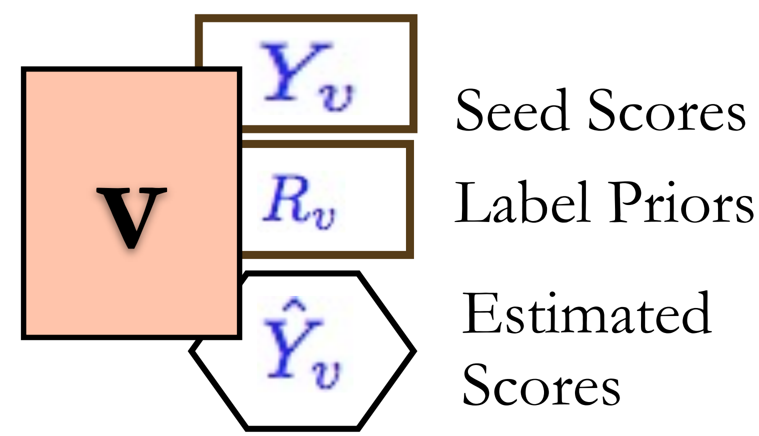
[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\overset{\text{Match Seeds (soft)}}{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}} + \mu_1 \overset{\text{Smooth}}{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2}} + \overset{\text{Match Priors (Regularizer)}}{\boxed{\mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2}} \right]$$

- m labels, $+1$ dummy label

for *none-of-the-above* label

- $\boldsymbol{M} =$ ...ed weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



Modified Adsorption (MAD)

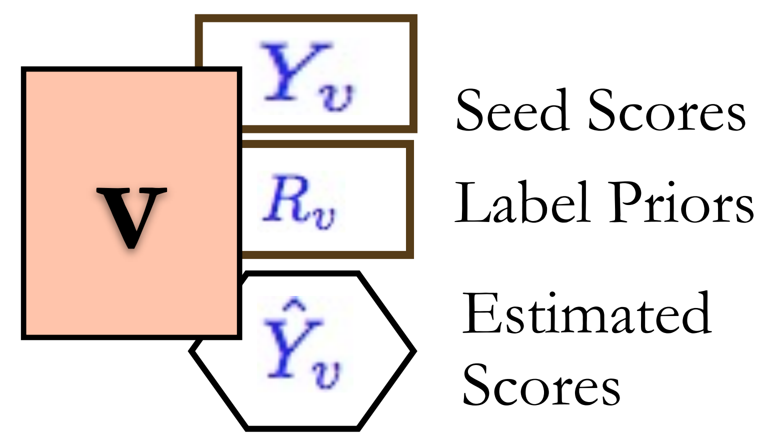
[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\mathbf{Y}}} \sum_{l=1}^{m+1} \left[\overset{\text{Match Seeds (soft)}}{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}} + \mu_1 \overset{\text{Smooth}}{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2}} + \overset{\text{Match Priors (Regularizer)}}{\boxed{\mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2}} \right]$$

- m labels, $+1$ dummy label

for *none-of-the-above* label

- $\boldsymbol{M} =$ ed weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



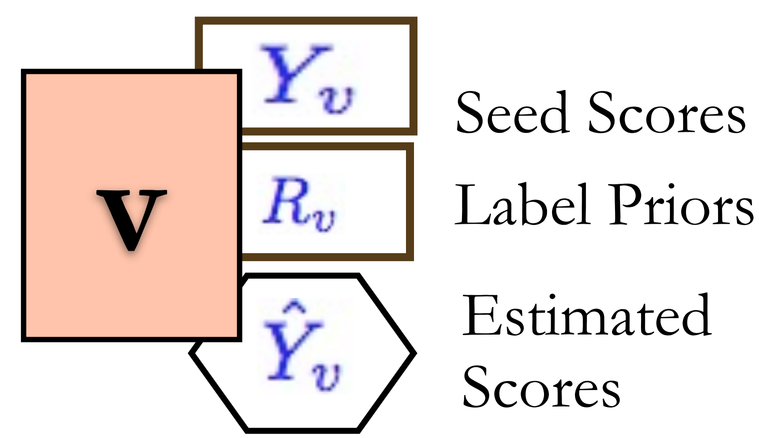
MAD has extra regularization compared to LP-ZGL [Zhu et al, ICML 03]; similar to QC [Bengio et al, 2006]

Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\overset{\text{Match Seeds (soft)}}{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}} + \mu_1 \overset{\text{Smooth}}{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2}} + \overset{\text{Match Priors (Regularizer)}}{\boxed{\mu_2\|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2}} \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} =$ for *none-of-the-above* label ed weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



MAD's Objective is Convex

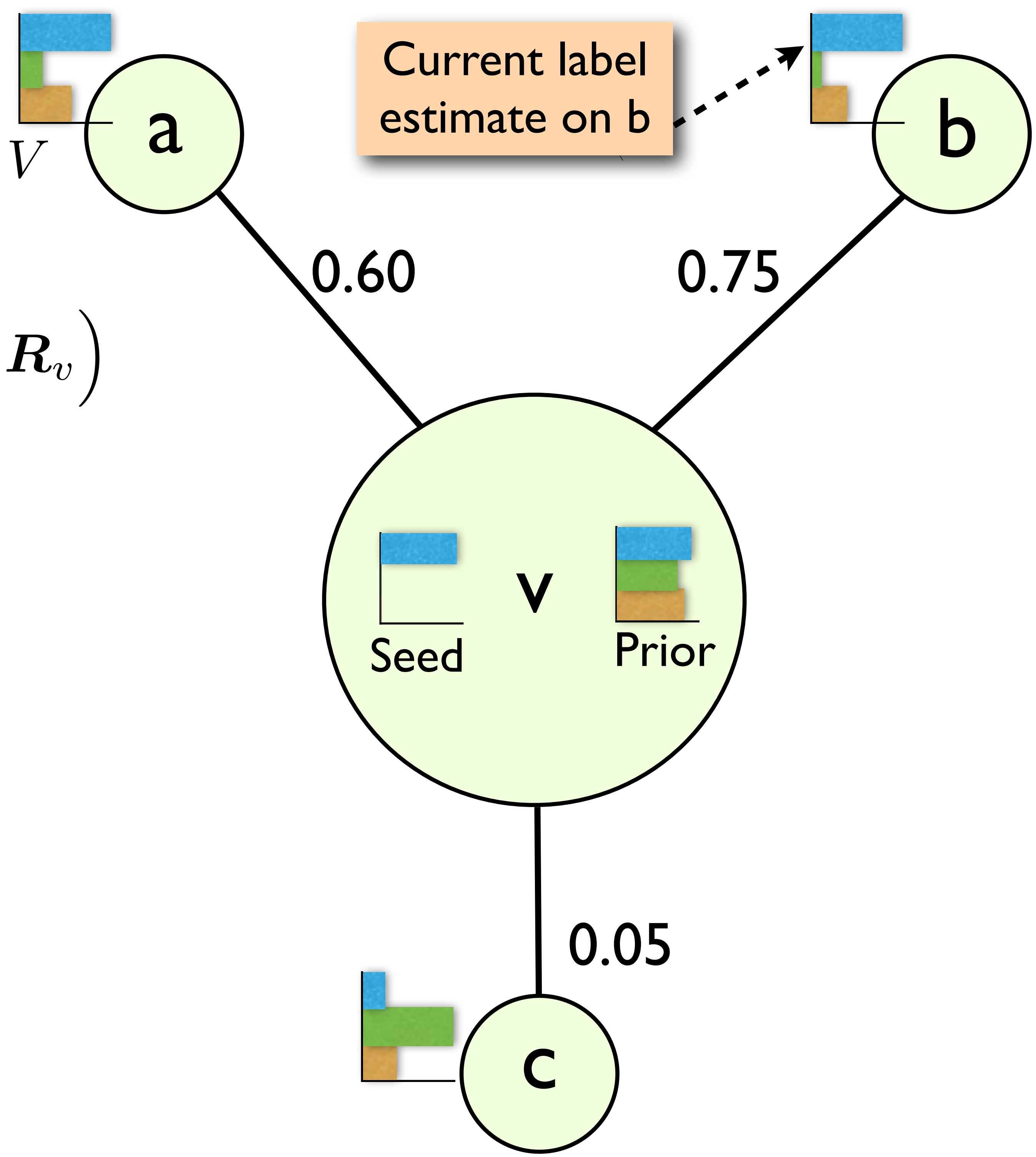
MAD has extra regularization compared to LP-ZGL [Zhu et al, ICML 03]; similar to QC [Bengio et al, 2006]

Solving MAD Objective

- Can be solved using matrix inversion (like in LP)
  - but matrix inversion is expensive (cubic)
- Instead solved exactly using a system of linear equations
  - solved using Jacobi iterations
  - results in iterative updates
  - guaranteed convergence
  - see [Bengio et al., 2006] and
    [Talukdar and Crammer, ECML 2009] for details

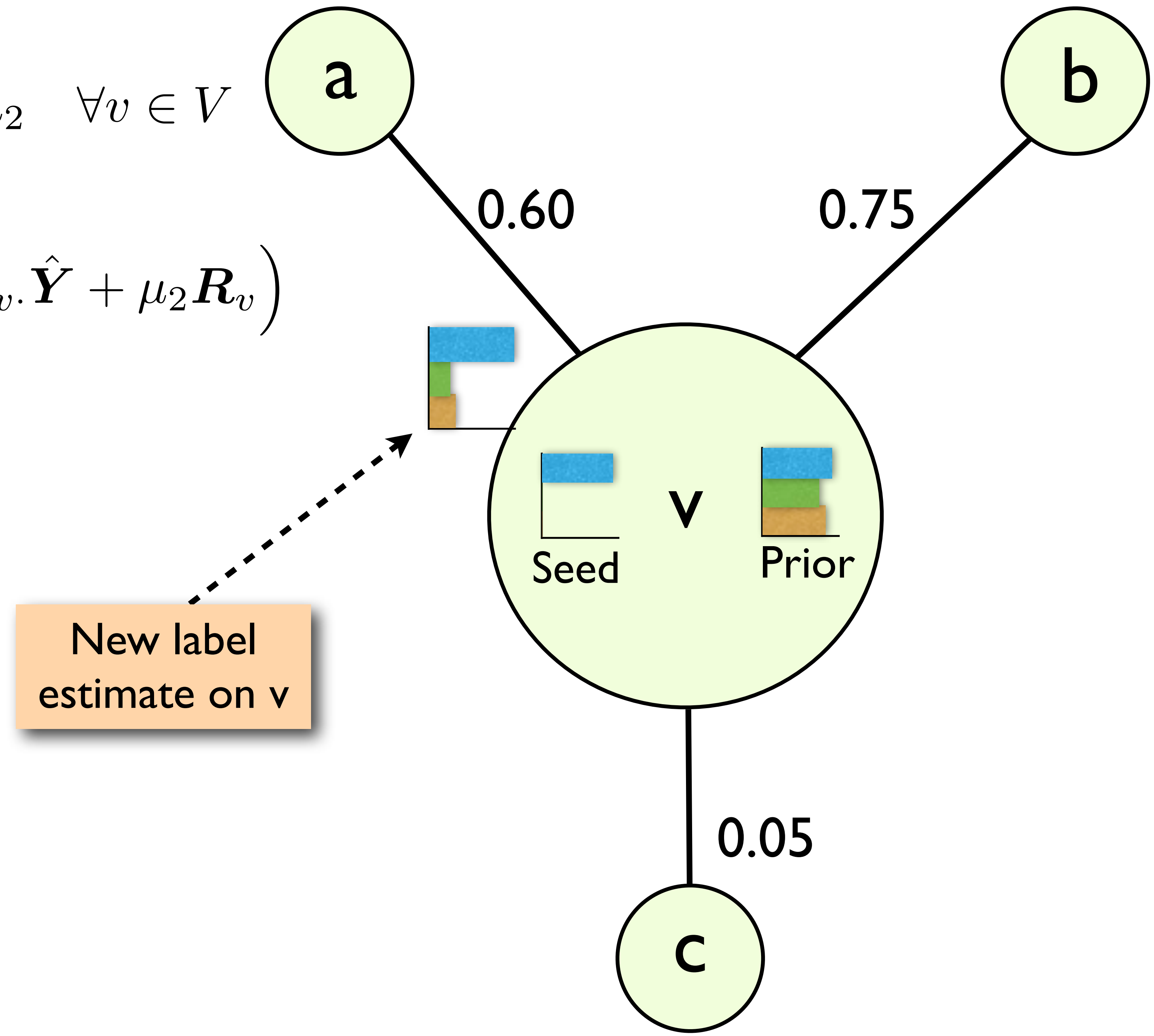
Solving MAD using Iterative Updates

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L|+1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal
$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$
$\boldsymbol{M} = \boldsymbol{W}' + \boldsymbol{W}'^{\dagger}$
$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$
repeat
  for all $v \in V$ **do**
    $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v}\left((\boldsymbol{SY})_v + \mu_1 \boldsymbol{M}_{v\cdot}\hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v\right)$
  end for
until convergence



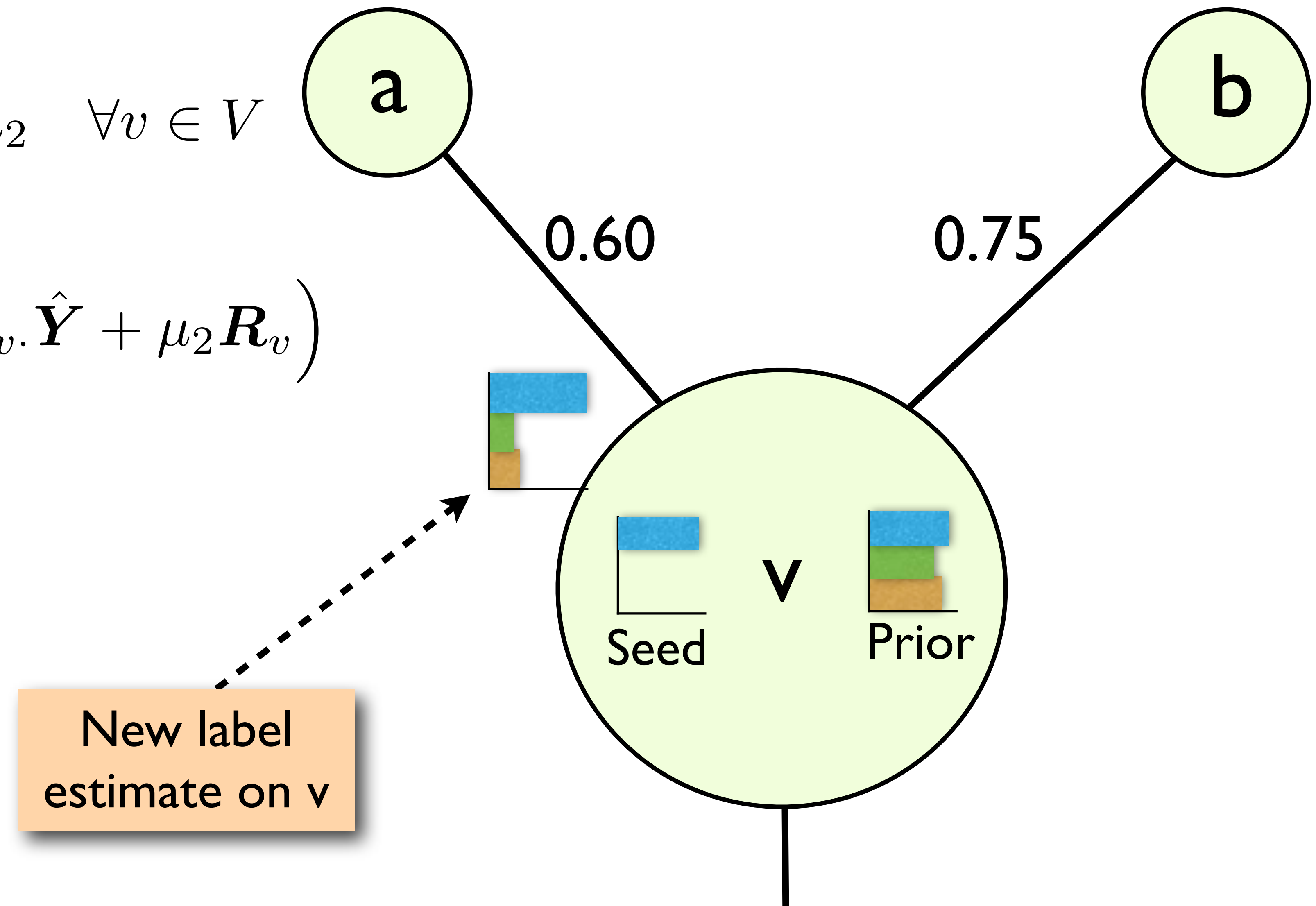
Solving MAD using Iterative Updates

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L|+1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal
$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$
$\boldsymbol{M} = \boldsymbol{W}' + \boldsymbol{W}'^{\dagger}$
$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$
repeat
  for all $v \in V$ **do**
    $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v}\left((\boldsymbol{SY})_v + \mu_1 \boldsymbol{M}_{v\cdot}\hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v\right)$
  end for
until convergence



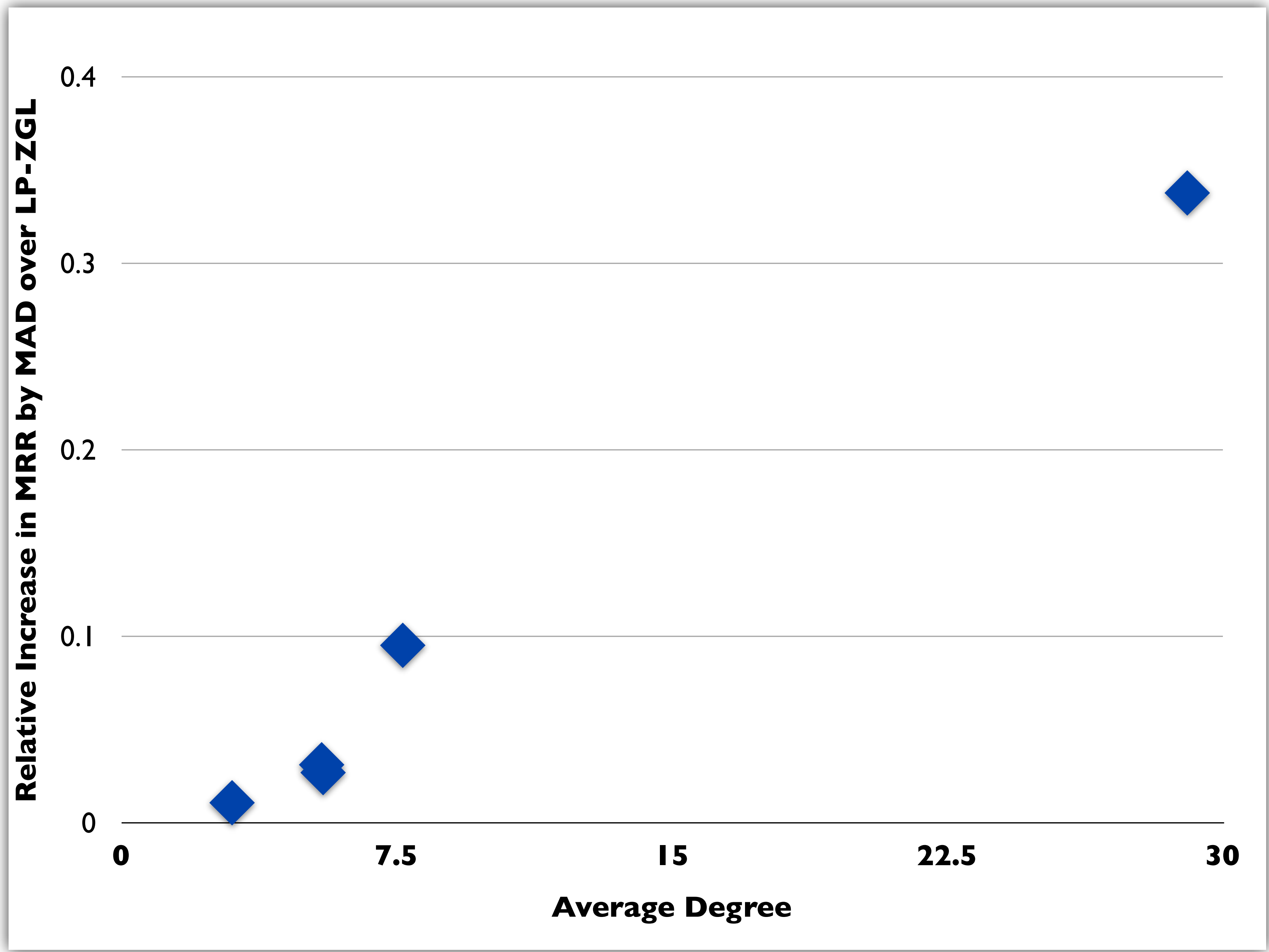
Solving MAD using Iterative Updates

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L|+1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal
$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$
$\boldsymbol{M} = \boldsymbol{W}' + \boldsymbol{W}^{\dagger}$
$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$
repeat
  for all $v \in V$ **do**
    $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v}\left((\boldsymbol{SY})_v + \mu_1 \boldsymbol{M}_{v\cdot}\hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v\right)$
  end for
until convergence



- Importance of a node can be discounted
- Easily Parallelizable: Scalable (more later)

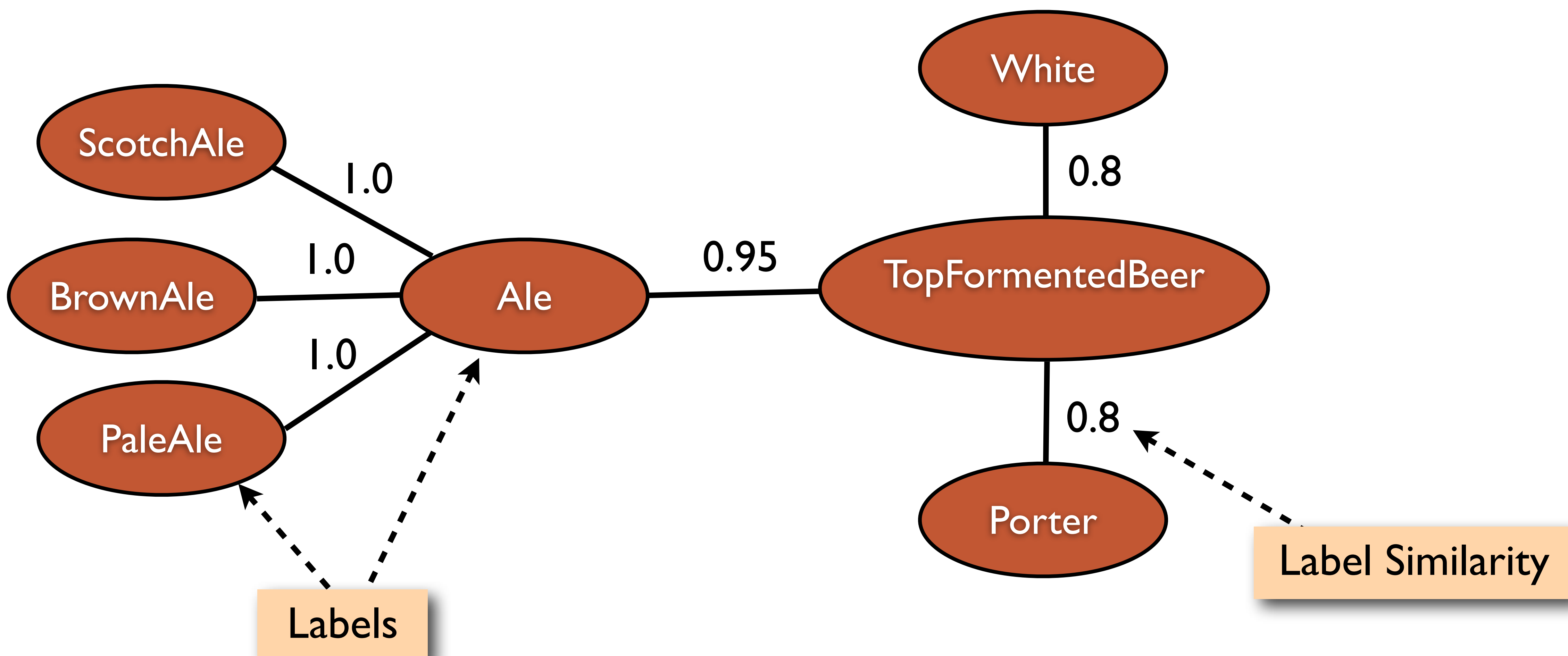
When is MAD most effective?



MAD is particularly effective in denser graphs, where there is greater need for regularization.

Extension to Dependent Labels

Labels are not always mutually exclusive



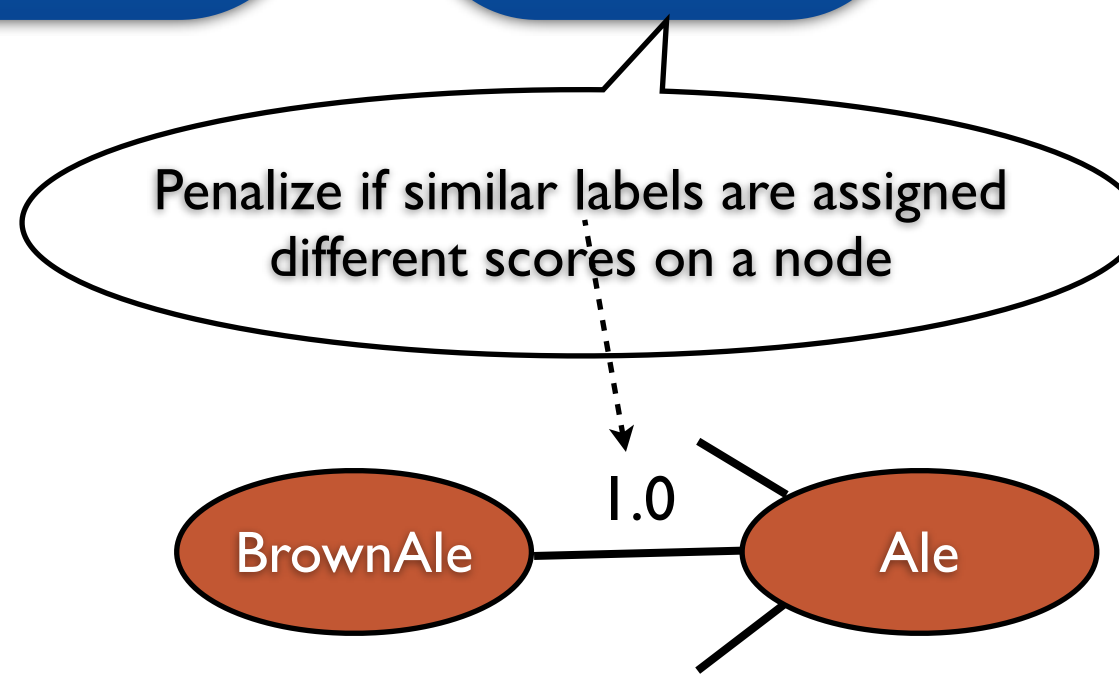
Label Graph

(over Beer Categories)

MAD with Dependent Labels (MADDL) [Talukdar and Crammer, ECML 2009]

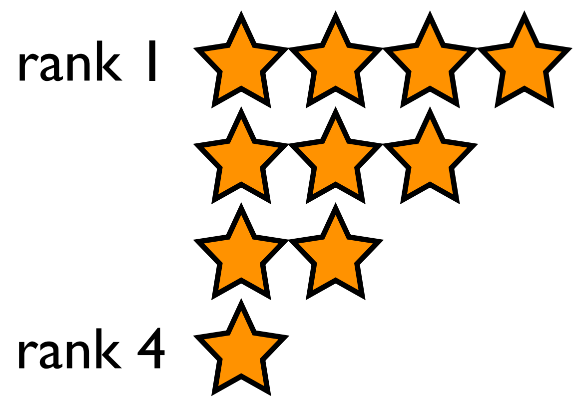
MADDL Objective

min Seed Label Loss (if any) + Edge Smoothness Loss + Label Prior Loss (e.g. prior on dummy label) + Dependent Label Loss



MADDL objective results in a scalable iterative update, with convergence guarantee.

Smooth Sentiment Ranking

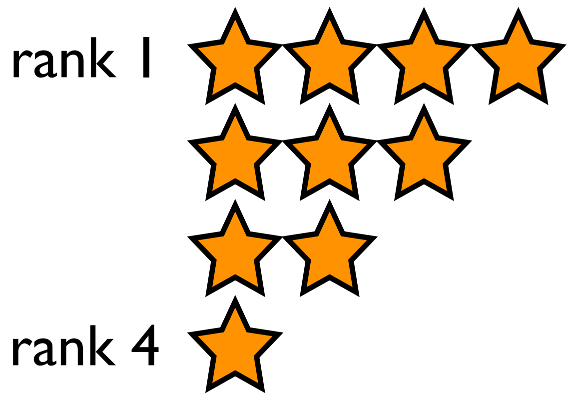


smooth predictions

Smooth Sentiment Ranking

Smooth Sentiment Ranking

Prefer



smooth predictions

over

non-smooth predictions

Smooth Sentiment Ranking

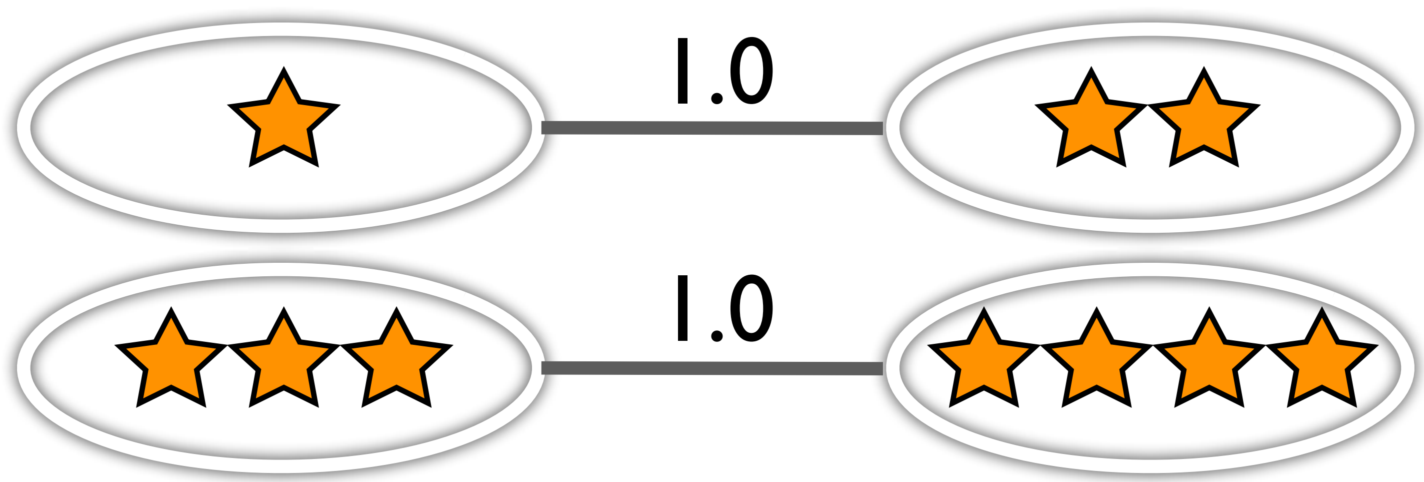
Prefer

smooth predictions

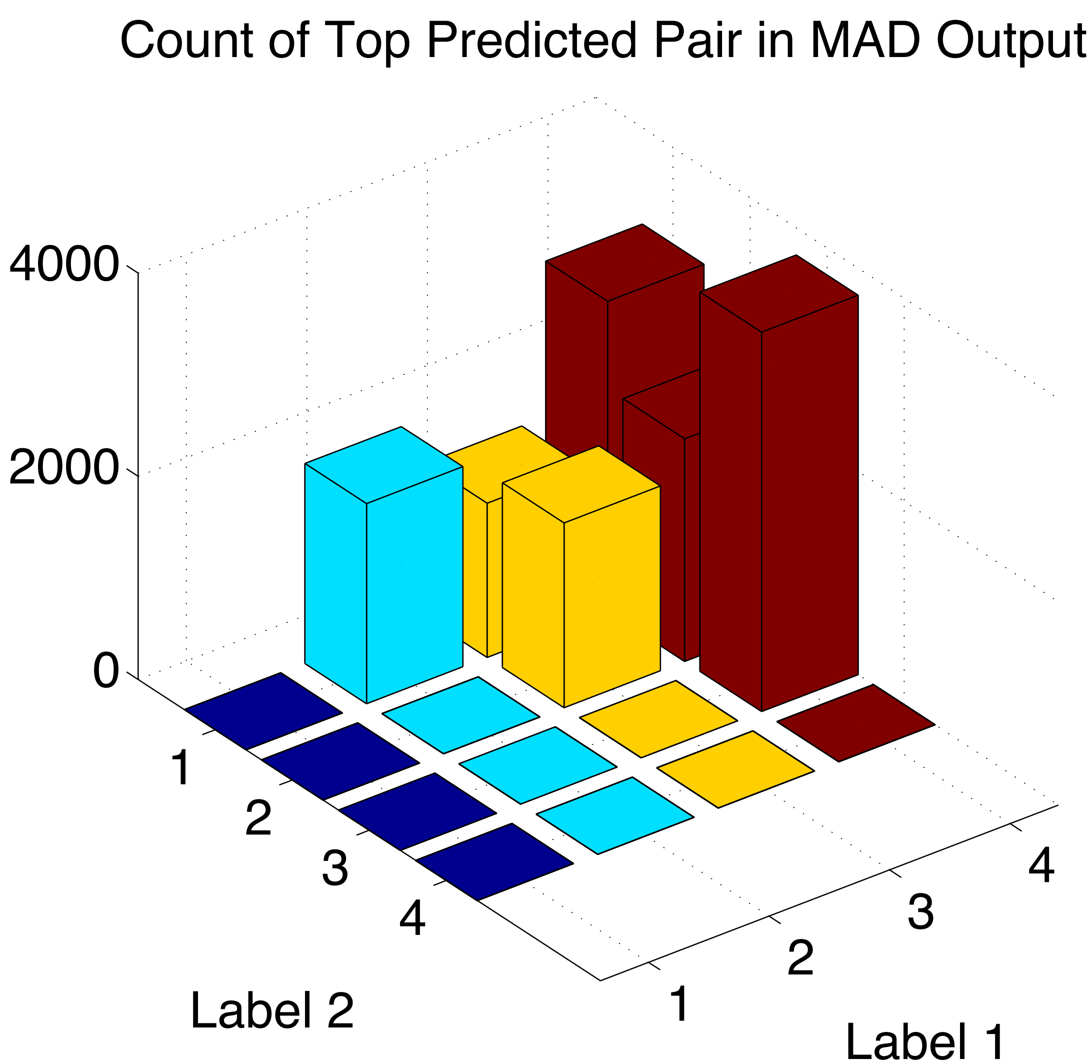
over

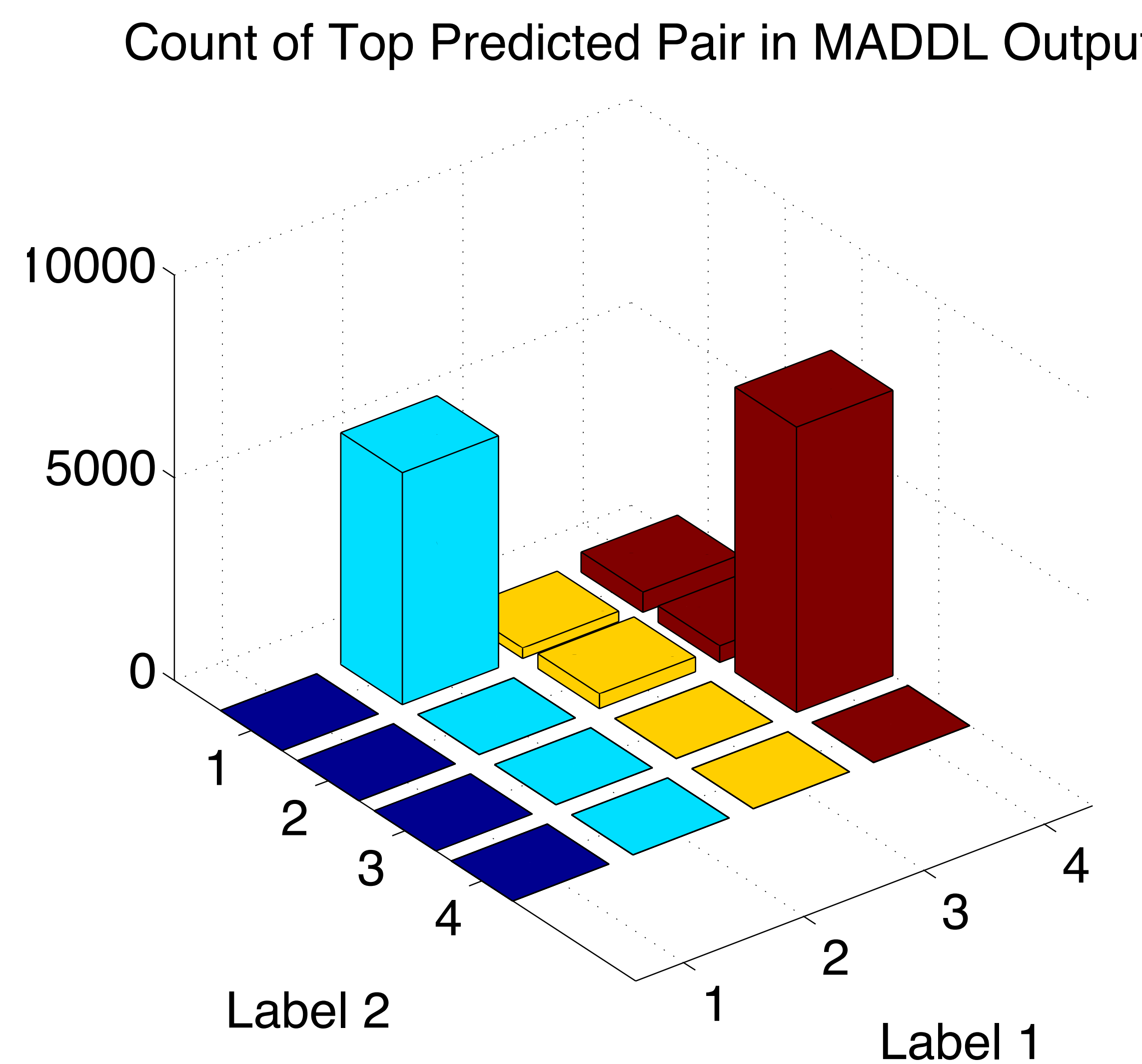
non-smooth predictions

MADDL Label Constraints



Smooth Sentiment Ranking





MADDL generates smoother ranking, while preserving accuracy of prediction.

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Manifold Regularization
[Belkin et al., JMLR 2006]

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \gamma_A ||f||_K^2 + \beta \ f^T L f$$

Loss Function
(e.g., soft margin)

Laplacian of graph
over labeled and
unlabeled data

Trains an <u>inductive</u> classifier (e.g., SVM) which can generalize to unseen instances

Manifold Regularization
[Belkin et al., JMLR 2006]

Training Data
Loss

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \gamma_A ||f||_K^2 + \beta\ f^T L f$$

Loss Function
(e.g., soft margin)

Laplacian of graph
over labeled and
unlabeled data

Trains an <u>inductive</u> classifier (e.g., SVM) which can generalize to unseen instances

Manifold Regularization

[Belkin et al., JMLR 2006]

Training Data Loss

Regulairzer (e.g., L2)

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \boxed{\gamma_A ||f||_K^2} + \beta\ f^T L f$$

Loss Function (e.g., soft margin)

Laplacian of graph over labeled and unlabeled data

Trains an <u>inductive</u> classifier (e.g., SVM) which can generalize to unseen instances

Manifold Regularization
[Belkin et al., JMLR 2006]

Training Data Loss

Regulairzer (e.g., L2)

Smoothness Regularizer

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \boxed{\gamma_A ||f||_K^2} + \boxed{\beta\ f^T L f}$$

Loss Function (e.g., soft margin)

Laplacian of graph over labeled and unlabeled data

Trains an <u>inductive</u> classifier (e.g., SVM) which can generalize to unseen instances

Spectral Graph Transduction
[Joachims, ICML 2003]

- Approximation to normalized graph cut with constraints
- Performs spectral analysis (finds eigenvalues and eigenfunctions) of the normalized Laplacian
- Code: http://sgt.joachims.org/

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2010]

C_{KL}

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} D_{KL}(r_i||p_i) + \mu \sum_{i,j} w_{ij} D_{KL}(p_i||p_j) - \nu \sum_{i=1}^{n} H(p_i)$$

$$\text{s.t.} \ \sum_{y} p_i(y) = 1, \ p_i(y) \geq 0, \ \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Normalization Constraint

KL Divergence

$$D_{KL}(p_i||p_j) = \sum_{y} p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = -\sum_{y} p_i(y) \log p_i(y)$$

C_{KL} is convex (with non-negative edge weights and hyper-parameters)

MP is related to Information Regularization [Corduneanu and Jaakkola, 2003]

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2010]

C_{KL}

Divergence on seed nodes

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i || p_i)} + \mu \sum_{i,j} w_{ij} D_{KL}(p_i || p_j) - \nu \sum_{i=1}^{n} H(p_i)$$

$$\text{s.t.} \sum_y p_i(y) = 1,\ p_i(y) \geq 0,\ \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Normalization Constraint

KL Divergence

$$D_{KL}(p_i || p_j) = \sum_y p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = - \sum_y p_i(y) \log p_i(y)$$

C_{KL} is convex (with non-negative edge weights and hyper-parameters)

MP is related to Information Regularization [Corduneanu and Jaakkola, 2003]

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2010]

C_{KL}

Divergence on seed nodes

Smoothness (divergence across edge)

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i||p_i)} + \mu \sum_{i,j} \boxed{w_{ij} D_{KL}(p_i||p_j)} - \nu \sum_{i=1}^{n} H(p_i)$$

$$\text{s.t.} \sum_y p_i(y) = 1,\ \ p_i(y) \geq 0,\ \ \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Normalization Constraint

KL Divergence

$$D_{KL}(p_i||p_j) = \sum_y p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = -\sum_y p_i(y) \log p_i(y)$$

C_{KL} is convex (with non-negative edge weights and hyper-parameters)

MP is related to Information Regularization [Corduneanu and Jaakkola, 2003]

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2010]

C_{KL}

Divergence on seed nodes

Smoothness (divergence across edge)

Entropic Regularizer

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i || p_i)} + \mu \sum_{i,j} \boxed{w_{ij} D_{KL}(p_i || p_j)} - \boxed{\nu \sum_{i=1}^{n} H(p_i)}$$

$$\text{s.t.} \sum_{y} p_i(y) = 1, \; p_i(y) \geq 0, \; \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Normalization Constraint

KL Divergence

$$D_{KL}(p_i || p_j) = \sum_{y} p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = -\sum_{y} p_i(y) \log p_i(y)$$

C_{KL} **is convex** (with non-negative edge weights and hyper-parameters)

MP is related to Information Regularization [Corduneanu and Jaakkola, 2003]

Solving MP Objective

- For ease of optimization, reformulate MP objective:

CMP

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

New probability measure, one for each vertex, similar to p_i

$$w'_{ij} = w_{ij} + \alpha \times \delta(i, j)$$

Encourages agreement between p_i and q_i

$$\operatorname*{argmin}_{p \in \Delta^n} C_{KL}(p) = \lim_{\alpha \to \infty} \operatorname*{argmin}_{p,q \in \Delta^n} C_{MP}(p, q)$$

CMP is also convex

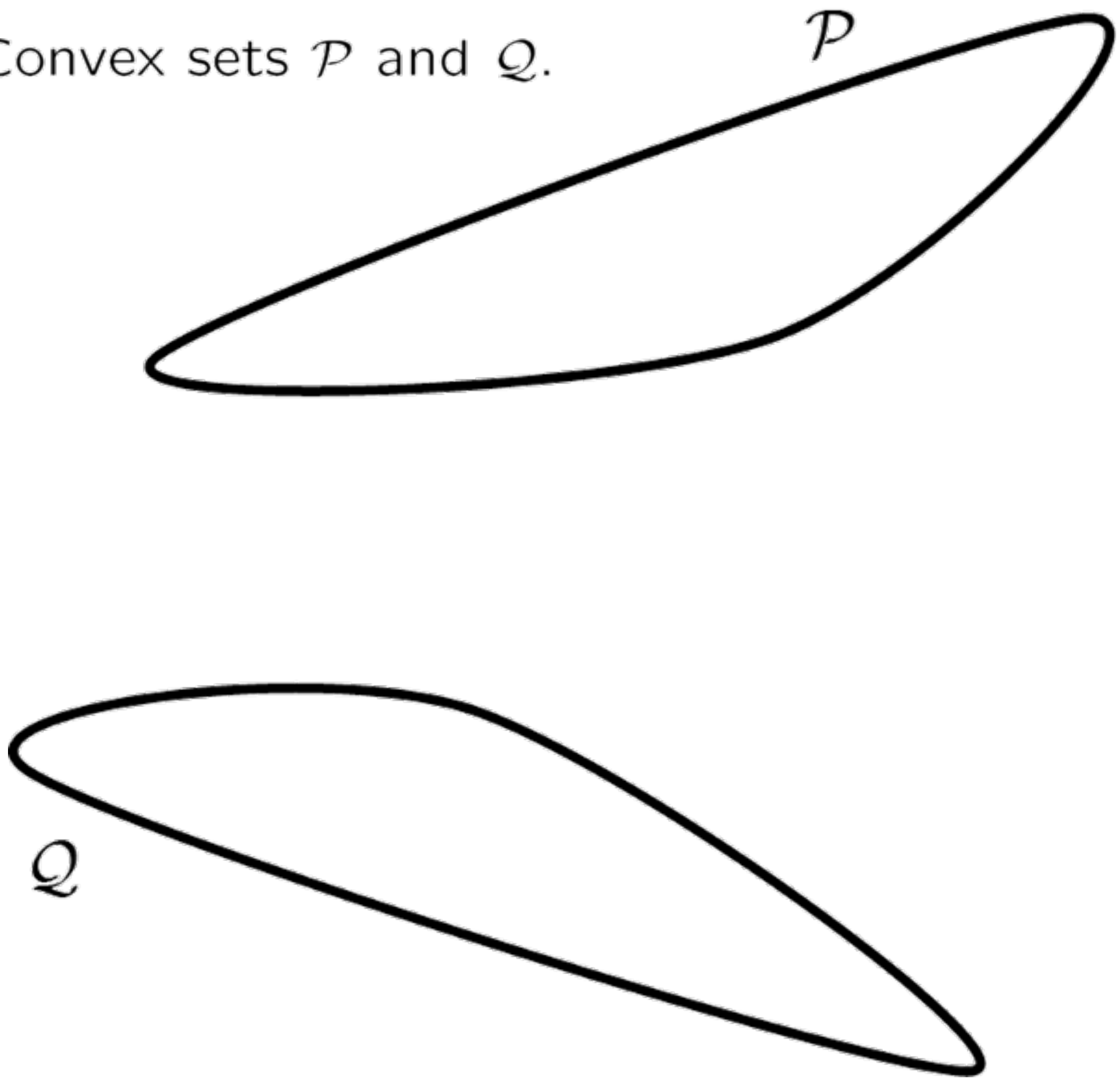
(with non-negative edge weights and hyper-parameters)

CMP can be solved using Alternating Minimization (AM)

Alternating Minimization

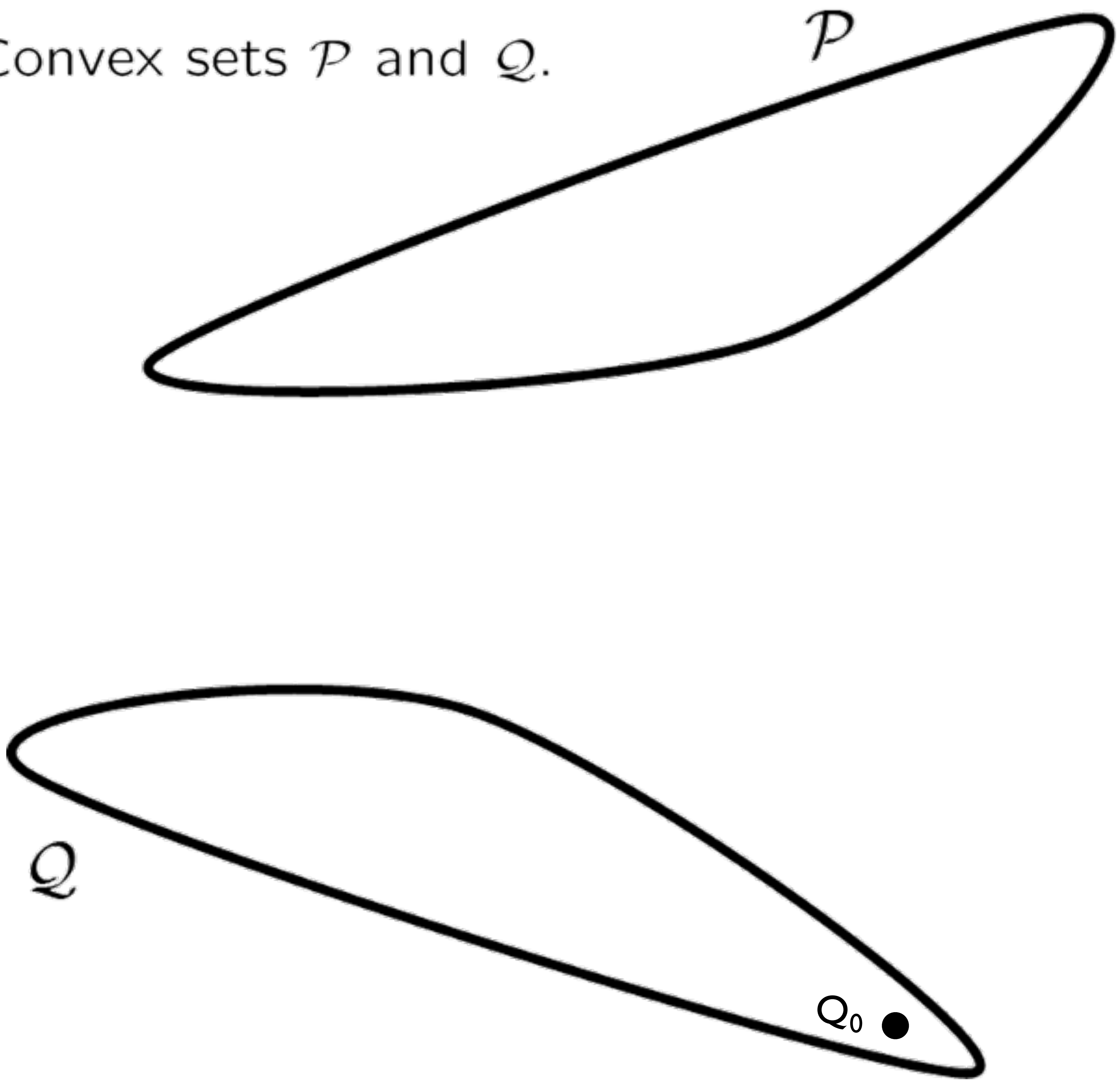
Convex sets $\mathcal{P}$ and $\mathcal{Q}$.

Given distance $d(P, Q)$
with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.



Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.

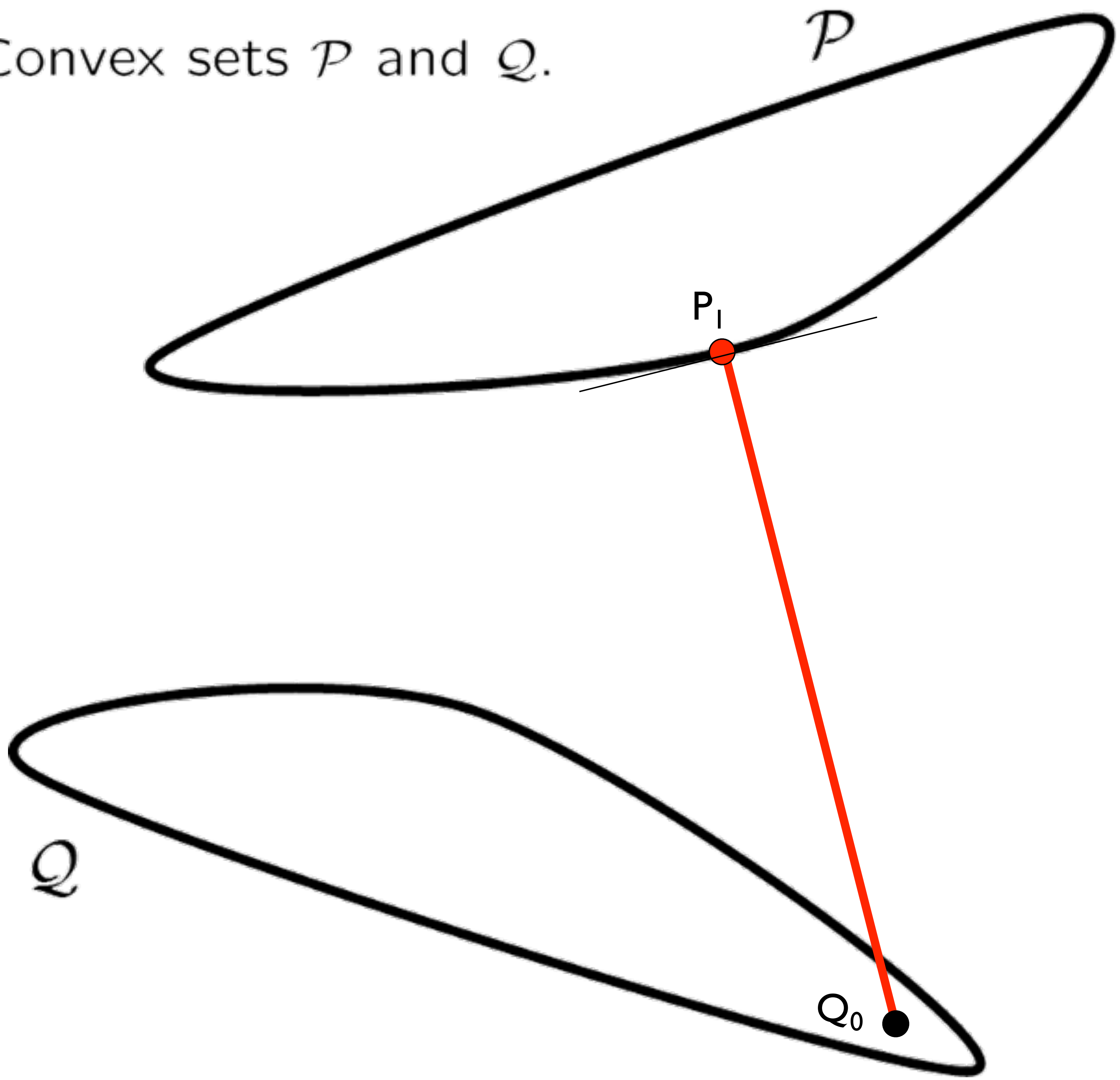


Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



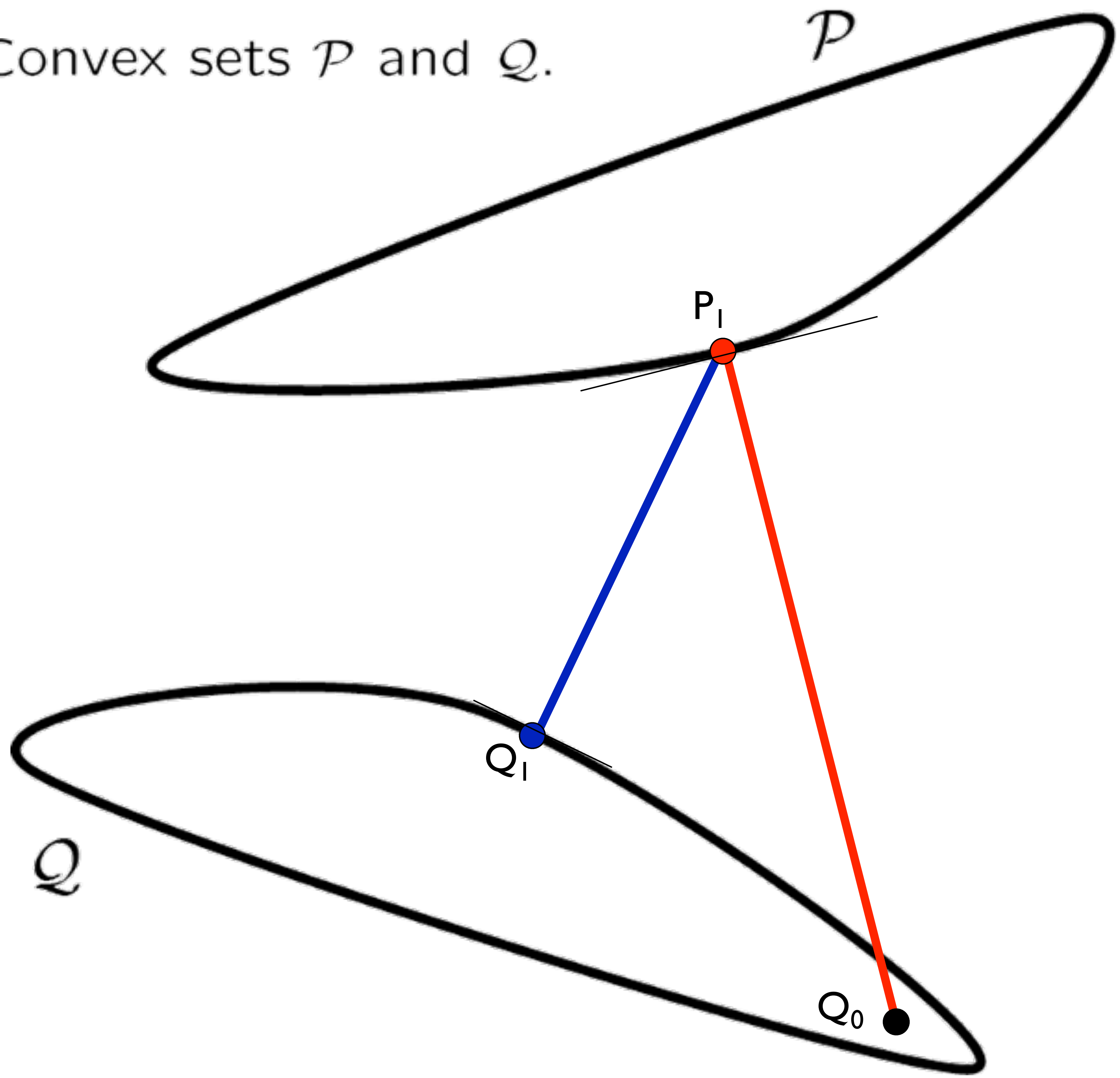
Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$P_1 = \operatorname*{argmin}_P d(P, Q_0)$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

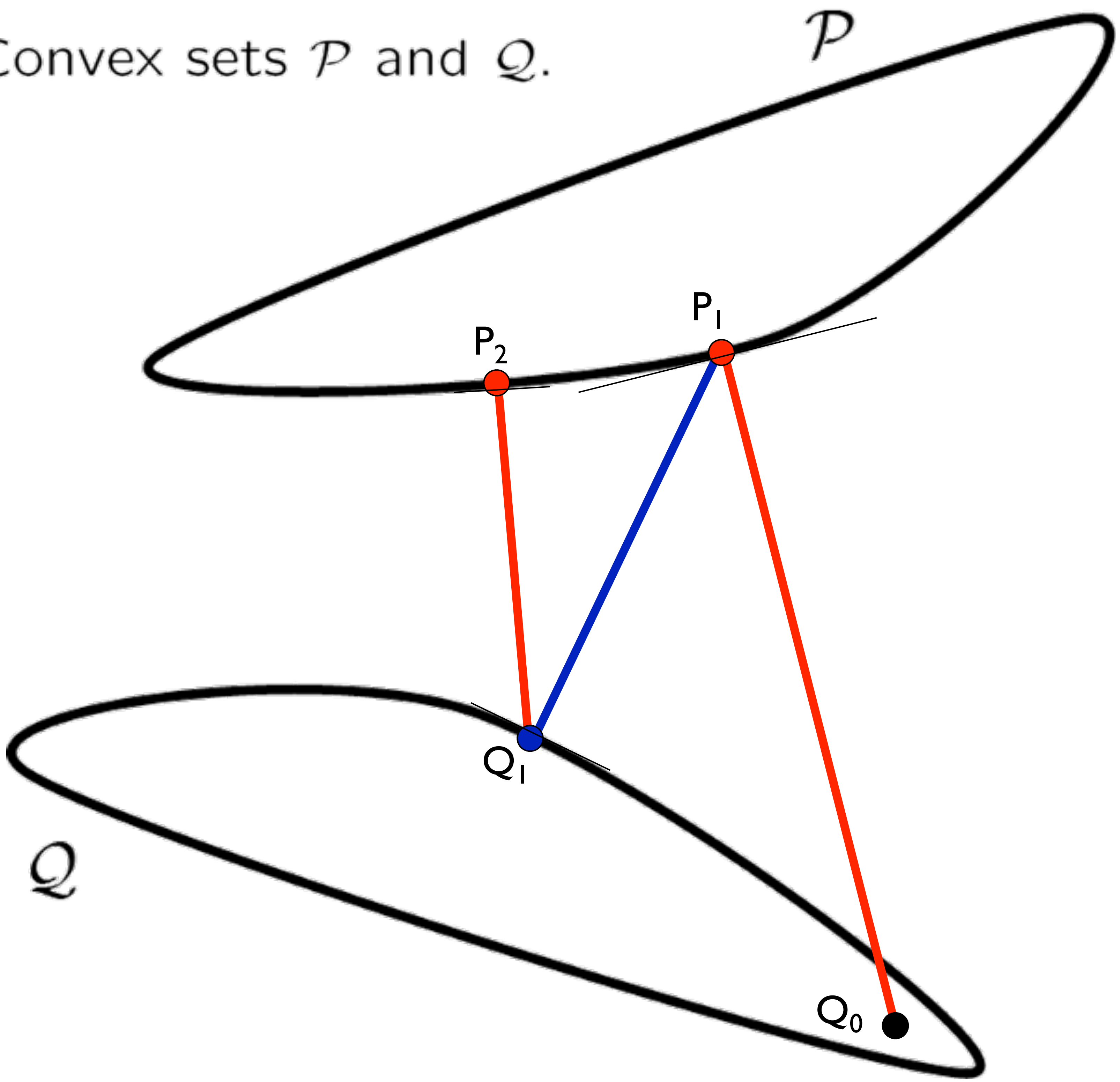
Start with $Q_0 \in \mathcal{Q}$

$P_1 = \underset{P}{\text{argmin}}\, d(P, Q_0)$

$Q_1 = \underset{Q}{\text{argmin}}\, d(P_1, Q)$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

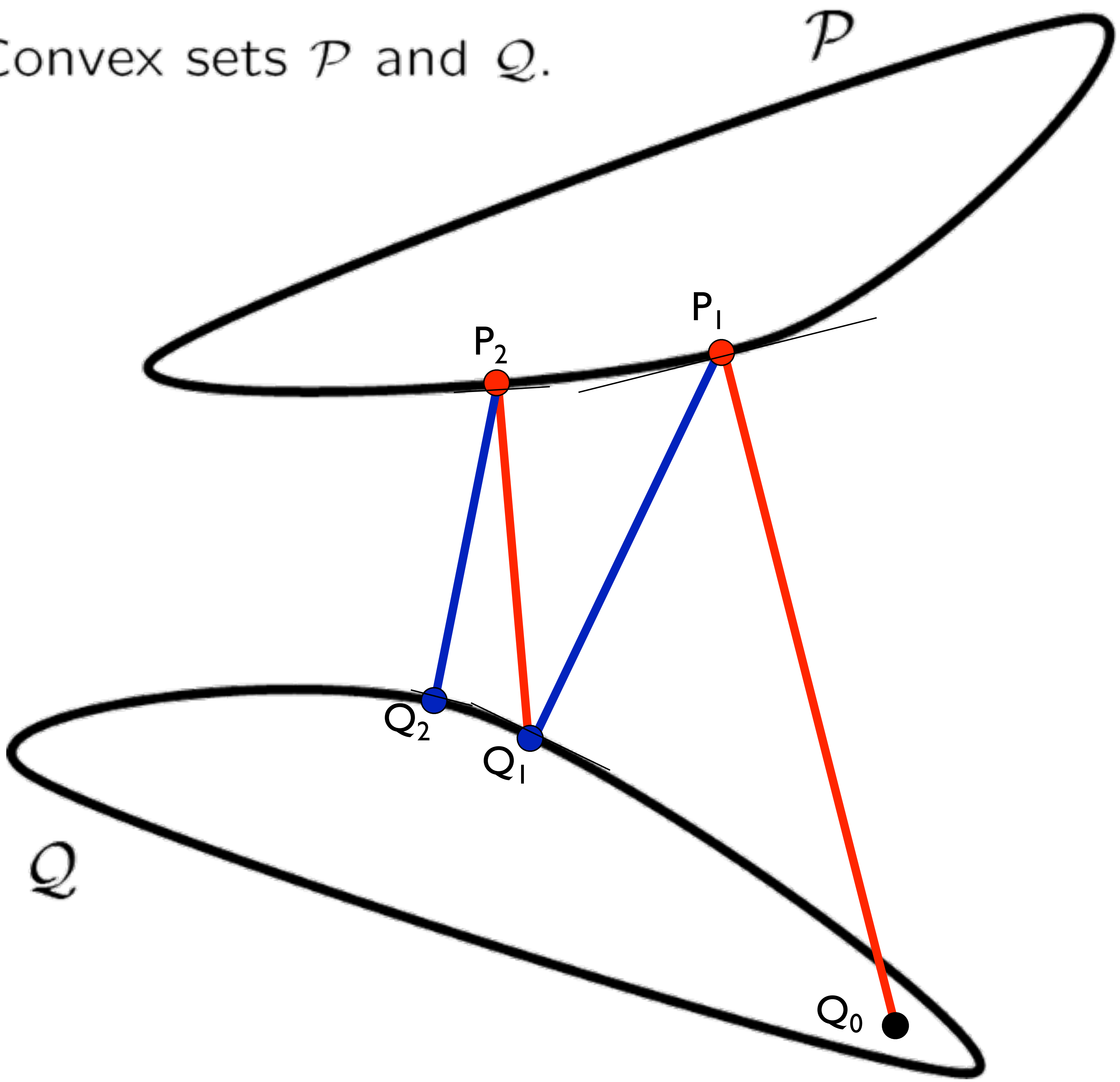
$$P_1 = \underset{P}{\text{argmin}}\, d(P, Q_0)$$

$$Q_1 = \underset{Q}{\text{argmin}}\, d(P_1, Q)$$

$$P_2 = \underset{P}{\text{argmin}}\, d(P, Q_1)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\text{argmin}}\ d(P, Q_0)$$

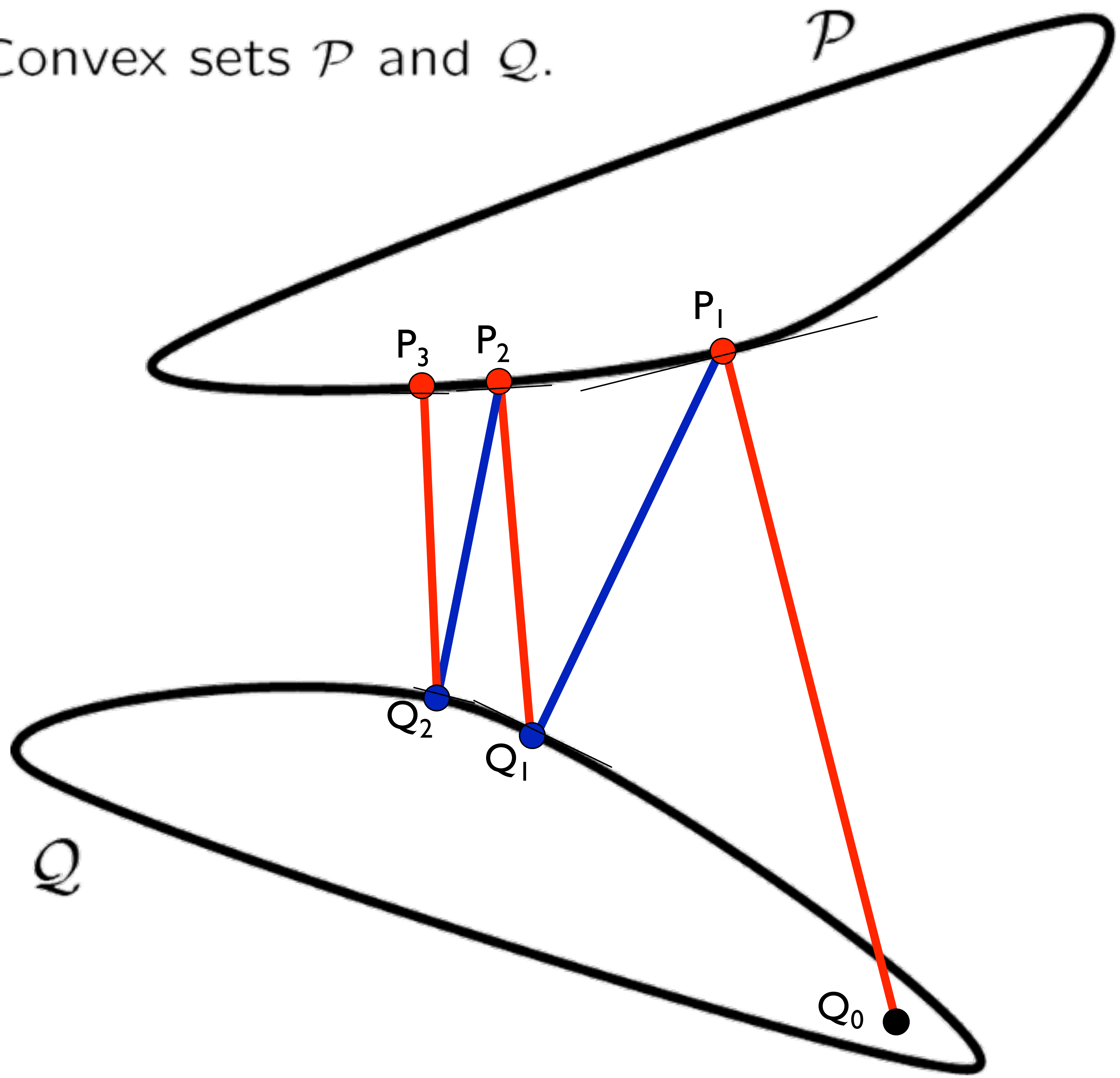
$$Q_1 = \underset{Q}{\text{argmin}}\ d(P_1, Q)$$

$$P_2 = \underset{P}{\text{argmin}}\ d(P, Q_1)$$

$$Q_2 = \underset{Q}{\text{argmin}}\ d(P_2, Q)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\, d(P, Q_0)$$

$$Q_1 = \underset{Q}{\operatorname{argmin}}\, d(P_1, Q)$$

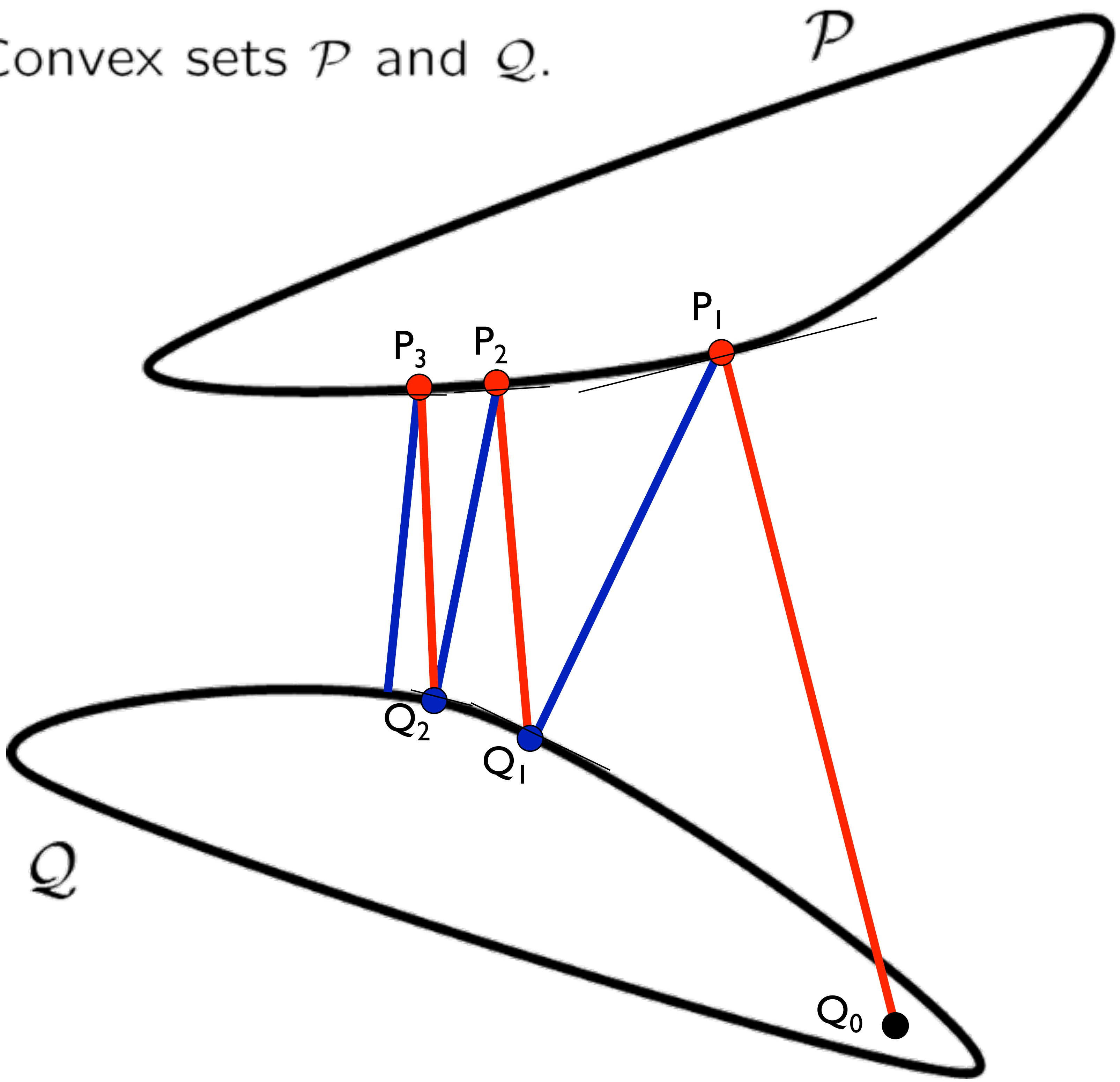
$$P_2 = \underset{P}{\operatorname{argmin}}\, d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\, d(P_2, Q)$$

$$P_3 = \underset{P}{\operatorname{argmin}}\, d(P, Q_2)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P,Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$P_1 = \underset{P}{\text{argmin}}\ d(P, Q_0)$

$Q_1 = \underset{Q}{\text{argmin}}\ d(P_1, Q)$

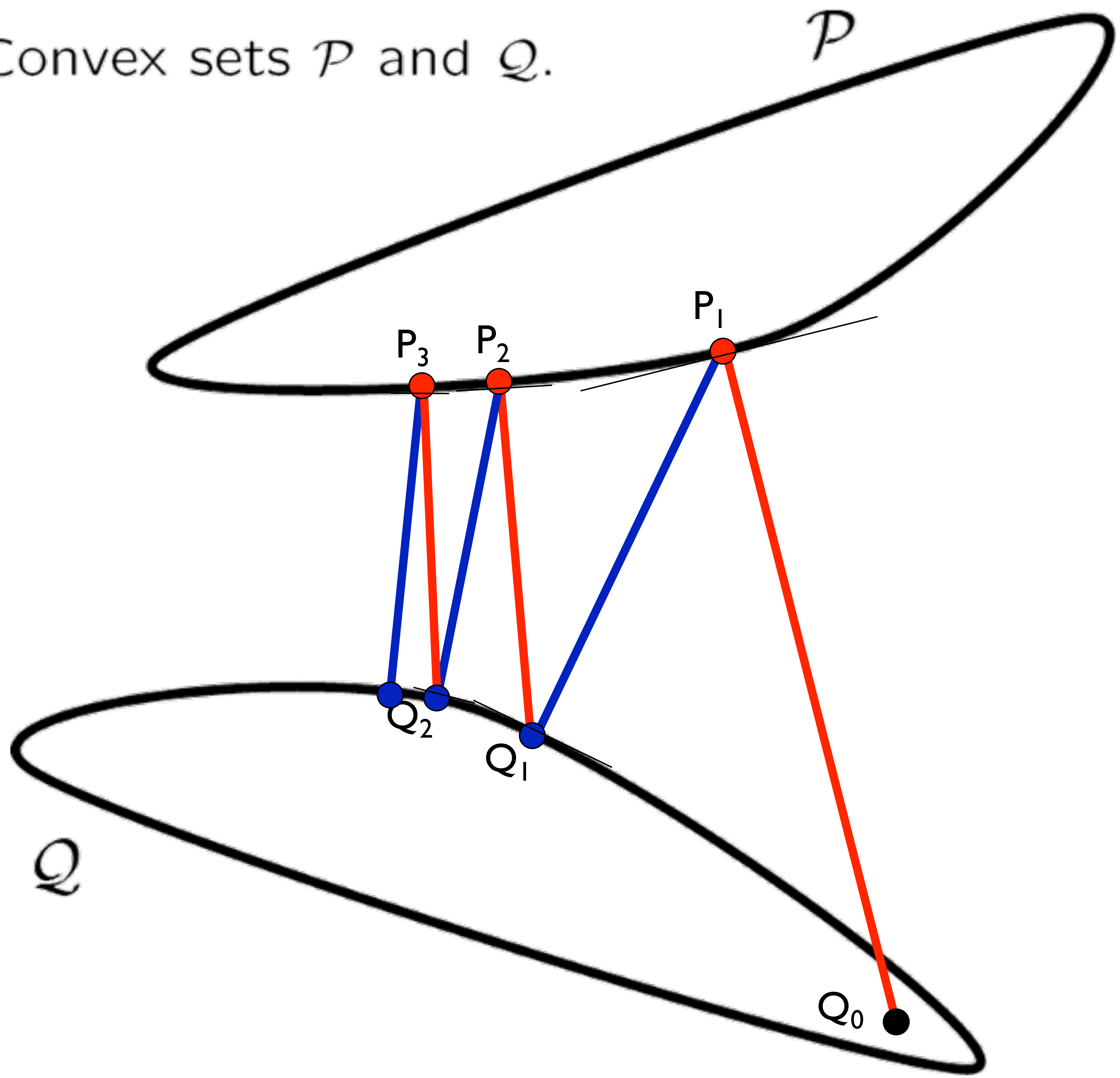
$P_2 = \underset{P}{\text{argmin}}\ d(P, Q_1)$

$Q_2 = \underset{Q}{\text{argmin}}\ d(P_2, Q)$

$P_3 = \underset{P}{\text{argmin}}\ d(P, Q_2)$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P,Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\ d(P, Q_0)$$

$$Q_1 = \underset{Q}{\operatorname{argmin}}\ d(P_1, Q)$$

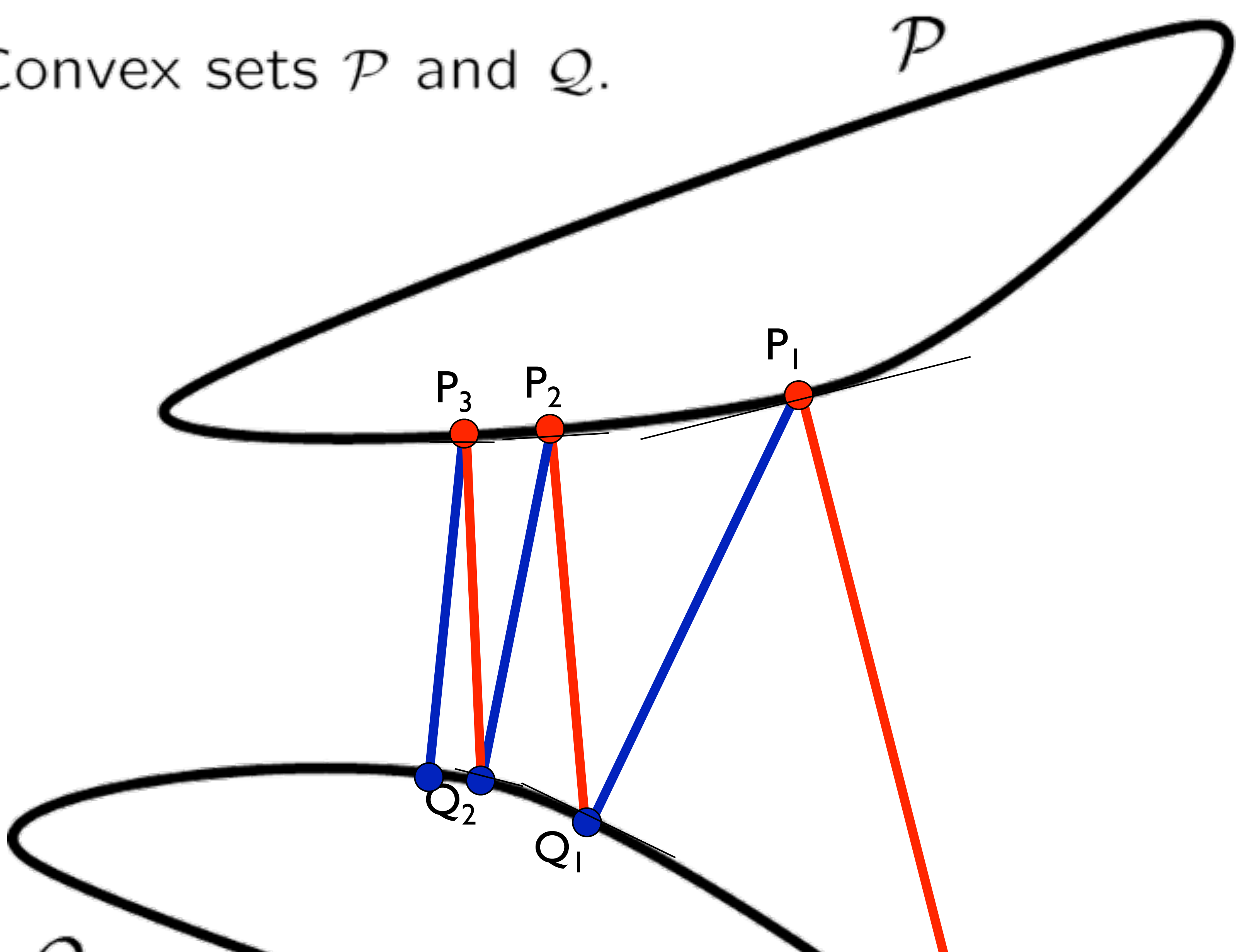
$$P_2 = \underset{P}{\operatorname{argmin}}\ d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\ d(P_2, Q)$$

$$P_3 = \underset{P}{\operatorname{argmin}}\ d(P, Q_2)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\ d(P, Q_0)$$

$$Q_1 = \underset{Q}{\operatorname{argmin}}\ d(P_1, Q)$$

$$P_2 = \underset{P}{\operatorname{argmin}}\ d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\ d(P_2, Q)$$

C_{MP} satisfies the necessary conditions for AM to converge [Subramanya and Bilmes, JMLR 2010]

Why AM?

Criteria	MOM	AM
Iterative	YES	YES
Learning Rate	Armijo Rule	None
Number of Hyper-parameters	7	1 (α)
Test for Convergence	Requires Tuning	Automatic
Update Equations	Not Intuitive	Intuitive and easily Parallelized

Table 1: There are two ways to solving the proposed objective, namely, the popular numerical optimization tool method of multipliers (MOM), and the proposed approach based on alternating minimization (AM). This table compares the two approaches on various fronts.

$$p_i^{(n)}(y) = \frac{\exp\{\frac{\mu}{\gamma_i}\sum_j w'_{ij}\log q_j^{(n-1)}(y)\}}{\sum_y \exp\{\frac{\mu}{\gamma_i}\sum_j w'_{ij}\log q_j^{(n-1)}(y)\}}$$

$$q_i^{(n)}(y) = \frac{r_i(y)\delta(i \leq l) + \mu\sum_j w'_{ji}p_j^{(n)}(y)}{\delta(i \leq l) + \mu\sum_j w'_{ji}}$$

where $\gamma_i = \nu + \mu\sum_j w'_{ij}$

Performance of SSL Algorithms

	COIL						OPT					
l	10	20	50	80	100	150	10	20	50	80	100	150
k-NN	34.5	53.9	66.9	77.9	79.2	83.5	79.6	83.9	85.5	90.5	92.0	93.8
SGT	40.1	61.2	78.0	88.5	89.0	89.9	90.4	90.6	91.4	94.7	**97.4**	**97.4**
LapRLS	**49.2**	61.4	78.4	80.1	84.5	87.8	89.7	**91.2**	92.3	96.1	**97.6**	**97.3**
SQ-Loss-I	**48.9**	63.0	**81.0**	87.5	89.0	90.9	**92.2**	90.2	**95.9**	**97.2**	**97.3**	**97.7**
MP	47.7	**65.7**	78.5	**89.6**	**90.2**	**91.1**	90.6	**90.8**	94.7	**96.6**	**97.0**	**97.1**

Comparison of accuracies for different number of labeled samples across COIL (6 classes) and OPT (10 classes) datasets

Graph SSL can be effective when the data satisfies manifold assumption. More results and discussion in Chapter 21 of the SSL Book (Chapelle et al.)

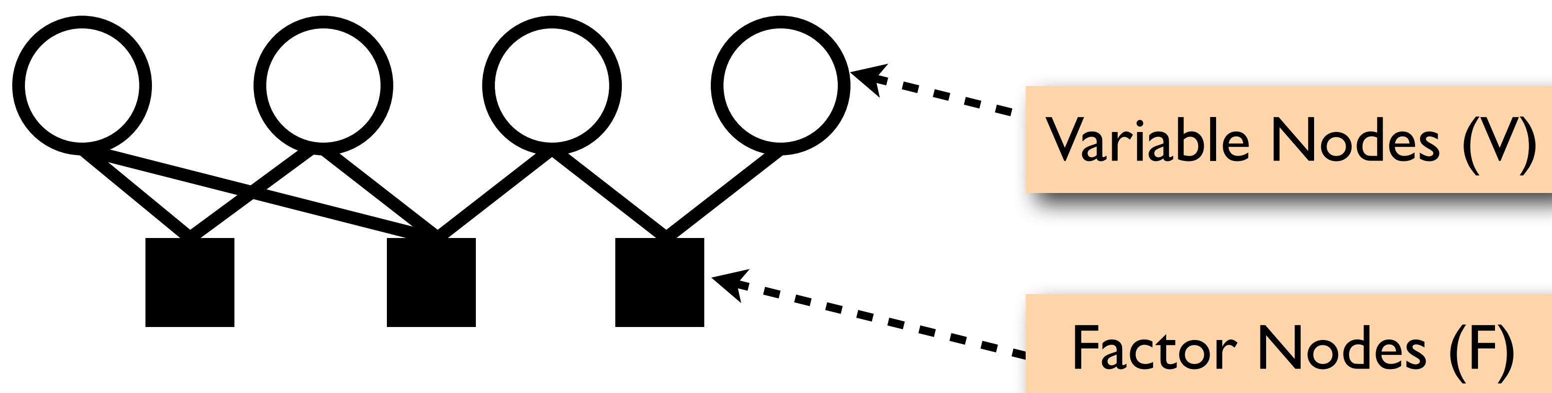
Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Background: Factor Graphs
[Kschischang et al., 2001]

Factor Graph

- bipartite graph
- variable nodes (e.g., label distribution on a node)
- factor nodes: fitness function over variable assignment

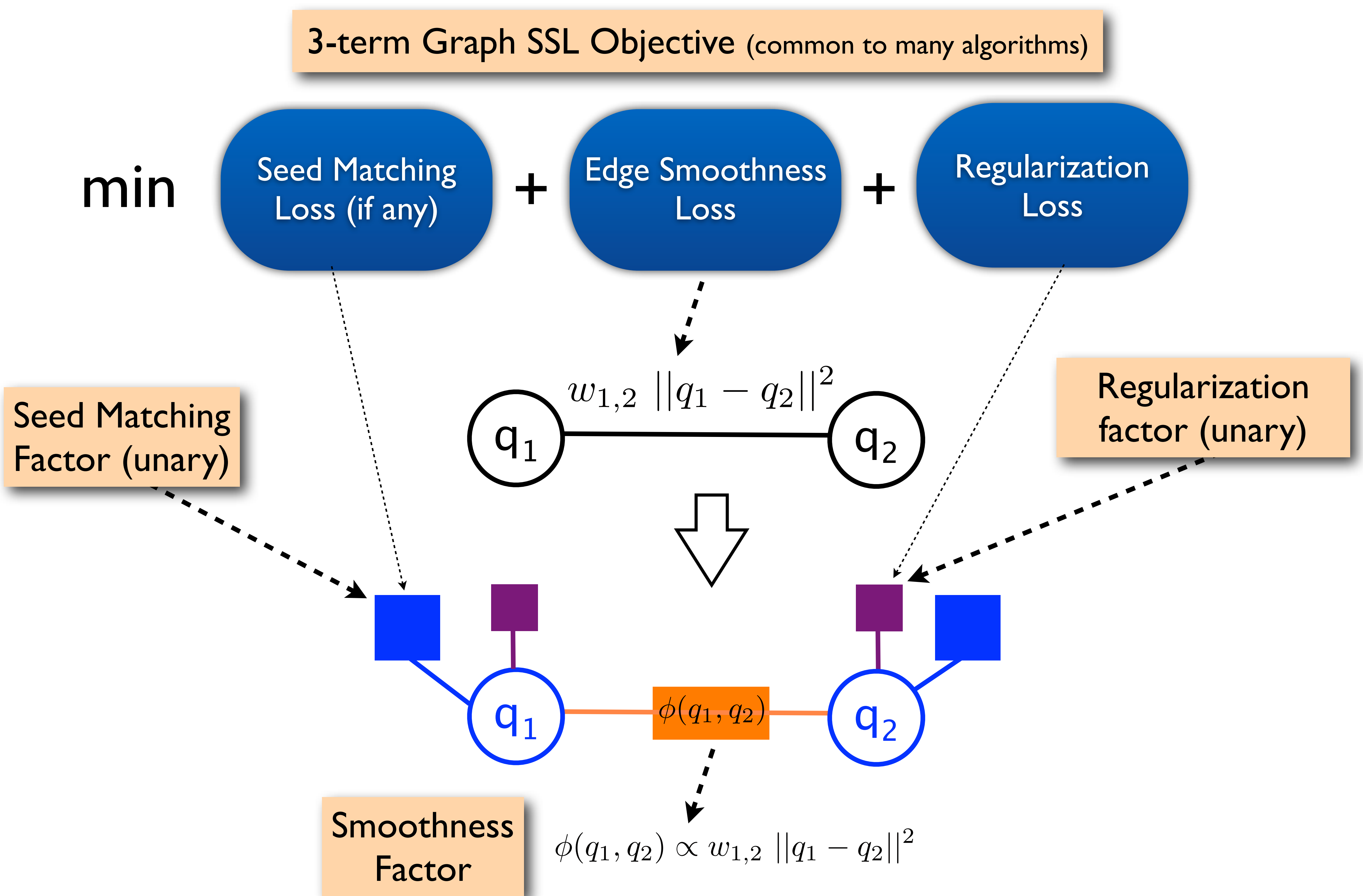


Distribution over all variables' values

$$\log P\left(\{v\}_{v\in V}\right) = -\log Z + \sum_{f\in F} \log \alpha_f\left(\{v\}_{(v,f)\in E}\right)$$

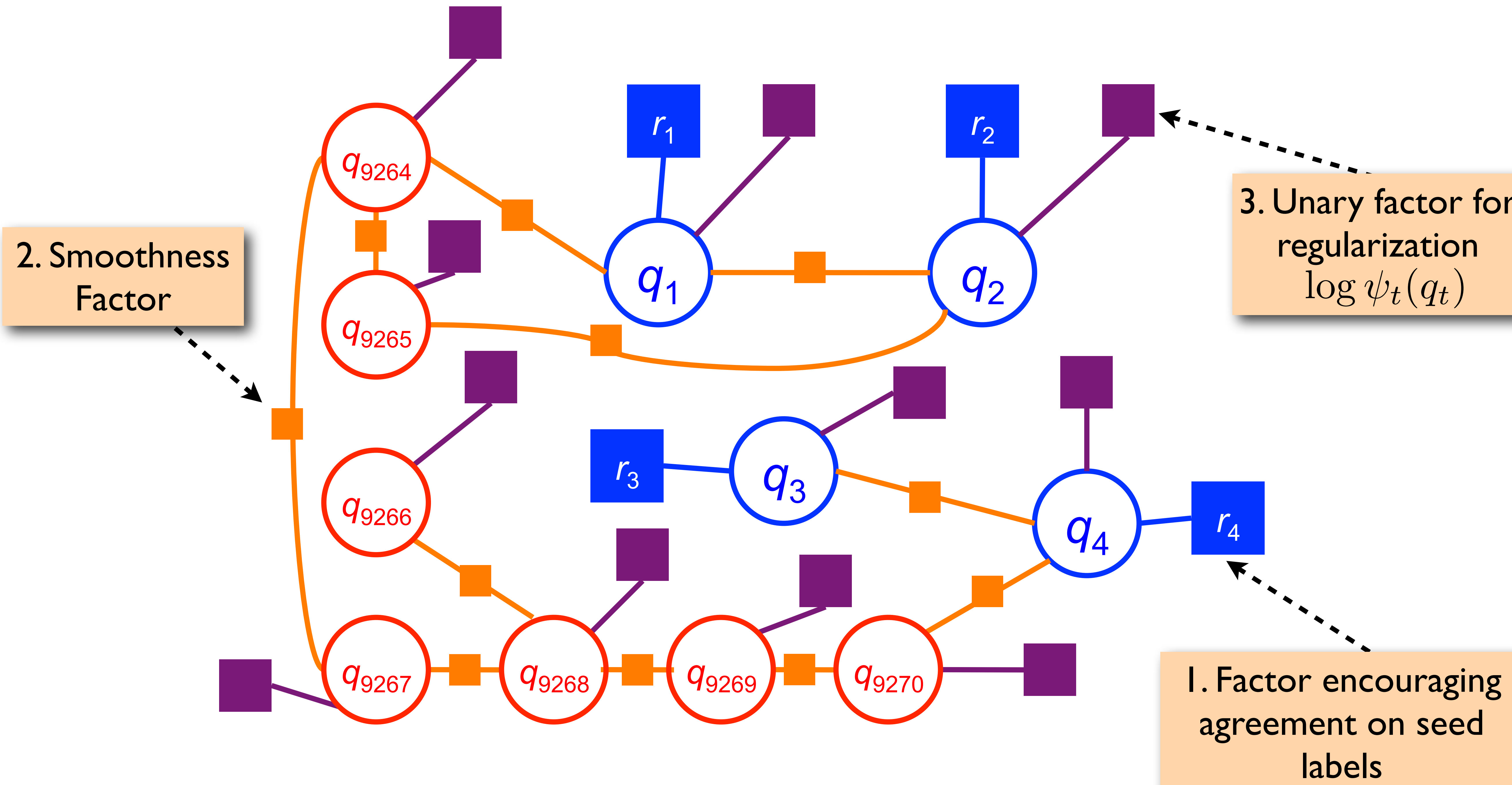
variables connected to factor f

Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]



Factor Graph Interpretation

[Zhu et al., ICML 2003][Das and Smith, NAACL 2012]



Label Propagation with Sparsity

Enforce through sparsity inducing unary factor

Lasso (Tibshirani, 1996) $\log \psi_t(q_t) = -\lambda \|q_t\|_1$

Elitist Lasso (Kowalski and Torrésani, 2009)

$$\log \psi_t(q_t) = -\lambda \left(\|q_t\|_1\right)^2$$

For more details, see [Das and Smith, NAACL 2012]

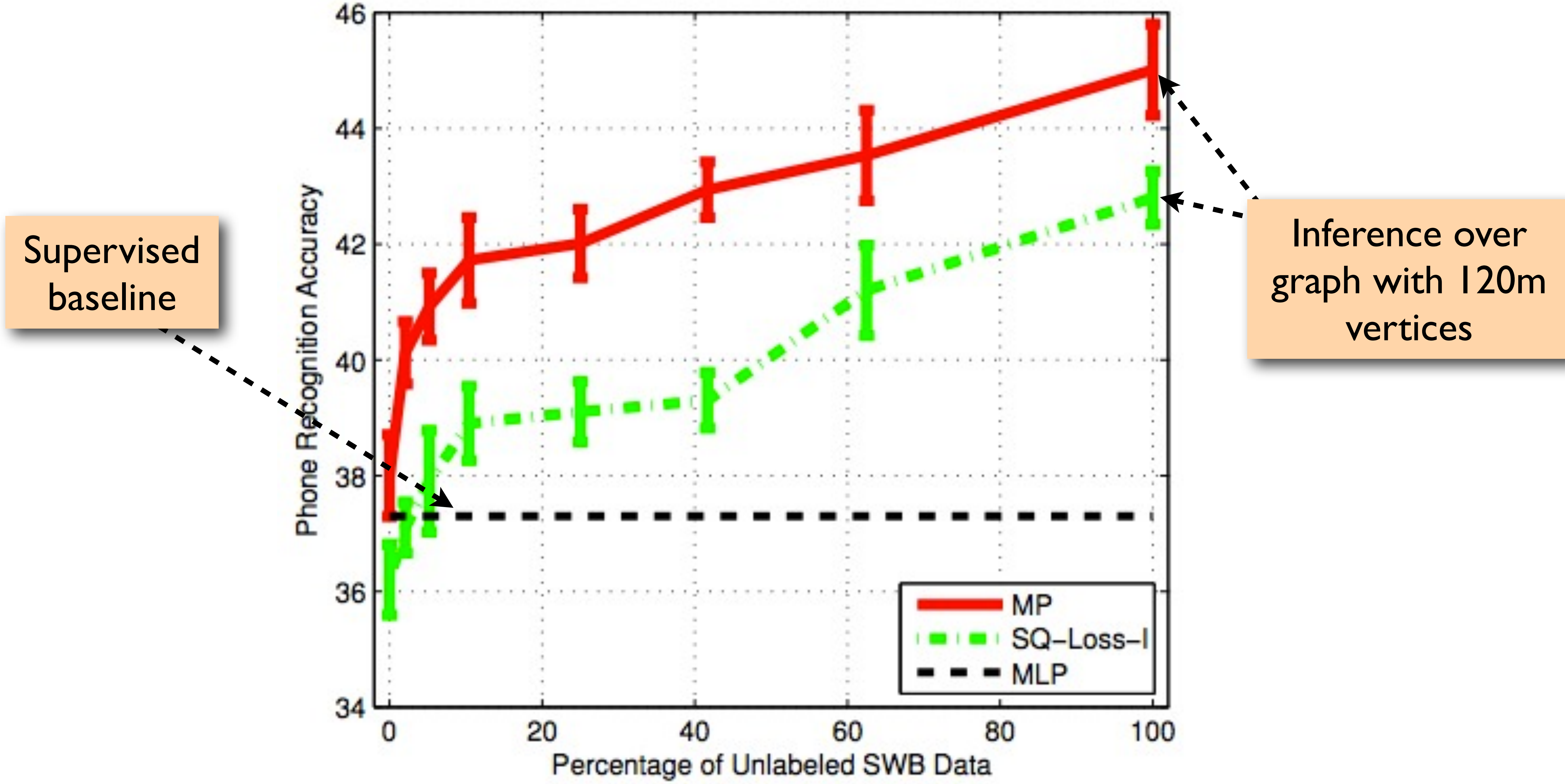
Other Graph-SSL Methods

- SSL on Directed Graphs
  - [Zhou et al, NIPS 2005], [Zhou et al., ICML 2005]
- Learning with dissimilarity edges
  - [Goldberg et al., AISTATS 2007]
- Graph Transduction using Alternating Minimization
  - [Wang et al., ICML 2008]
- Graph as regularizer for Multi-Layered Perceptron
  - [Karlen et al., ICML 2008], [Malkin et al., Interspeech 2009]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

More (Unlabeled) Data is Better Data



Challenges with large unlabeled data:

- Constructing graph from large data
- Scalable inference over large graphs

[Subramanya & Bilmes, JMLR 2011]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

Scalability Issues (I)

Graph Construction

- Brute force (exact) k-NN too expensive (quadratic)
- Approximate nearest neighbor using kd-tree [Friedman et al., 1977]
- Approximate Nearest Neighbor library (http://www.cs.umd.edu/˜mount/)

Scalability Issues (II)

Label Inference

- Sub-sample the data
  - Construct graph over a subset of a unlabeled data [Delalleau et al., AISTATS 2005]
  - Sparse Grids [Garcke & Griebel, KDD 2001]

Scalability Issues (II)

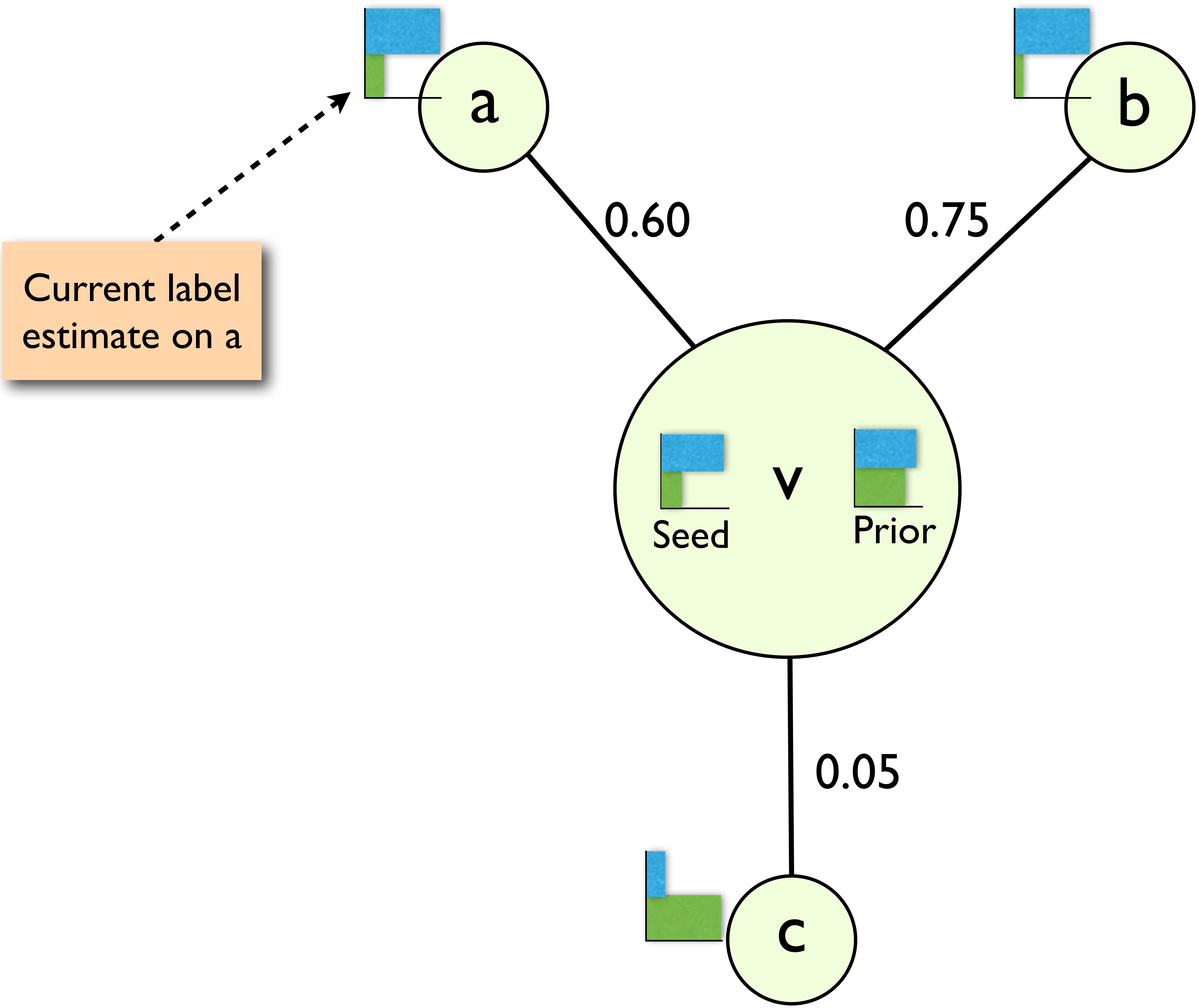
Label Inference

- Sub-sample the data
  - Construct graph over a subset of a unlabeled data [Delalleau et al., AISTATS 2005]
  - Sparse Grids [Garcke & Griebel, KDD 2001]
- How about using more compute? (next section)
  - Symmetric multi-processor (SMP)
  - Distributed Computer

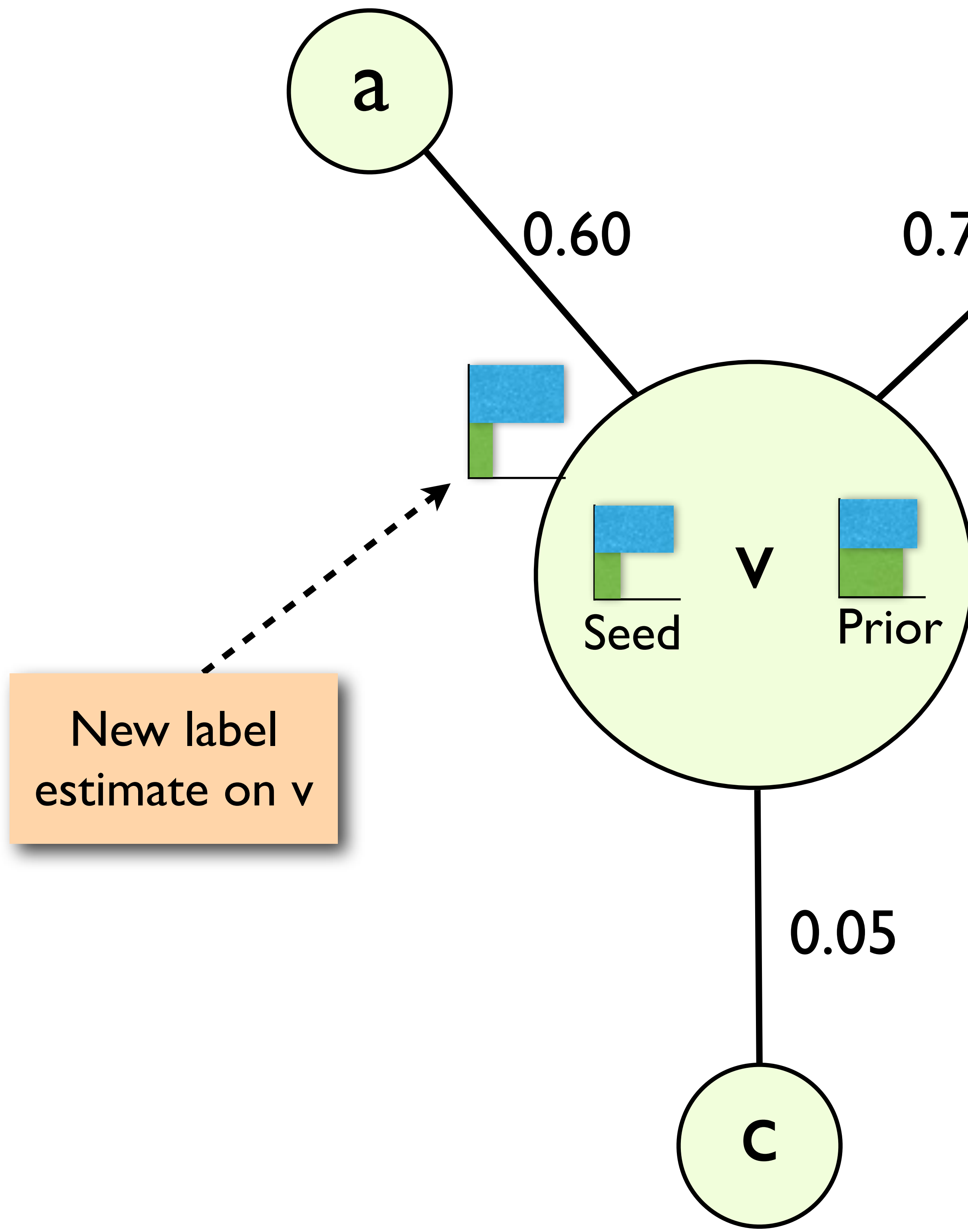
Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering [Subramanya & Bilmes, JMLR 2011; Bilmes & Subramanya, 2011]
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

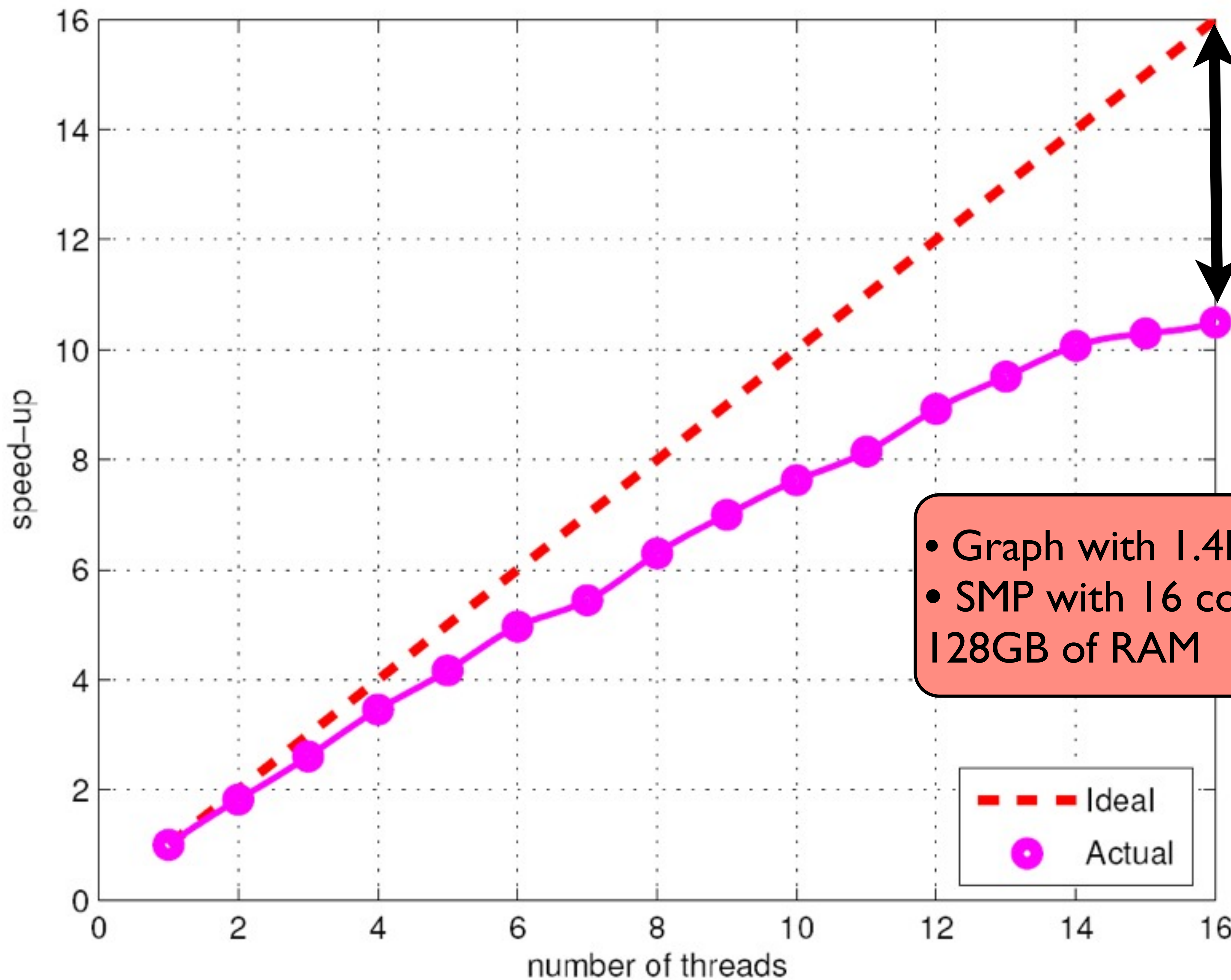
Label Update using Message Passing



Label Update using Message Passing



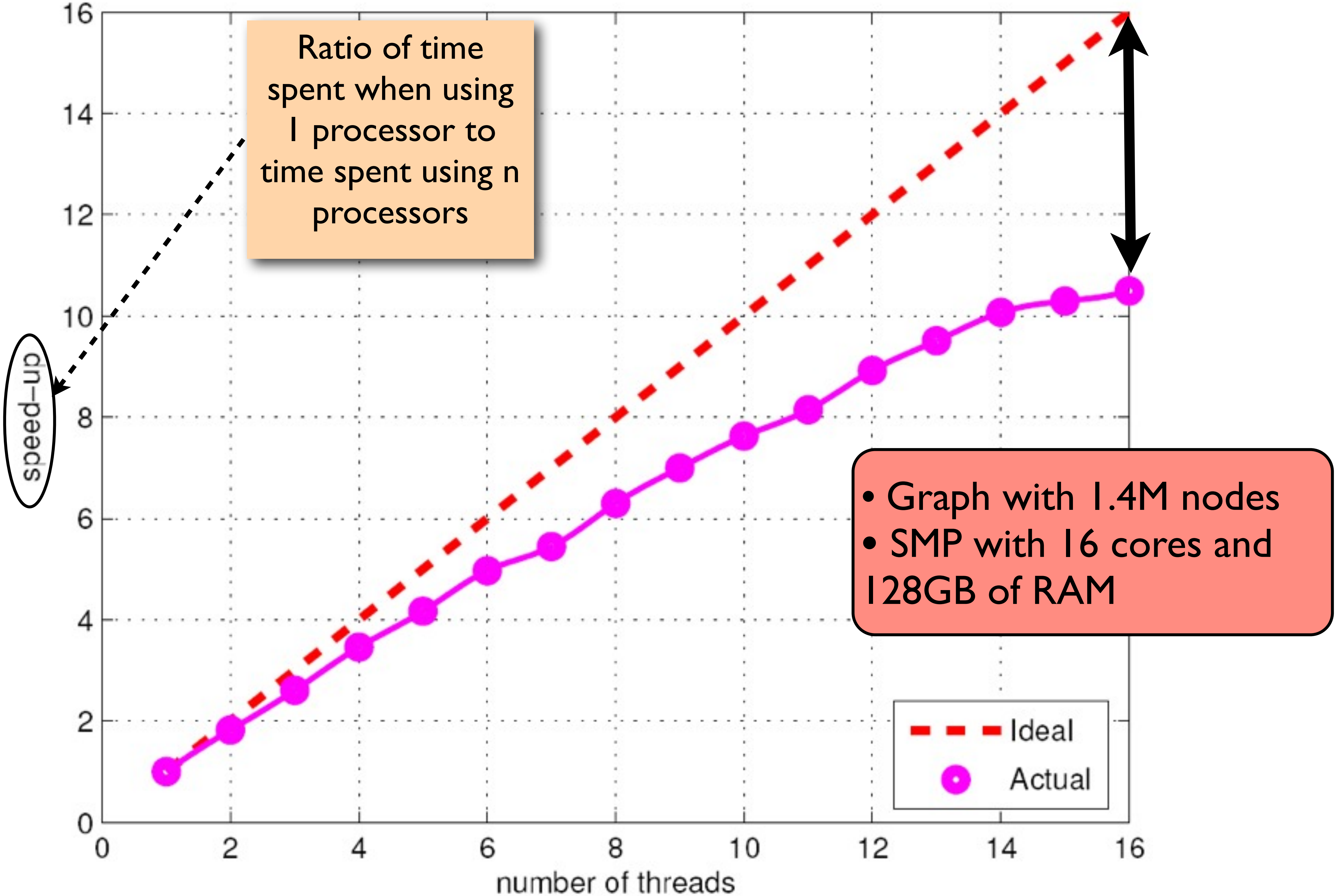
Speed-up on SMP



- Graph with 1.4M nodes
- SMP with 16 cores and 128GB of RAM

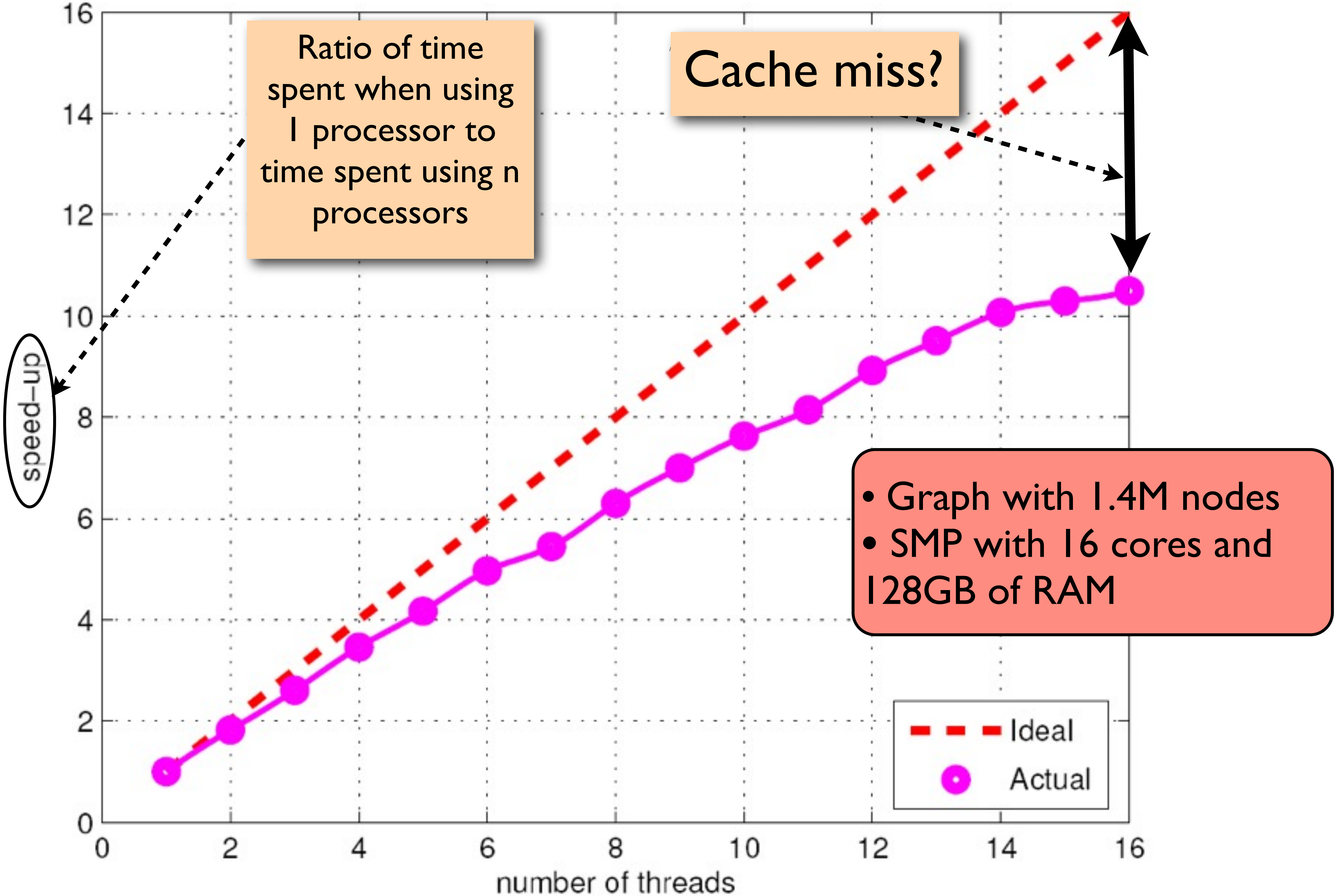
[Subramanya & Bilmes, JMLR, 2011]

Speed-up on SMP



[Subramanya & Bilmes, JMLR, 2011]

Speed-up on SMP



[Subramanya & Bilmes, JMLR, 2011]

Node Reordering Algorithm

Input: Graph G = (V, E)

Result: Node ordered graph

1. Select an arbitrary node v
2. while unselected nodes remain do
   2.1. select an unselected node v` from among the neighbors' neighbors of v that has maximum overlap with v' neighbors
   2.2. mark v` as selected
   2.3. set v to v`

[Subramanya & Bilmes, JMLR, 2011]

Node Reordering Algorithm

Input: Graph G = (V, E)

Result: Node ordered graph

1. Select an arbitrary node v
2. while unselected nodes remain do

    2.1. select an unselected node v` from among the neighbors' neighbors of v that has maximum overlap with v' neighbors

    2.2. mark v` as selected

    2.3. set v to v`

[Subramanya & Bilmes, JMLR, 2011]

Node Reordering Algorithm

Input: Graph G = (V, E)

Result: Node ordered graph

1. Select an arbitrary node v
2. while unselected nodes remain do

   2.1. select an unselected node v` from among the neighbors' neighbors of v that has maximum overlap with v' neighbors

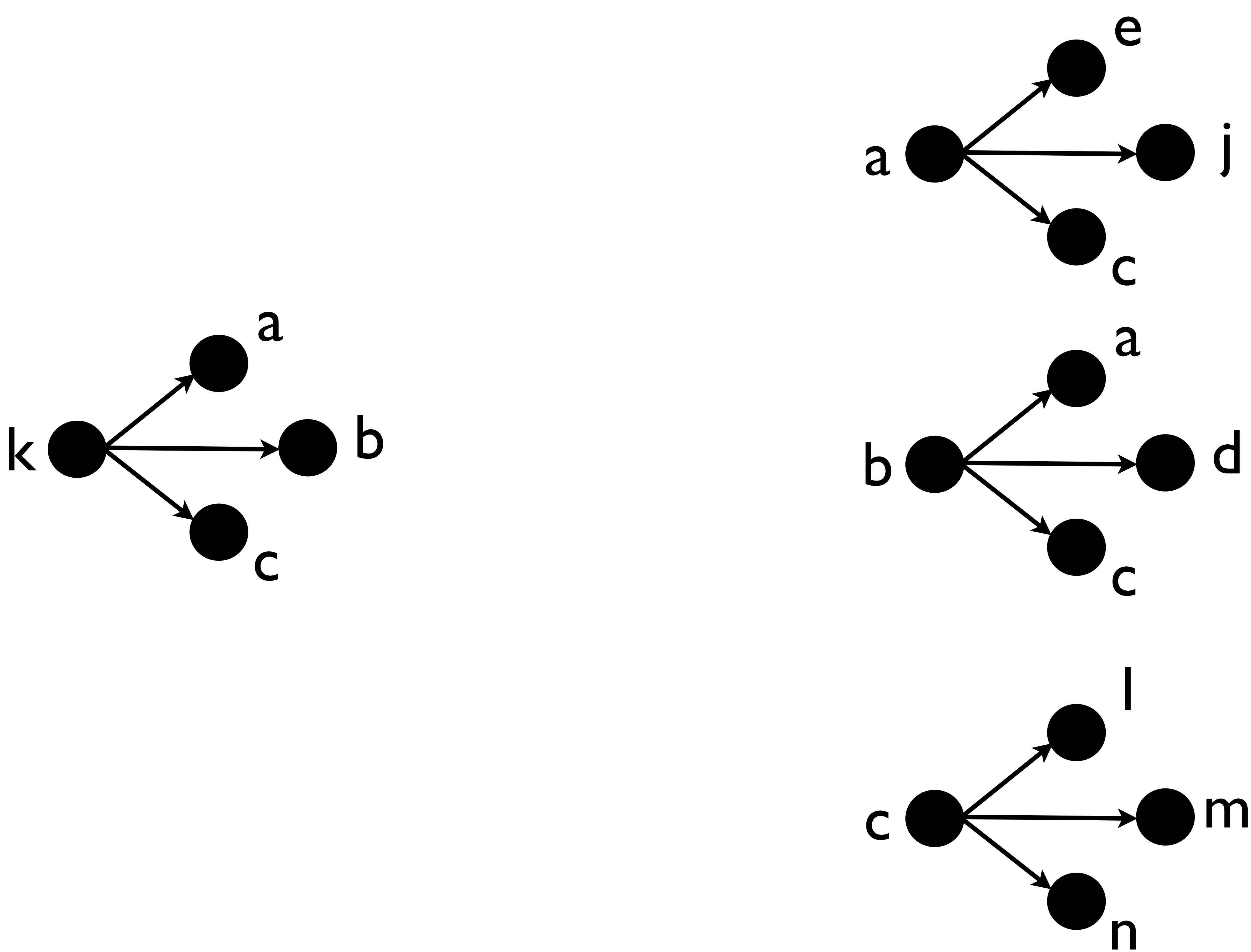
   2.2. mark v` as selected

   2.3. set v to v`

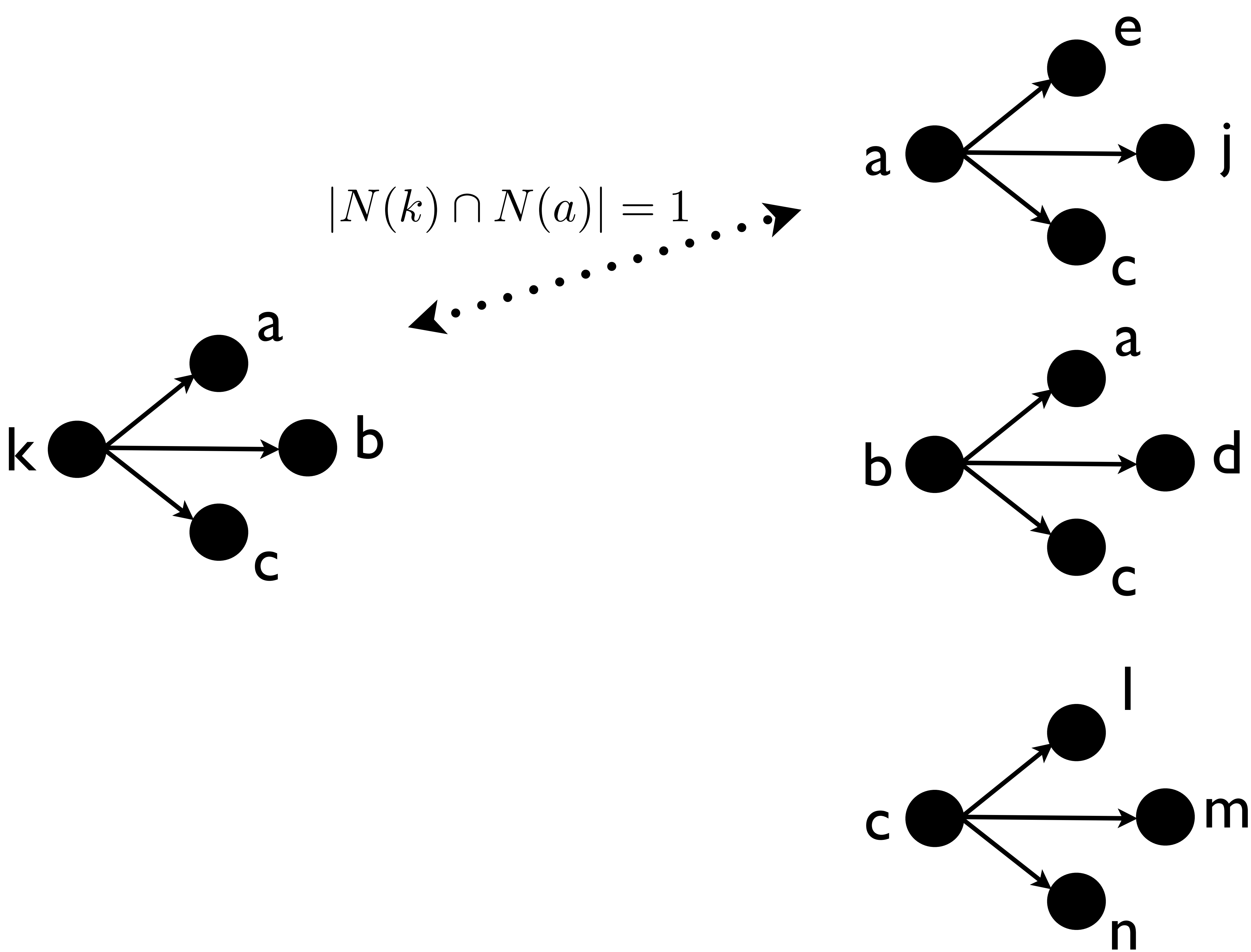
Exhaustive for sparse (e.g., k-NN) graphs

[Subramanya & Bilmes, JMLR, 2011]

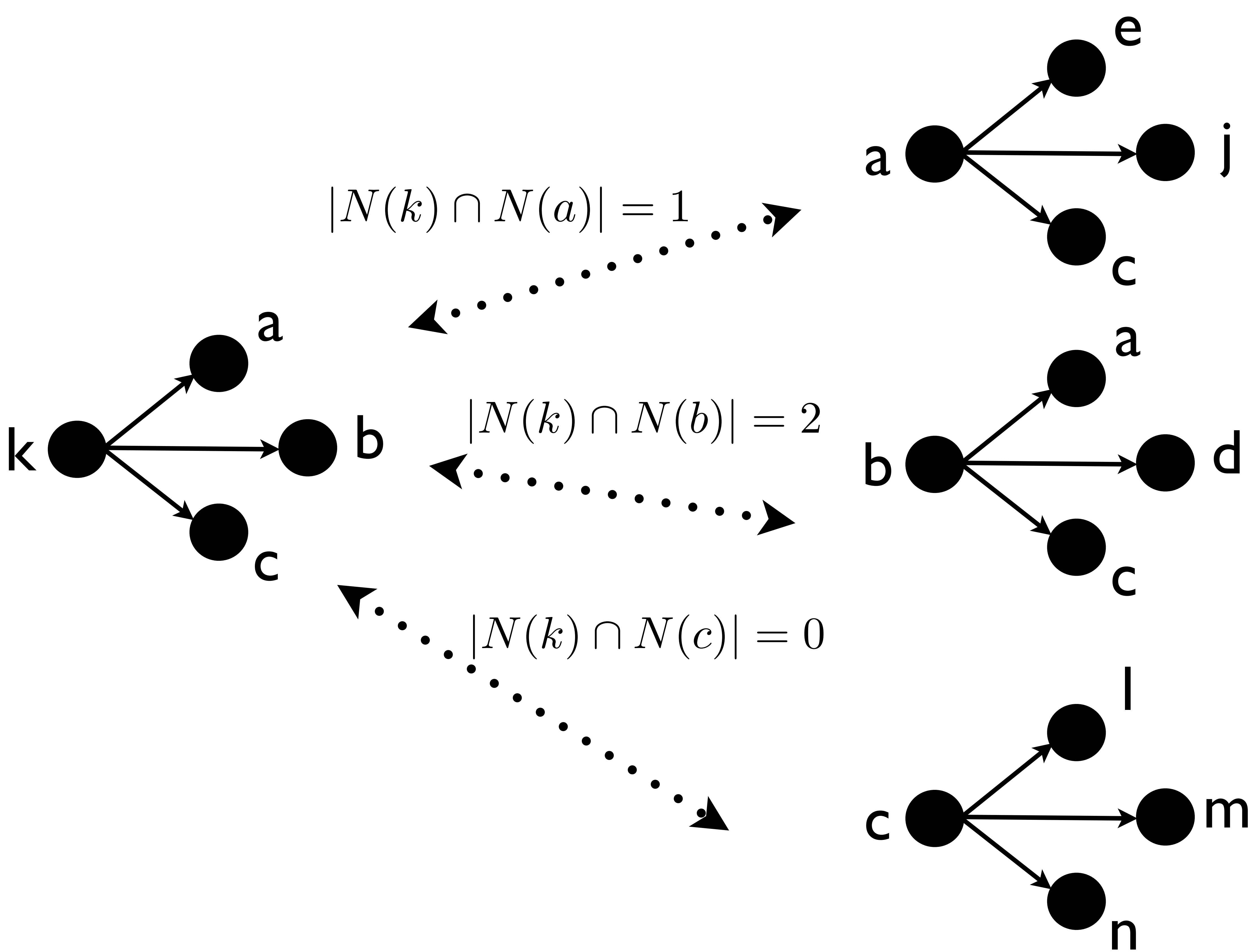
Node Reordering Algorithm : Intuition



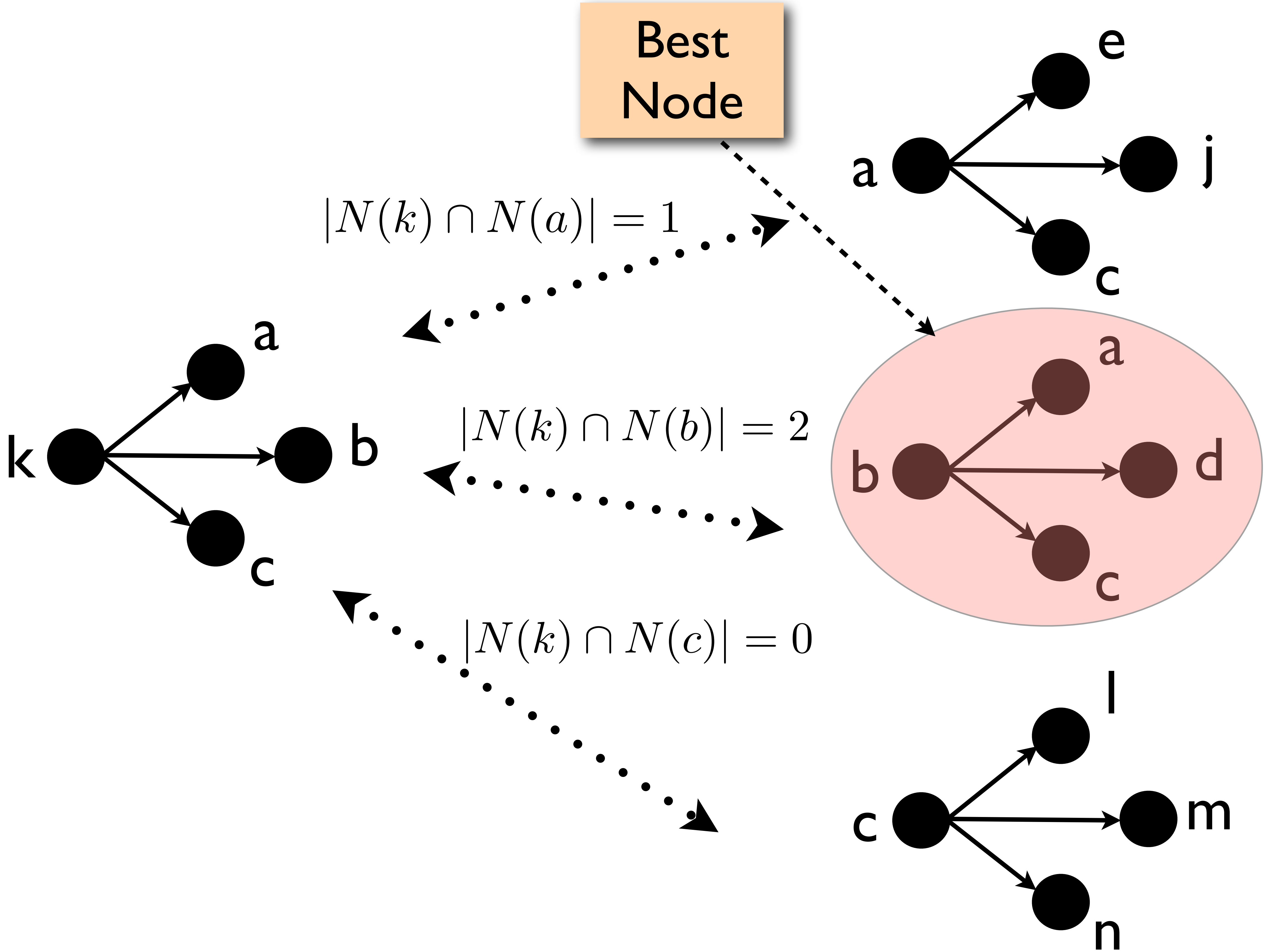
Node Reordering Algorithm : Intuition



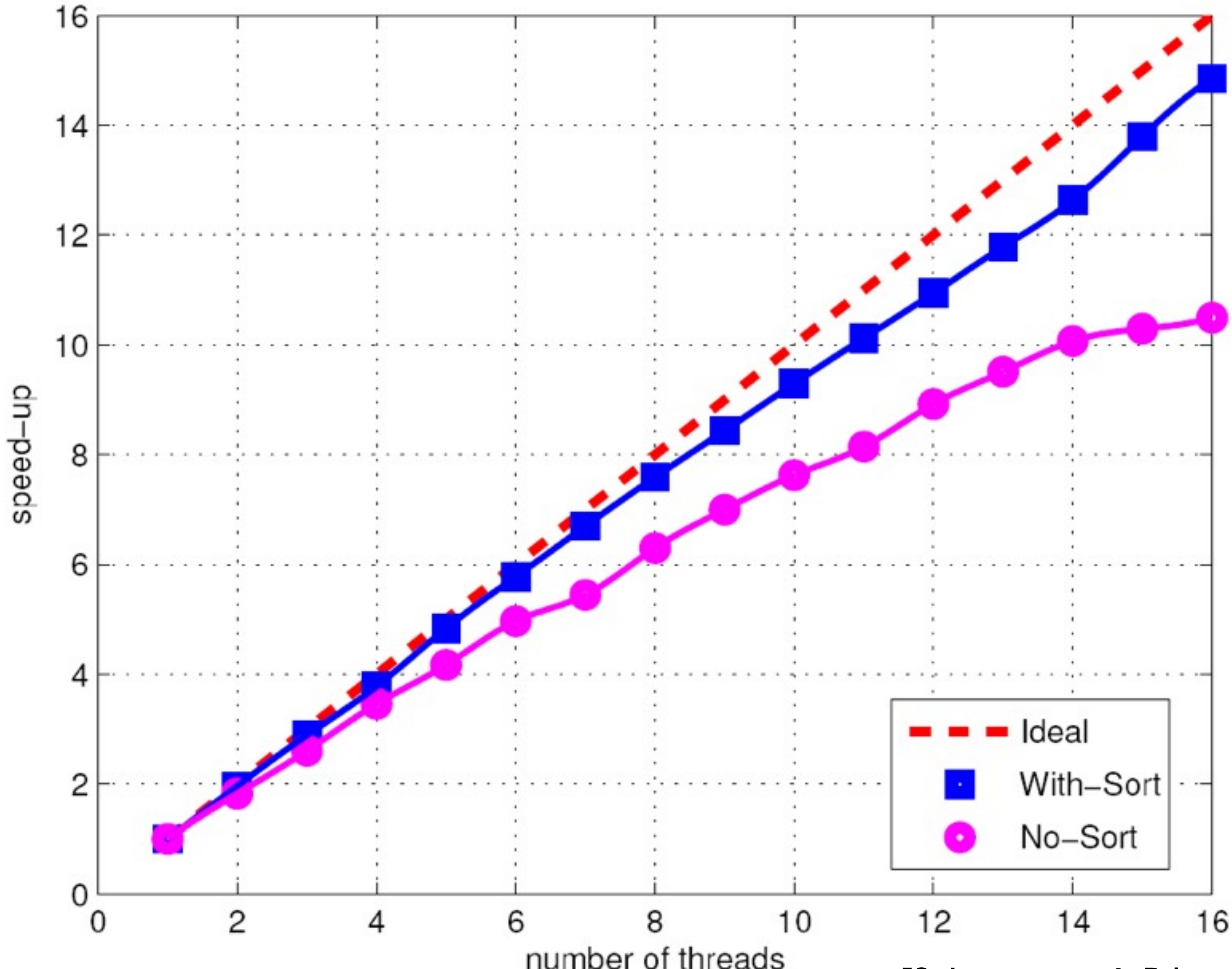
Node Reordering Algorithm : Intuition



Node Reordering Algorithm : Intuition



Speed-up on SMP after Node Ordering

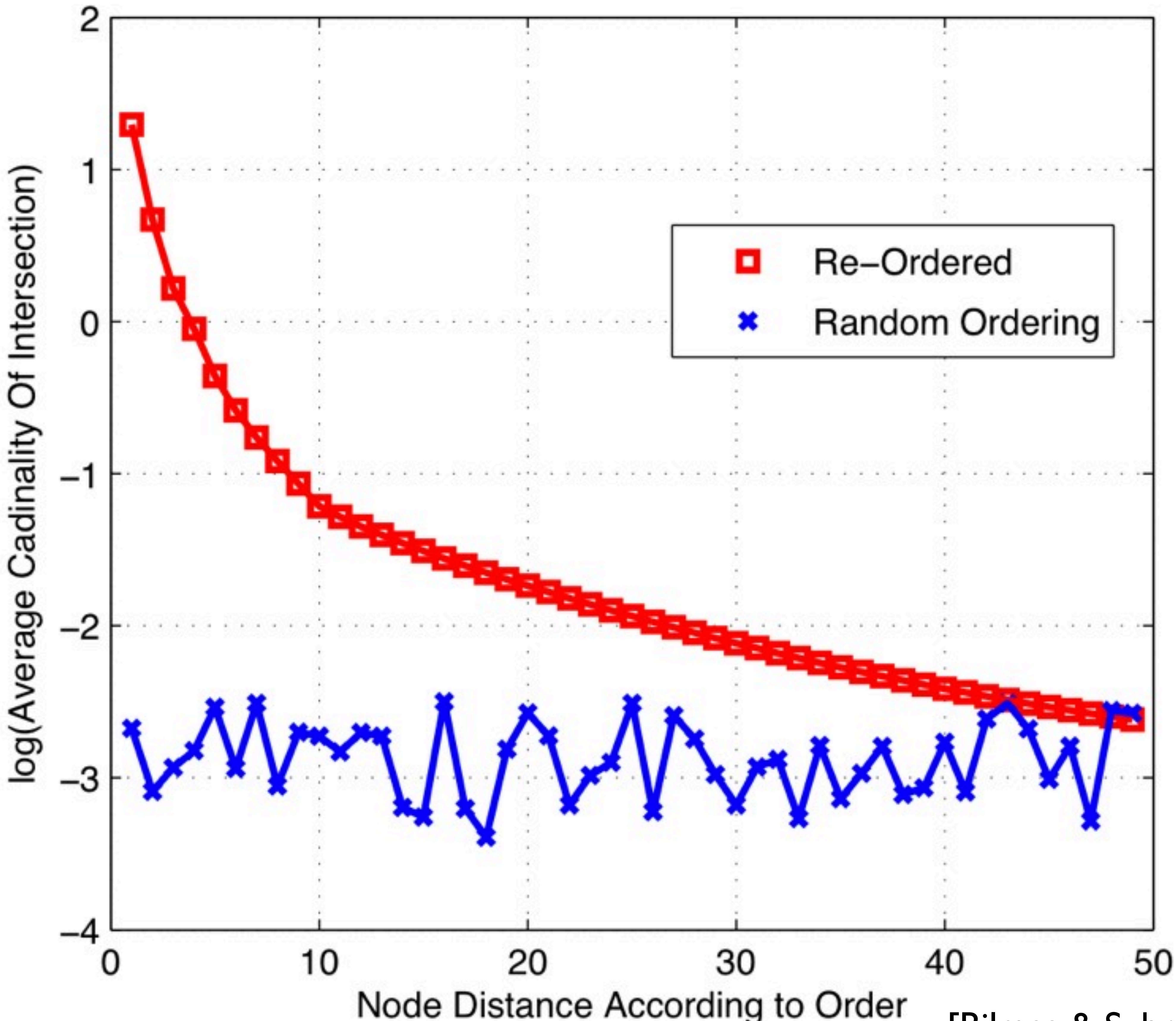


[Subramanya & Bilmes, JMLR, 2011]

Distributed Processing

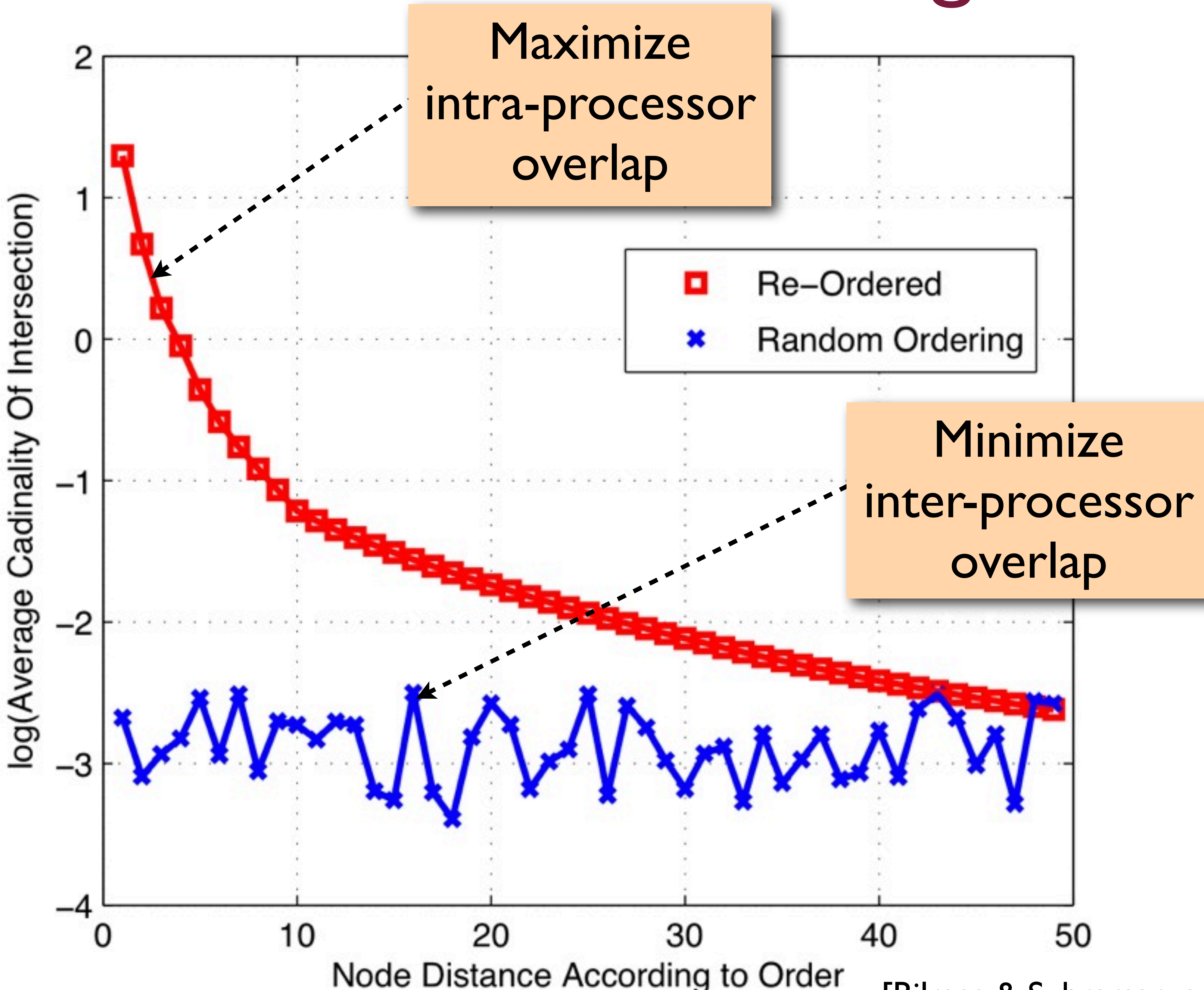
- **Maximize** overlap between consecutive nodes within the same machine
- **Minimize** overlap across machines (reduce inter machine communication)

Distributed Processing



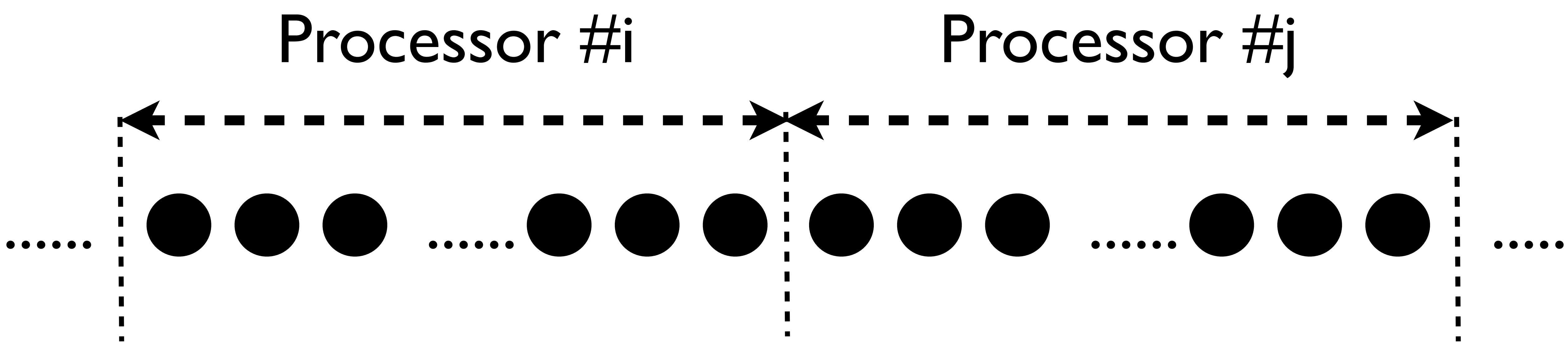
[Bilmes & Subramanya, 2011]

Distributed Processing

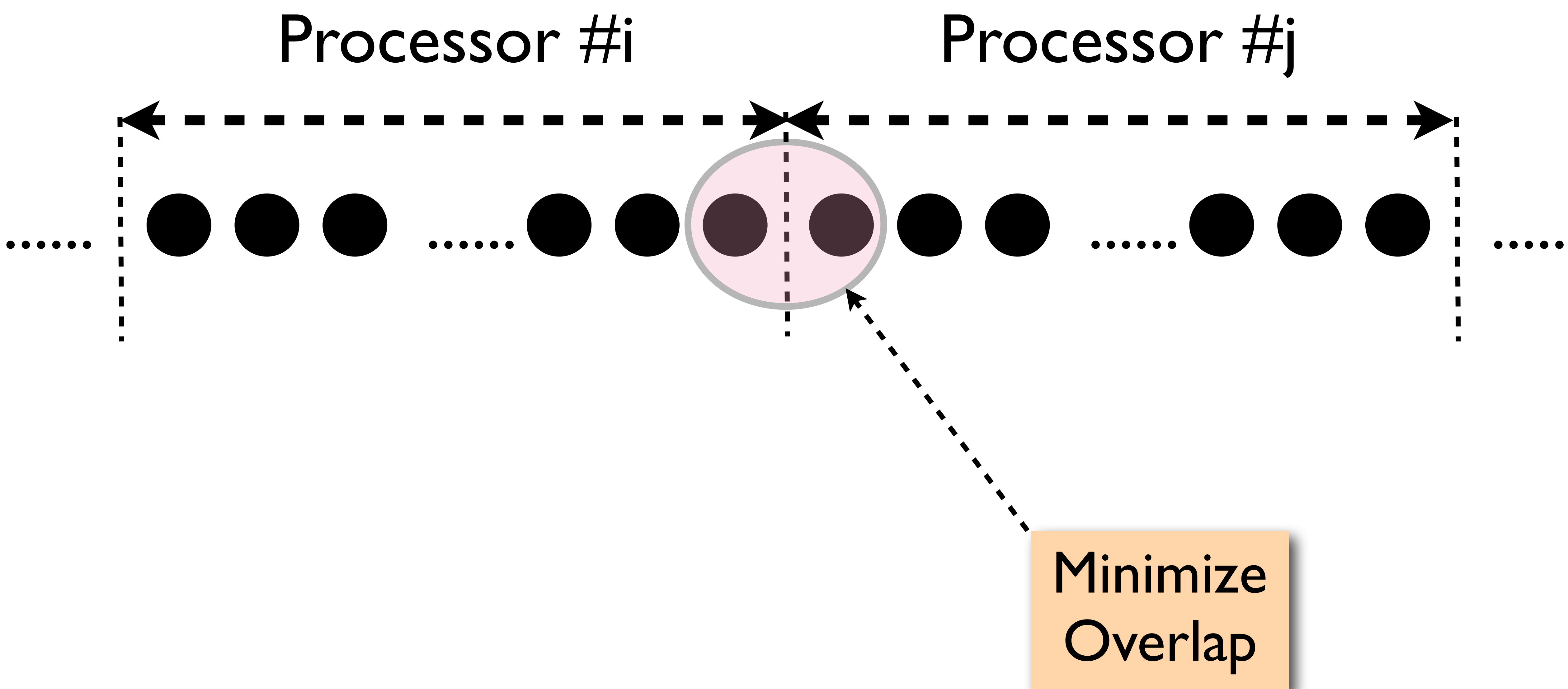


[Bilmes & Subramanya, 2011]

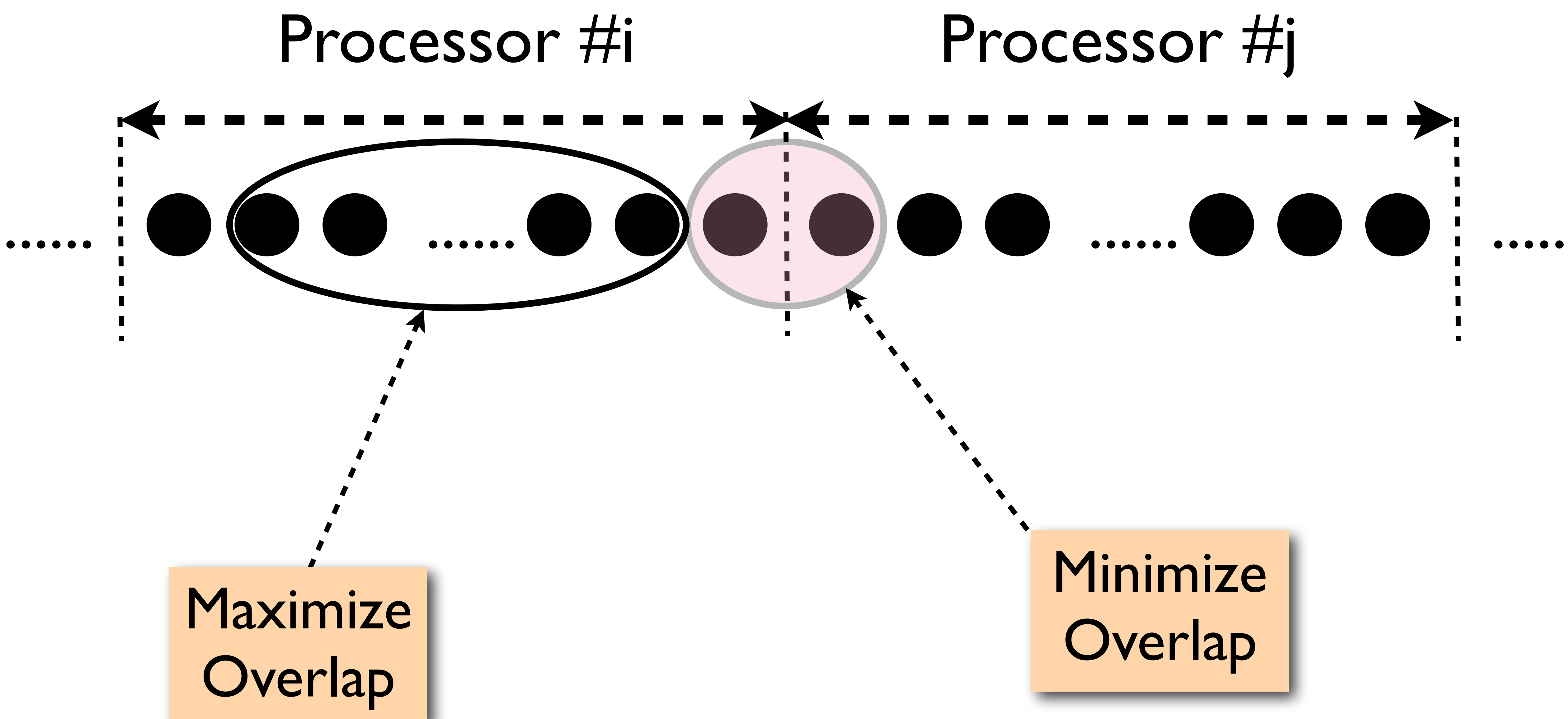
Node reordering for Distributed Computer



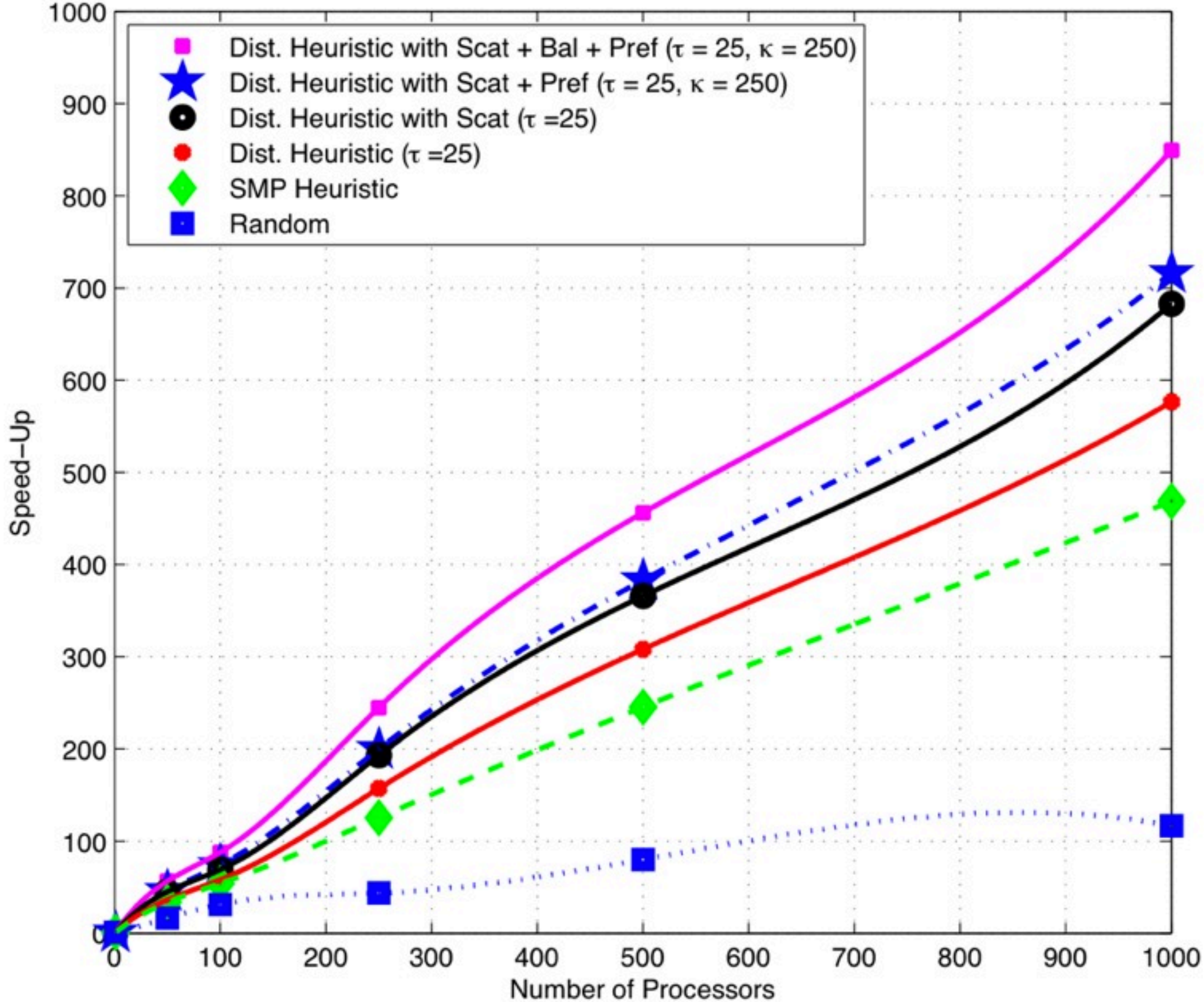
Node reordering for Distributed Computer



Node reordering for Distributed Computer

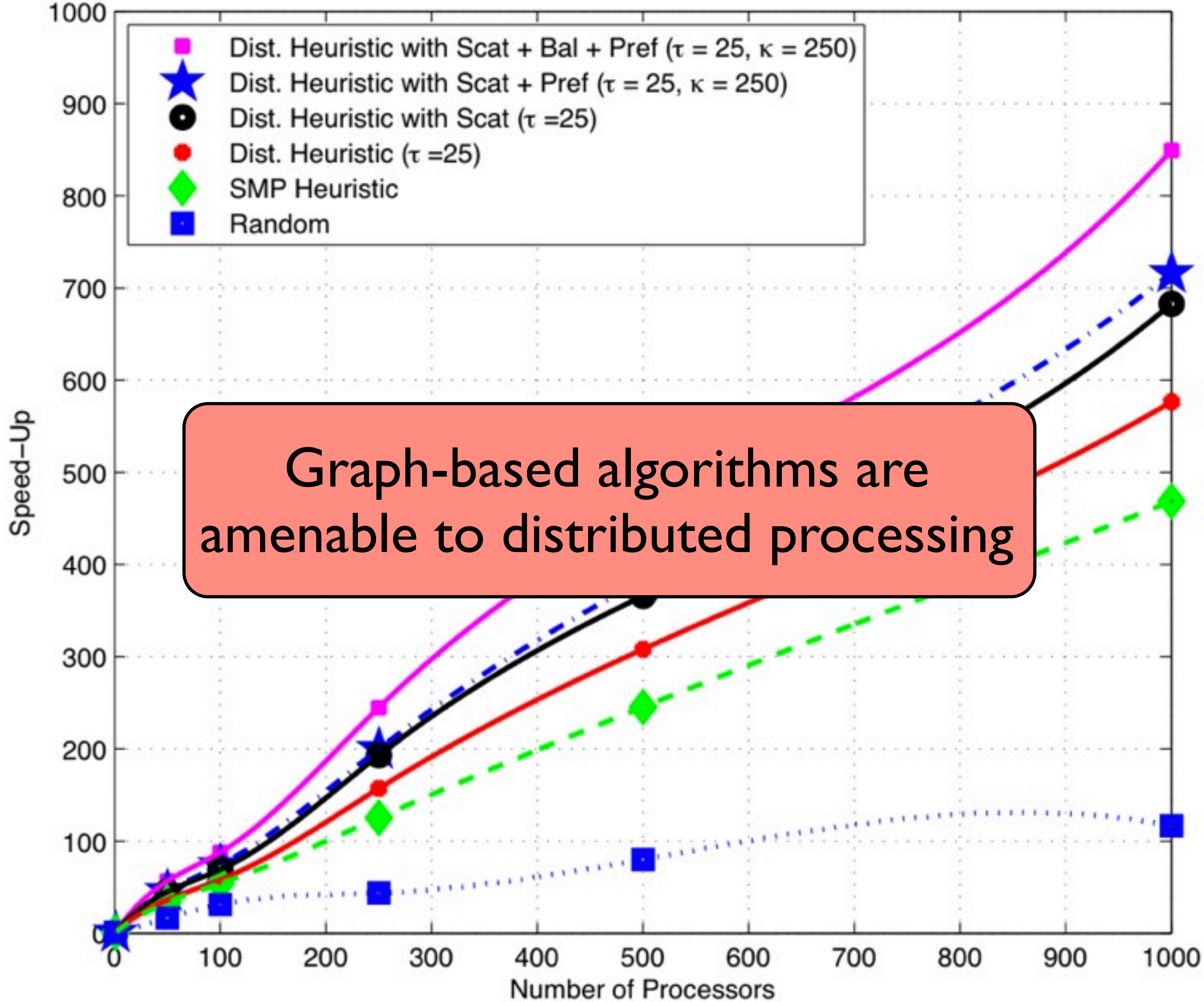


Distributed Processing Results



[Bilmes & Subramanya, 2011]

Distributed Processing Results

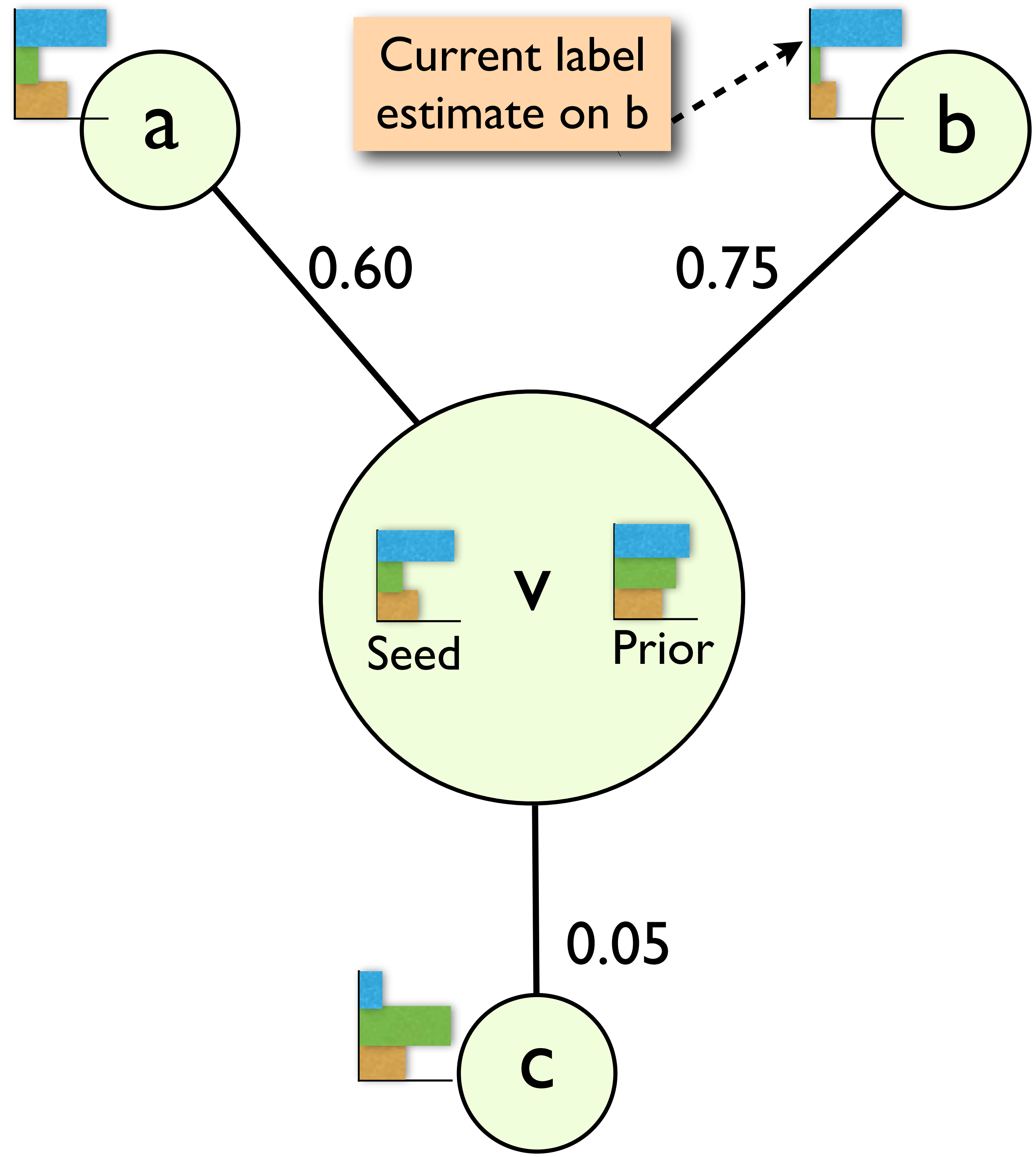


[Bilmes & Subramanya, 2011]

Outline

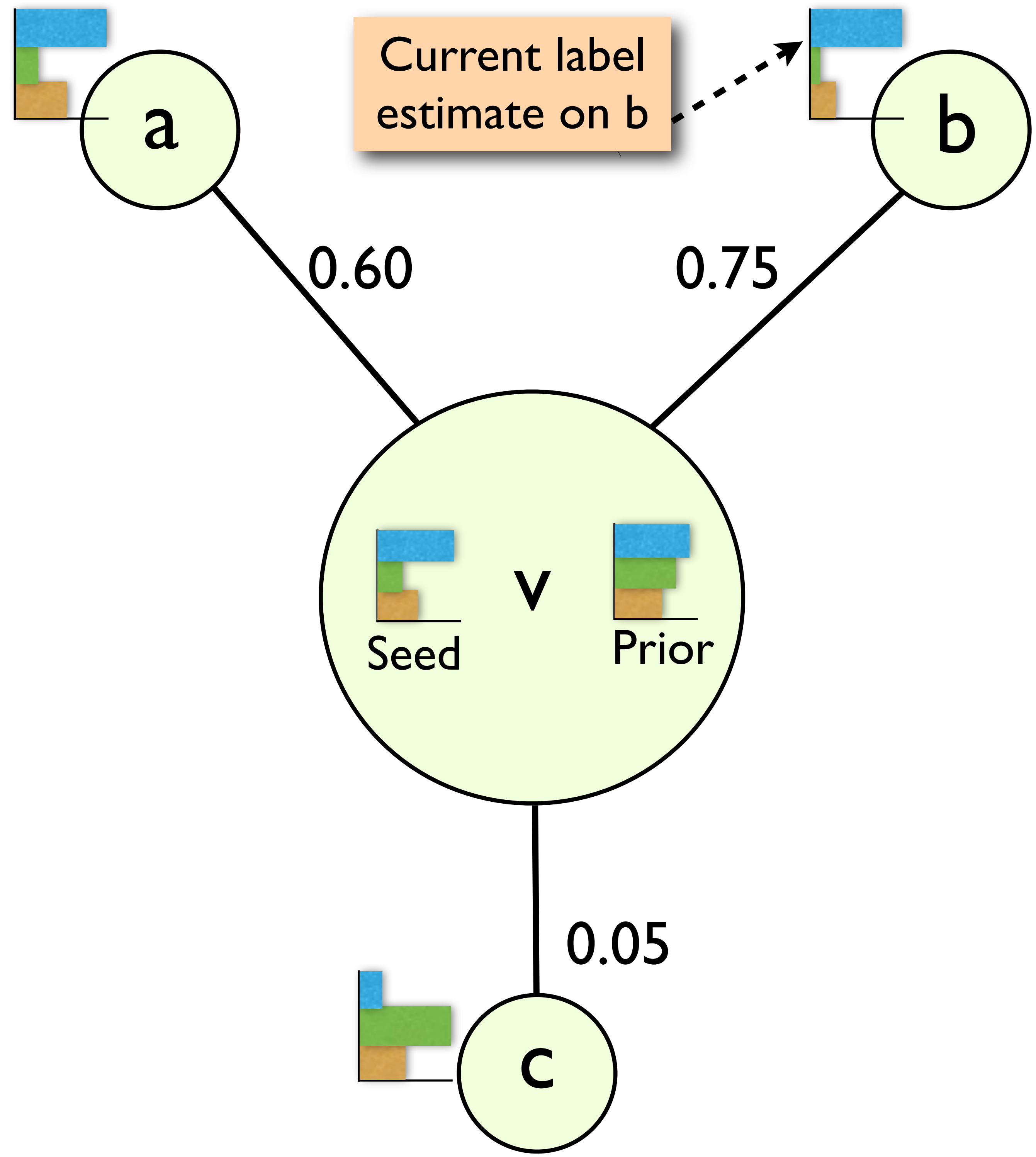
- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

MapReduce Implementation of MAD



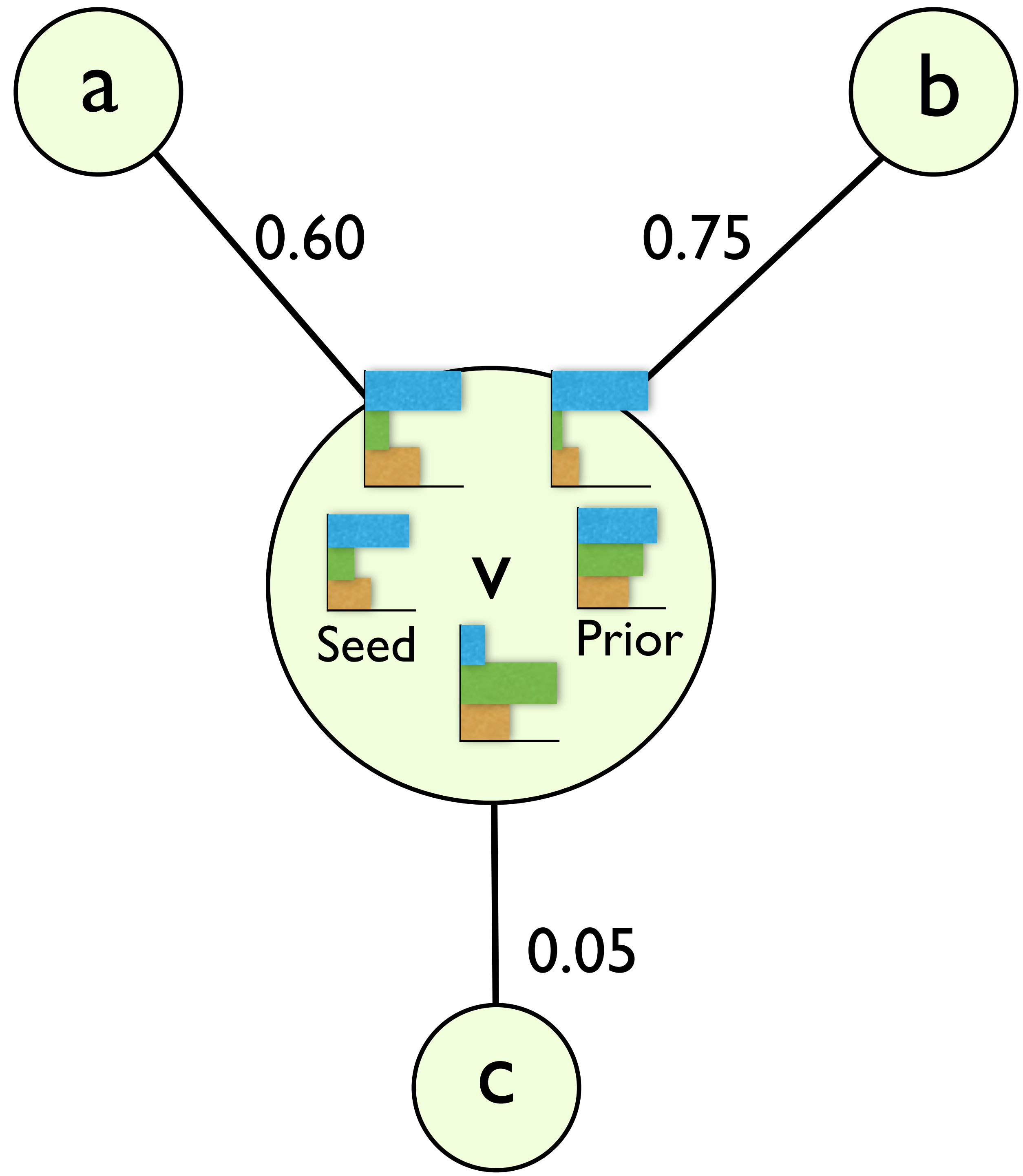
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors



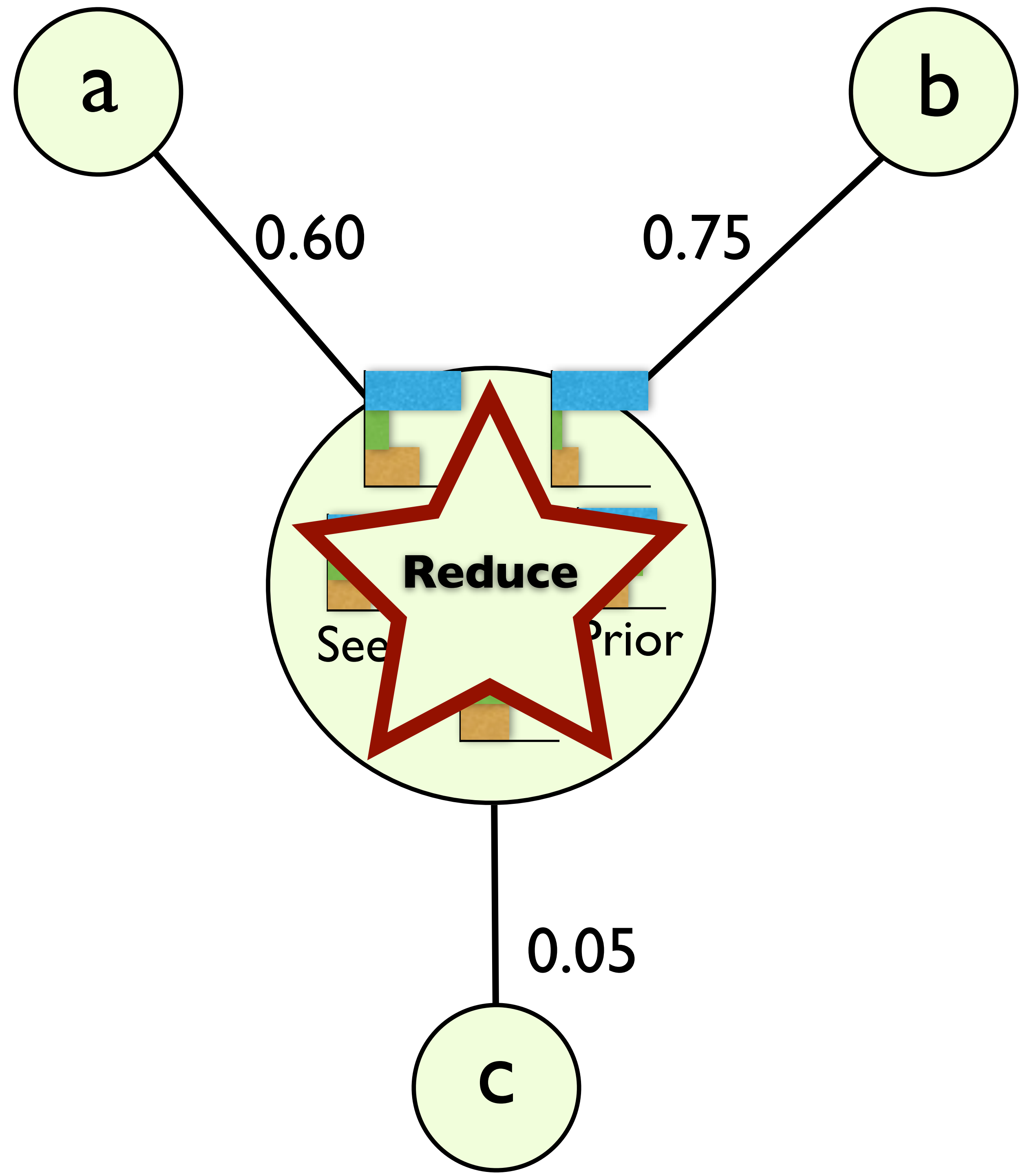
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors



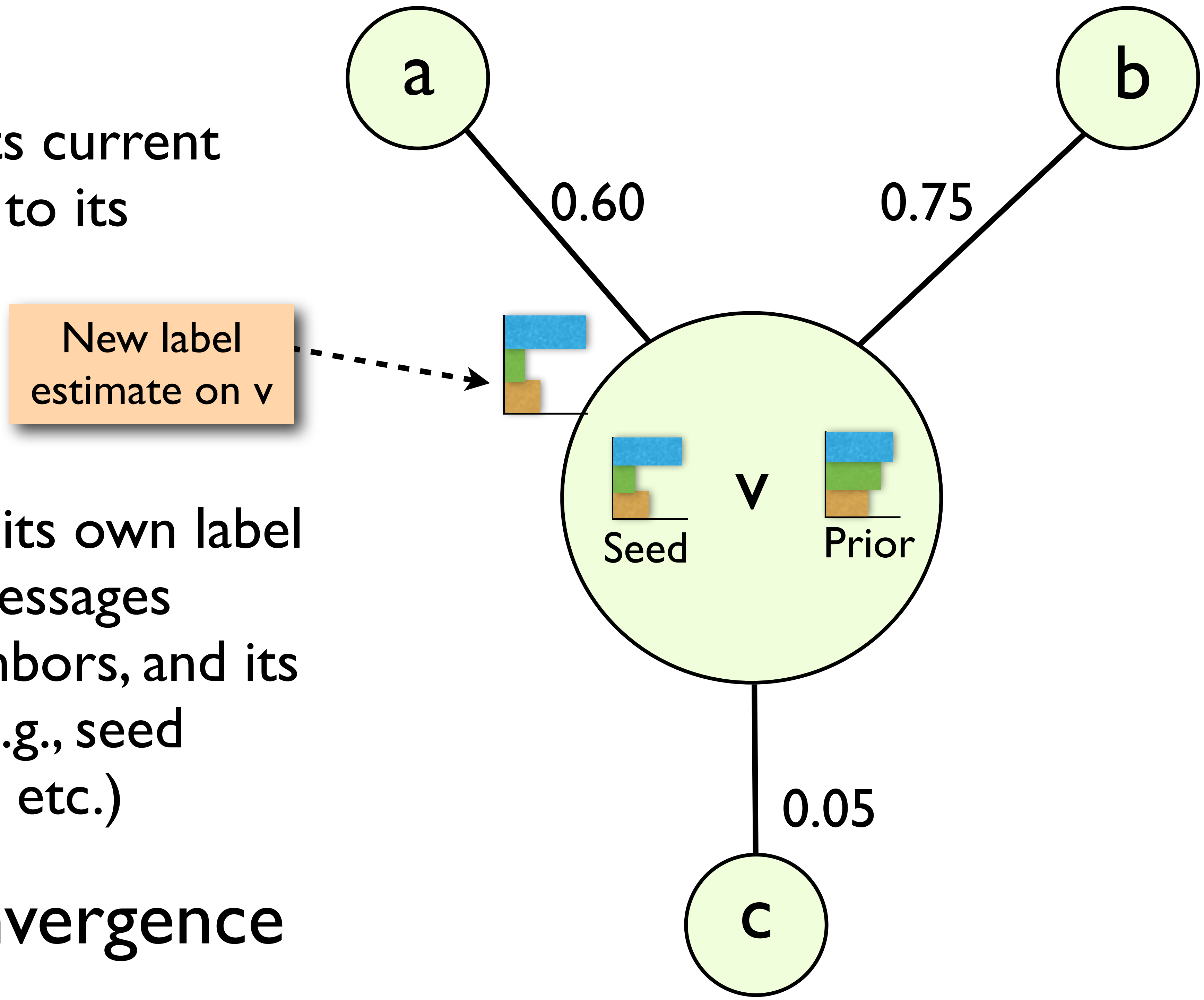
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node updates its own label assignment using messages received from neighbors, and its own information (e.g., seed labels, reg. penalties etc.)
- Repeat until convergence



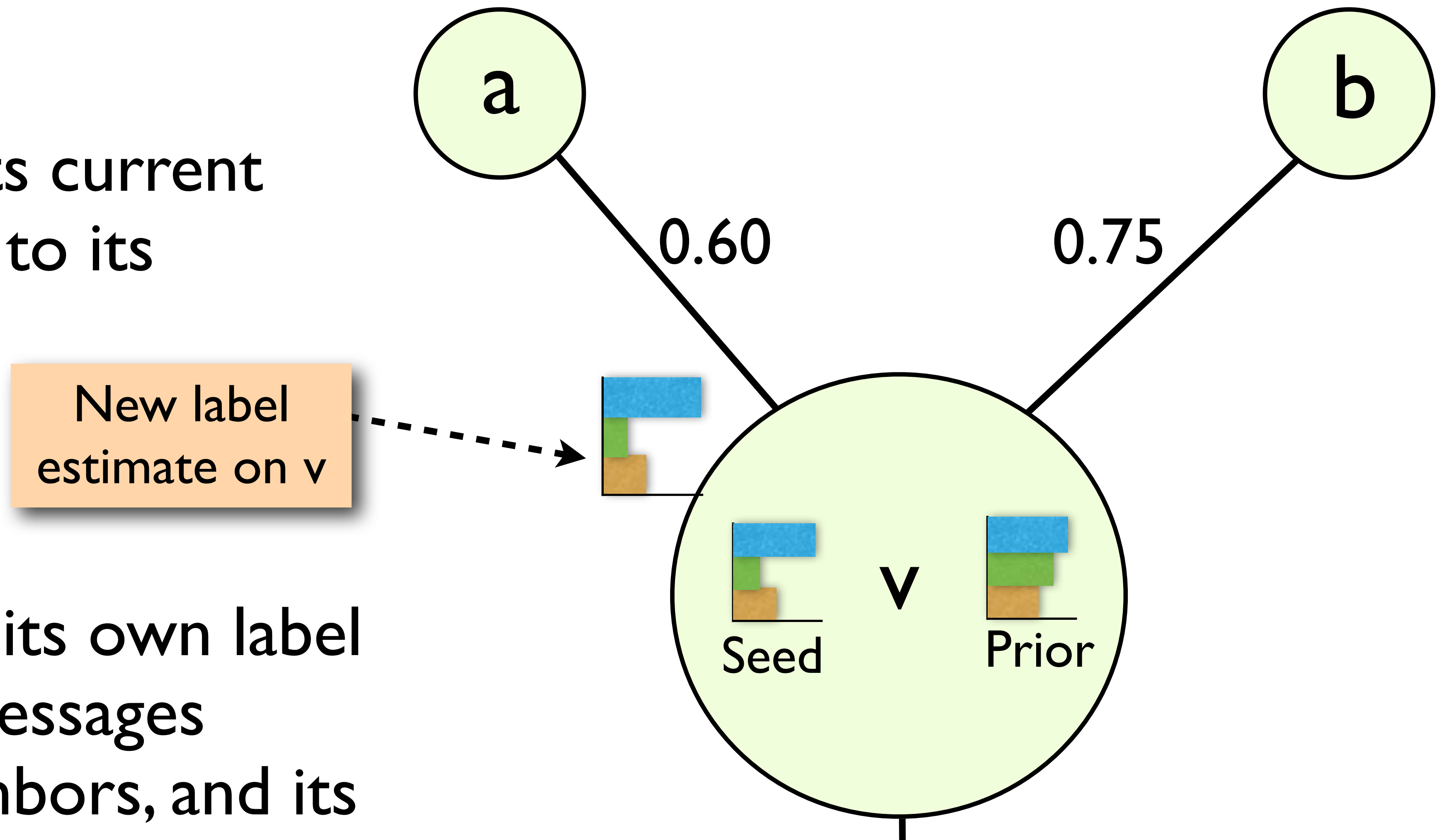
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node updates its own label assignment using messages received from neighbors, and its own information (e.g., seed labels, reg. penalties etc.)
- Repeat until convergence



MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node updates its own label assignment using messages received from neighbors, and its own[illegible]
    label[illegible]
- Repe[illegible]



Code in Junto Label Propagation Toolkit

(includes Hadoop-based implementation)

http://code.google.com/p/junto/

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Text Categorization
  - Sentiment Analysis
  - Class Instance Acquisition
  - POS Tagging
  - MultiLingual POS Tagging
  - Semantic Parsing
- Conclusion & Future Work

Problem Description & Motivation

- Given a document (e.g., web page, news article), assign it to a fixed number of semantic categories (e.g., sports, politics, entertainment)
- <u>Multi-label</u> problem
- Training supervised models requires large amounts of labeled data [Dumais et al., 1998]

Corpora

- **Reuters** [Lewis, et al., 1978]
  - Newswire
  - About 20K document with 135 categories. Use top 10 categories (e.g., "earnings", "acquistions", "wheat", "interest") and label the remaining as "other"

Corpora

- **Reuters** [Lewis, et al., 1978]
  - Newswire
  - About 20K document with 135 categories. Use top 10 categories (e.g., "earnings", "acquistions", "wheat", "interest") and label the remaining as "other"
- **WebKB** [Bekkerman, et al., 2003]
  - 8K webpages from 4 academic domains
  - Categories include "course", "department", "faculty" and "project"

Feature Extraction

Showers continued throughout the week in
the Bahia cocoa zone, alleviating the drought since early January and improving prospects for the coming temporao, ...

Document

[Lewis, et al., 1978]

Feature Extraction

Showers continued throughout the week in the Bahia cocoa zone, alleviating the drought since early January and improving prospects for the coming temporao, ...

Document
[Lewis, et al., 1978]

Showers continued week Bahia cocoa zone alleviating drought early January improving prospects coming temporao, ...

Stop-word Removal

Feature Extraction

Showers continued throughout the week in the Bahia cocoa zone, alleviating the drought since early January and improving prospects for the coming temporao, ...

Document
[Lewis, et al., 1978]

→

Showers continued week Bahia cocoa zone alleviating drought early January improving prospects coming temporao, ...

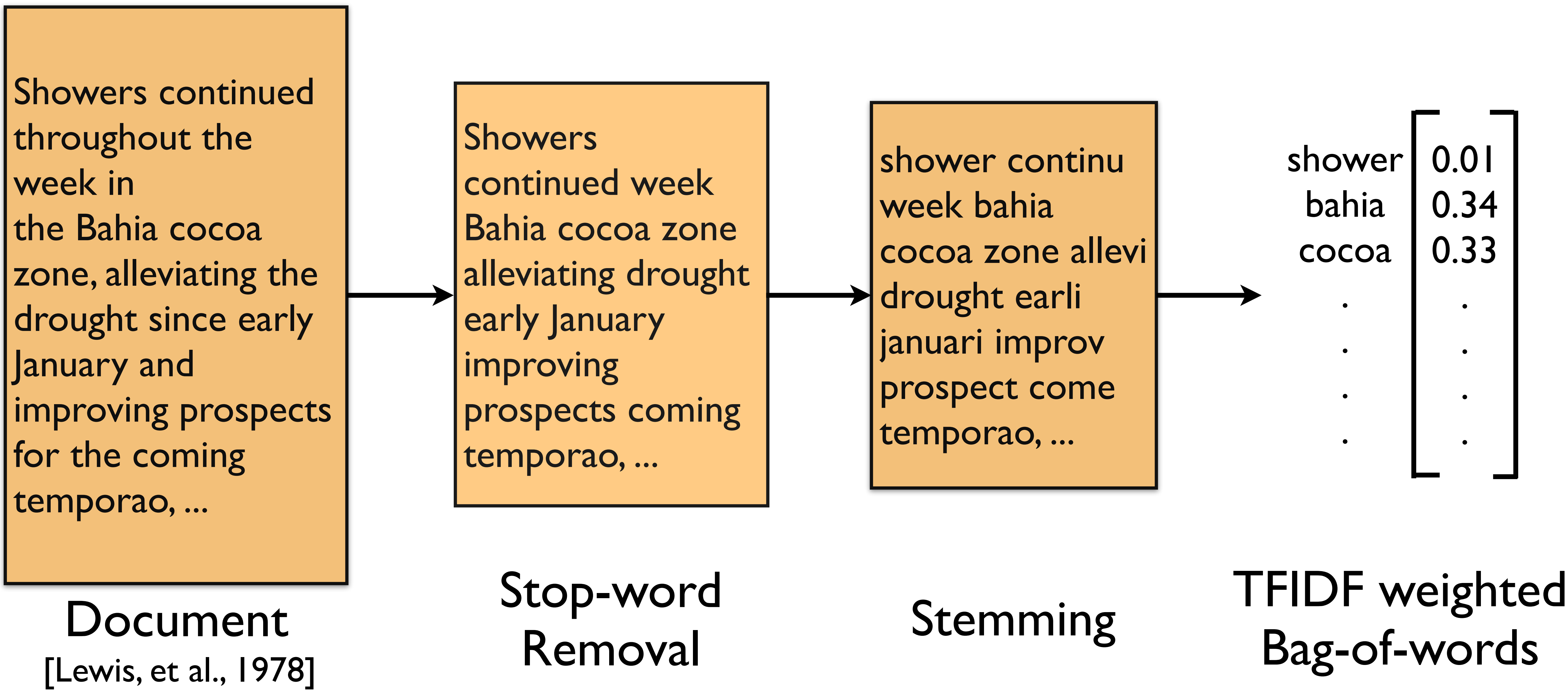
Stop-word Removal

→

shower continu week bahia cocoa zone allevi drought earli januari improv prospect come temporao, ...

Stemming

Feature Extraction



Results

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

Precision-recall break even point (PRBEP)

Results

Support Vector Machine (Supervised)

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

Precision-recall break even point (PRBEP)

Results

Transductive SVM [Joachims 1999]

Support Vector Machine (Supervised)

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

Precision-recall break even point (PRBEP)

Results

Transductive SVM [Joachims 1999]

Spectral Graph Transduction (SGT) [Joachims 2003]

Label Propagation [Zhu & Ghahramani 2002]

Measure Propagation [Subramanya & Bilmes 2008]

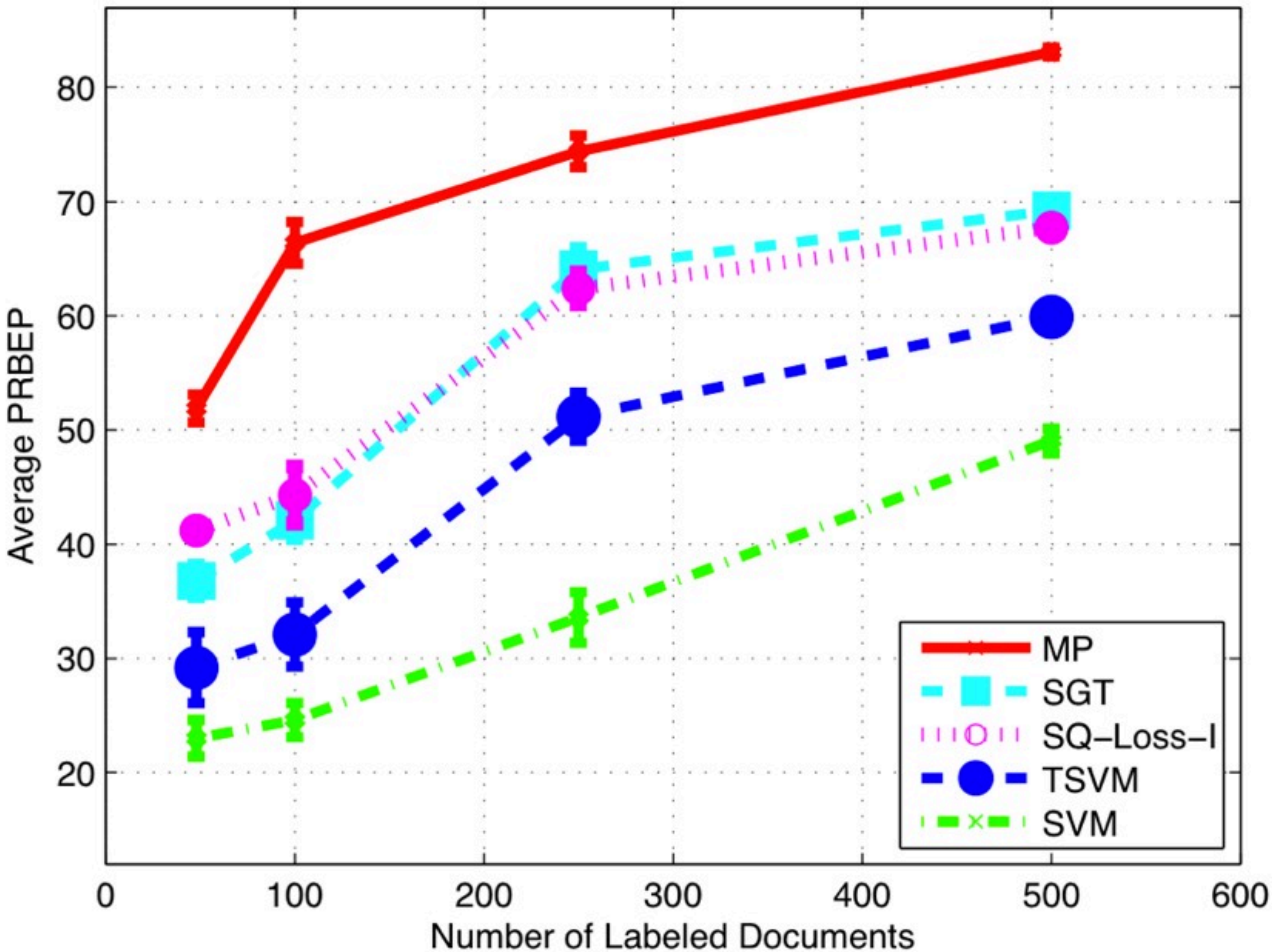
Support Vector Machine (Supervised)

Modified Adsorption [Talukdar & Crammer 2009]

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

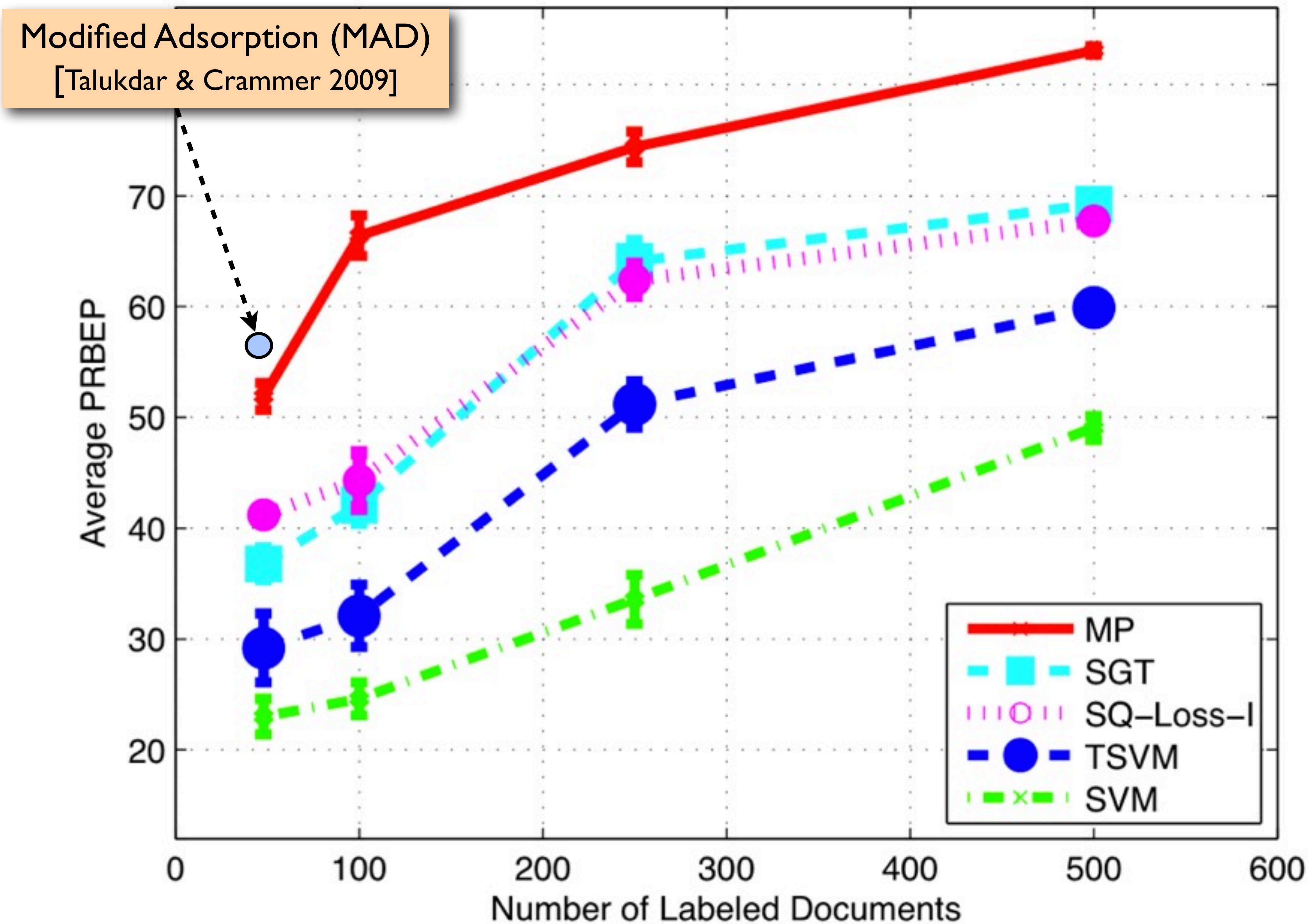
Precision-recall break even point (PRBEP)

Results on WebKB



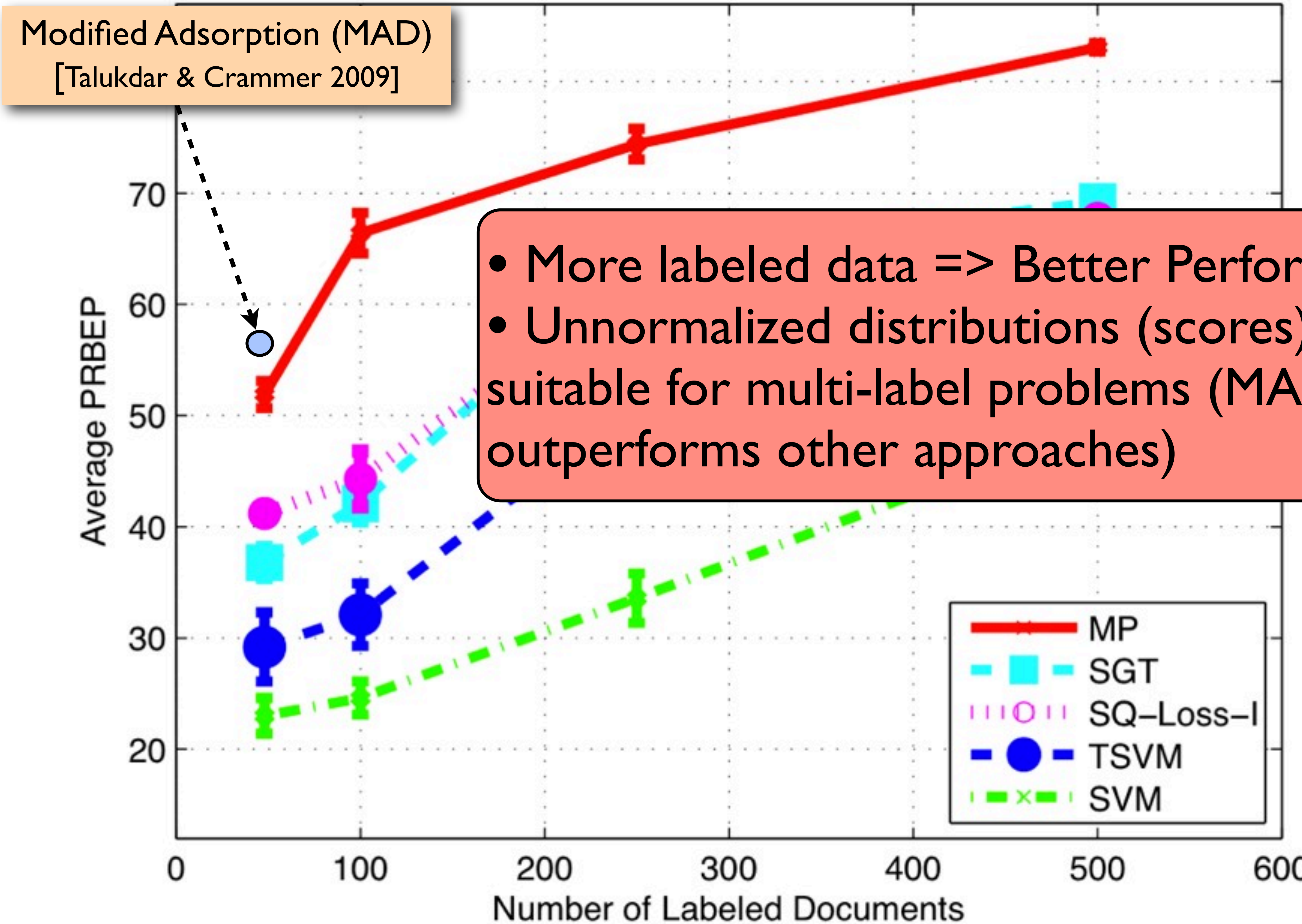
[Subramanya & Bilmes, EMNLP 2008]

Results on WebKB



[Subramanya & Bilmes, EMNLP 2008]

Results on WebKB



• More labeled data => Better Performance
• Unnormalized distributions (scores) more suitable for multi-label problems (MAD outperforms other approaches)

[Subramanya & Bilmes, EMNLP 2008]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Text Categorization
  - Sentiment Analysis
  - Class Instance Acquisition
  - POS Tagging
  - MultiLingual POS Tagging
  - Semantic Parsing
- Conclusion & Future Work

Problem Description

- fortunately, they managed to do it in an interesting and funny way.
- he is one of the most exciting martial artists on the big screen.
- the romance was enchanting.

- A woman in peril. A confrontation. An explosion. The end. Yawn. Yawn. Yawn.
- don’t go see this movie

Movie review dataset [Pang et al. EMNLP 2002]

Problem Description

- Given a document either
  - classify it as expressing a positive or negative sentiment or
  - assign a star rating
- Similar to text categorization
  - Can be solved using standard machine learning approaches [Pang, Lee & Vaidyanathan, EMNLP 2002]

Polarity Lexicons (I)

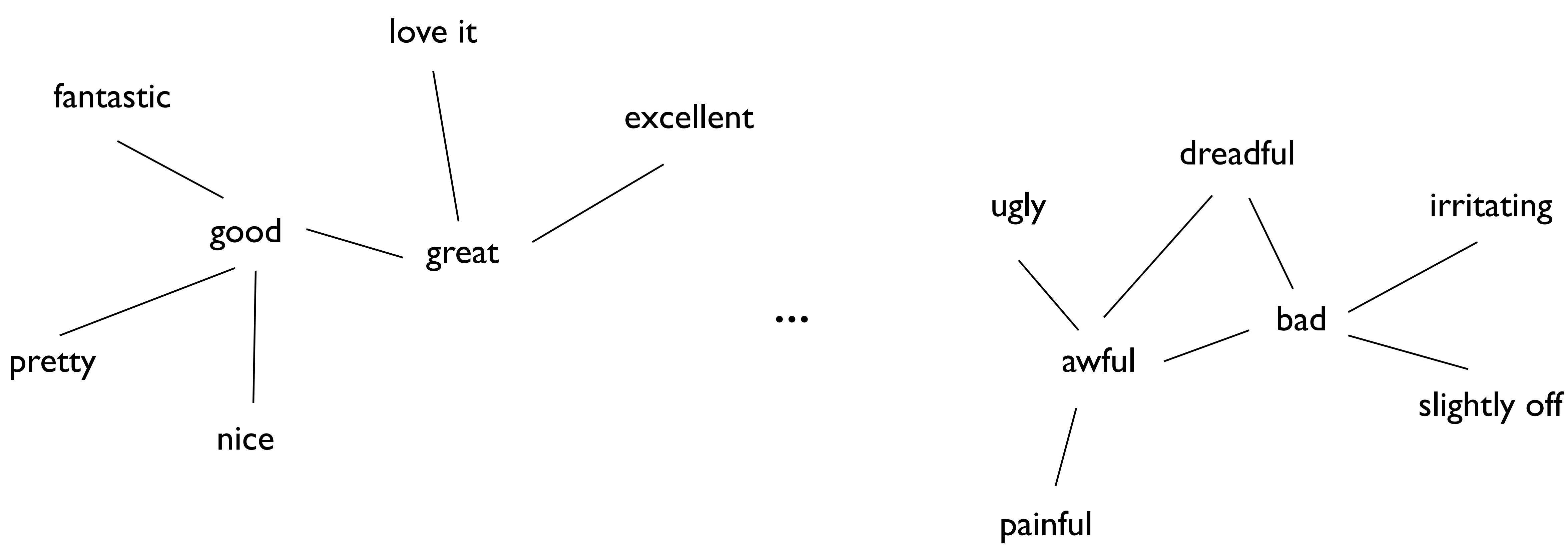
- Large lists of phrases that encode the polarity (positive or negative) of each phrase
  - Positive polarity: “enjoyable”, “breathtakingly”, “once in a life time”
  - Negative polarity: “bad”, “humorless”, “unbearable”, “out of touch”, “bumps in the road”
- Best results obtained by combining with machine learning approaches [Wilson et al., HLT-EMNLP 05; Blair-Goldensohn et al. 08; Choi & Cardie EMNLP 09]

Polarity Lexicons (II)

- Common strategy: start with two **small** seed sets
  - P: positive phrases, e.g., “great” “fantastic”
  - N: negative phrases, e.g., “awful”, “dreadful”
- Grow lexicons with graph propagation algorithms

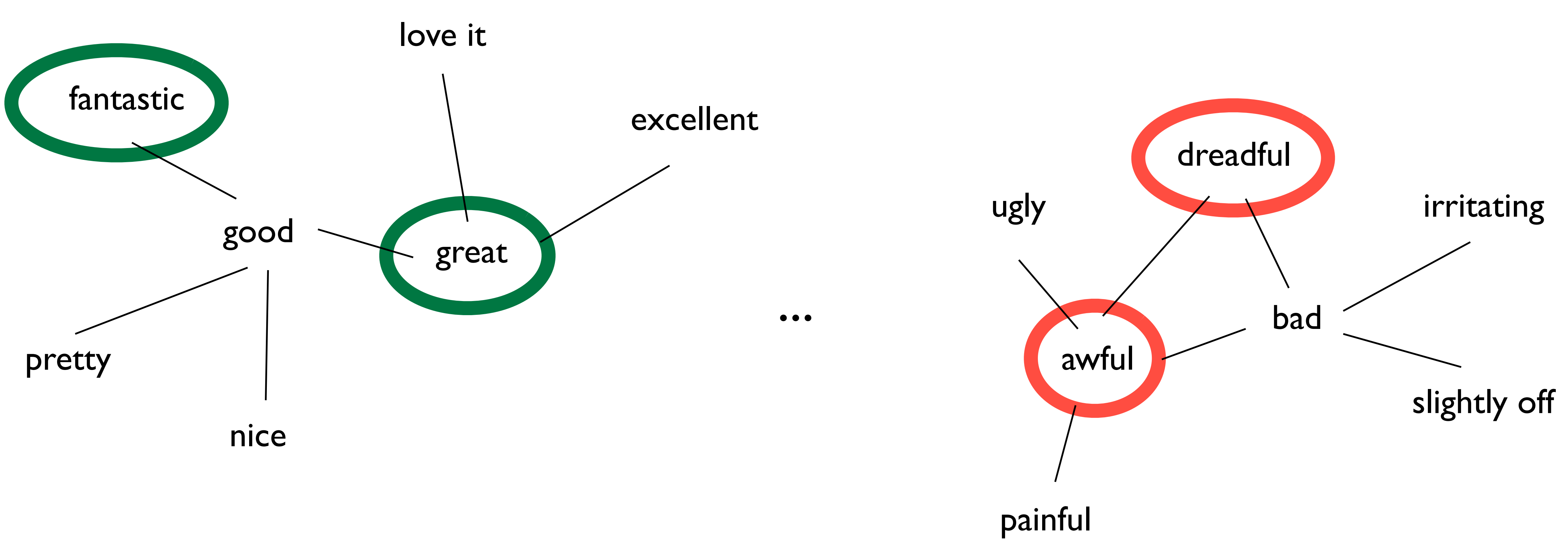
Polarity Lexicons (II)

- Common strategy: start with two **small** seed sets
  - P: positive phrases, e.g., "great" "fantastic"
  - N: negative phrases, e.g., "awful", "dreadful"
- Grow lexicons with graph propagation algorithms



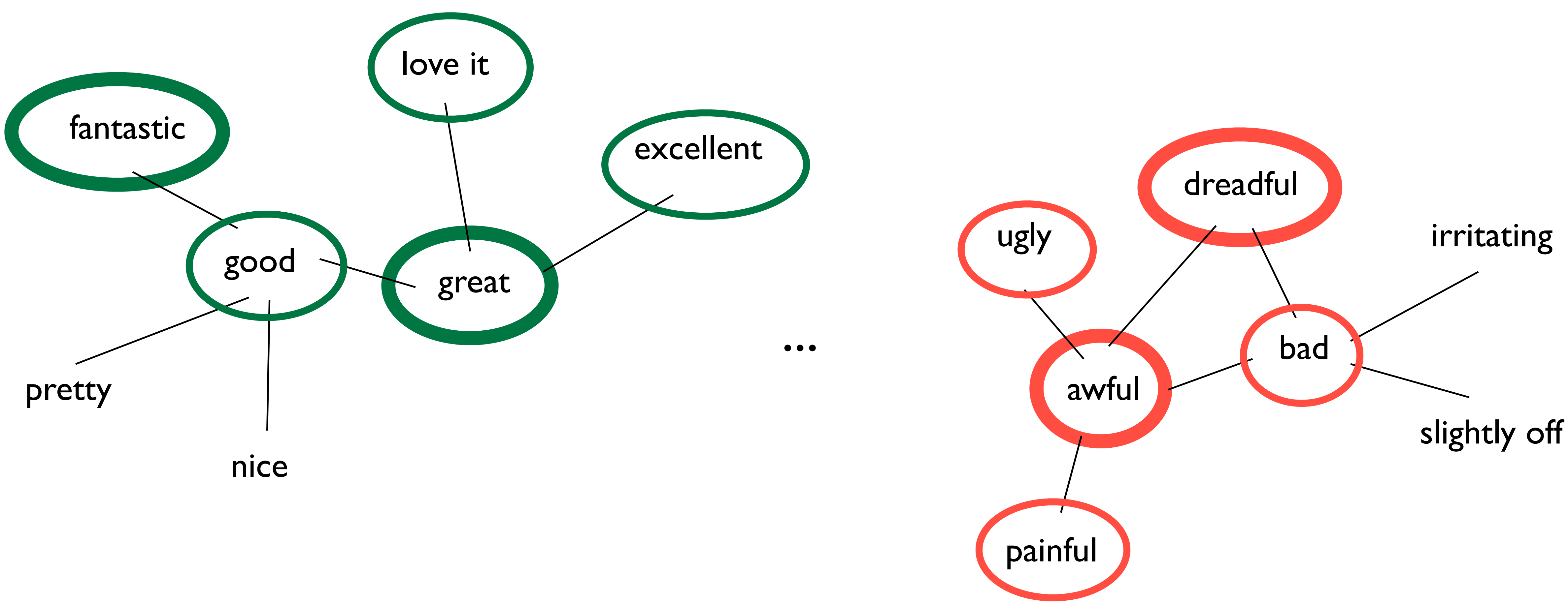
Polarity Lexicons (II)

- Common strategy: start with two **small** seed sets
  - P: positive phrases, e.g., “great” “fantastic”
  - N: negative phrases, e.g., “awful”, “dreadful”
- Grow lexicons with graph propagation algorithms



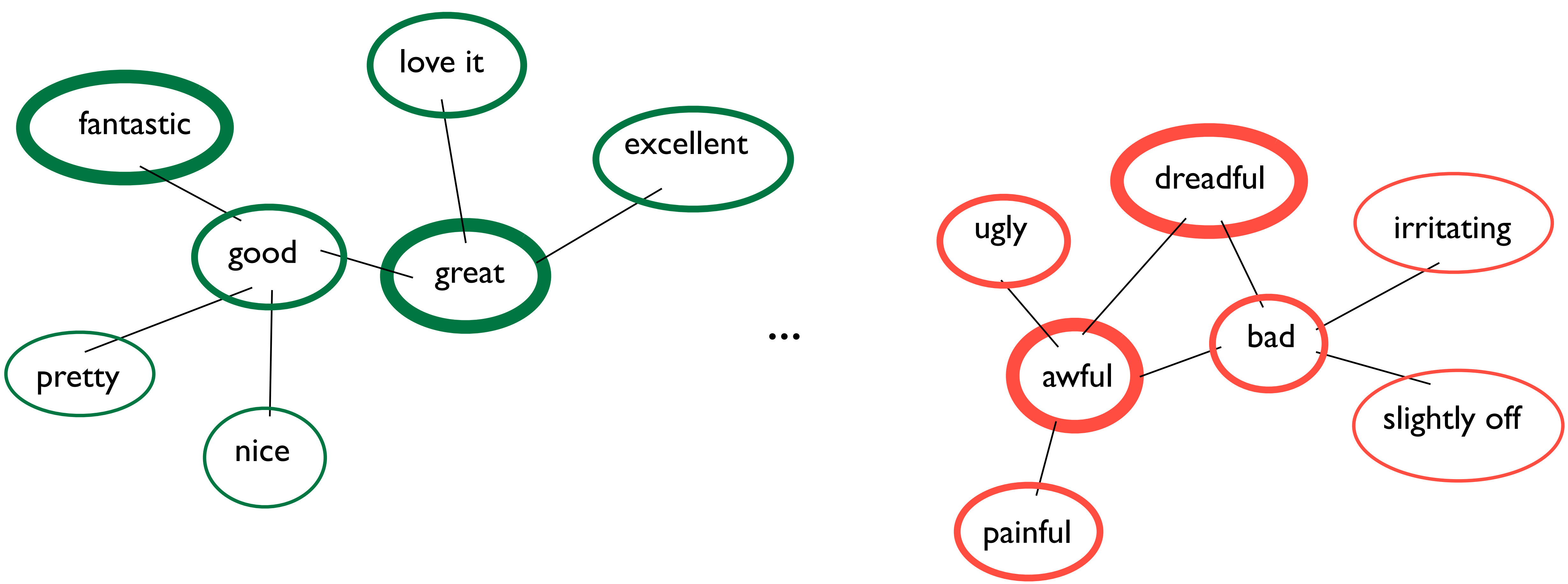
Polarity Lexicons (II)

- Common strategy: start with two **small** seed sets
  - P: positive phrases, e.g., "great" "fantastic"
  - N: negative phrases, e.g., "awful", "dreadful"
- Grow lexicons with graph propagation algorithms



Polarity Lexicons (II)

- Common strategy: start with two **small** seed sets
  - P: positive phrases, e.g., "great" "fantastic"
  - N: negative phrases, e.g., "awful", "dreadful"
- Grow lexicons with graph propagation algorithms



Graph Construction (I)

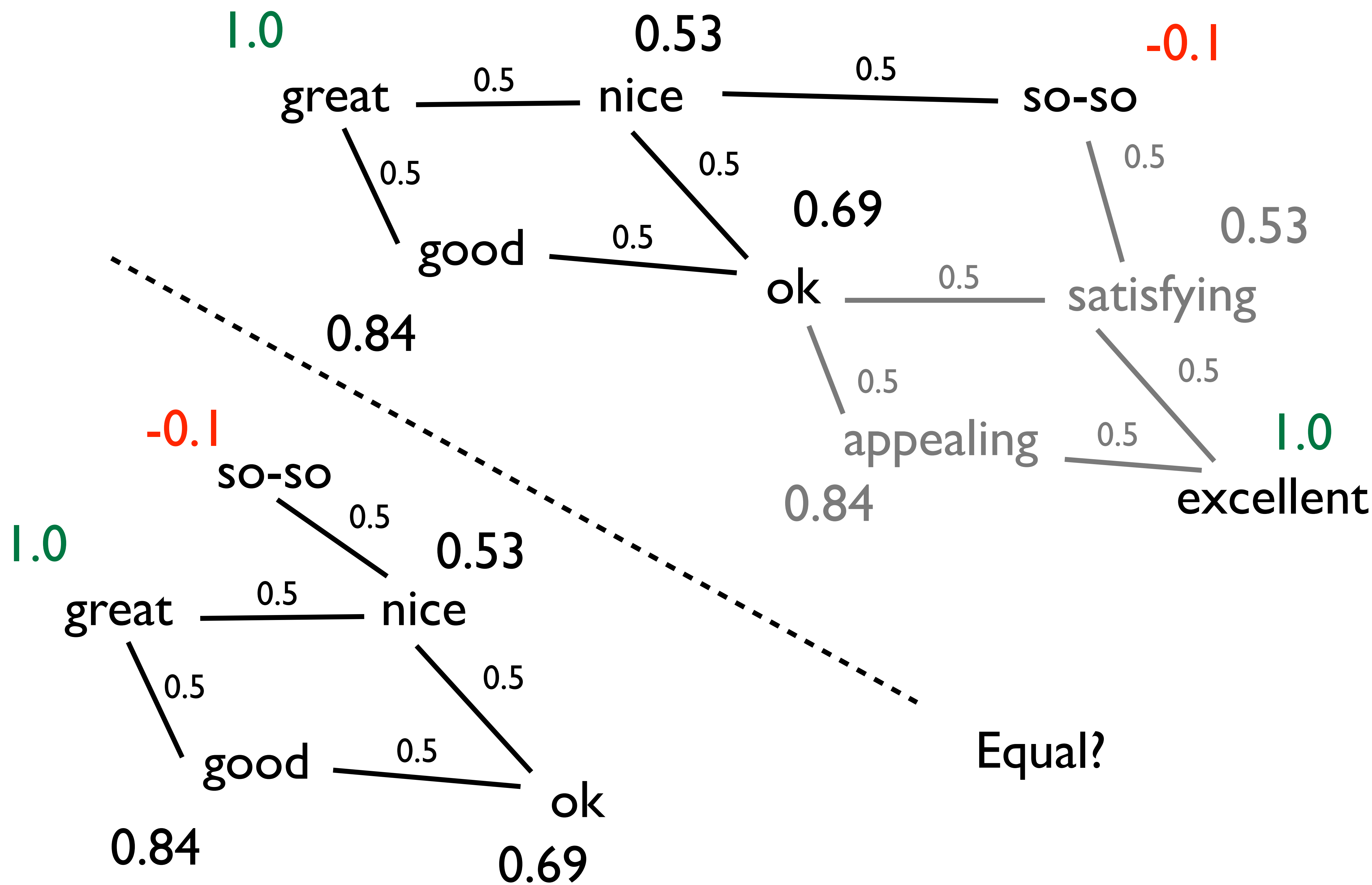
- WordNet [Hu & Liu, KDD 04; Kim & Hovy, ICCL 04; Blair-Goldensohn 08; Rao & Ravichandran EACL 09]
  - Defines synonyms, antonyms, hypernyms, etc.
  - Make edges between synonyms
  - Enforce constraints between antonyms
  - Issues
    - coverage
    - hard to find resources for all languages

Graph Construction (II)

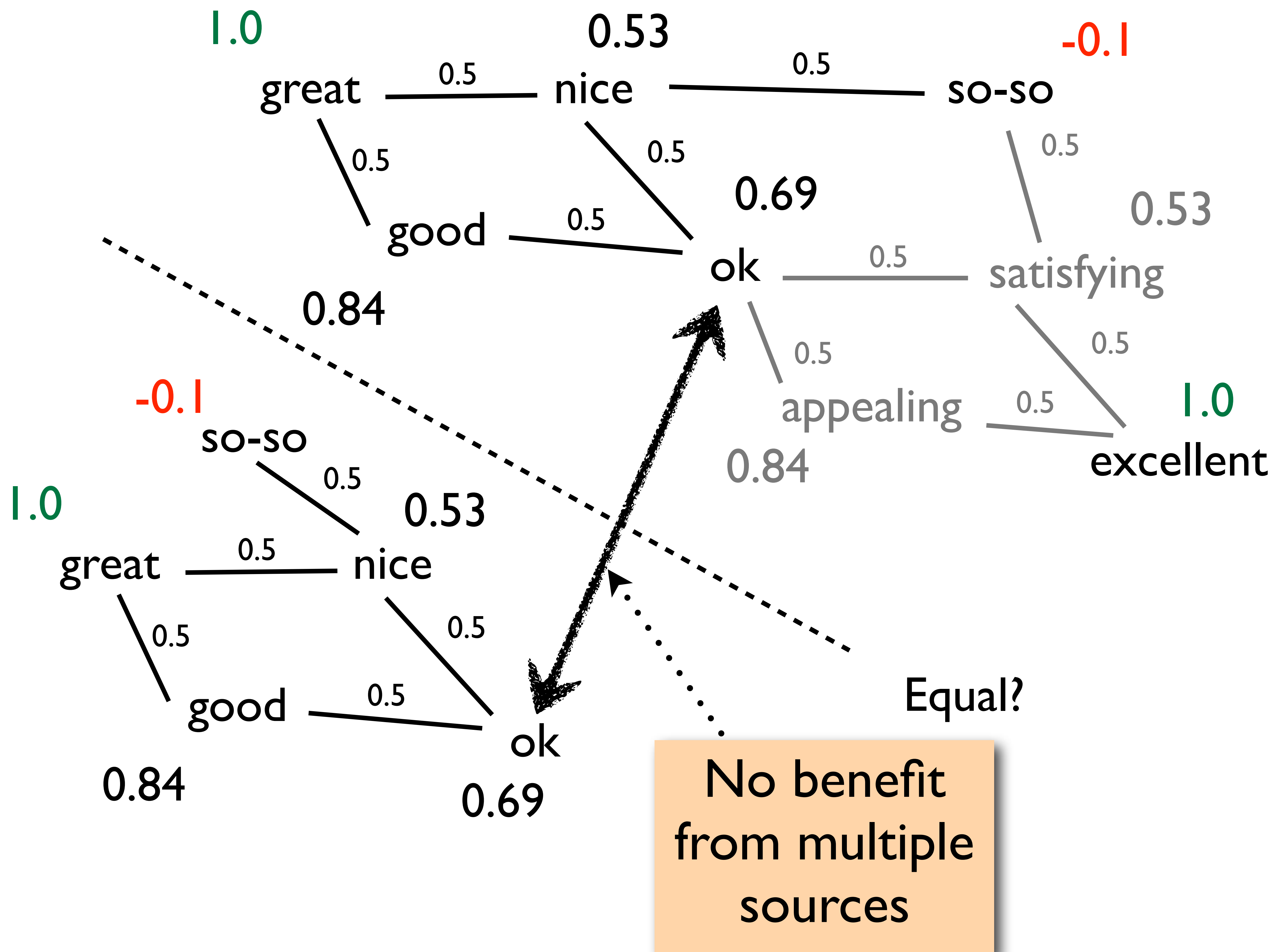
- Use web data!
- All n-grams (phrases) up to length 10 from 4 billion web pages
  - Pruned down to 20 million candidate phrases
  - Feature vector obtained by aggregating words that occurred in **local** context
- Graph is more "syntactic" than "semantic"

[Velikovich, et al., NAACL 2010]

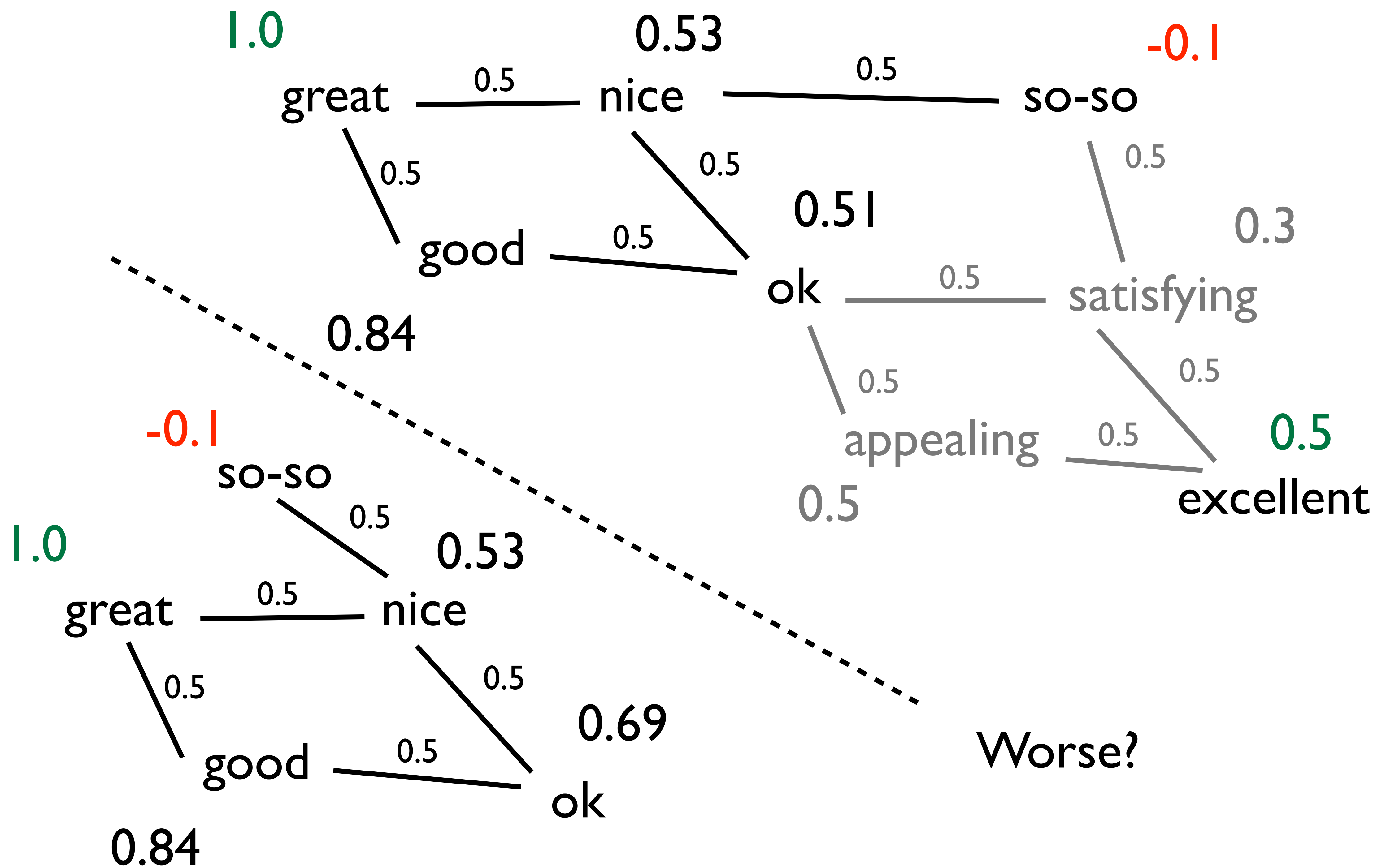
Graph Propagation (I)



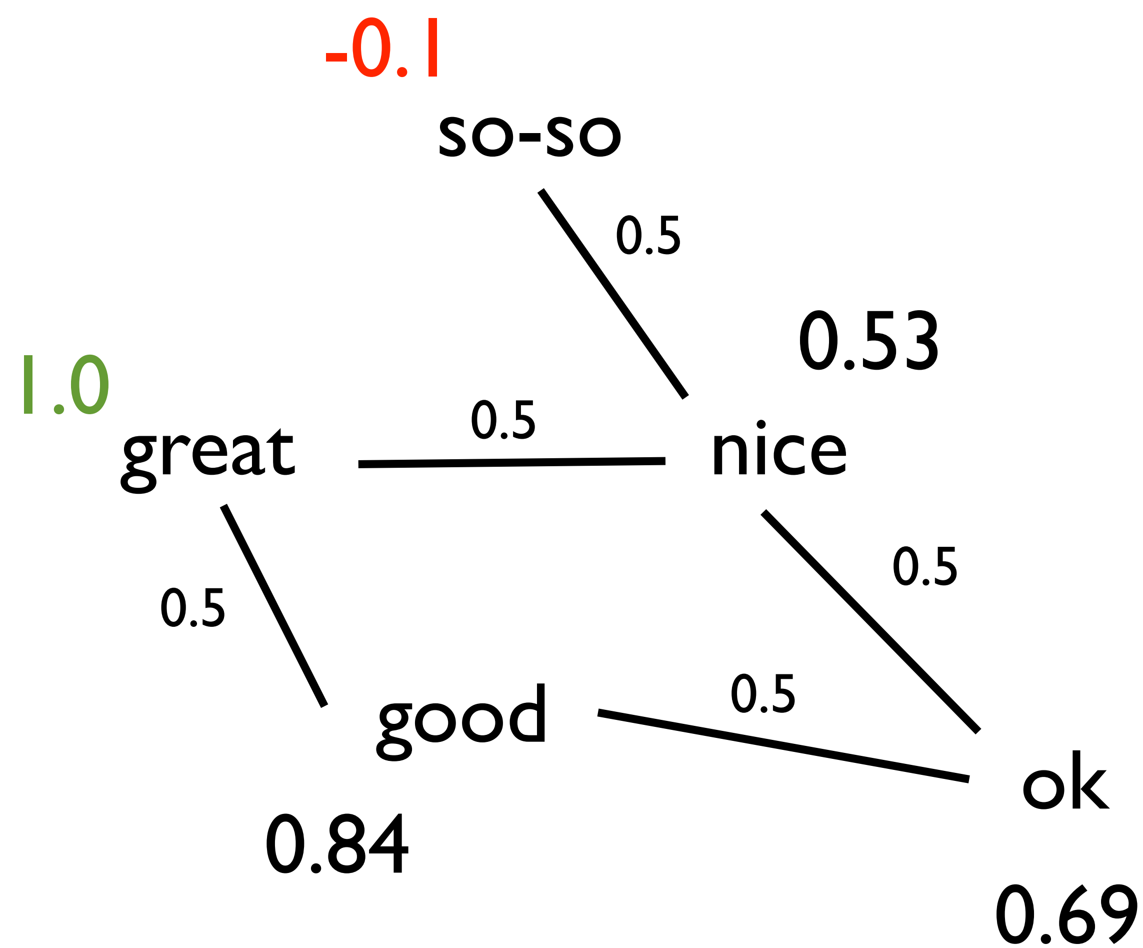
Graph Propagation (I)

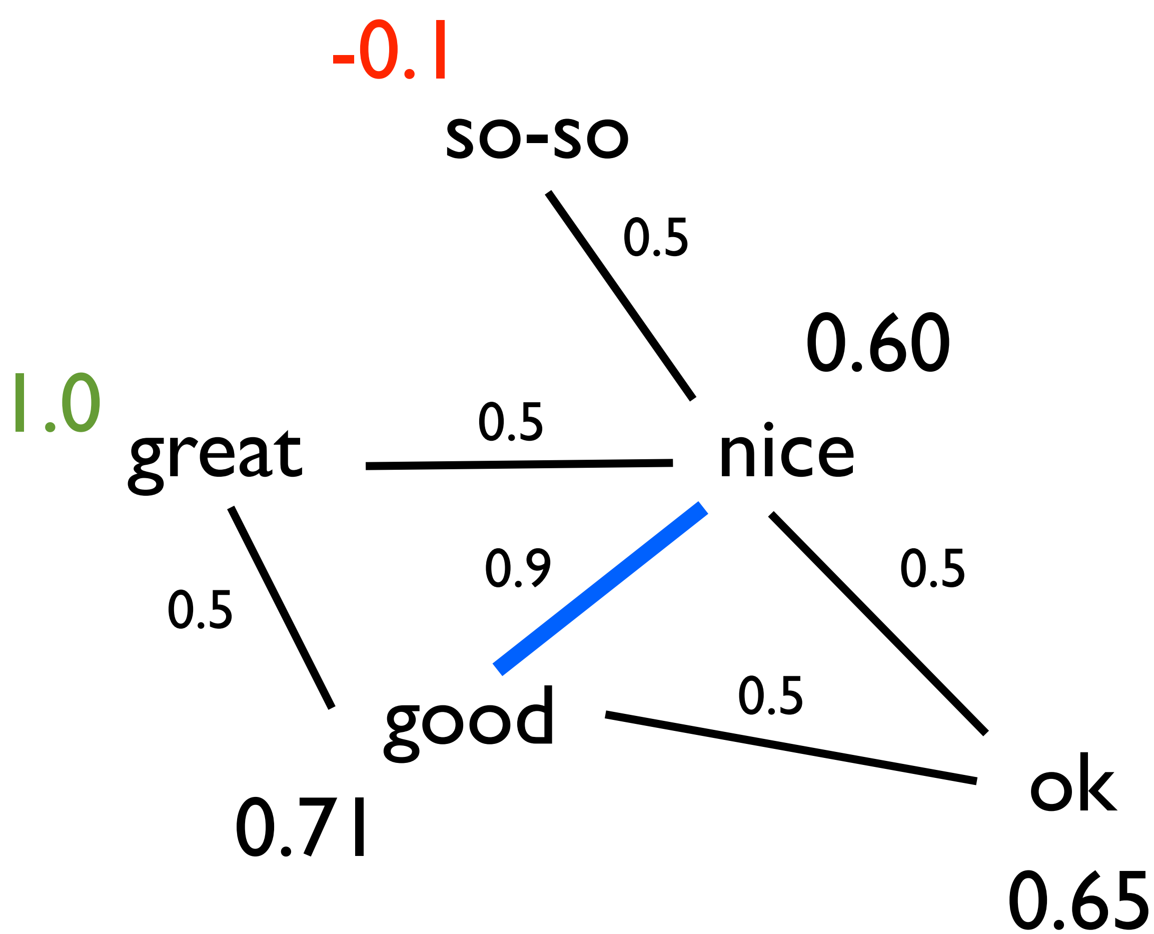


Graph Propagation (II)

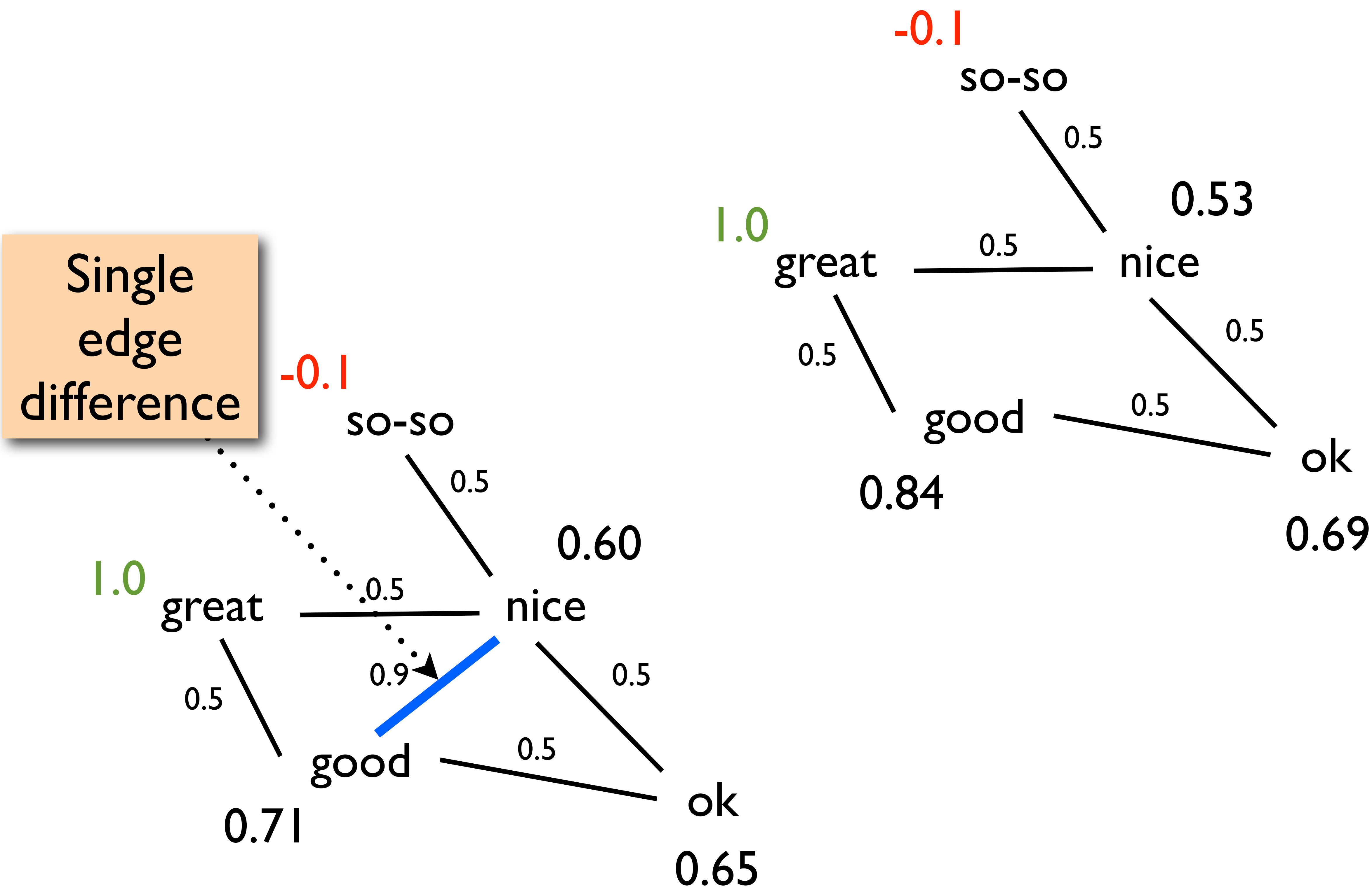


Graph Propagation (III)

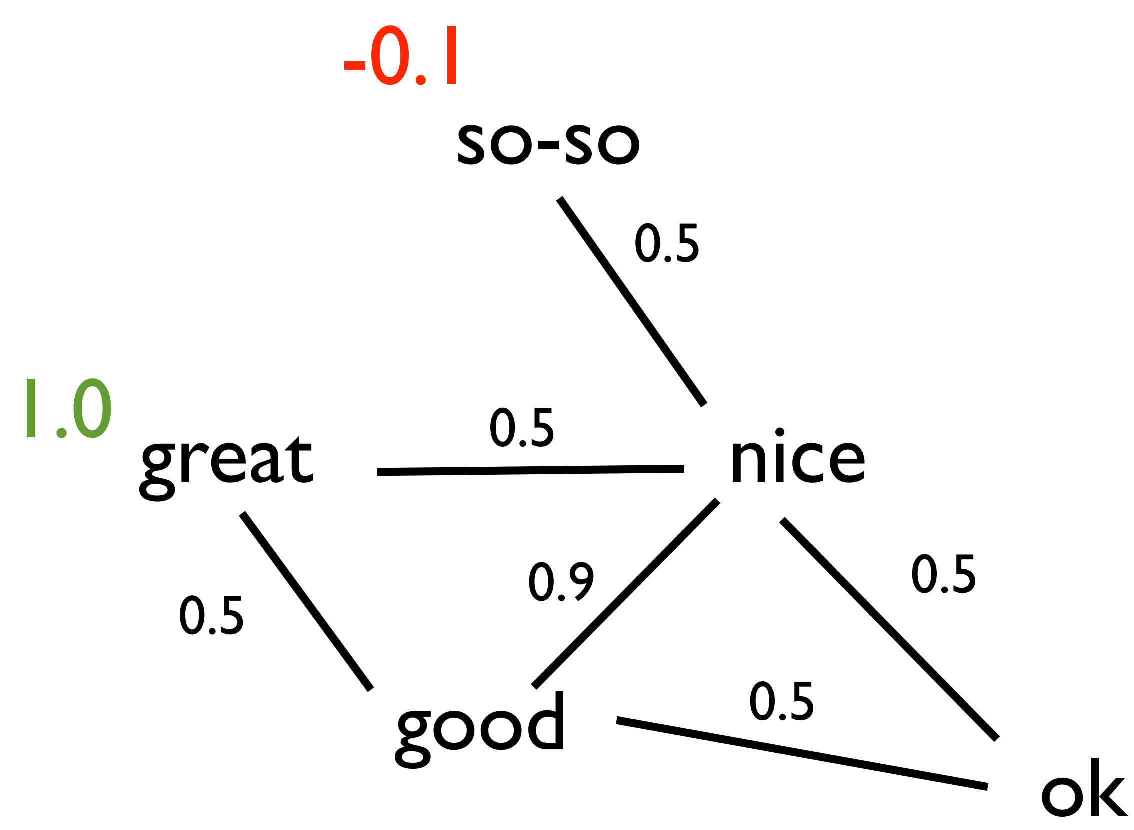




Graph Propagation (III)

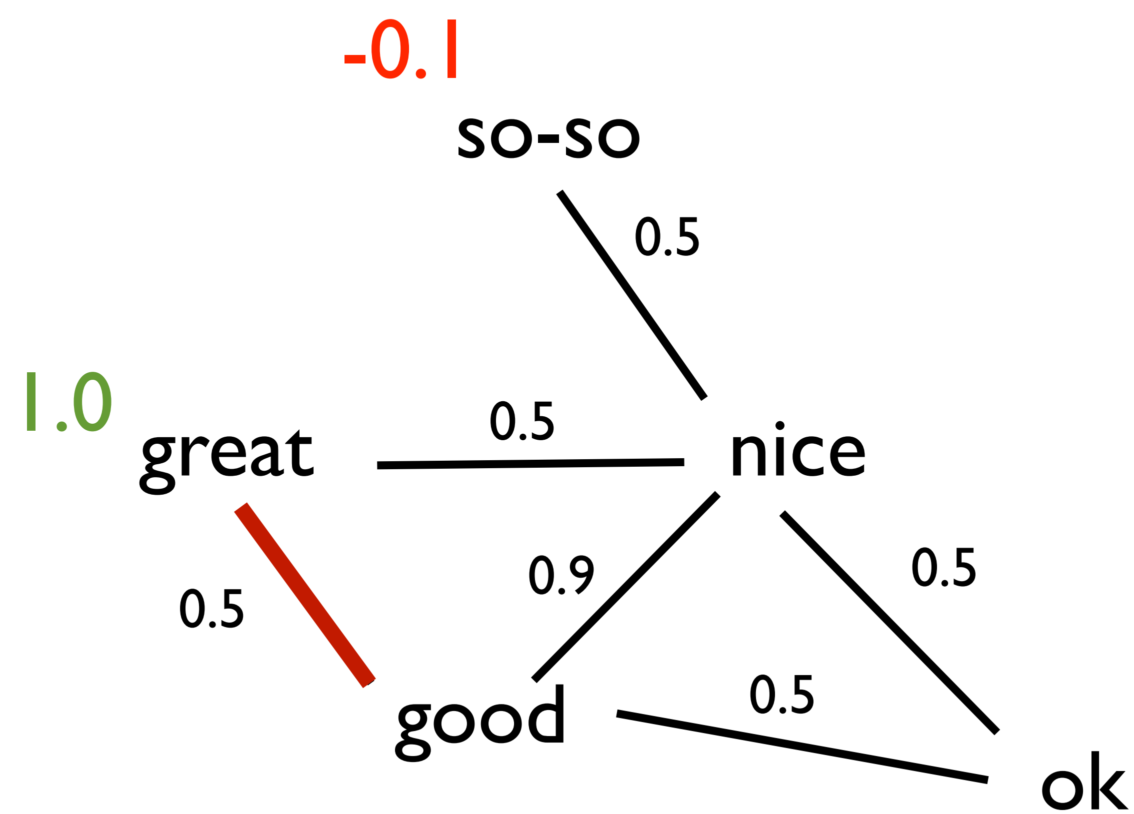


“Best Path to Seed” Propagation



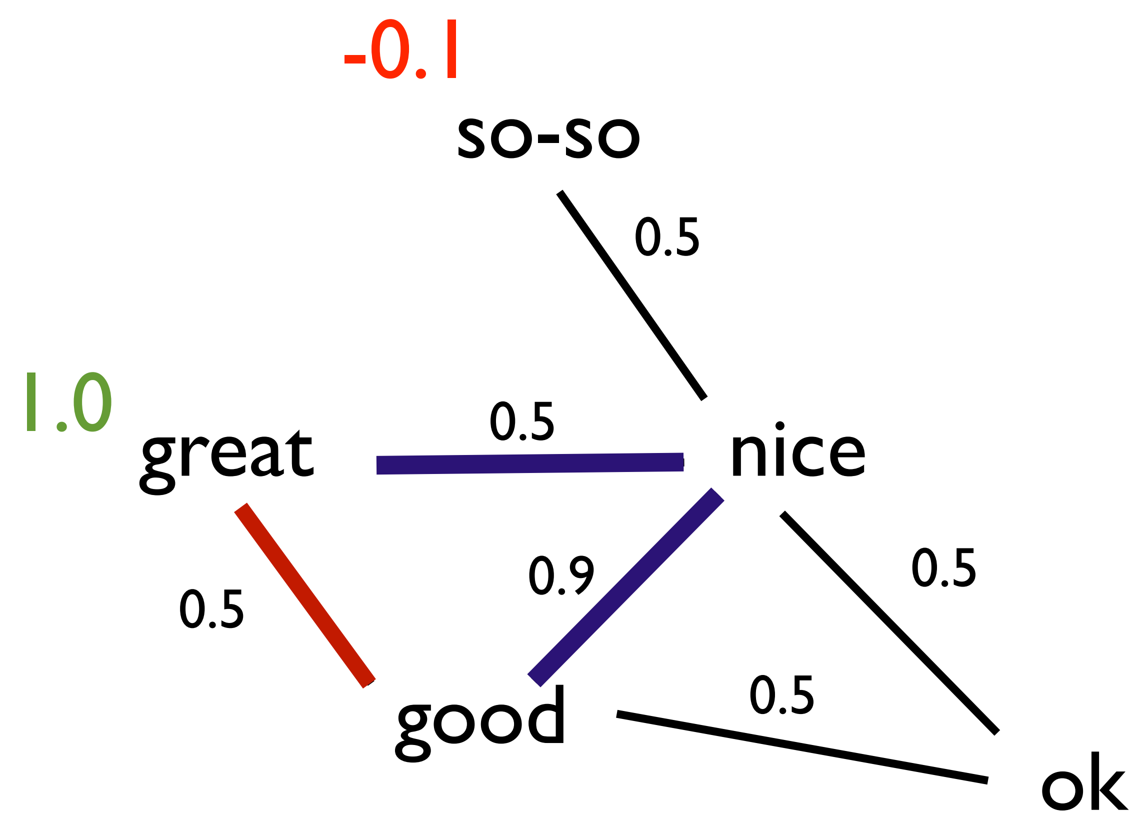
[Velikovich, et al., NAACL 2010]

“Best Path to Seed” Propagation



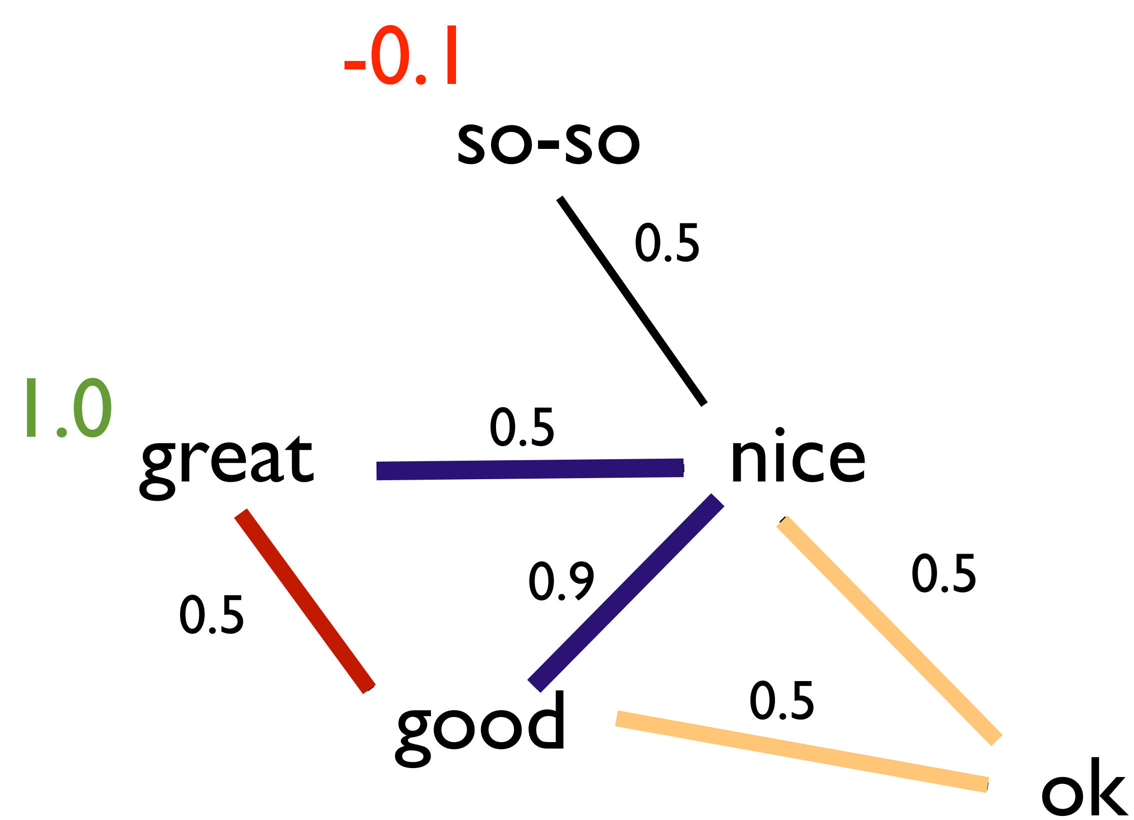
[Velikovich, et al., NAACL 2010]

“Best Path to Seed” Propagation



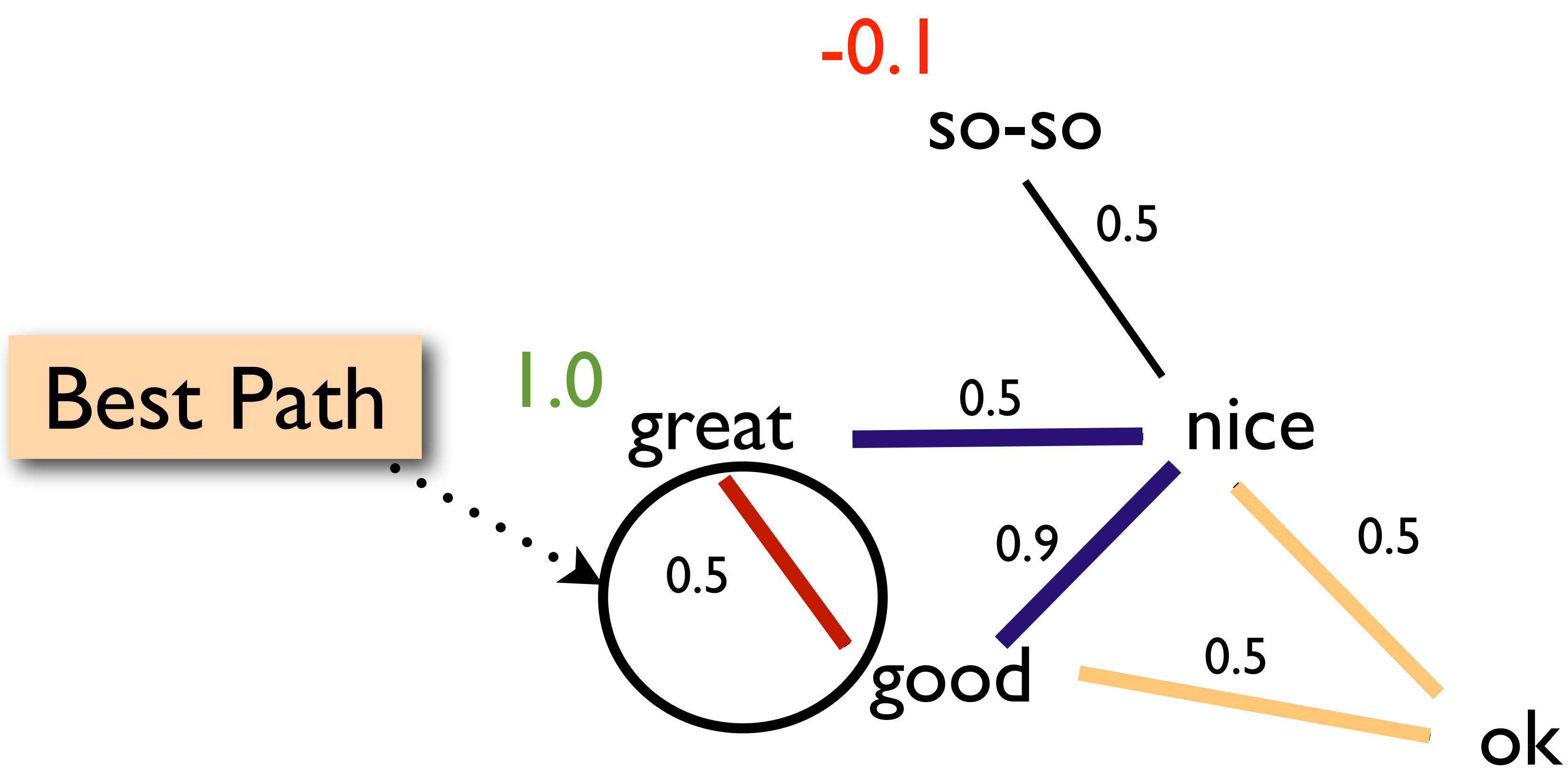
[Velikovich, et al., NAACL 2010]

“Best Path to Seed” Propagation



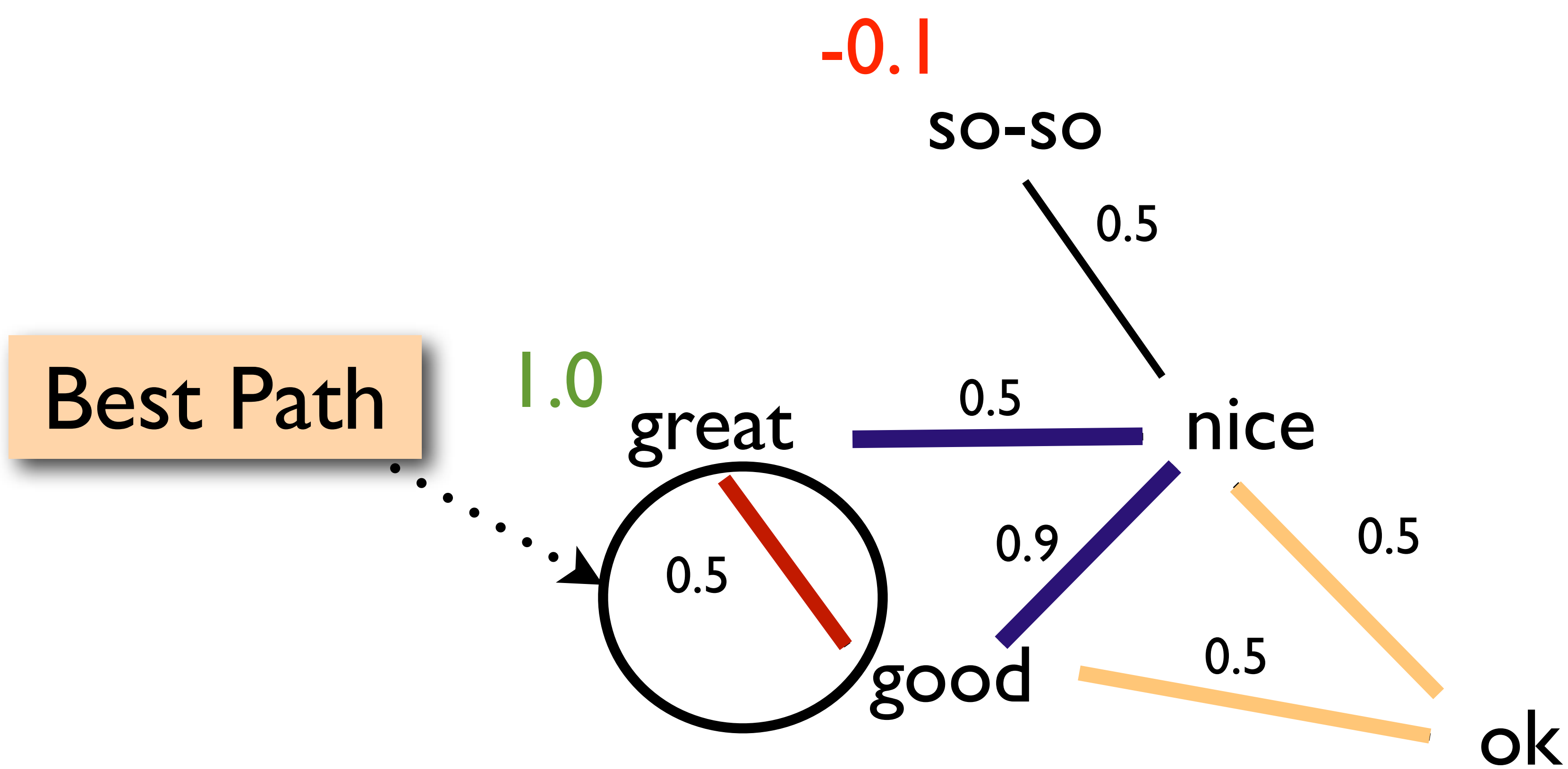
[Velikovich, et al., NAACL 2010]

"Best Path to Seed" Propagation



[Velikovich, et al., NAACL 2010]

“Best Path to Seed” Propagation



Key observation: **sentiment phrases** are those that have **short highly weighted paths** to **multiple seeds**

[Velikovich, et al., NAACL 2010]

Results

Lexicon	Phrases	Positive	Negative
Wilson et al. 2005	7,618	2,718	4,900
WordNet LP [Blair-Goldensohn et al. 07]	12,310	5,705	6,605
Web GP [Velikovich et al. 2010]	**178,104**	**90,337**	**87,767**

Size of the output lexicon

Results

Positive

What you'd expect
excellent
fabulous
beautiful
inspiring
awesome
plucky
ravishing
brilliant
nice
delightful
splendid
incredible
stupendous
comfortable

Spelling variations
loveable
nicee
niice
cooool
coooool
koool
kewl
cozy
cosy
sikk

Multi-word expressions
once in a life time
state - of - the - art
fail - safe operation
just what you need
just what the doctor ordered
out of this world
top of the line
melt in your mouth
snug as a bug
up to the job
out of the box
more good than bad

Negative

What you'd expect
bad
awful
terrible
dirty
repulsive
crappy
sucky
subpar
horrendous
miserable
lousy
abysmal
stupid
wretched

Vulgarity, ???
$#%! face
$#%!ed up
shut your $#%!ing mouth
complete bull$#%!
bladder spasms
green slime
vacuum of leadership
electro - static discharge
muttered under his breath
harm to the environment

Multi-word expressions
run of the mill
out of touch
over the hill
flash in the pan
bumps in the road
hit or miss
foaming at the mouth
dime a dozen
pie - in - the - sky
cast a pall over
sick to my stomach
pain in my ass

[Velikovich, et al., NAACL 2010]

Results

Ability to learn spelling variations and mistakes

Positive

Spelling variations
loveable
nicee
niice
cooool
coooool
koool
kewl
cozy
cosy
sikk

What you'd expect
excellent
fabulous
beautiful
inspiring
awesome
plucky
ravishing
brilliant
nice
delightful
splendid
incredible
stupendous
comfortable

Multi-word expressions
once in a life time
state - of - the - art
fail - safe operation
just what you need
just what the doctor ordered
out of this world
top of the line
melt in your mouth
snug as a bug
up to the job
out of the box
more good than bad

Negative

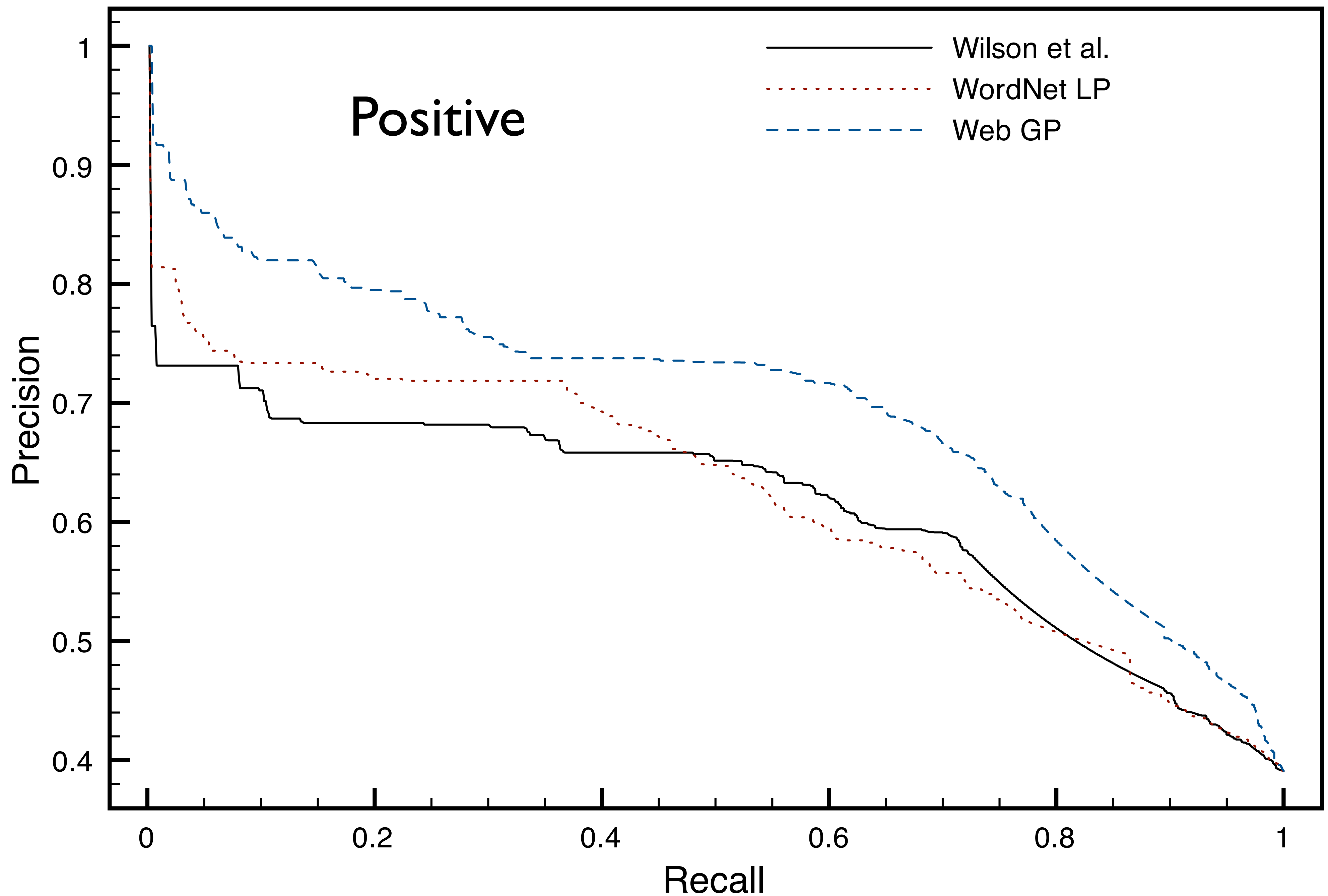
Vulgarity, ???
$#%! face
$#%!ed up
shut your $#%!ing mouth
complete bull$#%!
bladder spasms
green slime
vacuum of leadership
electro - static discharge
muttered under his breath
harm to the environment

What you'd expect
bad
awful
terrible
dirty
repulsive
crappy
sucky
subpar
horrendous
miserable
lousy
abysmal
stupid
wretched

Multi-word expressions
run of the mill
out of touch
over the hill
flash in the pan
bumps in the road
hit or miss
foaming at the mouth
dime a dozen
pie - in - the - sky
cast a pall over
sick to my stomach
pain in my ass

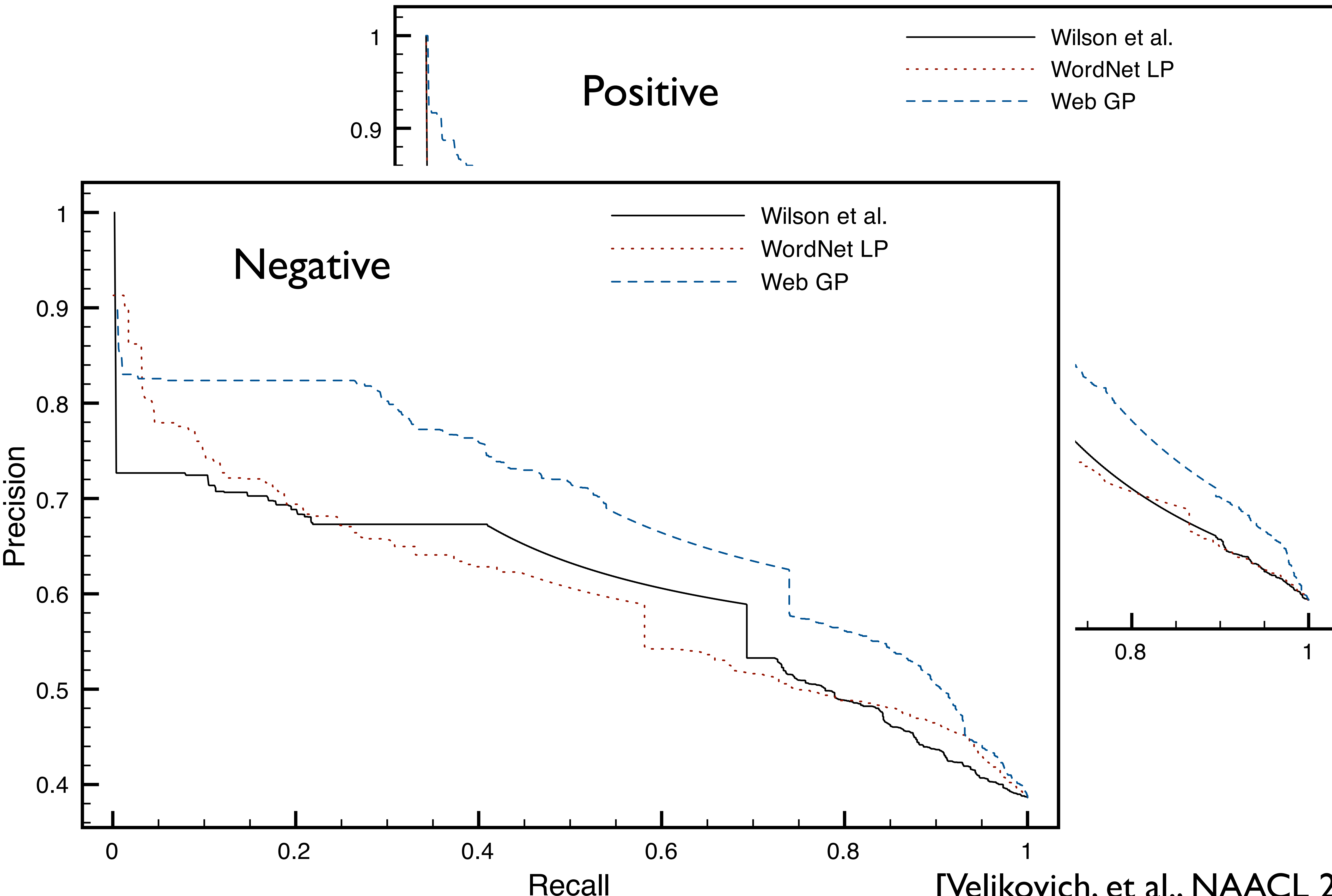
[Velikovich, et al., NAACL 2010]

Results



[Velikovich, et al., NAACL 2010]

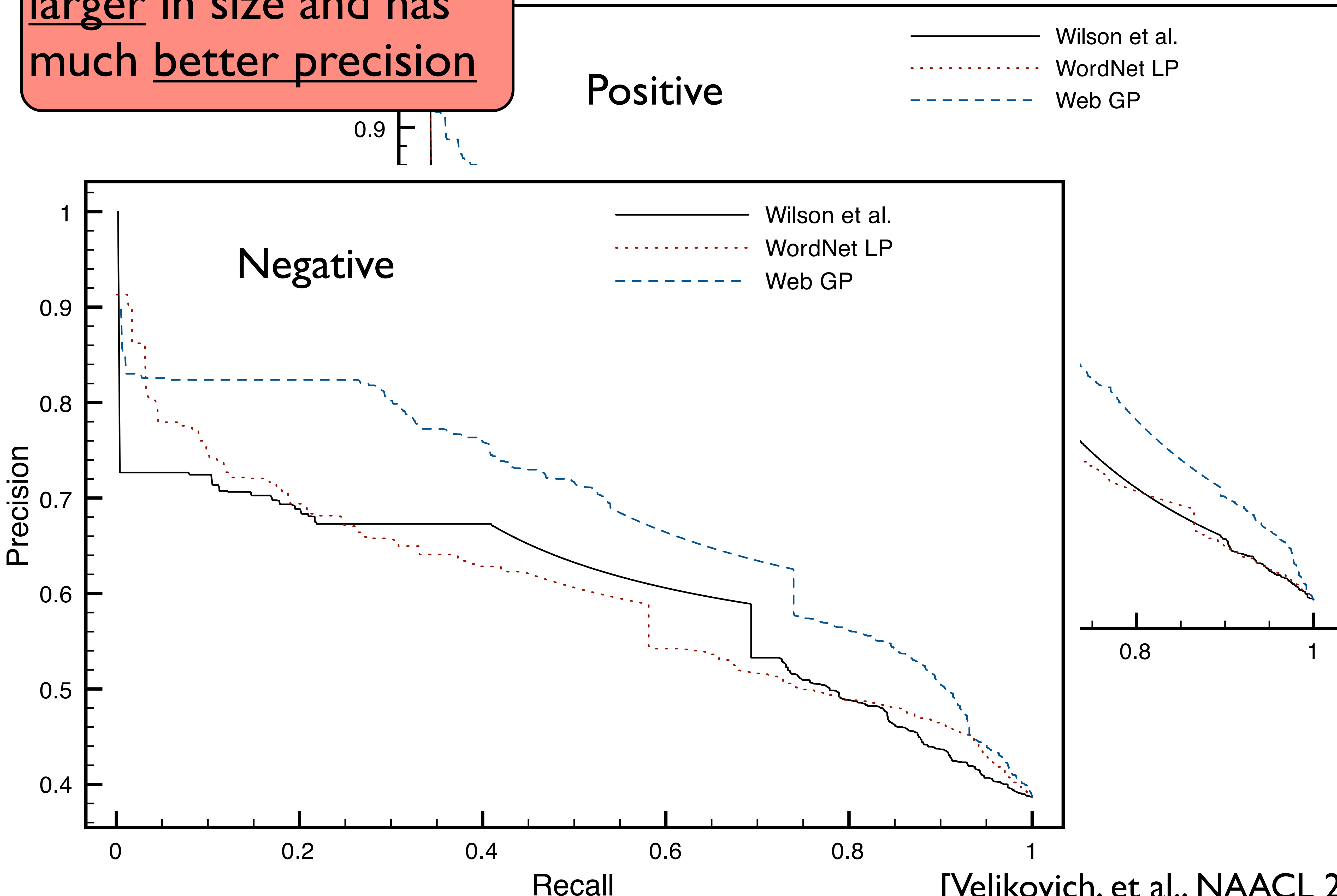
Results



[Velikovich, et al., NAACL 2010]

Results

Resulting lexicon is <u>larger</u> in size and has much <u>better precision</u>



[Velikovich, et al., NAACL 2010]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Text Categorization
  - Sentiment Analysis
  - Class Instance Acquisition [Talukdar et al., EMNLP 2008]
  - POS Tagging
  - MultiLingual POS Tagging
  - Semantic Parsing
- Conclusion & Future Work

Problem Description

- Given an entity, assign human readable descriptors to it
  - *Toyota* is a *car manufacturer*, *japanese company*, *multinational company*
  - *African countries* such as *Uganda* and *Angola*
- Large scale, open domain (> 100 classes)
- Applications
  - web search, advertising, etc.

Extraction Techniques

Extraction Techniques

....
What Other Musicians Would Fans of the Album Listen to:
Storytelling musicians come to mind. Musicians *such as* Johnny Cash, and Woodie Guthrie.

What is Distinctive About this Release?:
Every song on the album has its own unique sound. From the fast paced *That Texas Girl* to the acoustic ....

[van Durme and Pasca, AAAI 2008]

- Uses "<Class> *such as* <Instance>" patterns
- Extracts both class (musician) and instance (Johnny Cash)

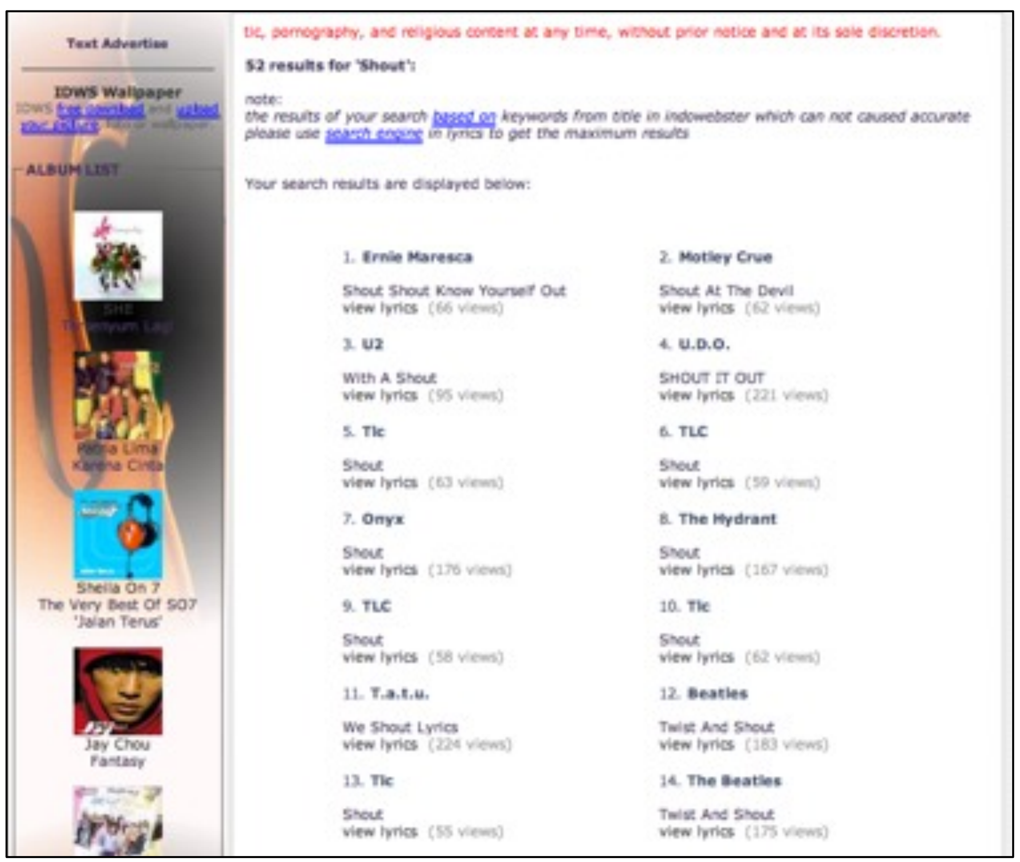
Extraction Techniques

....
What Other Musicians Would Fans of the Album Listen to:
Storytelling musicians come to mind. Musicians *such as* Johnny Cash, and Woodie Guthrie.

What is Distinctive About this Release?:
Every song on the album has its own unique sound. From the fast paced *That Texas Girl* to the acoustic ....

[van Durme and Pasca, AAAI 2008]

- Uses “<Class> *such as* <Instance>” patterns
- Extracts both class (musician) and instance (Johnny Cash)



Extractions from HTML lists and tables

- [Wang and Cohen, ICDM 2007]
- WebTables [Cafarella et al., VLDB 2008], 154 million HTML tables

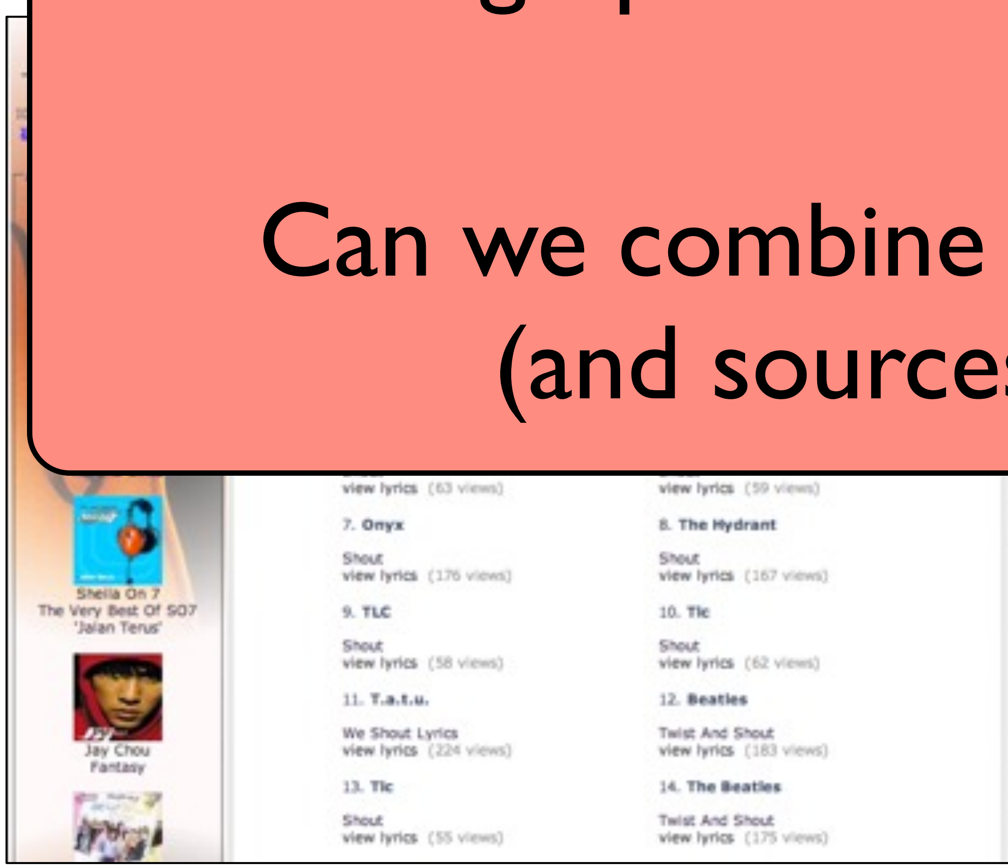
Extraction Techniques

....
What Other Musicians Would Fans of the Album Listen to:
Storytelling musicians come to mind. Musicians *such as* Johnny Cash, and Woodie Guthrie.

What is Distinctive About this Release?:

[van Durme and Pasca, AAAI 2008]

- Uses "<Class> *such as* <Instance>" patterns



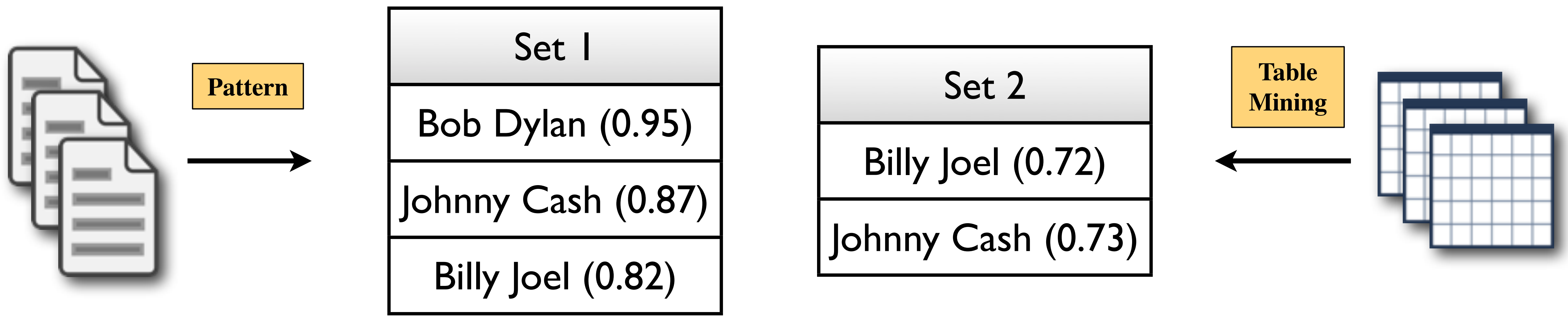
- WebTables [Cafarella et al., VLDB 2008], 154 million HTML tables

Pattern-based methods are usually tuned for high-precision, resulting in low coverage

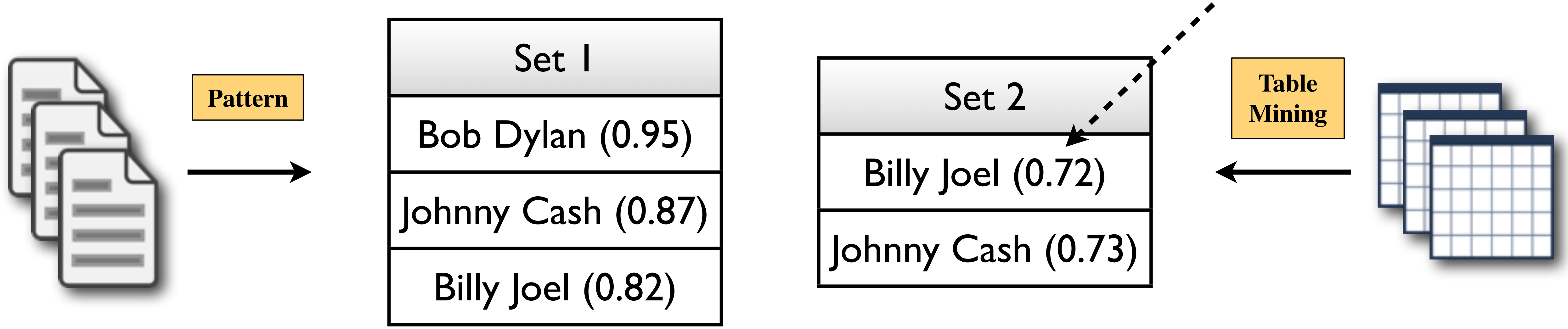
Can we combine extractions from all methods (and sources) to improve coverage?

Graph Construction

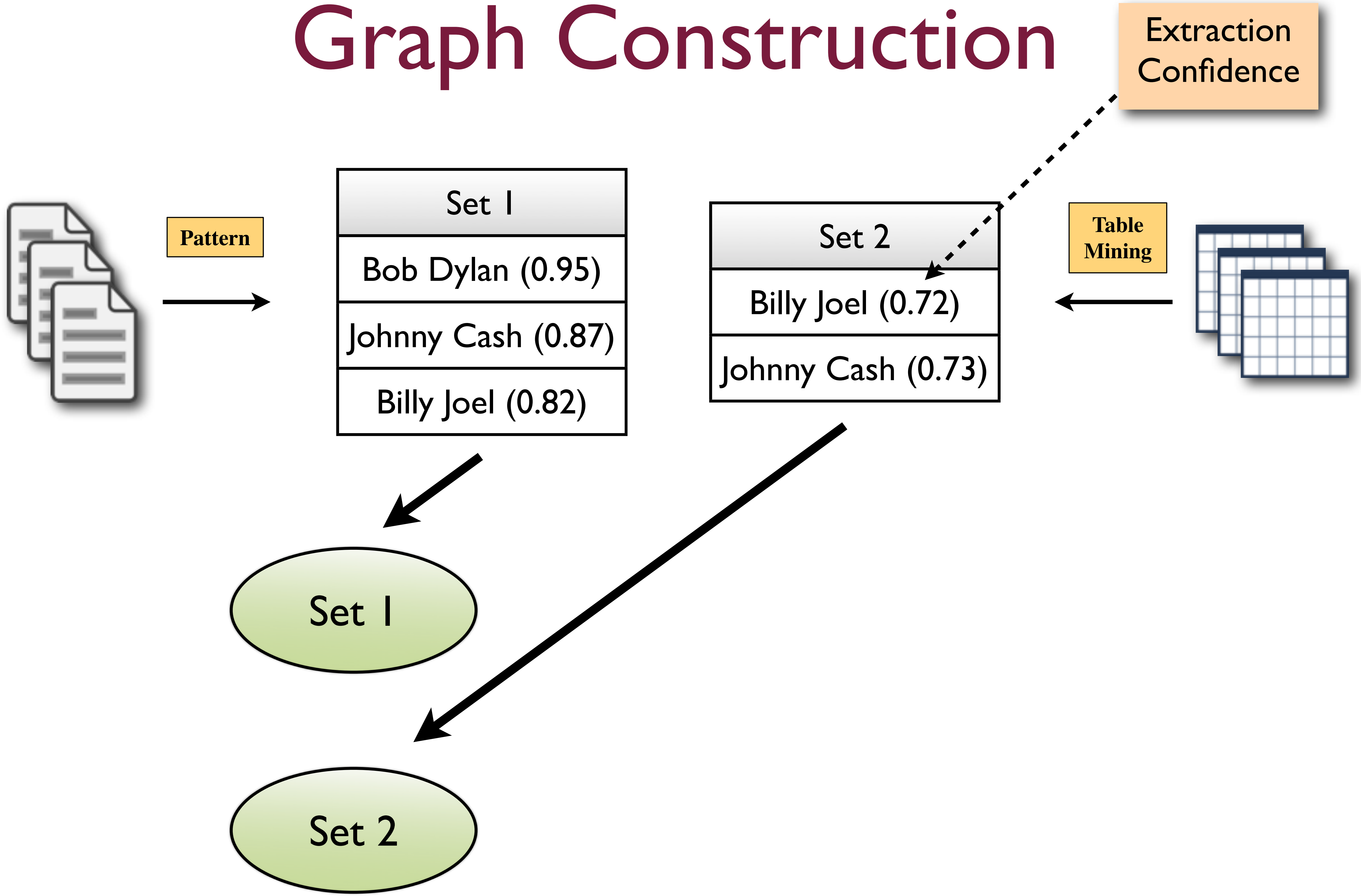
Graph Construction



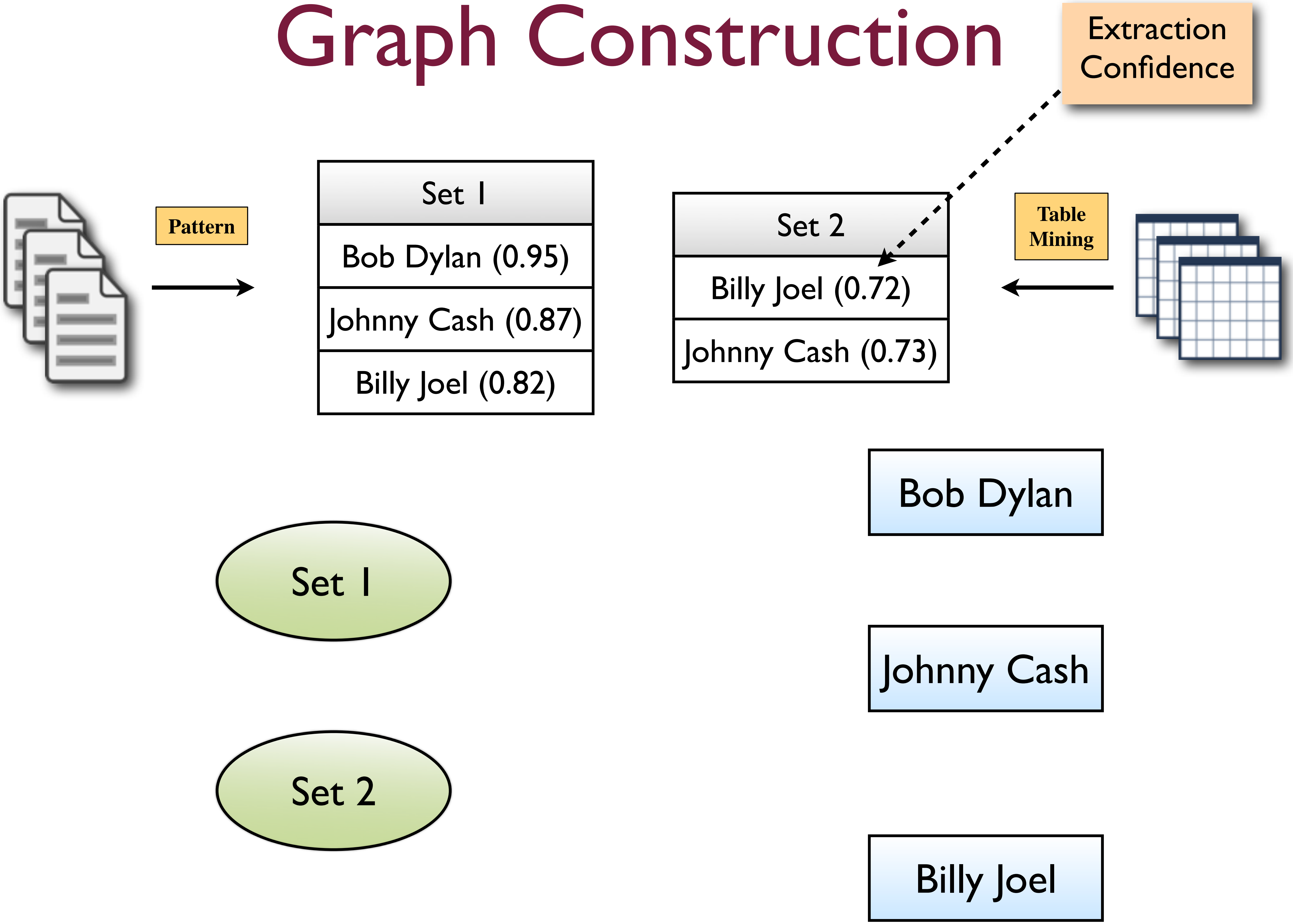
Graph Construction



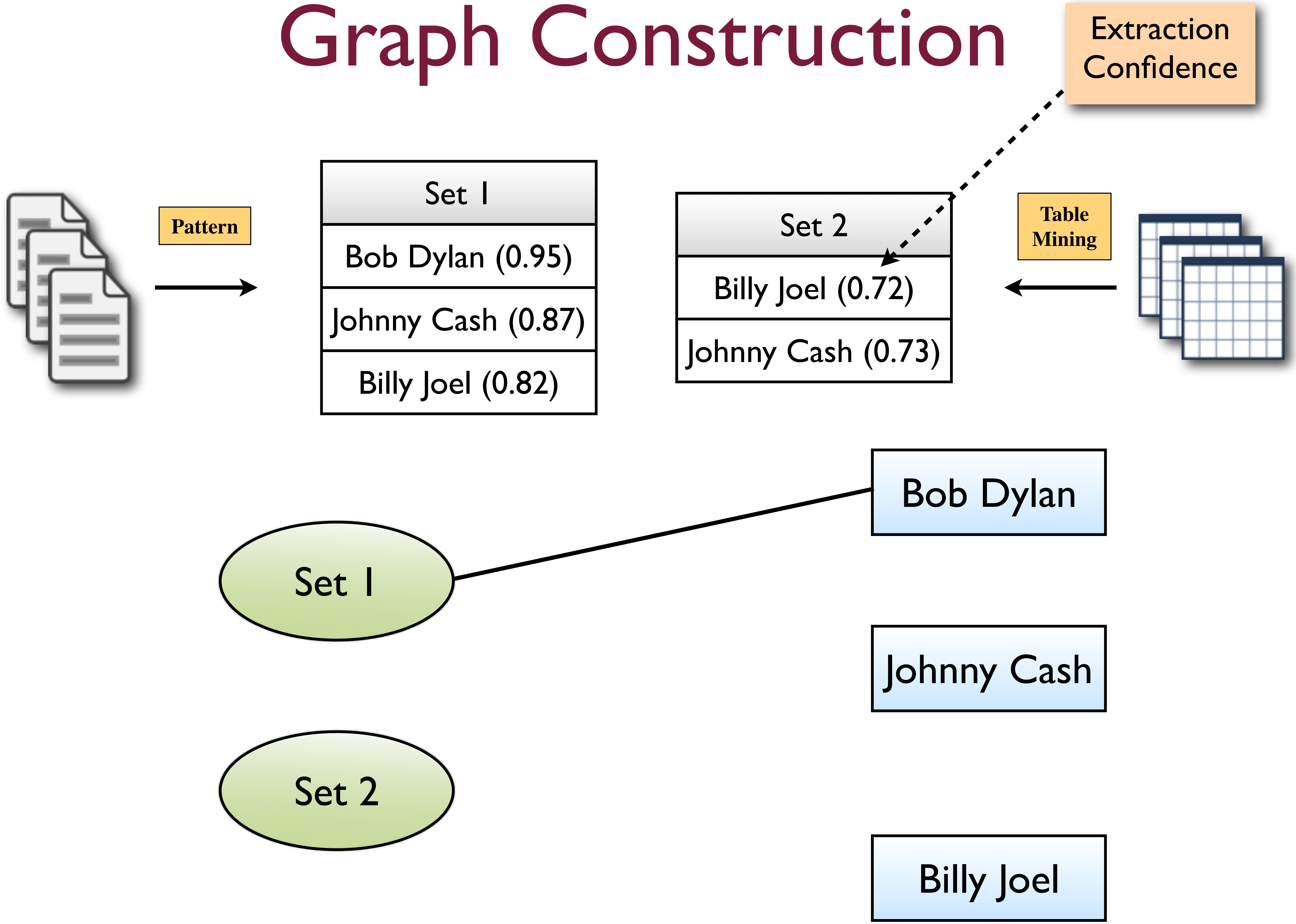
Graph Construction



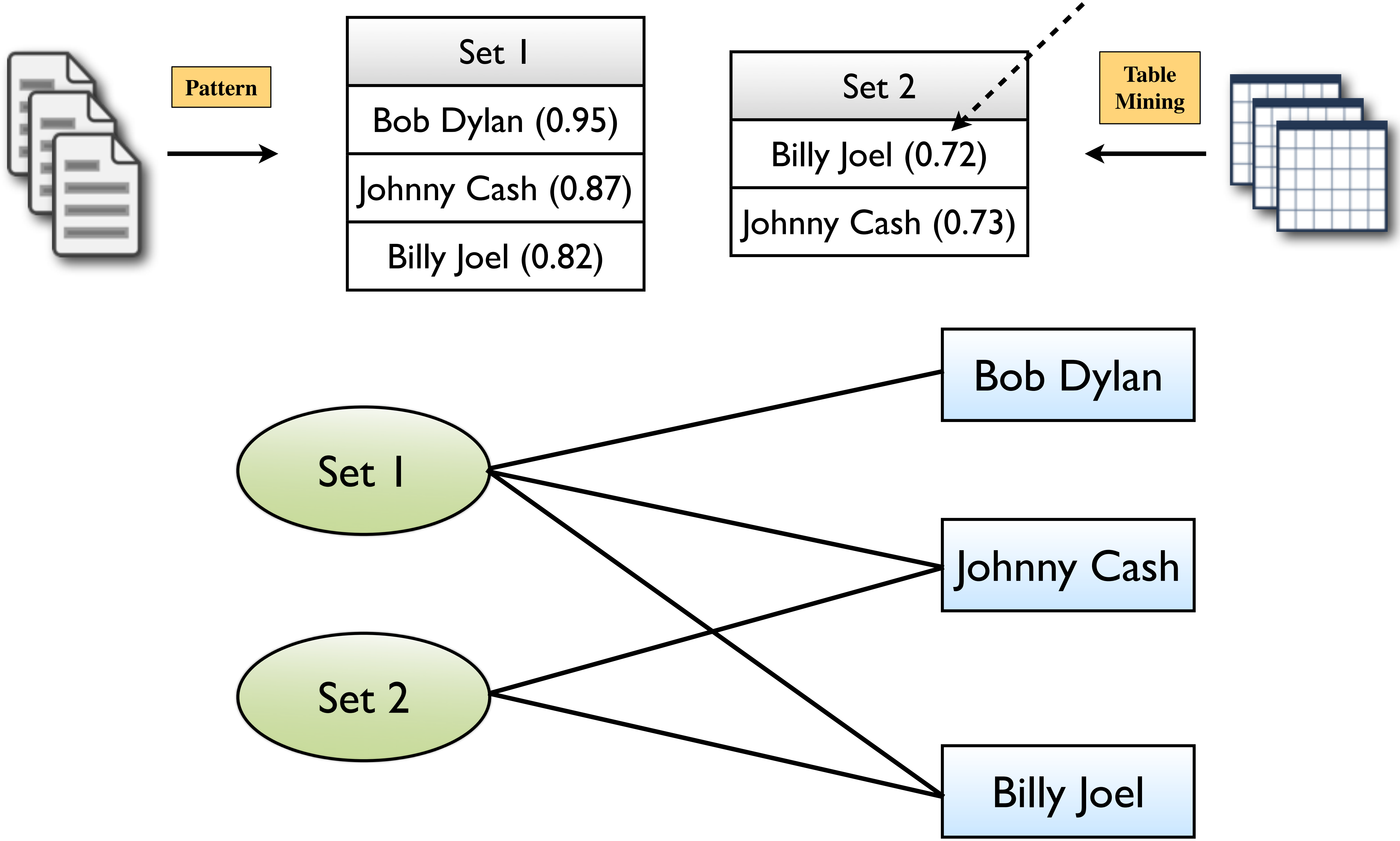
Graph Construction



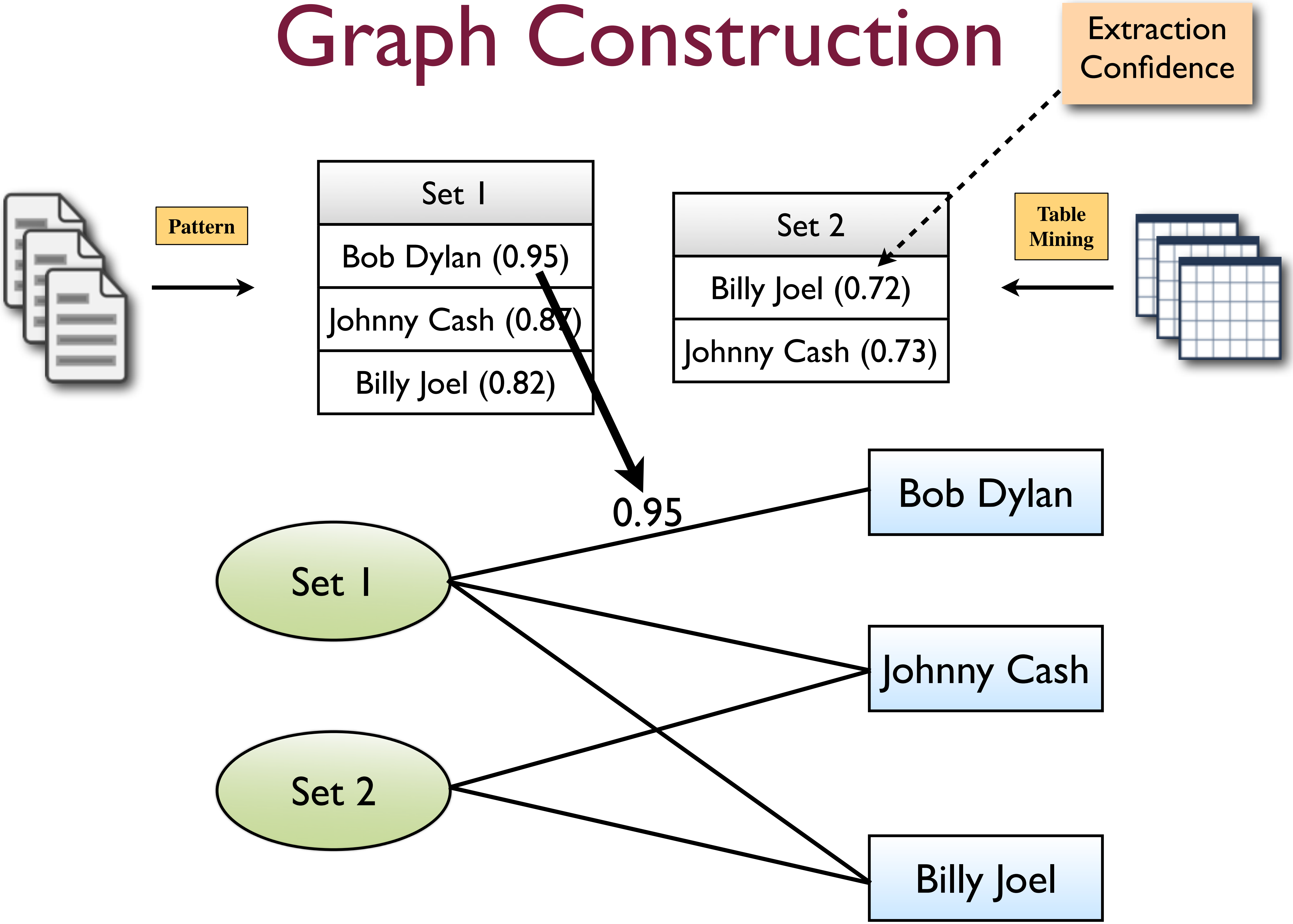
Graph Construction



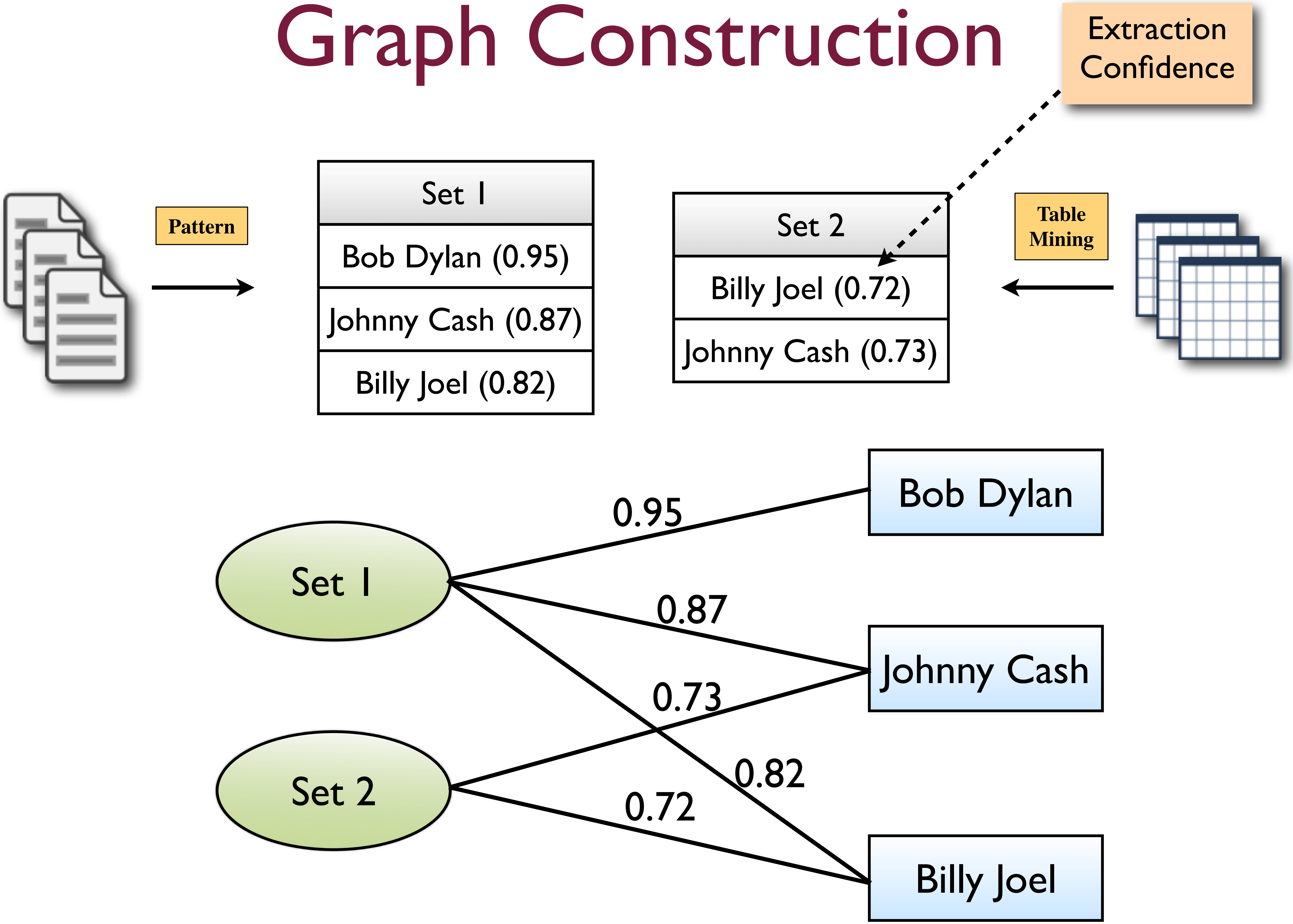
Graph Construction



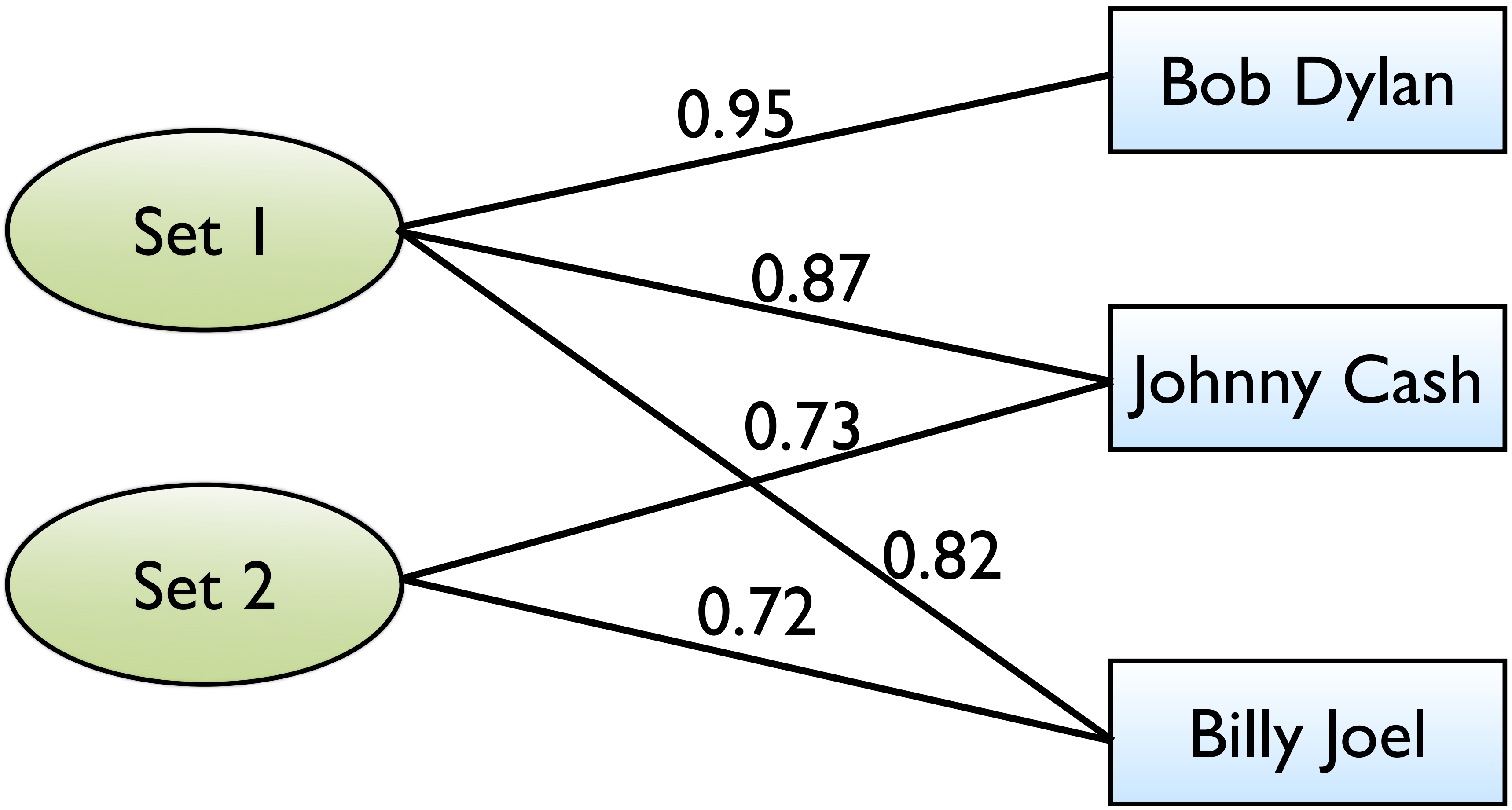
Graph Construction



Graph Construction



Graph Construction



- Bi-partite graph (not a k-NN graph)
- "Set" nodes encourage members of the set to have similar labels
- Natural way to represent extractions from many sources and methods

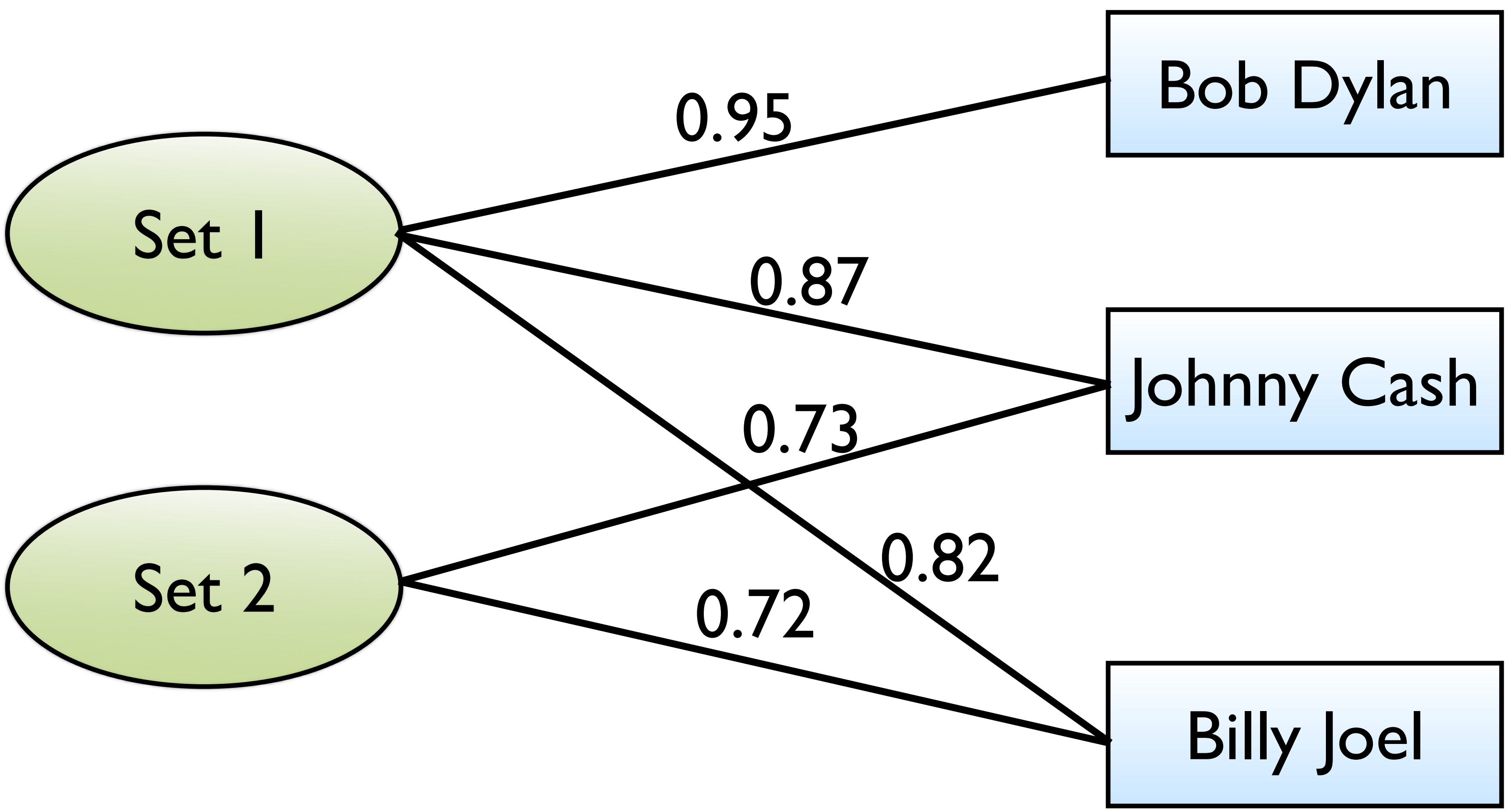
Goal

Pattern

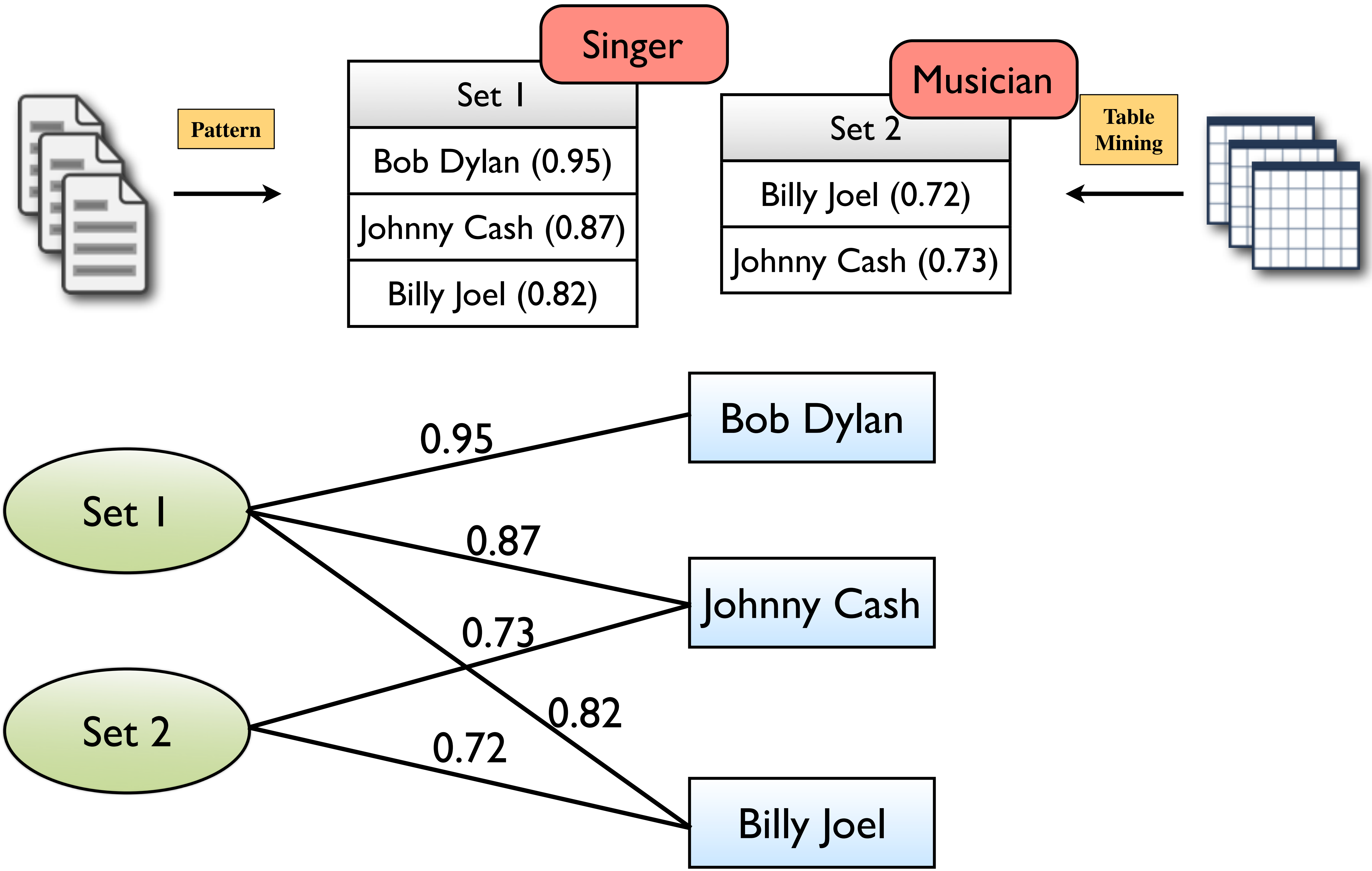
Set 1
Bob Dylan (0.95)
Johnny Cash (0.87)
Billy Joel (0.82)

Set 2
Billy Joel (0.72)
Johnny Cash (0.73)

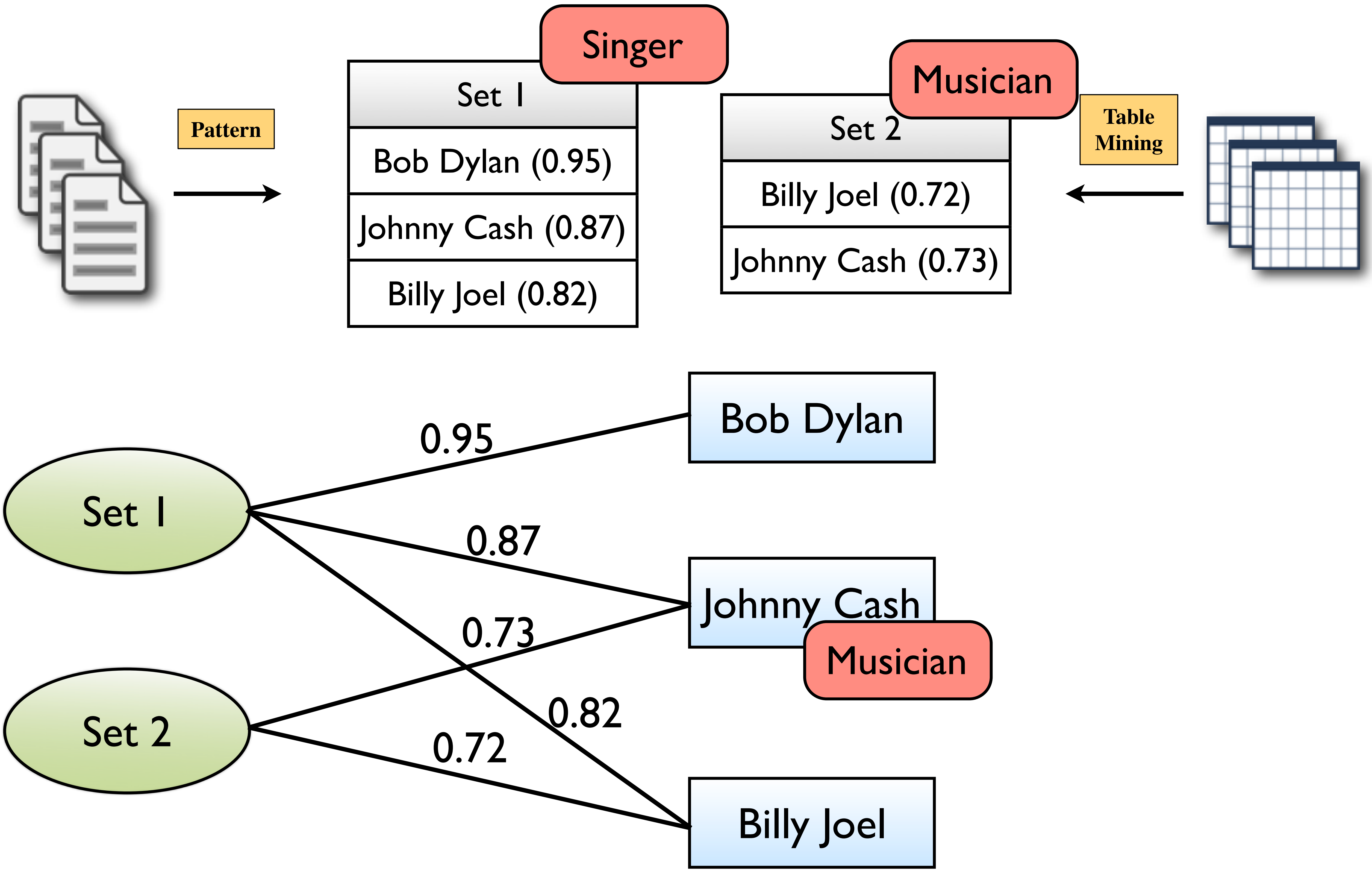
Table Mining



Goal

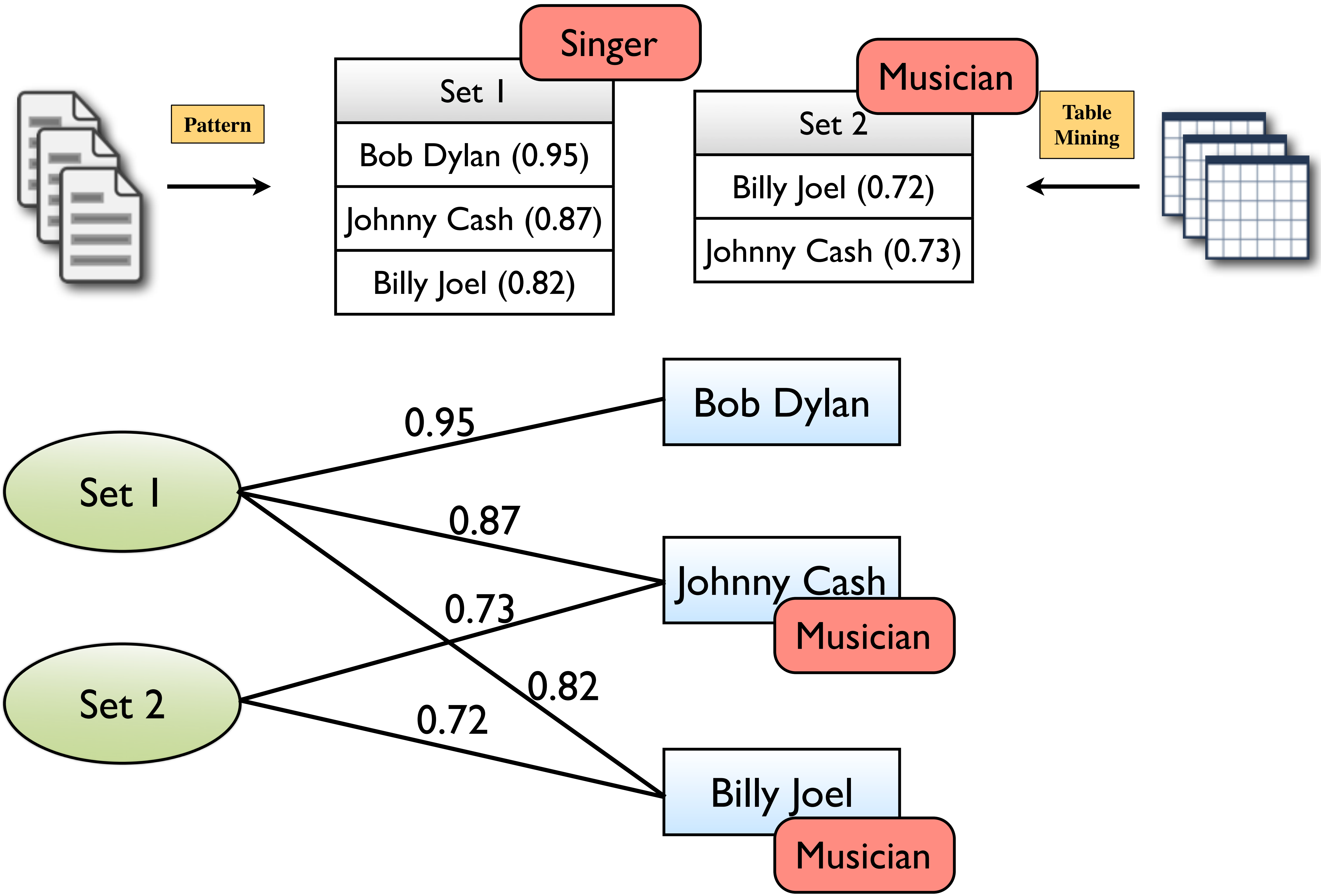

Singer
Musician
Pattern
Table Mining
Set 1
Bob Dylan (0.95)
Johnny Cash (0.87)
Billy Joel (0.82)
Set 2
Billy Joel (0.72)
Johnny Cash (0.73)
Set 1
Set 2
Bob Dylan
Johnny Cash
Billy Joel
0.95
0.87
0.73
0.82
0.72

Goal



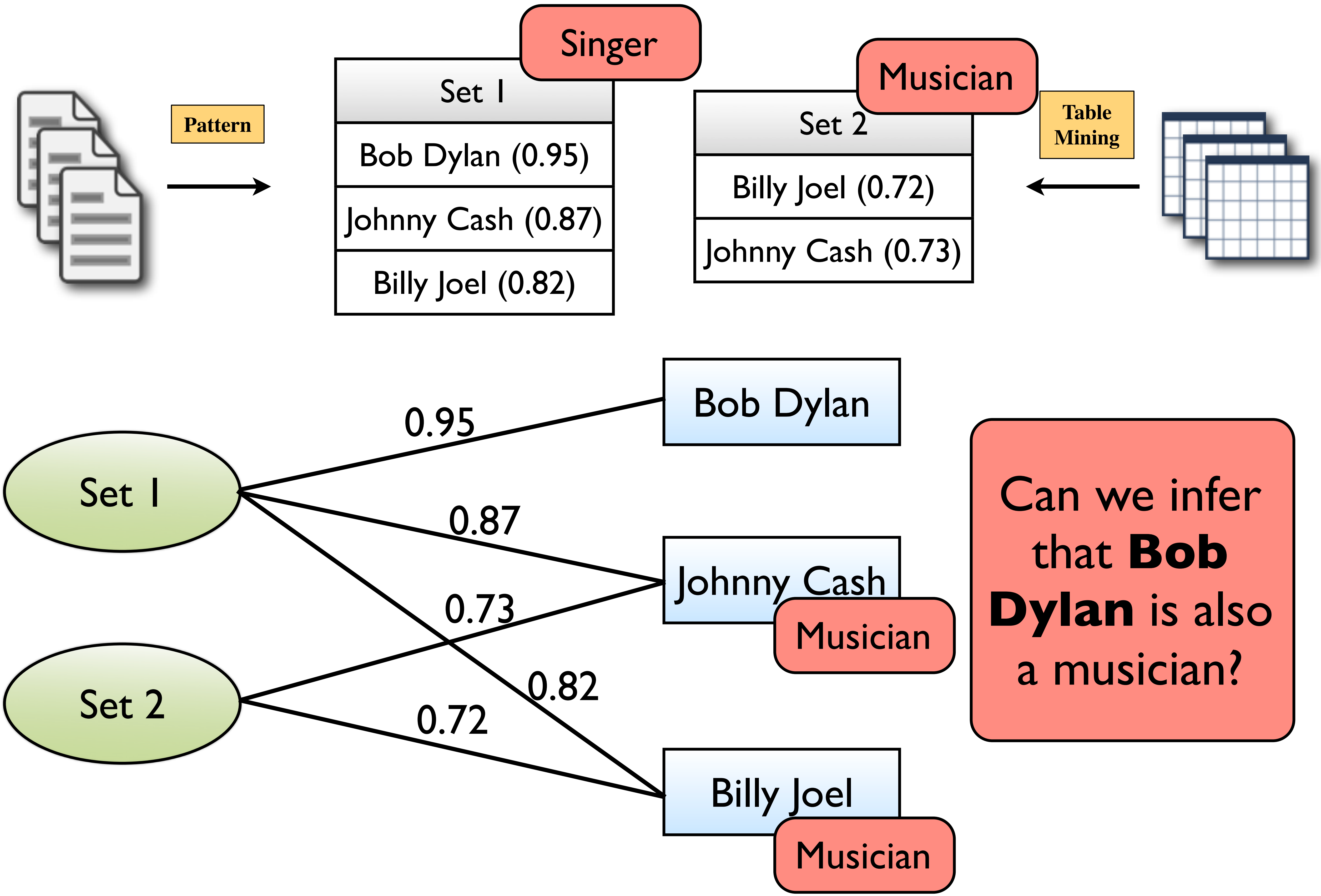
Singer
Set 1
Bob Dylan (0.95)
Johnny Cash (0.87)
Billy Joel (0.82)
Pattern
Musician
Set 2
Billy Joel (0.72)
Johnny Cash (0.73)
Table Mining
0.95
0.87
0.73
0.82
0.72
Set 1
Set 2
Bob Dylan
Johnny Cash
Musician
Billy Joel

Goal

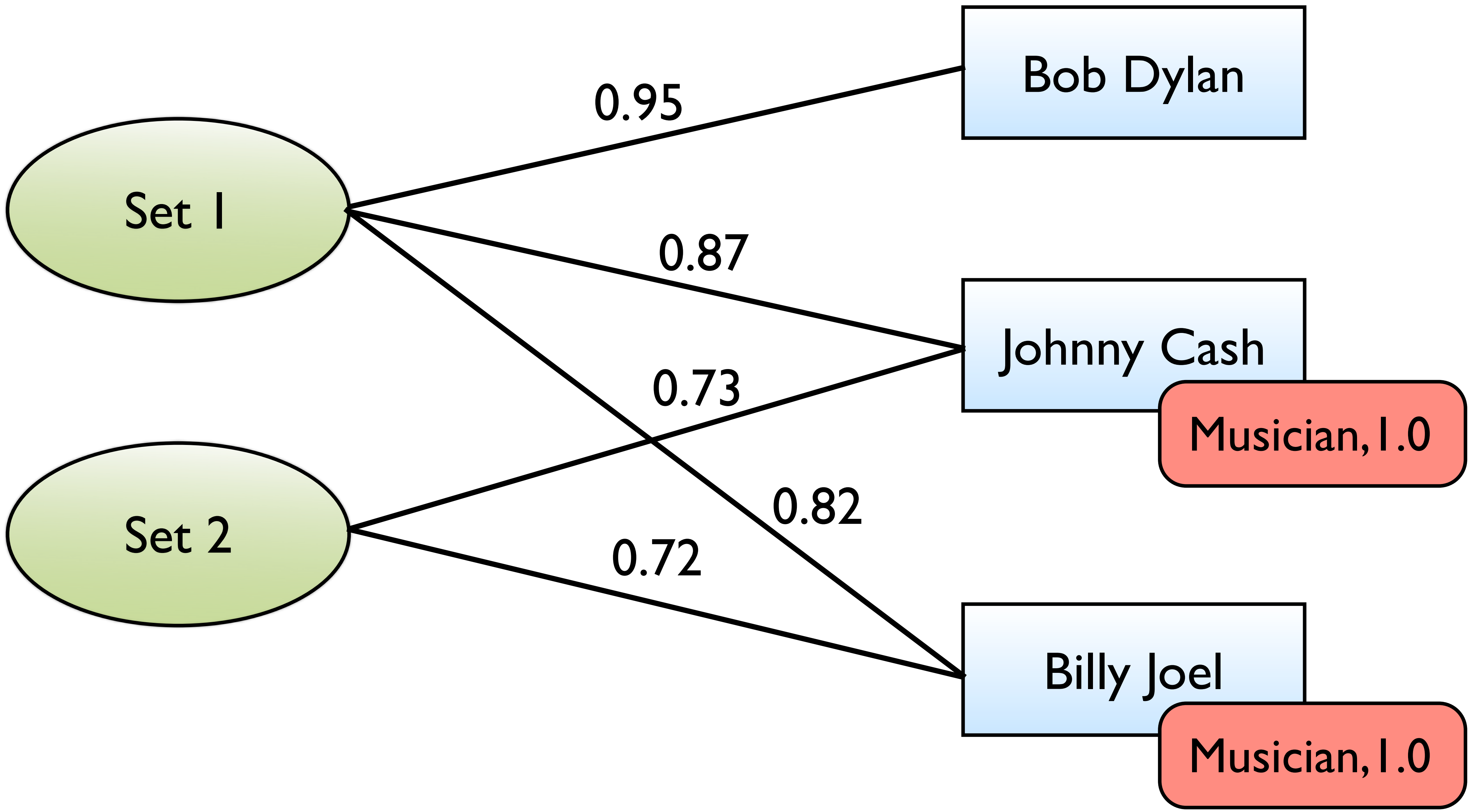


Pattern
Singer
Set 1
Bob Dylan (0.95)
Johnny Cash (0.87)
Billy Joel (0.82)
Musician
Set 2
Billy Joel (0.72)
Johnny Cash (0.73)
Table Mining
Set 1
Set 2
0.95
0.87
0.73
0.82
0.72
Bob Dylan
Johnny Cash
Musician
Billy Joel
Musician

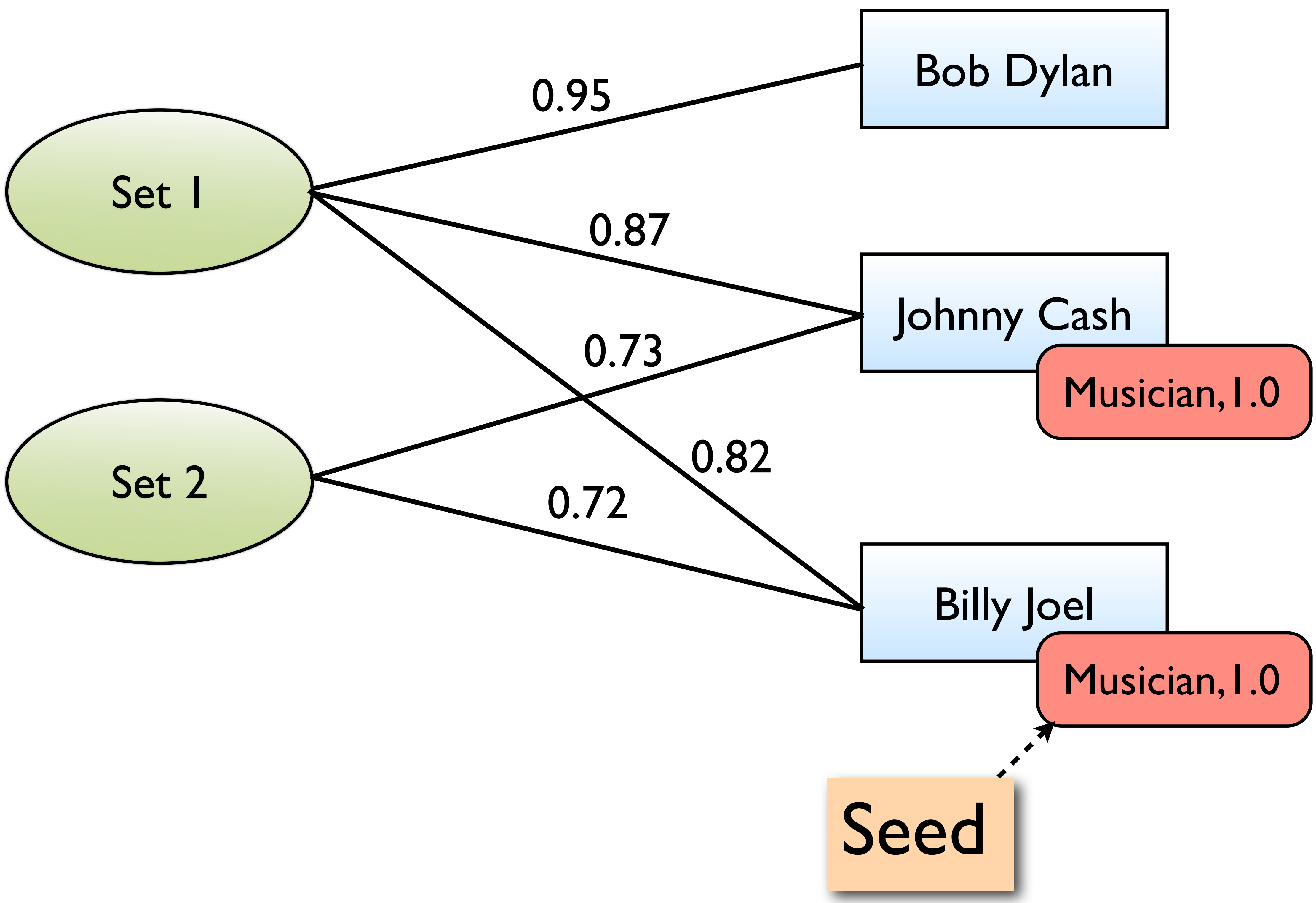
Goal


Singer
Set 1
Bob Dylan (0.95)
Johnny Cash (0.87)
Billy Joel (0.82)
Pattern
Musician
Set 2
Billy Joel (0.72)
Johnny Cash (0.73)
Table Mining
Set 1
Set 2
0.95
0.87
0.73
0.82
0.72
Bob Dylan
Johnny Cash
Musician
Billy Joel
Musician
Can we infer that **Bob Dylan** is also a musician?

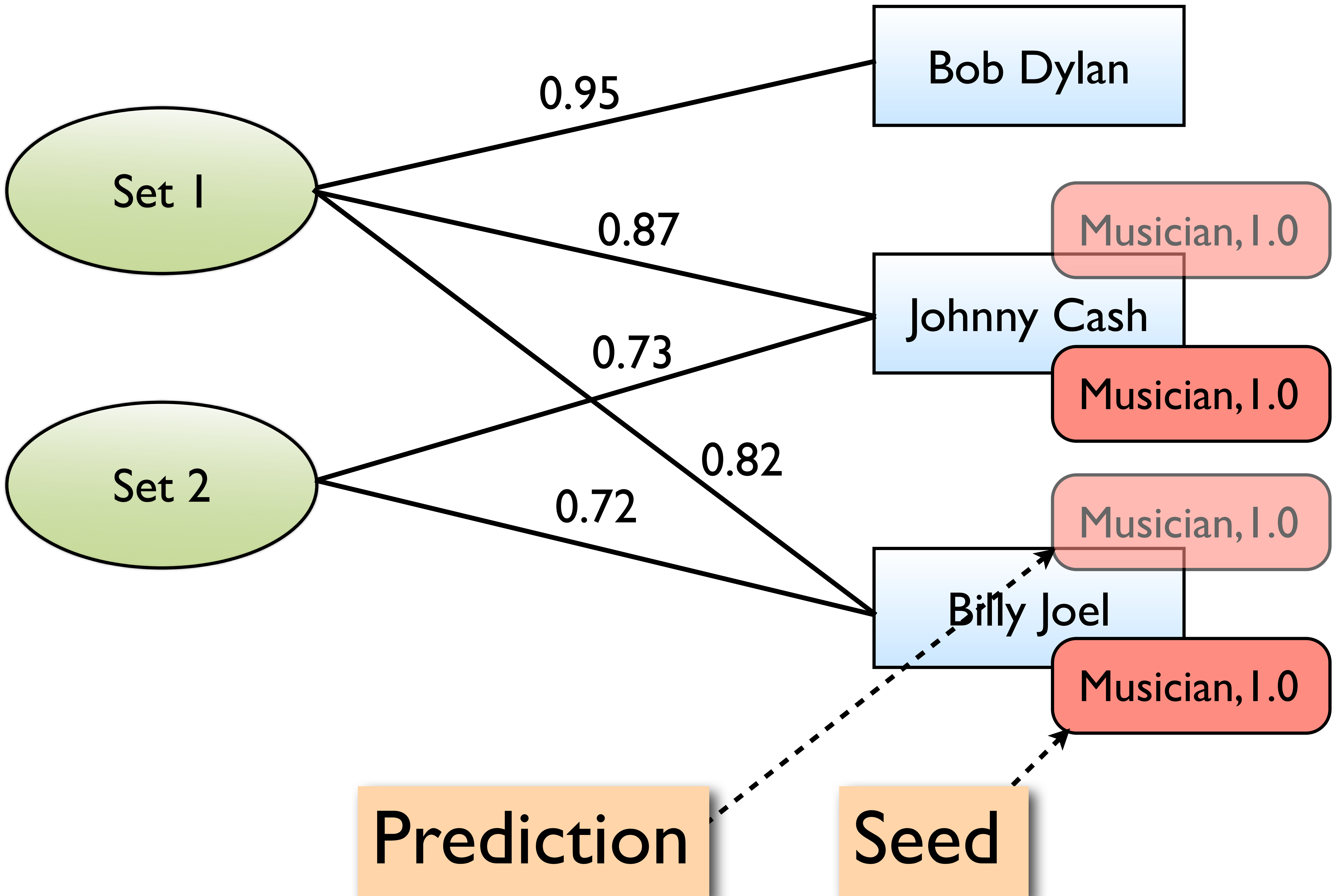
Graph Propagation



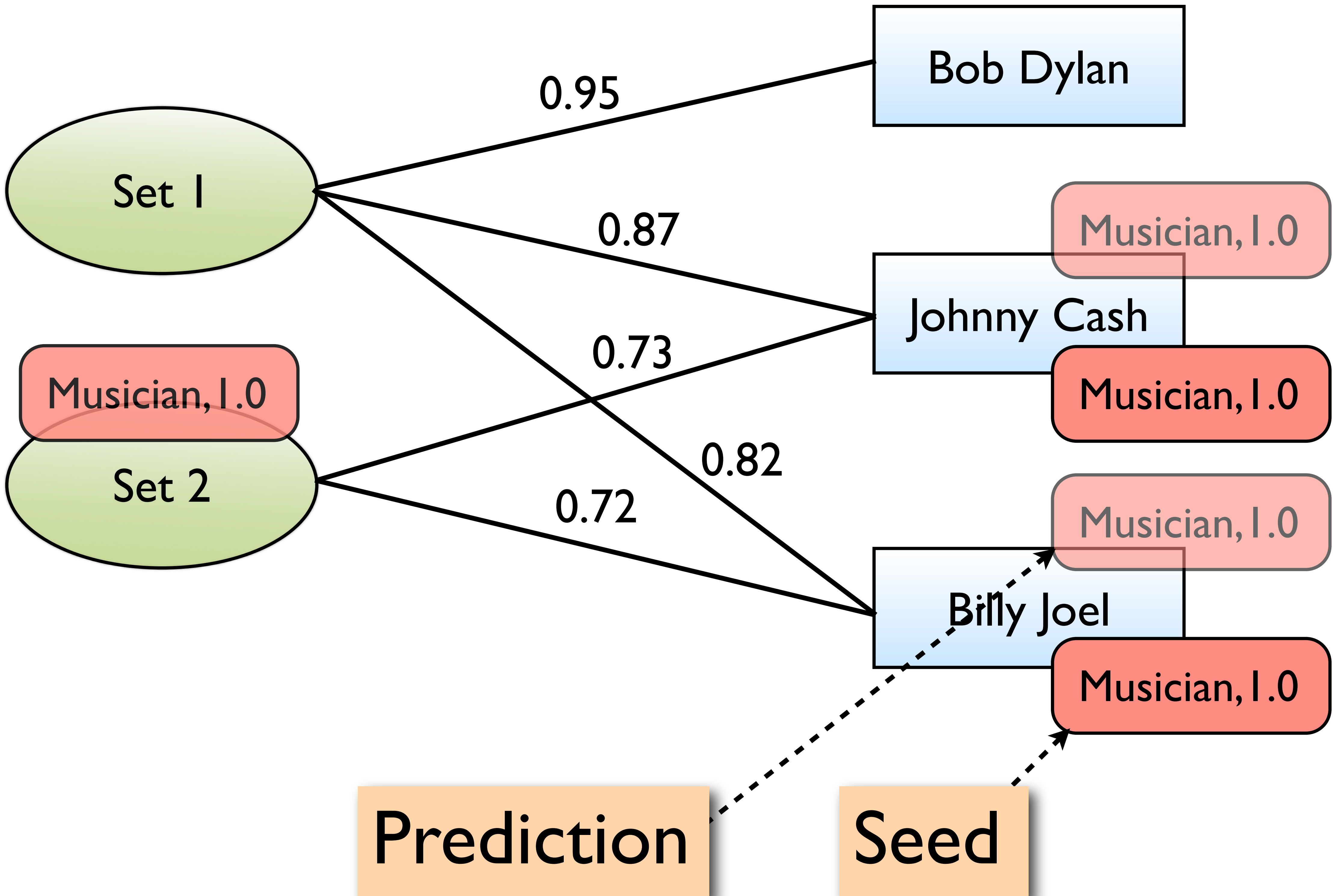
Graph Propagation



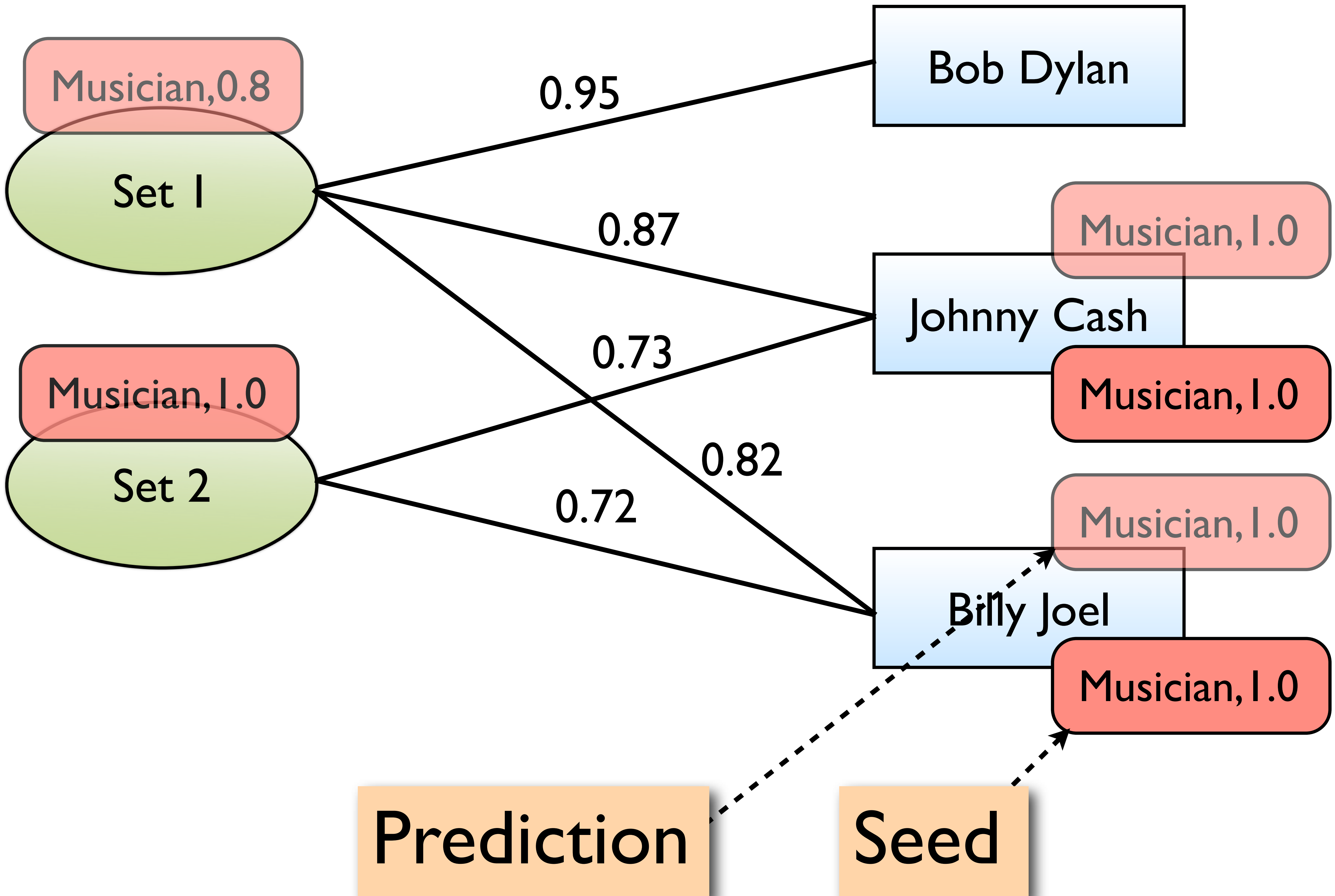
Graph Propagation



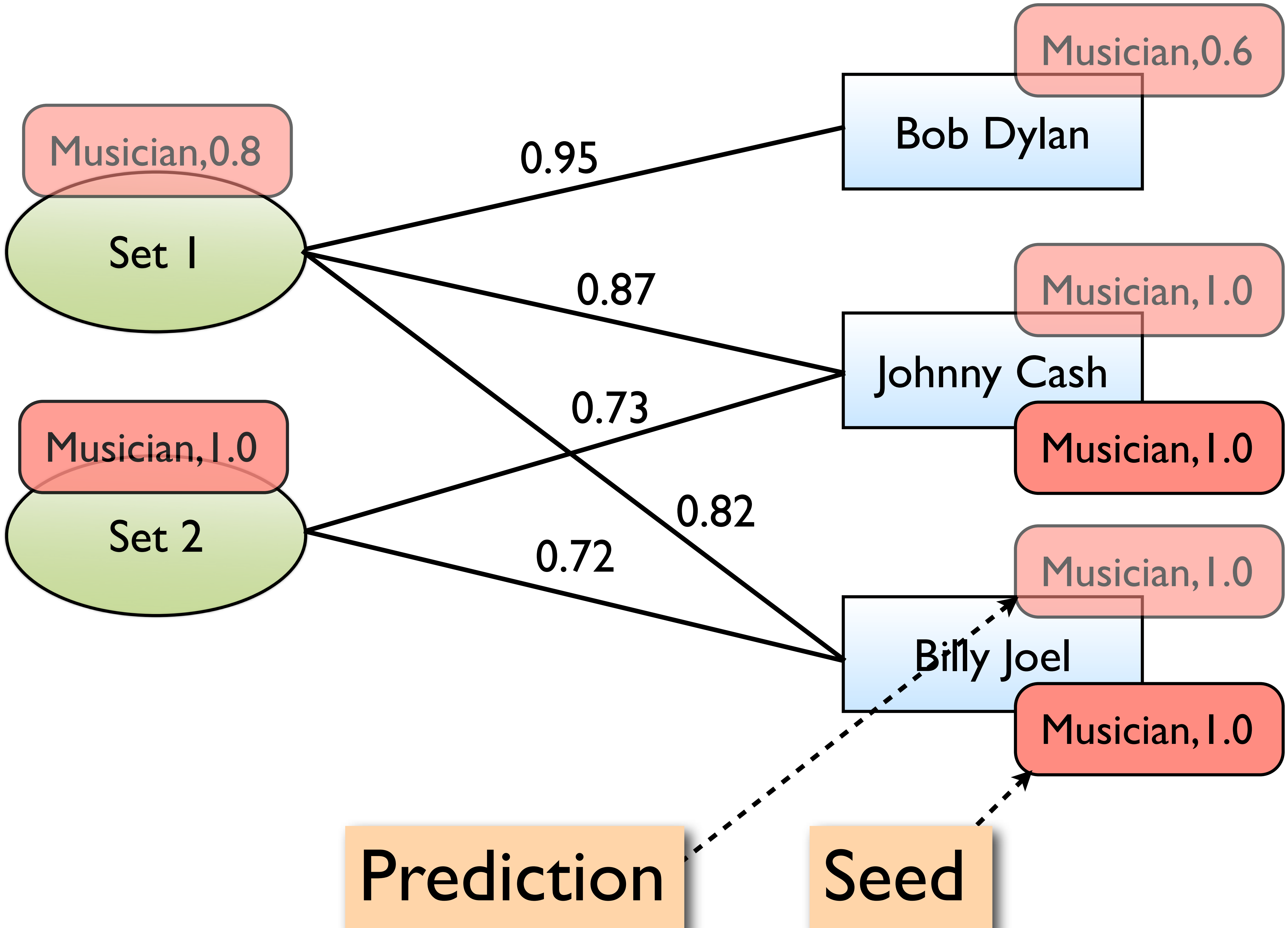
Graph Propagation



Graph Propagation



Graph Propagation



Evaluation Metric

Mean Reciprocal Rank

$$\mathrm{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\mathrm{rank}_v(\mathrm{class}(v))}$$

Evaluation Metric

Mean Reciprocal Rank

$$\mathrm{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\mathrm{rank}_v(\mathrm{class}(v))}$$

Linguist, 0.6

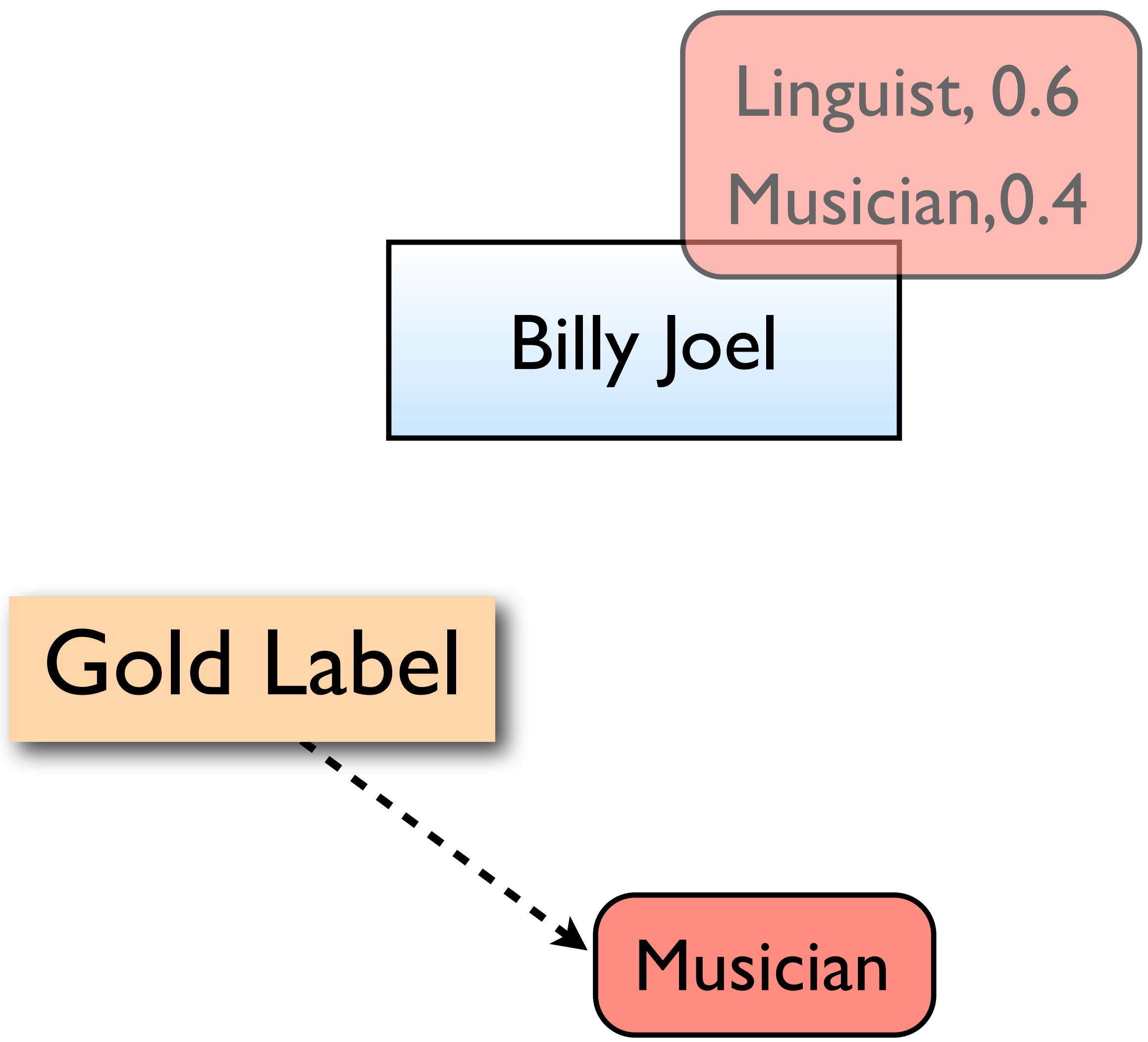
Musician,0.4

Billy Joel

Evaluation Metric

Mean Reciprocal Rank

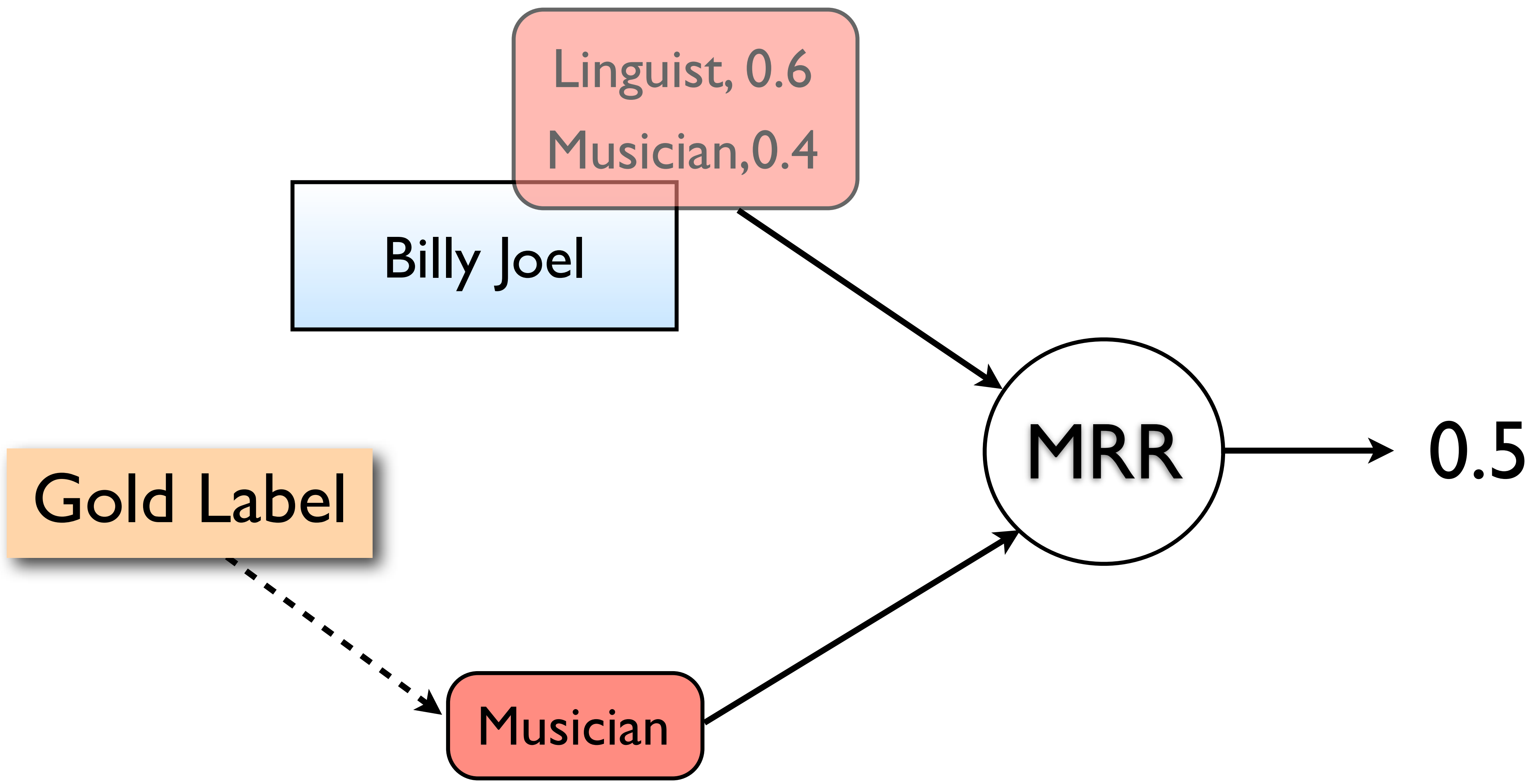
$$\mathrm{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\mathrm{rank}_v(\mathrm{class}(v))}$$



Evaluation Metric

Mean Reciprocal Rank

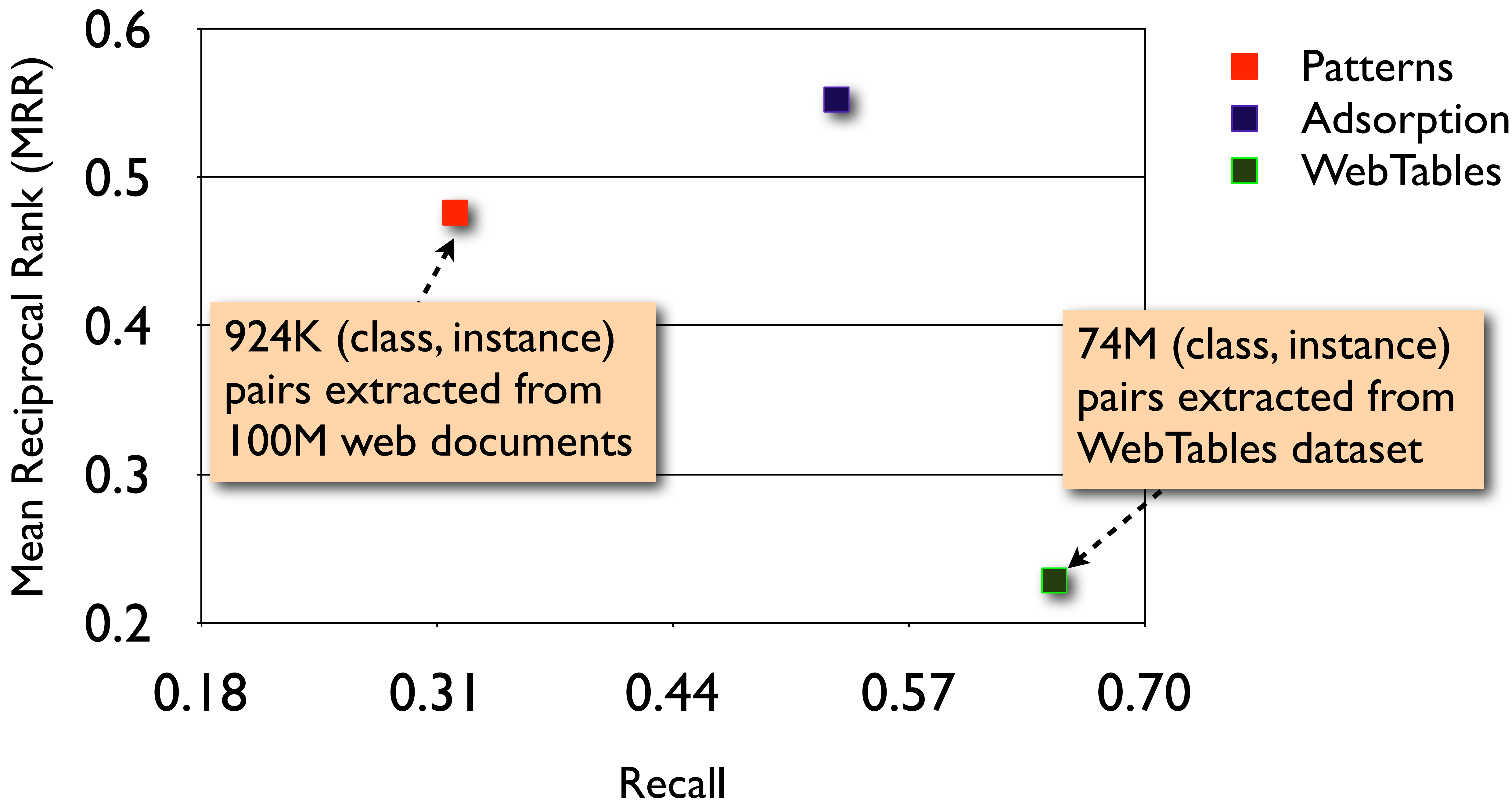
$$\mathrm{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\mathrm{rank}_v(\mathrm{class}(v))}$$



Extraction for Known Instances

Graph with 1.4m nodes, 75m edges used.

Evaluation against WordNet Dataset (38 classes, 8910 instances)

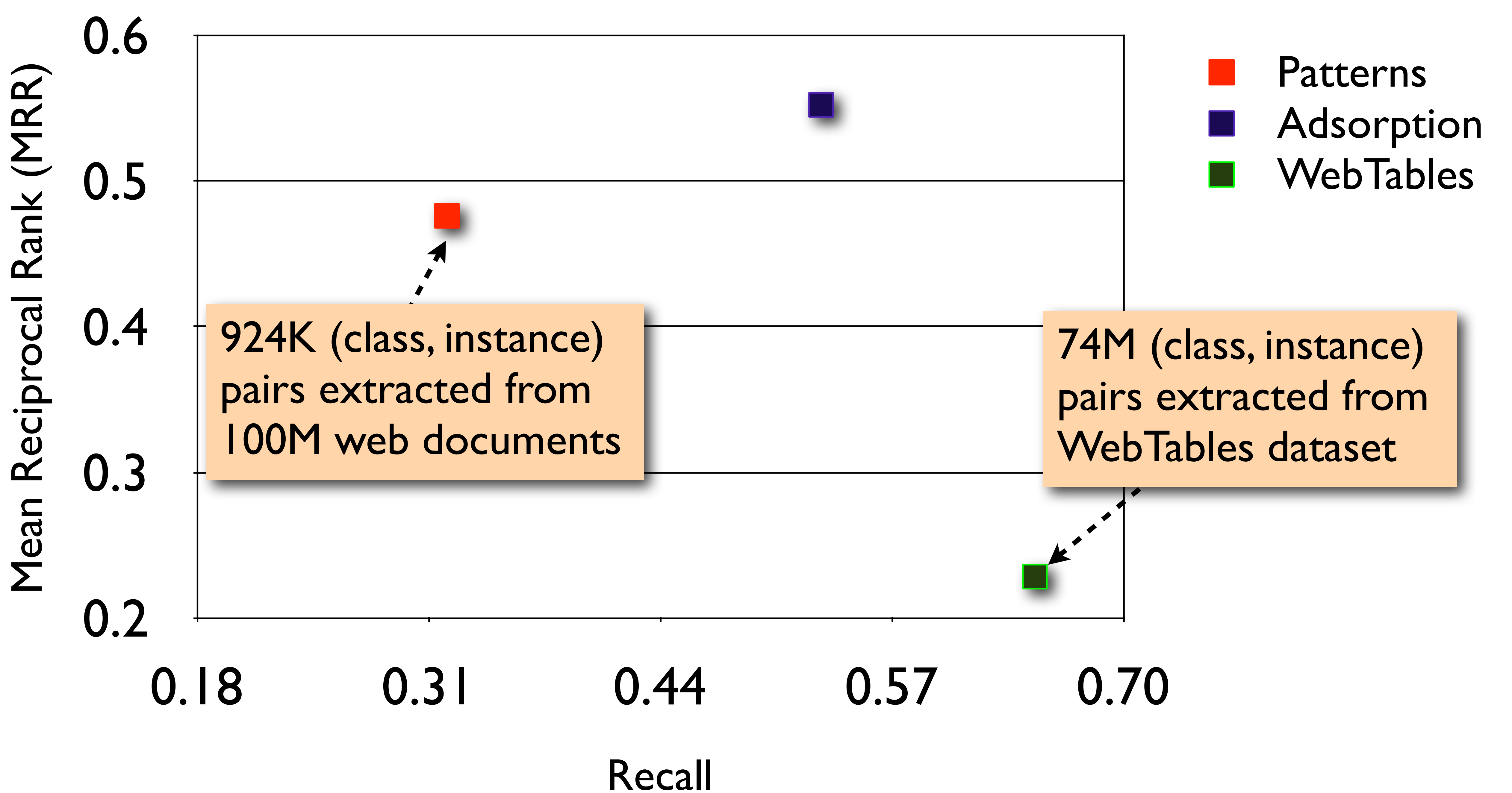


Extraction for Known Instances

Adsorption is able to assign ***better*** class labels to ***more*** instances.

Graph with 1.4m nodes, 75m edges used.

Evaluation against WordNet Dataset (38 classes, 8910 instances)



Extracted Pairs

Total classes: **9081**

Class	**A few non-seed Instances found by Adsorption**
Scientific Journals	Journal of Physics, Nature, Structural and Molecular Biology, Sciences Sociales et sante, Kidney and Blood Pressure Research, American Journal of Physiology-Cell Physiology, …
NFL Players	Tony Gonzales, Thabiti Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannan, …
Book Publishers	Small Night Shade Books, House of Ansari Press, Highwater Books, Distributed Art Publishers, Cooper Canyon Press, …

Extracted Pairs

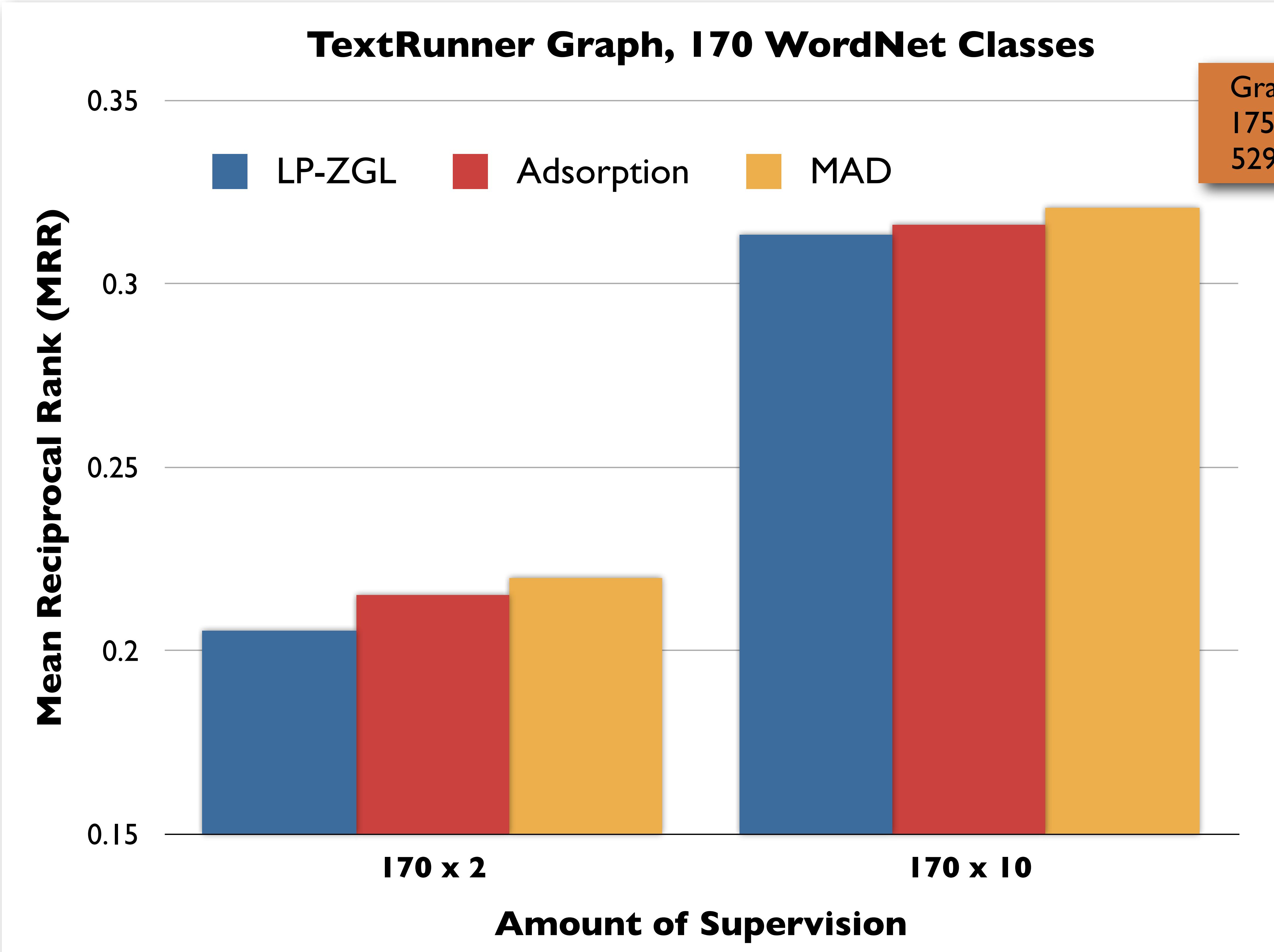
Total classes: **9081**

Class	**A few non-seed Instances found by Adsorption**
Scientific Journals	Journal of Physics, Nature, Structural and Molecular Biology, Sciences Sociales et sante, Kidney and Blood Pressure Research, American Journal of Physiology-Cell Physiology, …
NFL Players	Tony Gonzales, Thabiti Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannan, …
Book Publishers	Small Night Shade Books, House of Ansari Press, Highwater Books, Distributed Art Publishers, Cooper

Graph-based methods can easily handle large number of classes

Results

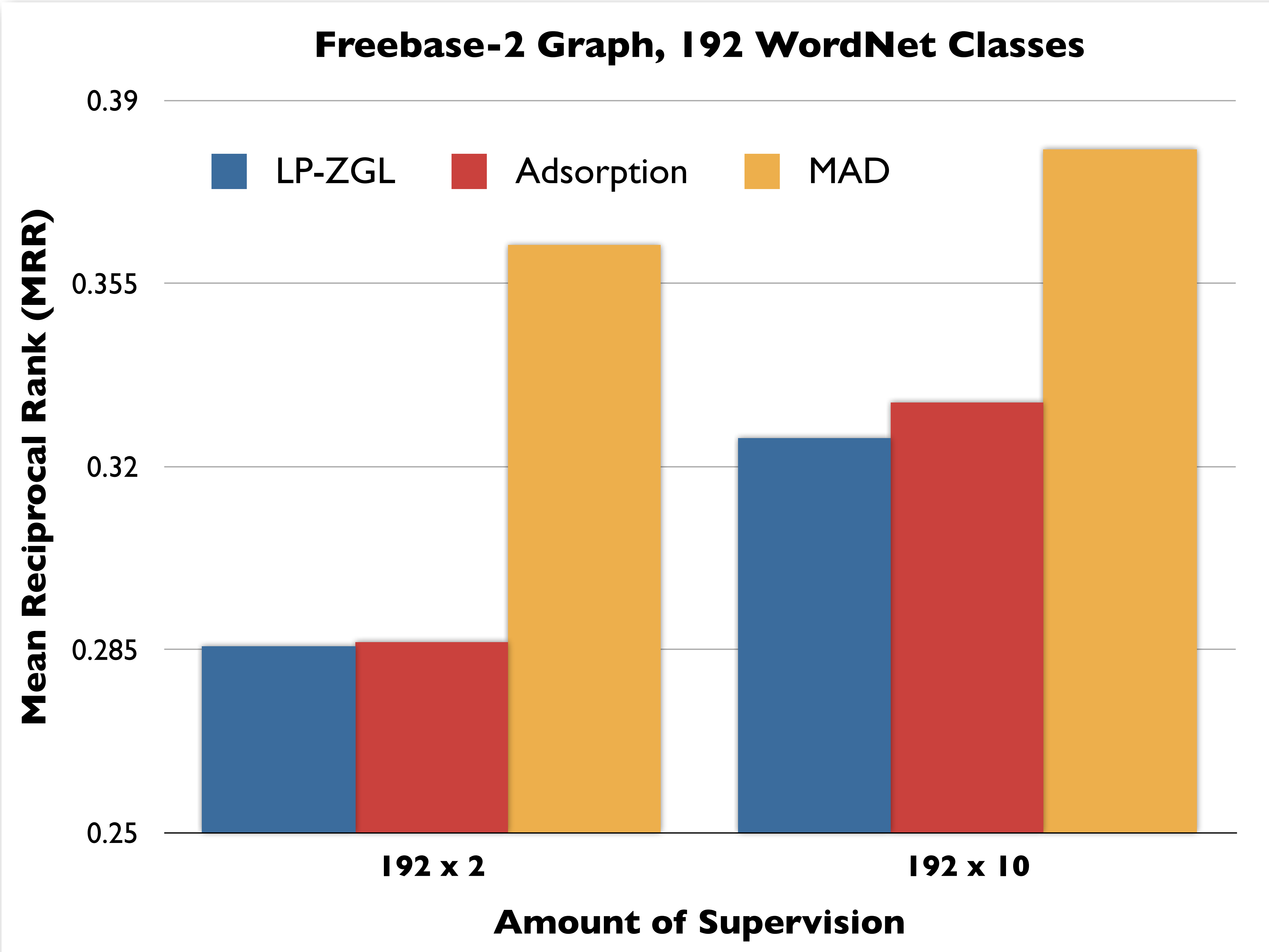
Data available @ http://www.talukdar.net/datasets/class_inst/



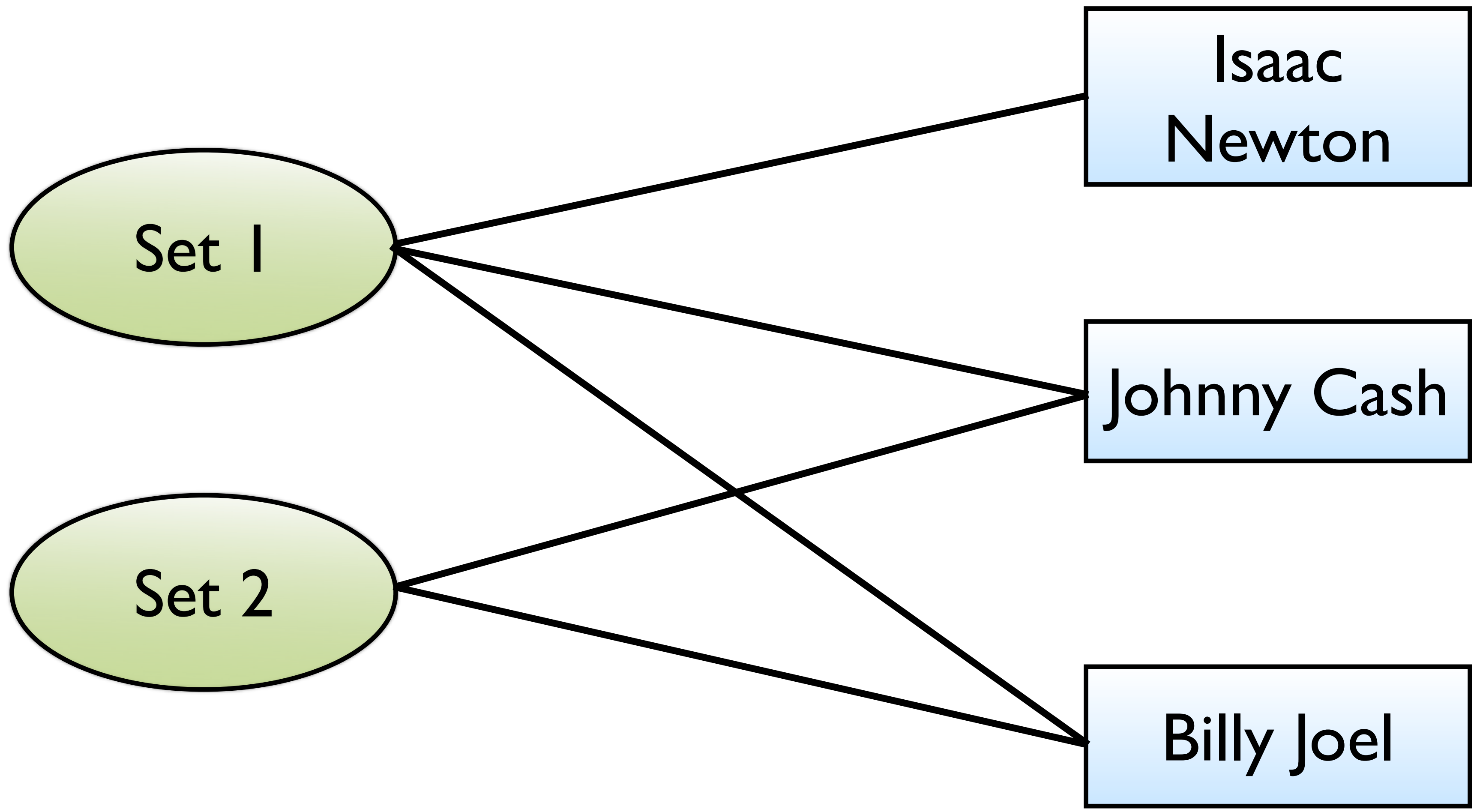
Graph with 175k nodes 529k edges

Results

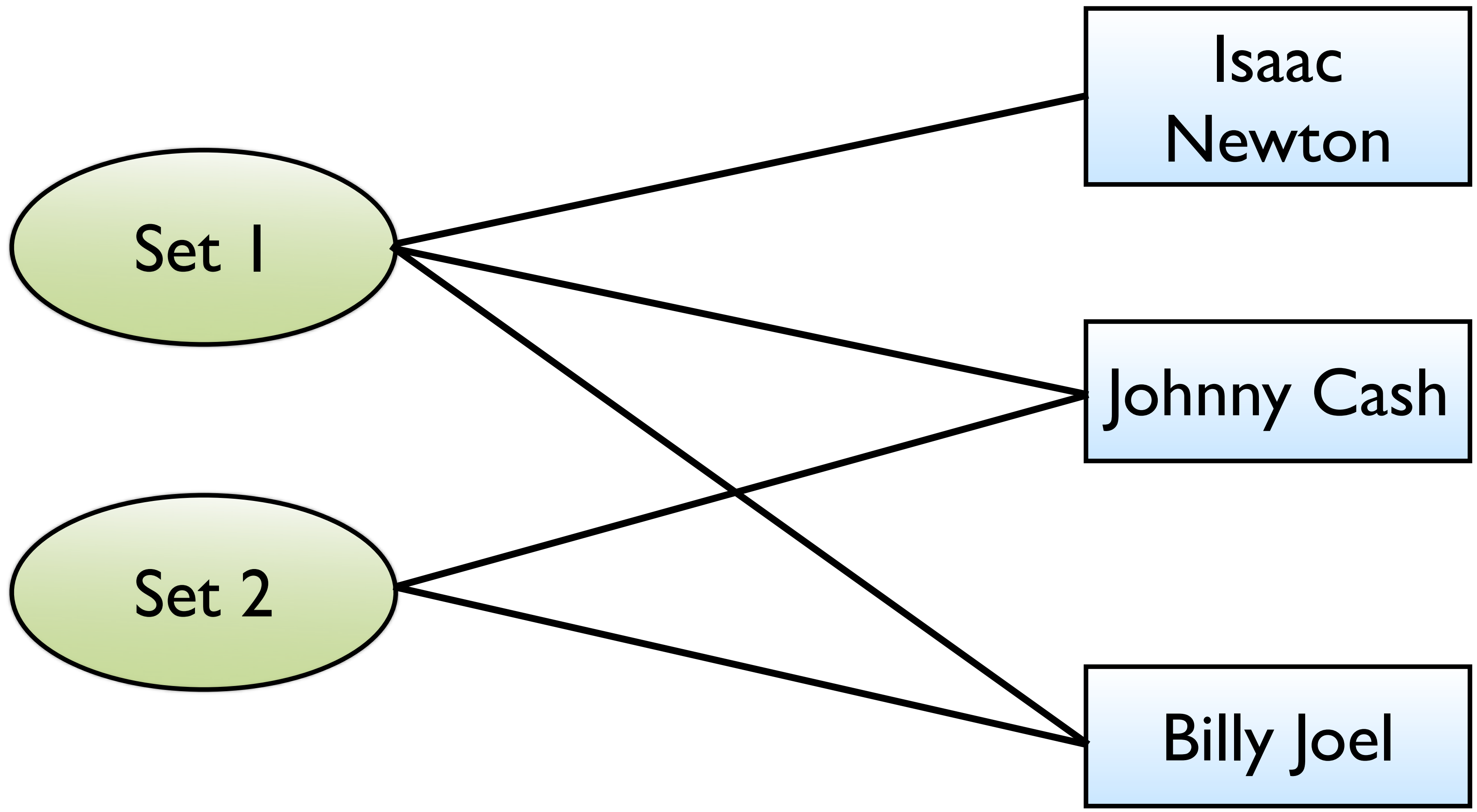
Graph with
303k nodes
2.3m edges



Semantic Constraints



Semantic Constraints

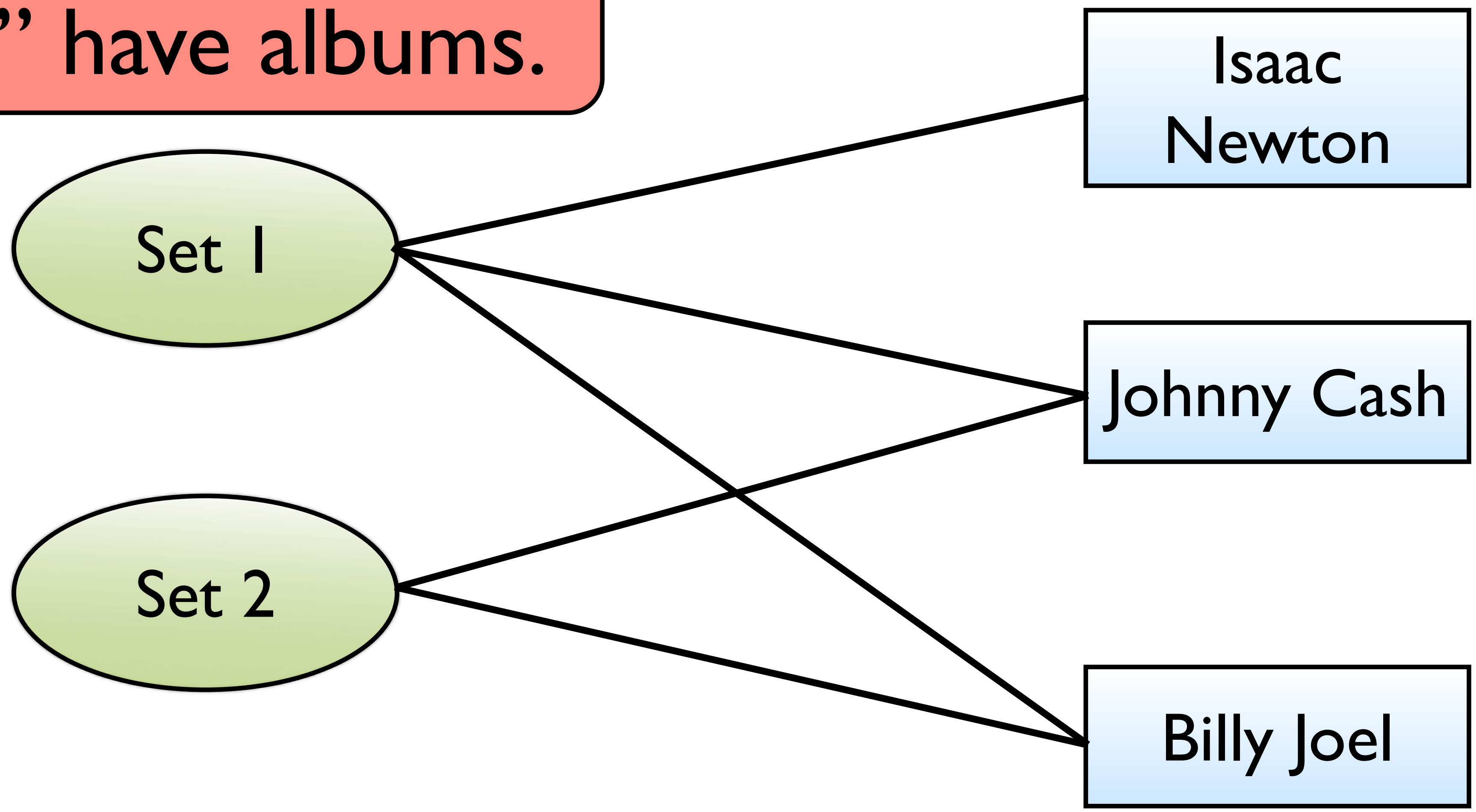


Suppose we knew that both "Johnny Cash" and "Billy Joel" have albums.

How do we encode this constraint?

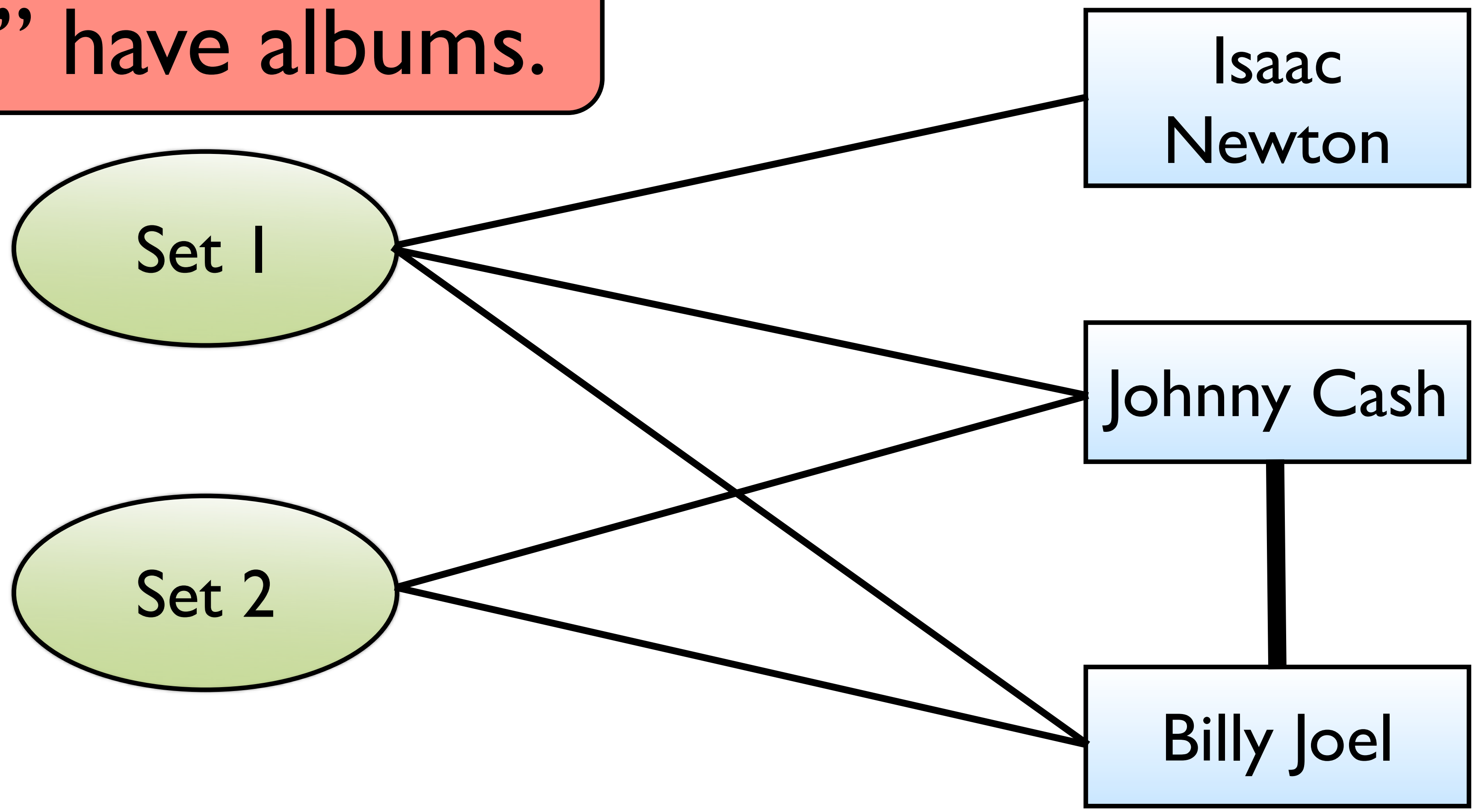
Solution (I)

Both "Johnny Cash" and "Billy Joel" have albums.



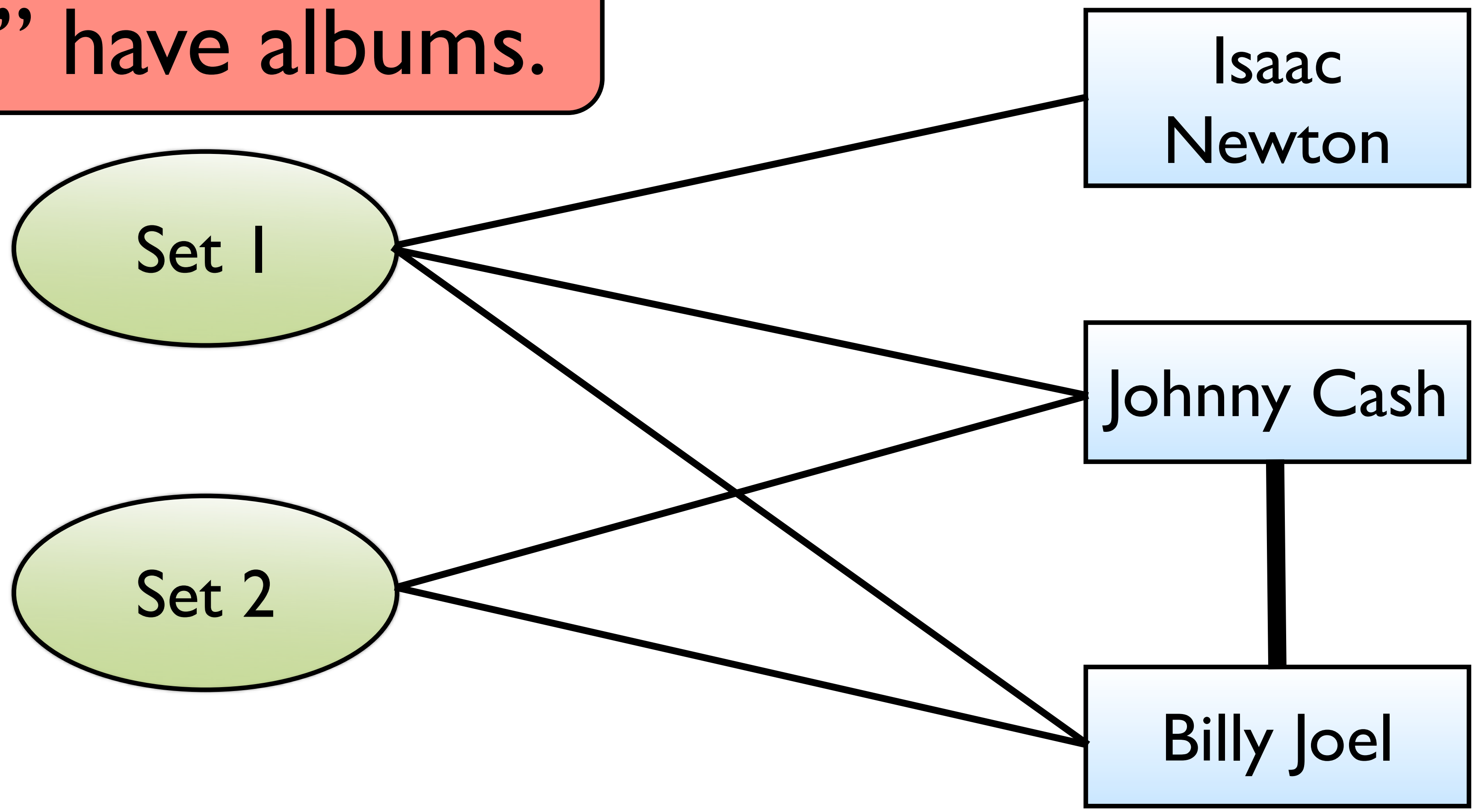
Solution (I)

Both "Johnny Cash" and "Billy Joel" have albums.



Solution (I)

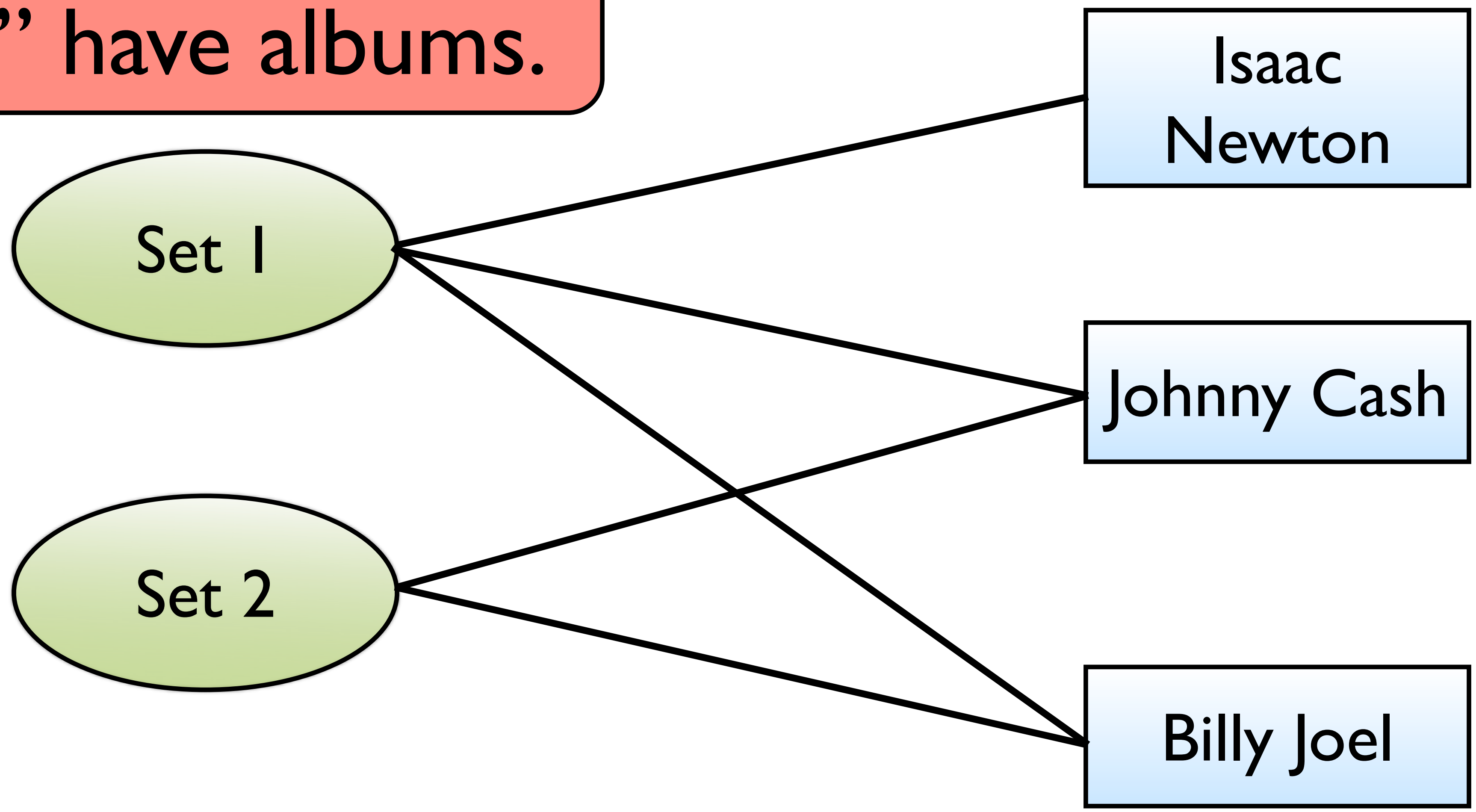
Both “Johnny Cash” and “Billy Joel” have albums.



- Graph is no longer bi-partite (not necessarily bad)
- Can lead to cliques of size of number of instances in the constraint (bad)

Solution (II)

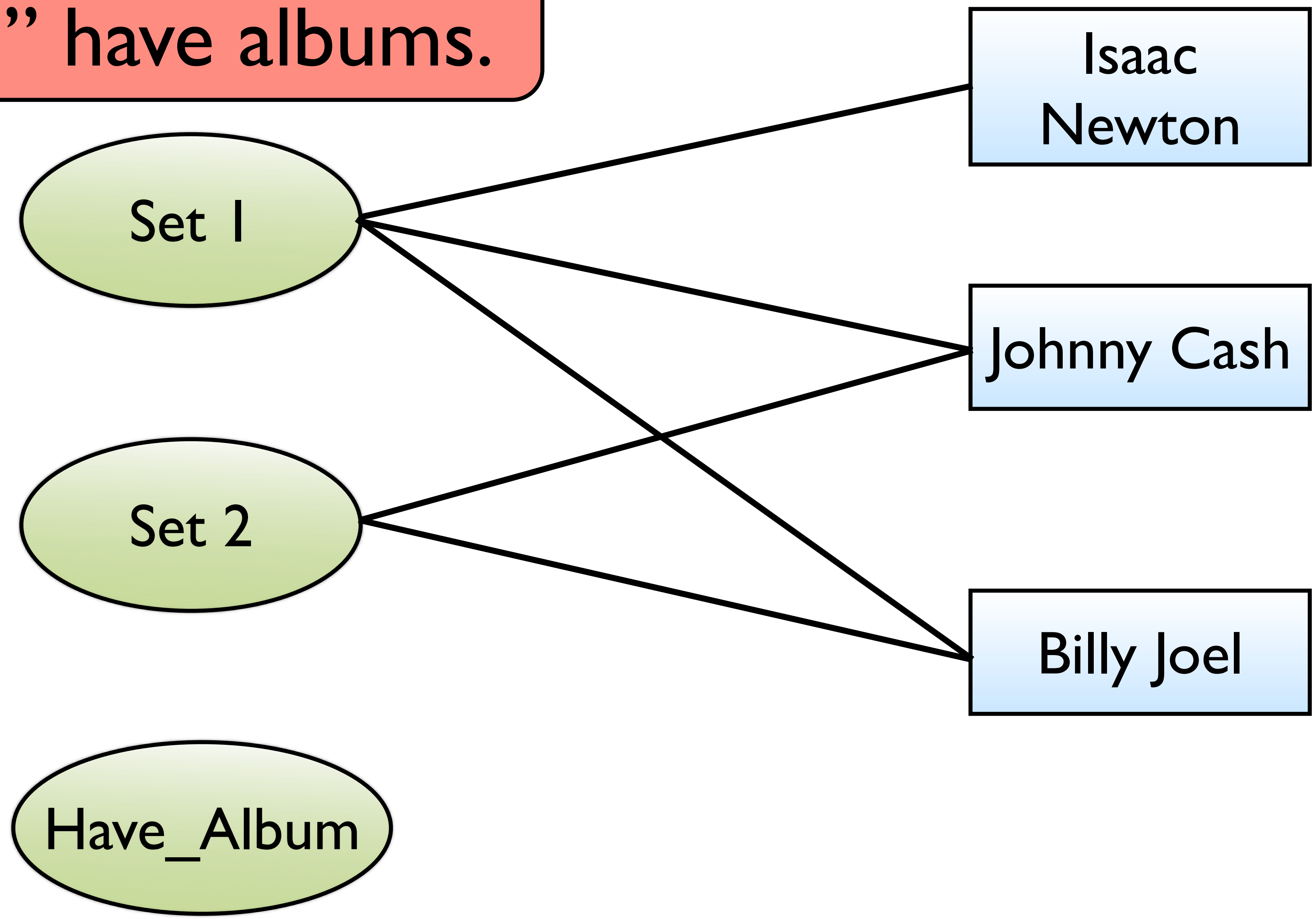
Both "Johnny Cash" and "Billy Joel" have albums.



[Talukdar & Periera, ACL 2010]

Solution (II)

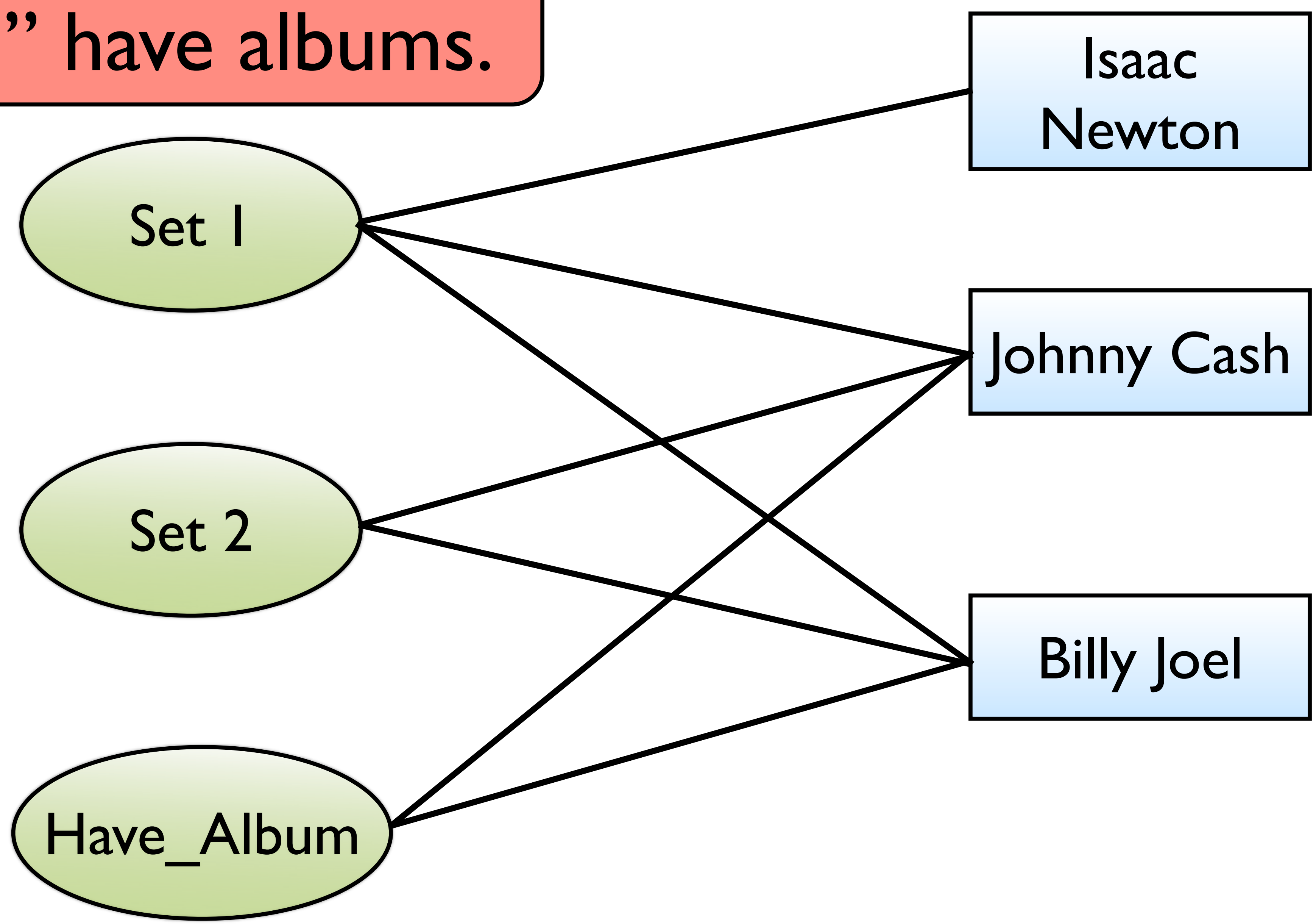
Both "Johnny Cash" and "Billy Joel" have albums.



[Talukdar & Periera, ACL 2010]

Solution (II)

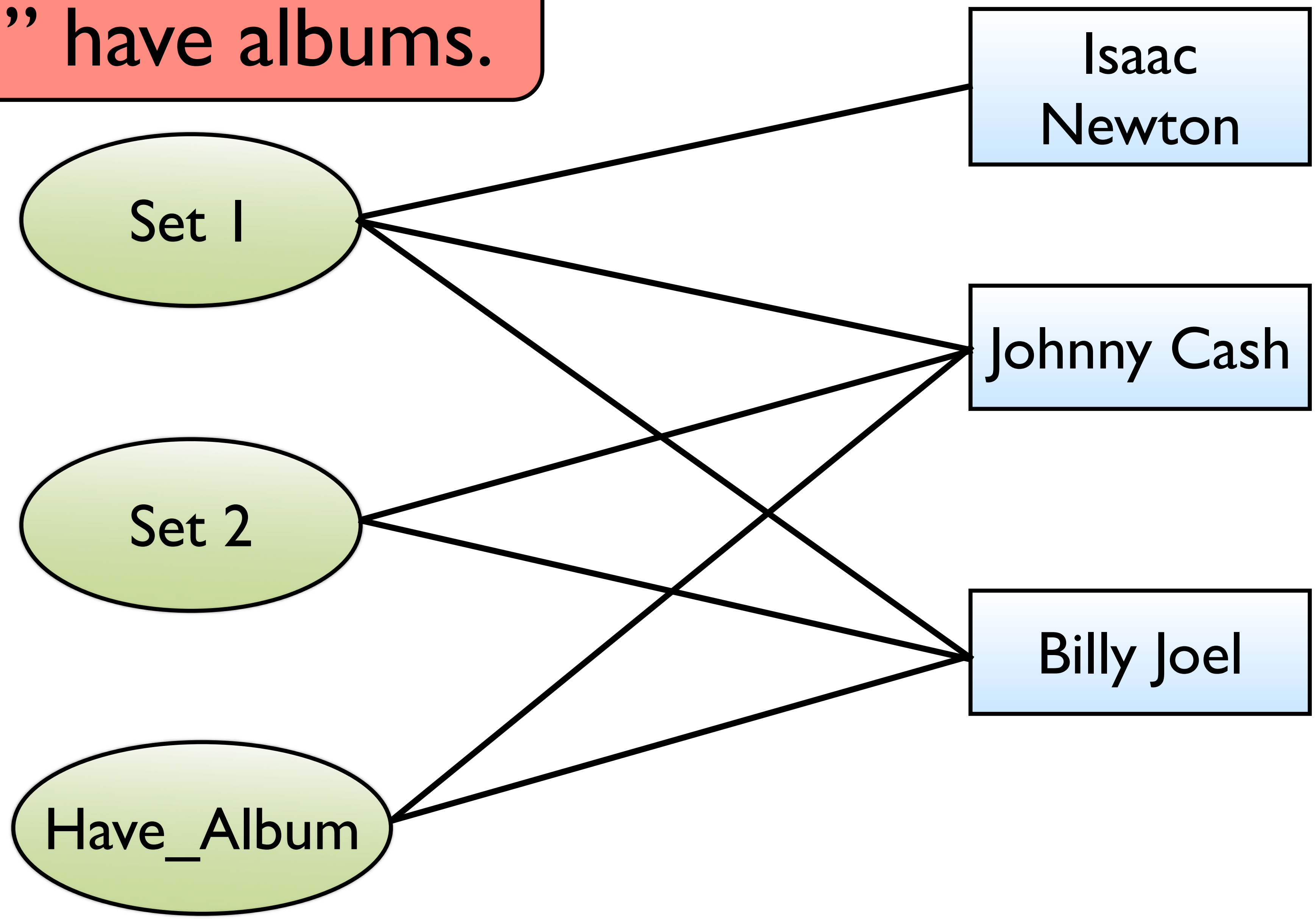
Both "Johnny Cash" and "Billy Joel" have albums.



[Talukdar & Periera, ACL 2010]

Solution (II)

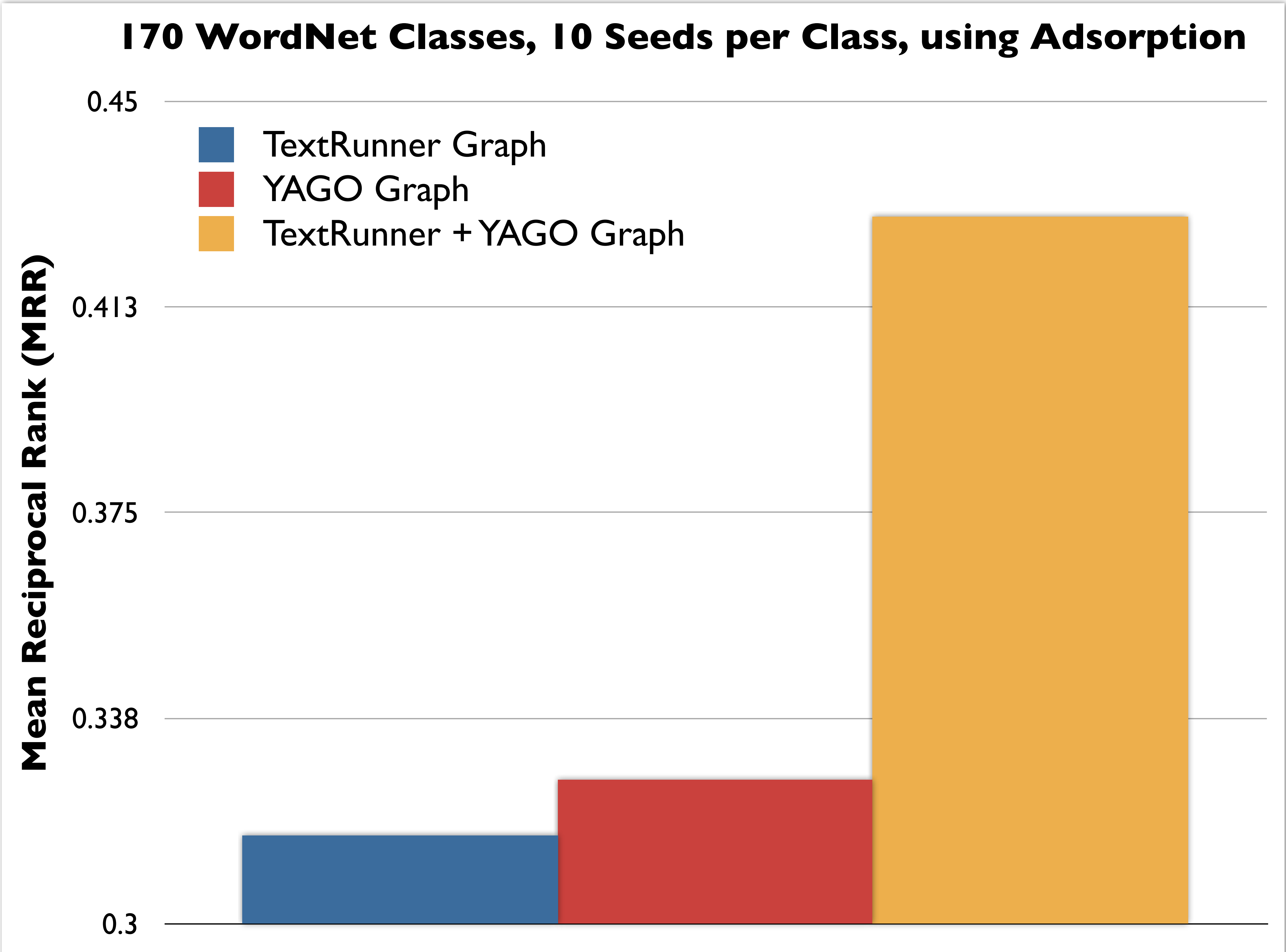
Both “Johnny Cash” and “Billy Joel” have albums.



Semantic Constraints may be easily encoded

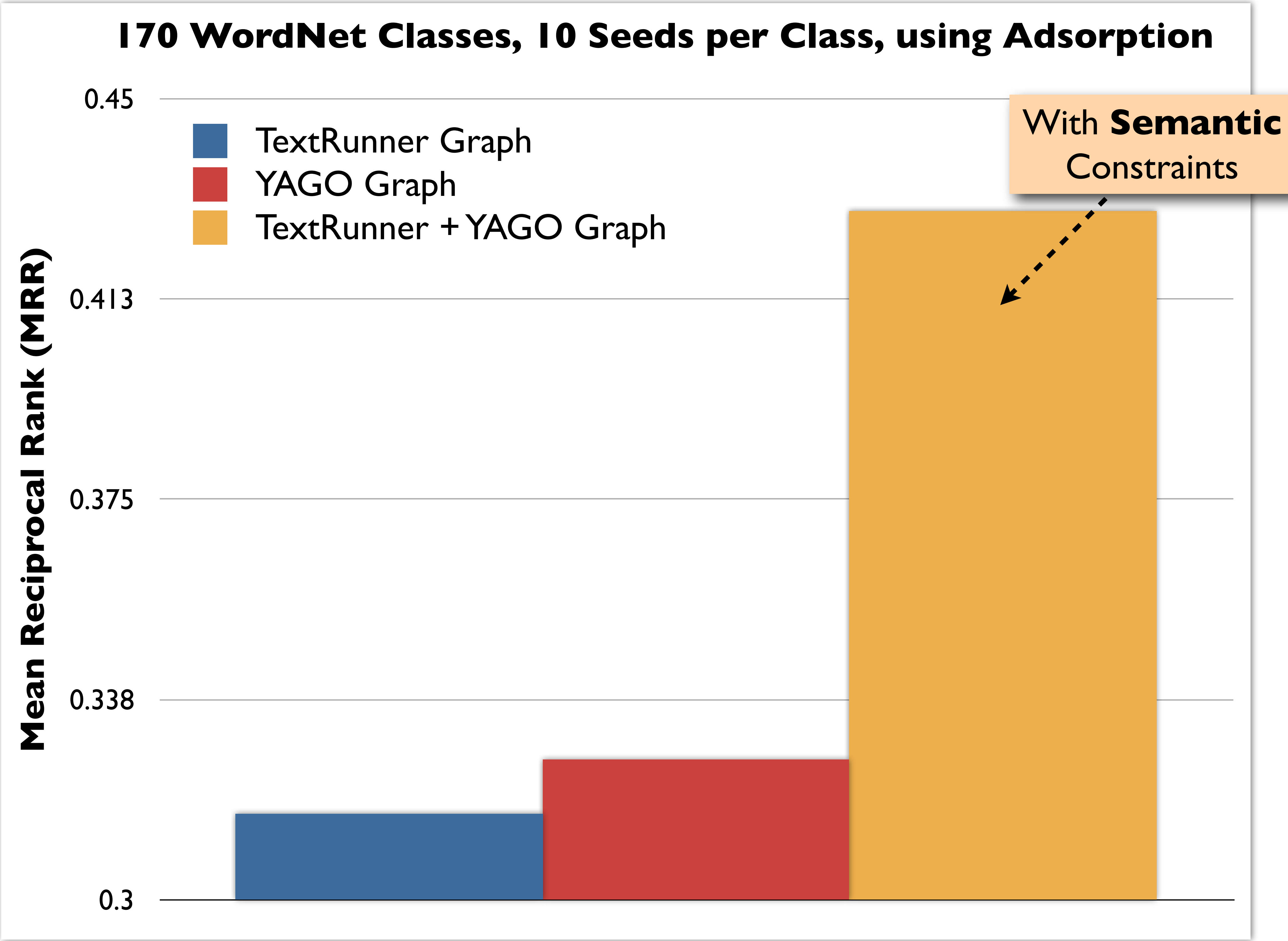
[Talukdar & Periera, ACL 2010]

Results with Semantic Constraints



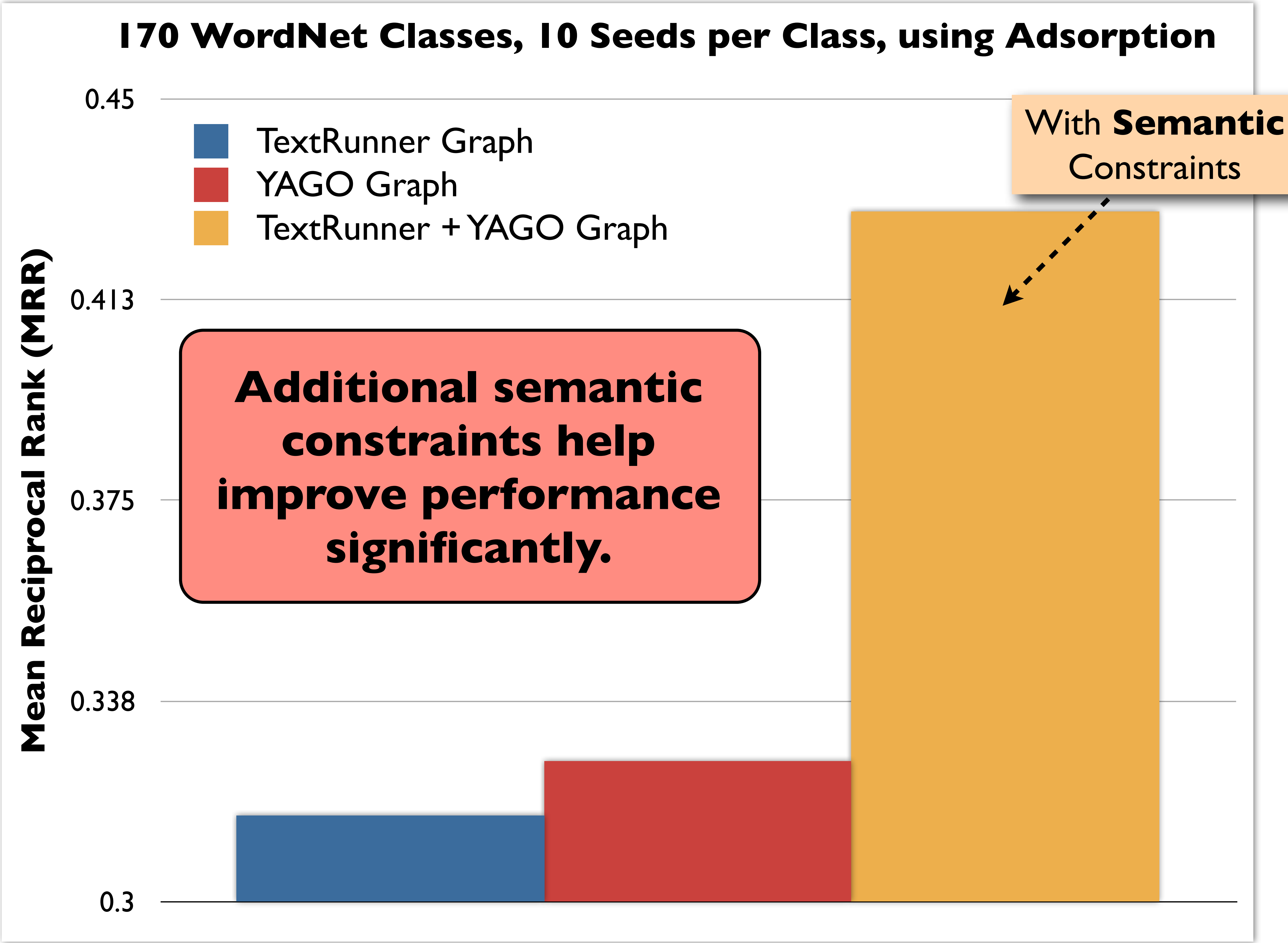
[Talukdar & Periera, ACL 2010]

Results with Semantic Constraints



[Talukdar & Periera, ACL 2010]

Results with Semantic Constraints

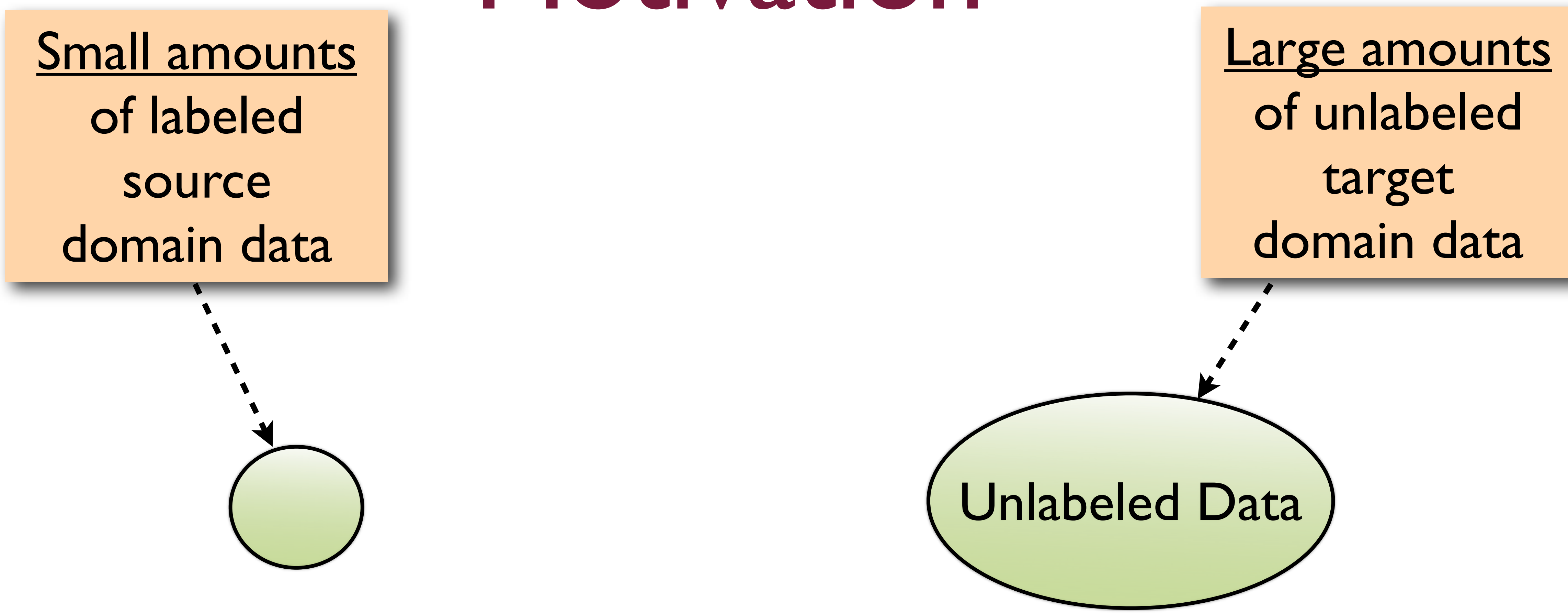


[Talukdar & Periera, ACL 2010]

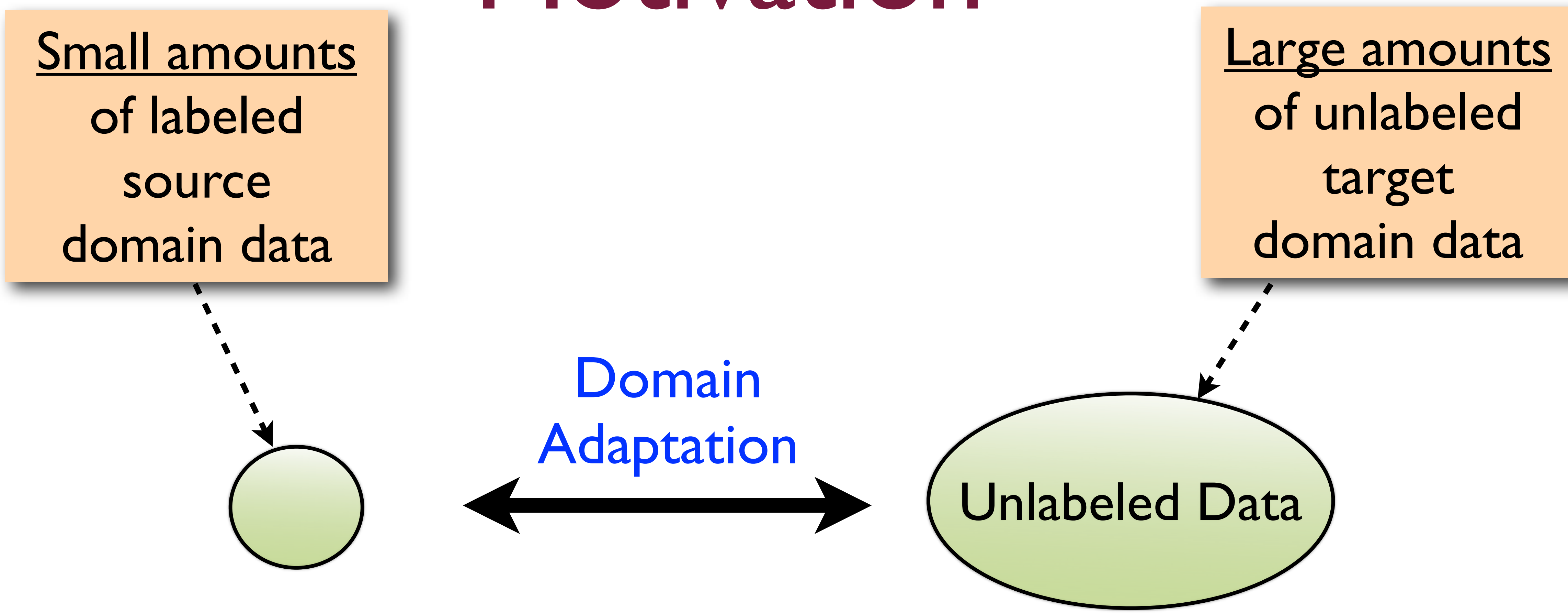
Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Text Categorization
  - Sentiment Analysis
  - Class Instance Acquisition
  - POS Tagging [Subramanya et. al., EMNLP 2008]
  - MultiLingual POS Tagging
  - Semantic Parsing
- Conclusion & Future Work

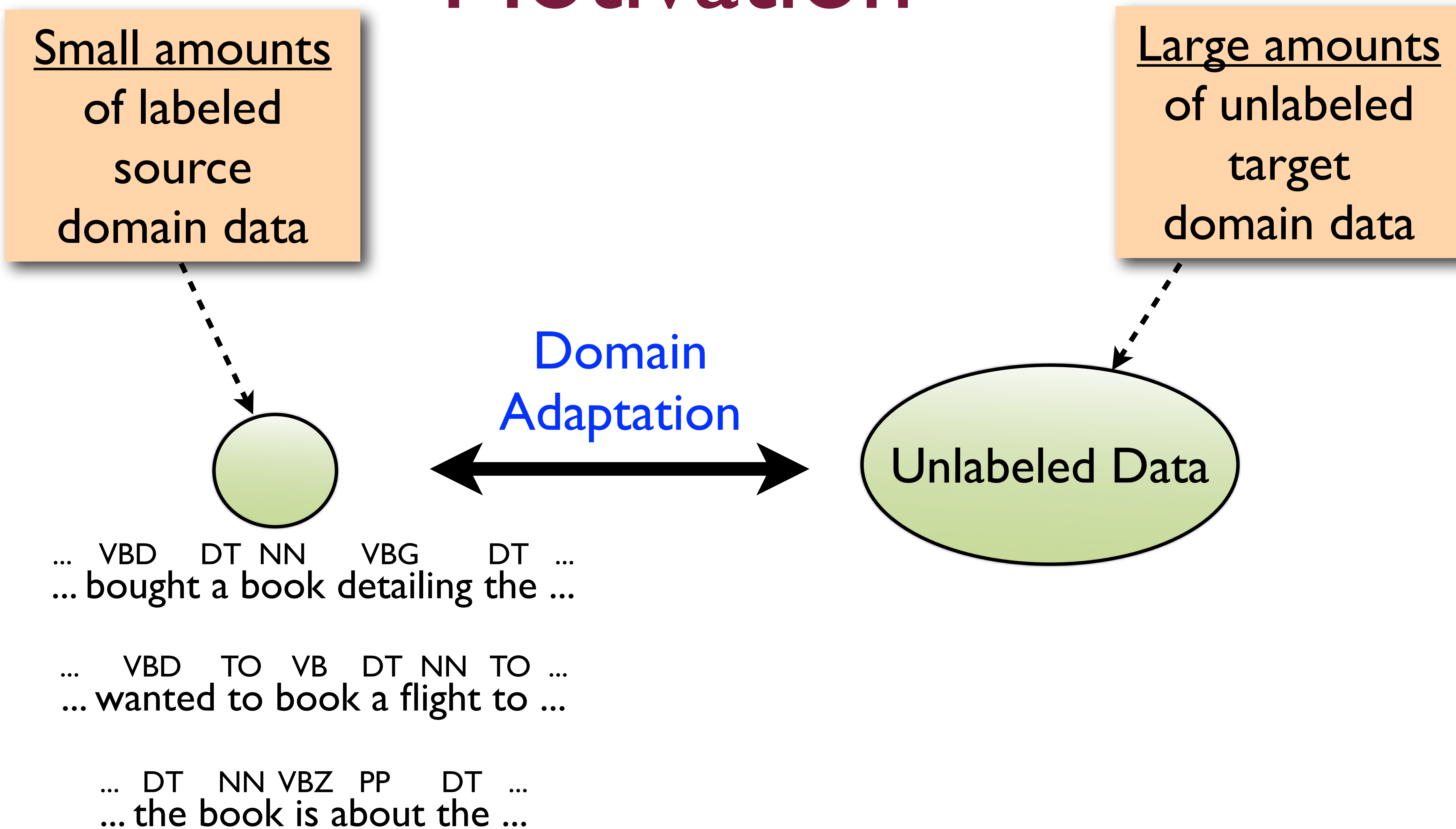
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Unlabeled Data

Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data

Motivation



Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...

Motivation

Small amounts of labeled source domain data

Large amounts of unlabeled target domain data

Domain Adaptation

Unlabeled Data

... VBD DT NN VBG DT ...
... bought a book detailing the ...

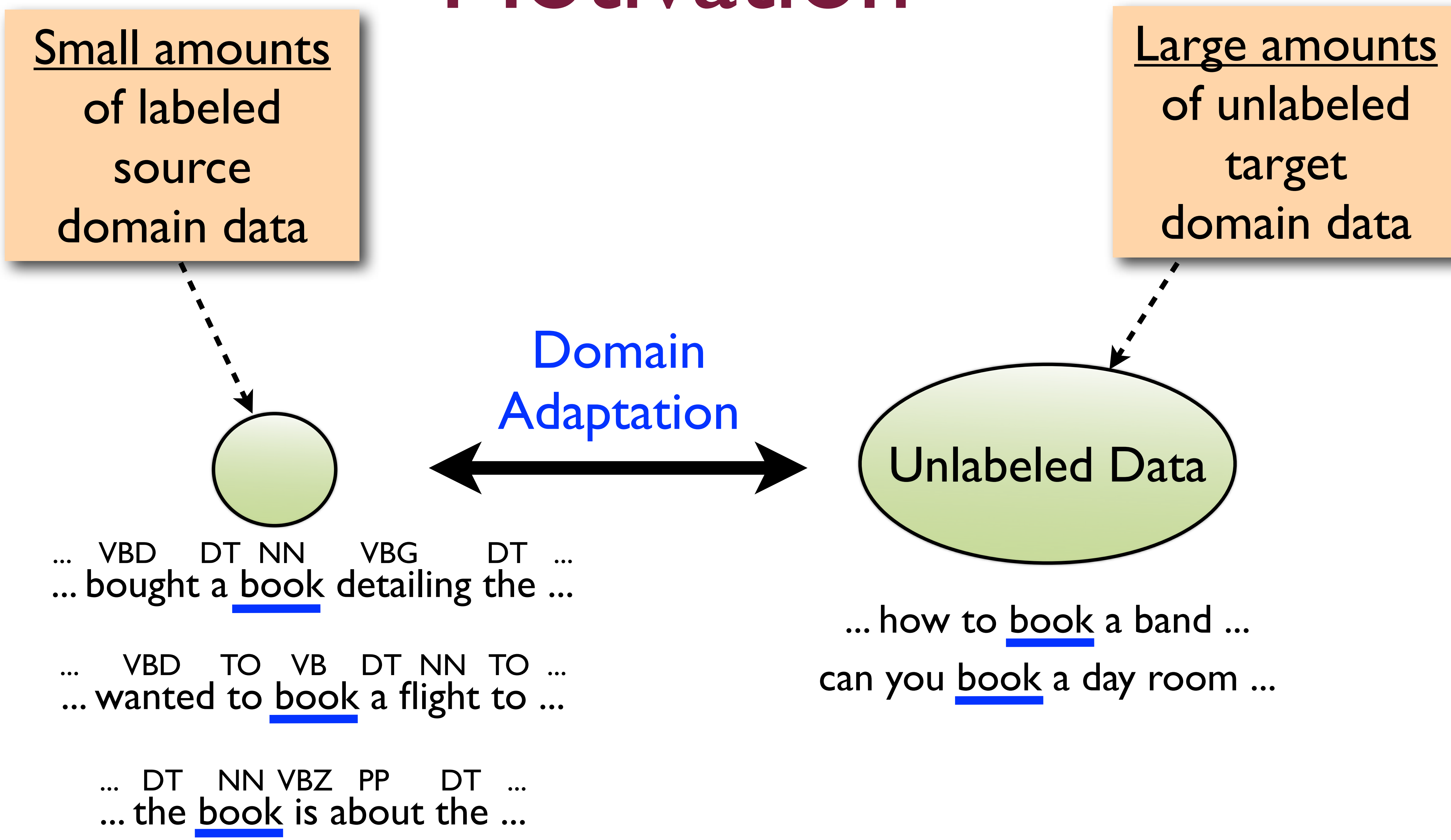
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...

... DT NN VBZ PP DT ...
... the book is about the ...

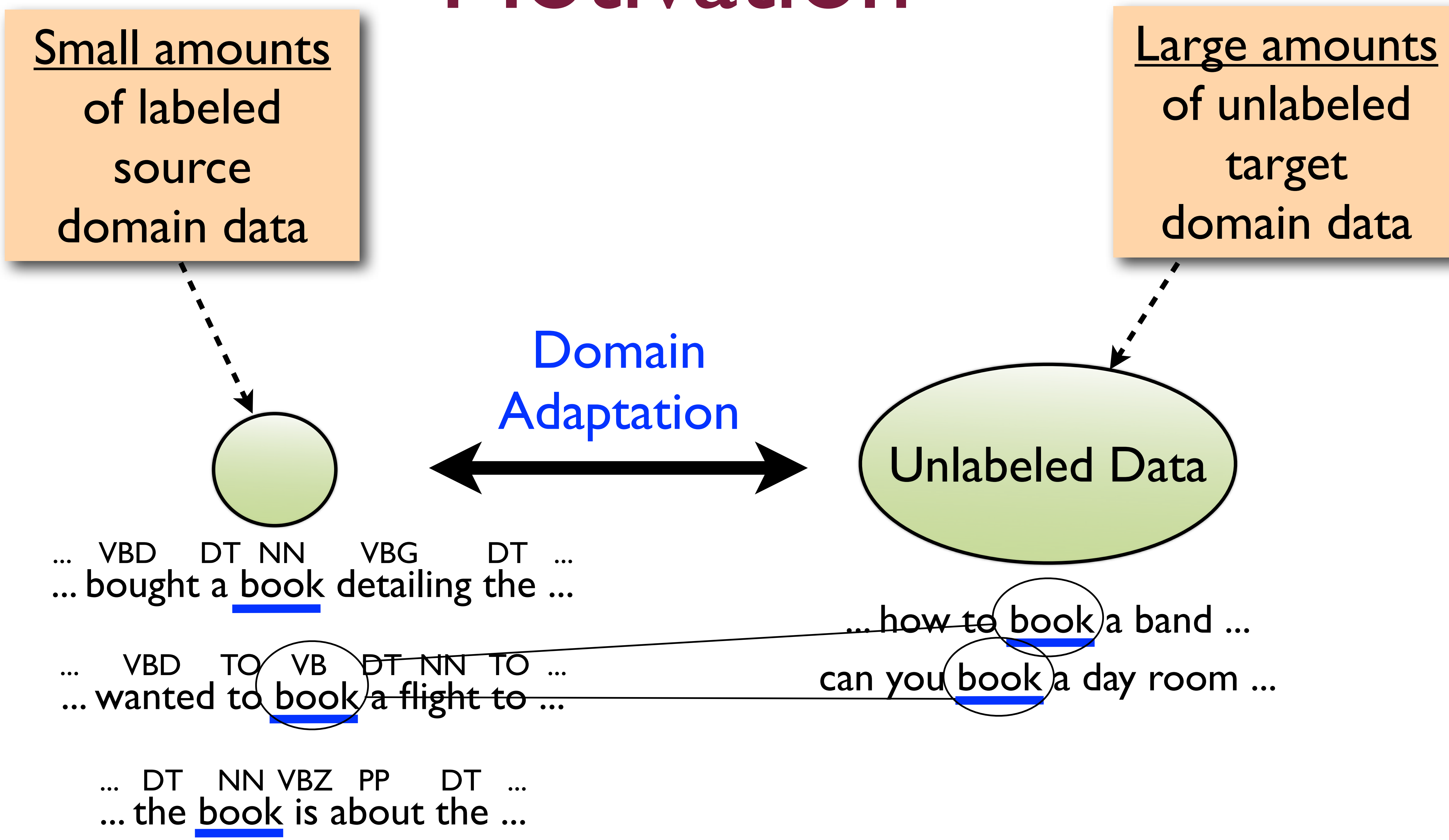
... how to book a band ...

can you book a day room ...

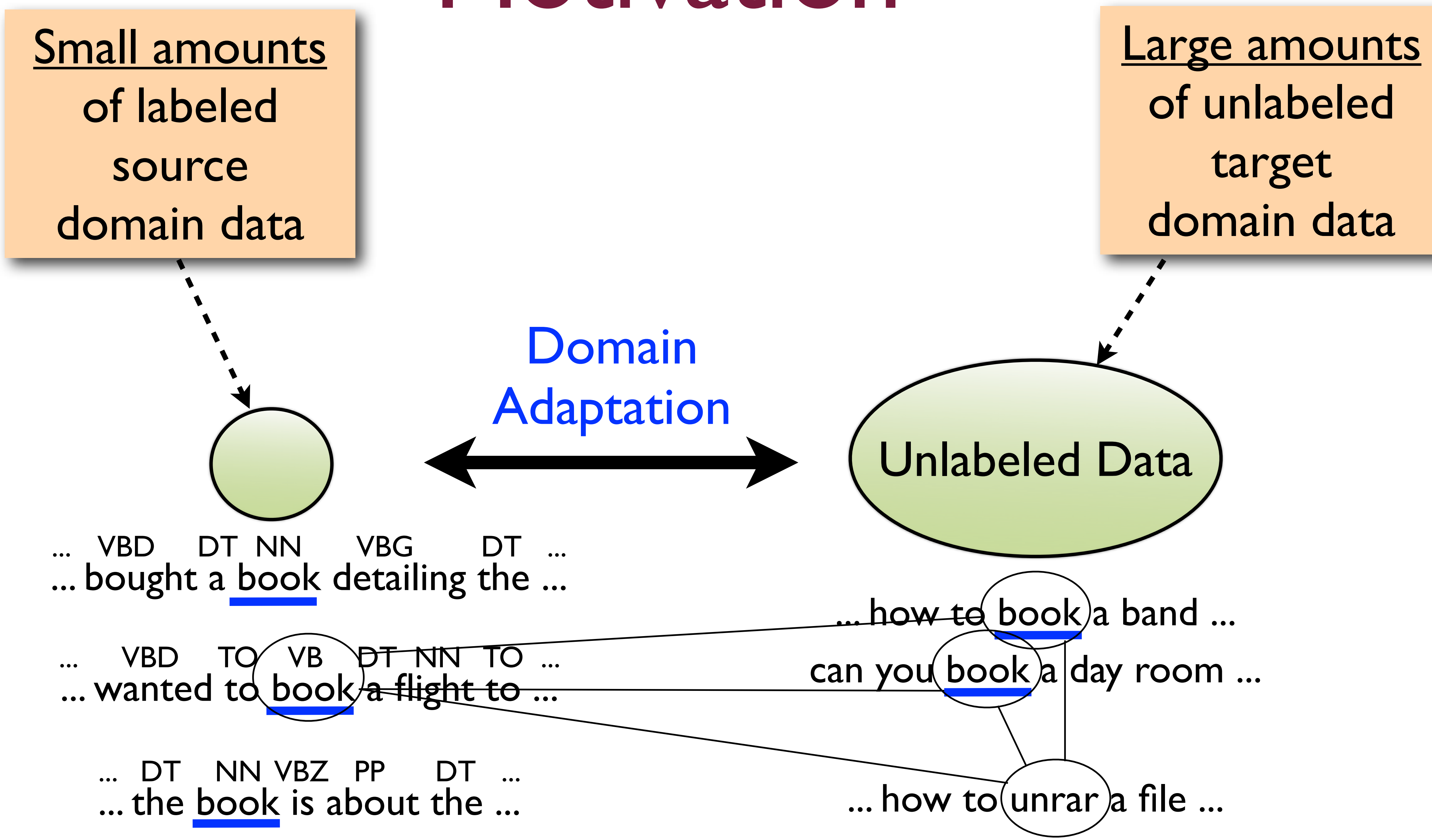
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...

Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...

Motivation



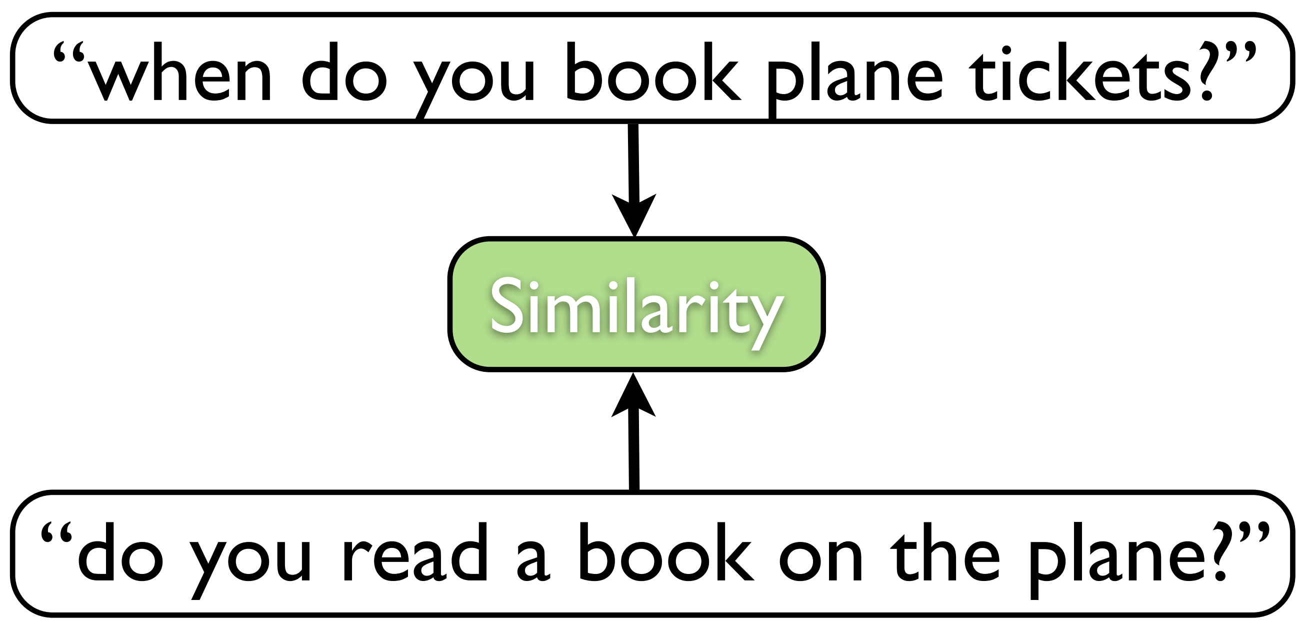
Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...
... how to unrar a file ...

Graph Construction (I)

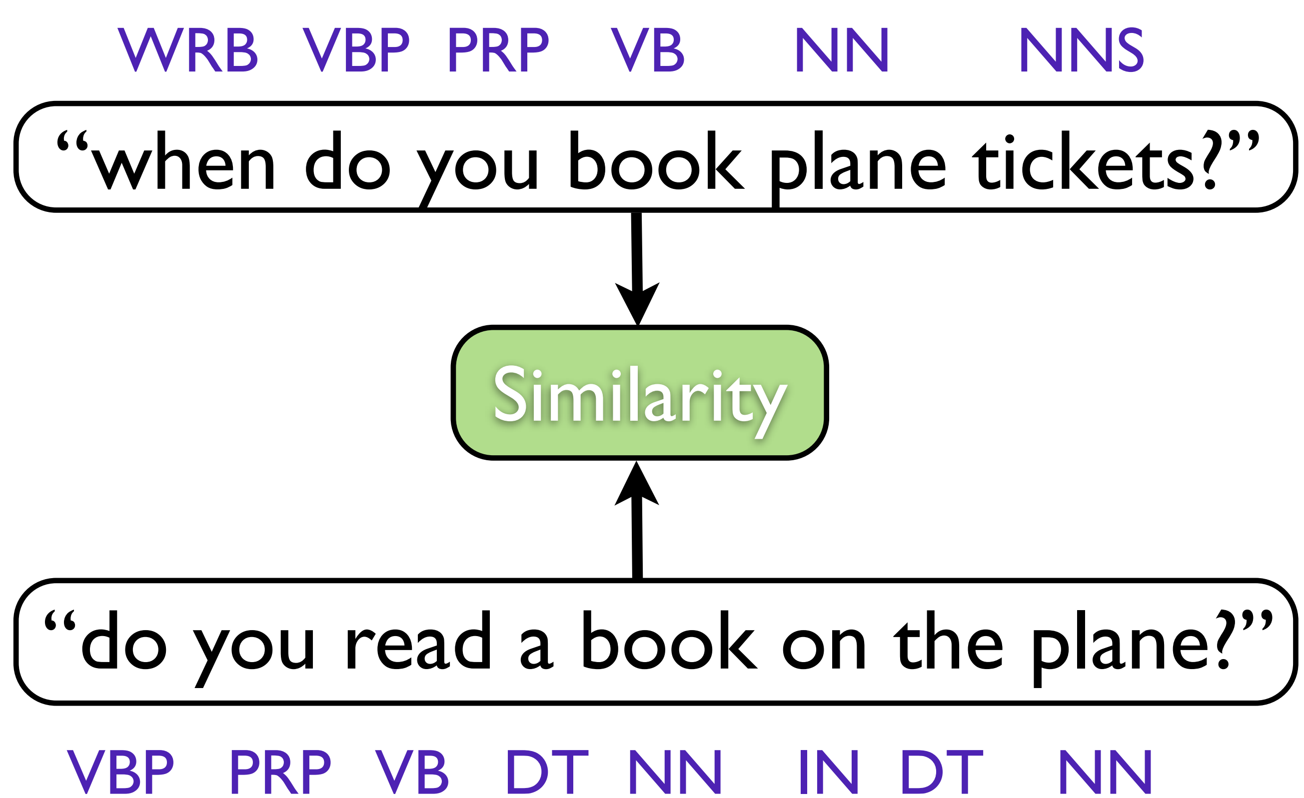
"when do you book plane tickets?"

"do you read a book on the plane?"

Graph Construction (I)



Graph Construction (I)



Graph Construction (II)

can you book a day room at hilton hawaiian village ?

what was the book that has no letter e ?

how much does it cost to book a band ?

how to get a book agent ?

Graph Construction (II)

can you book a day room at hilton hawaiian village ?

what was the book that has no letter e ?

how much does it cost to book a band ?

how to get a book agent ?

Graph Construction (II)

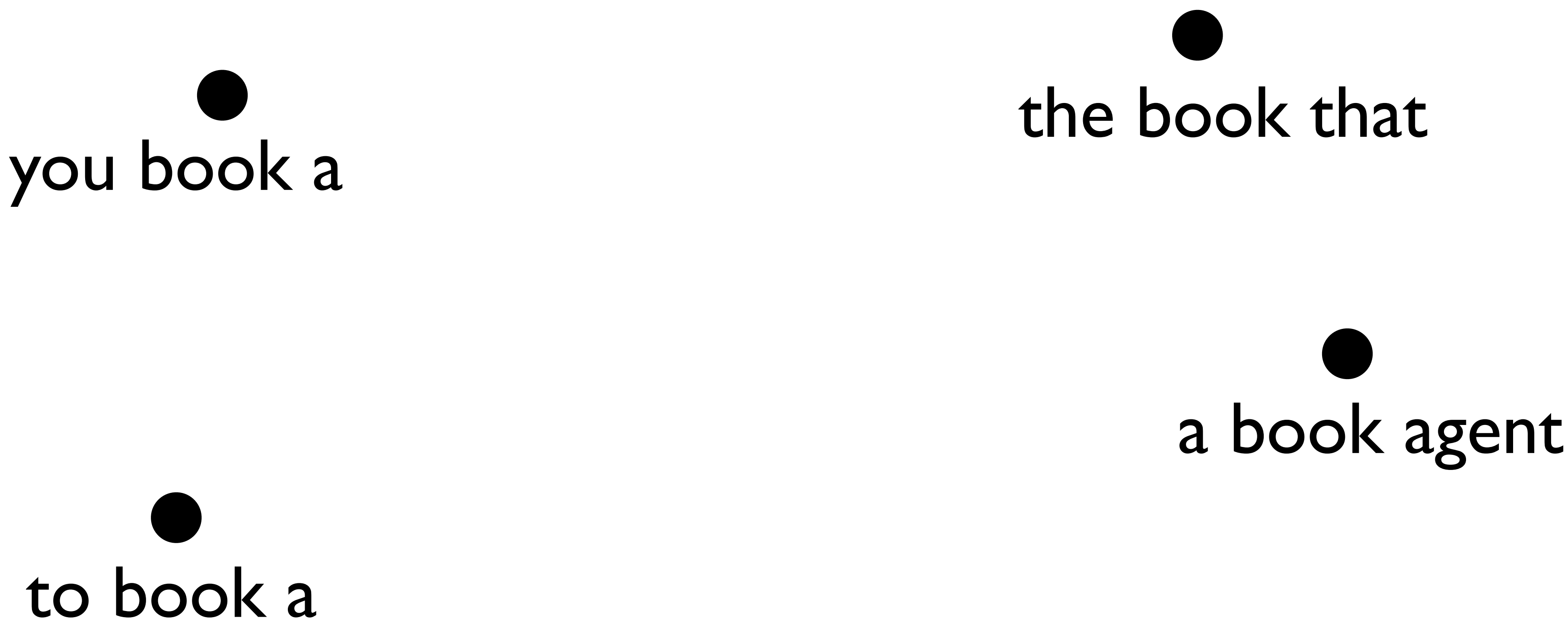
can you book a day room at hilton hawaiian village ?

what was the book that has no letter e ?

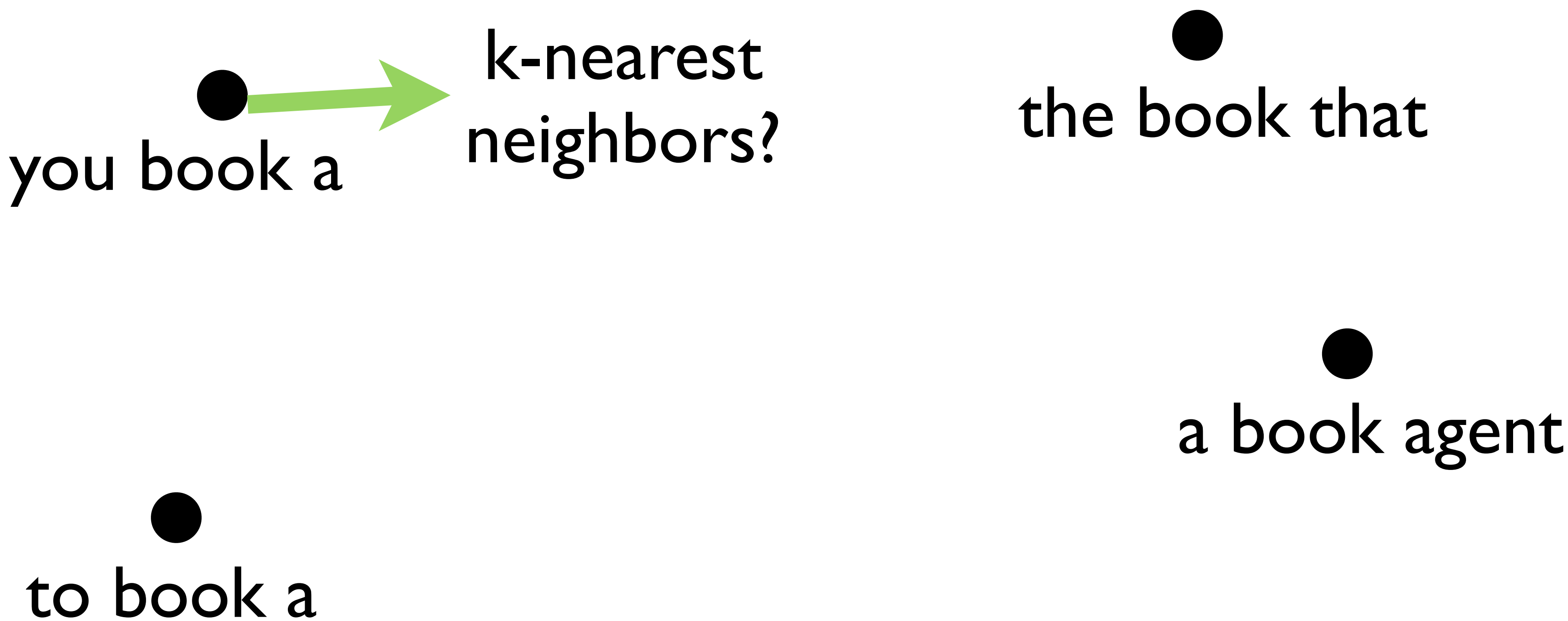
how much does it cost to book a band ?

how to get a book agent ?

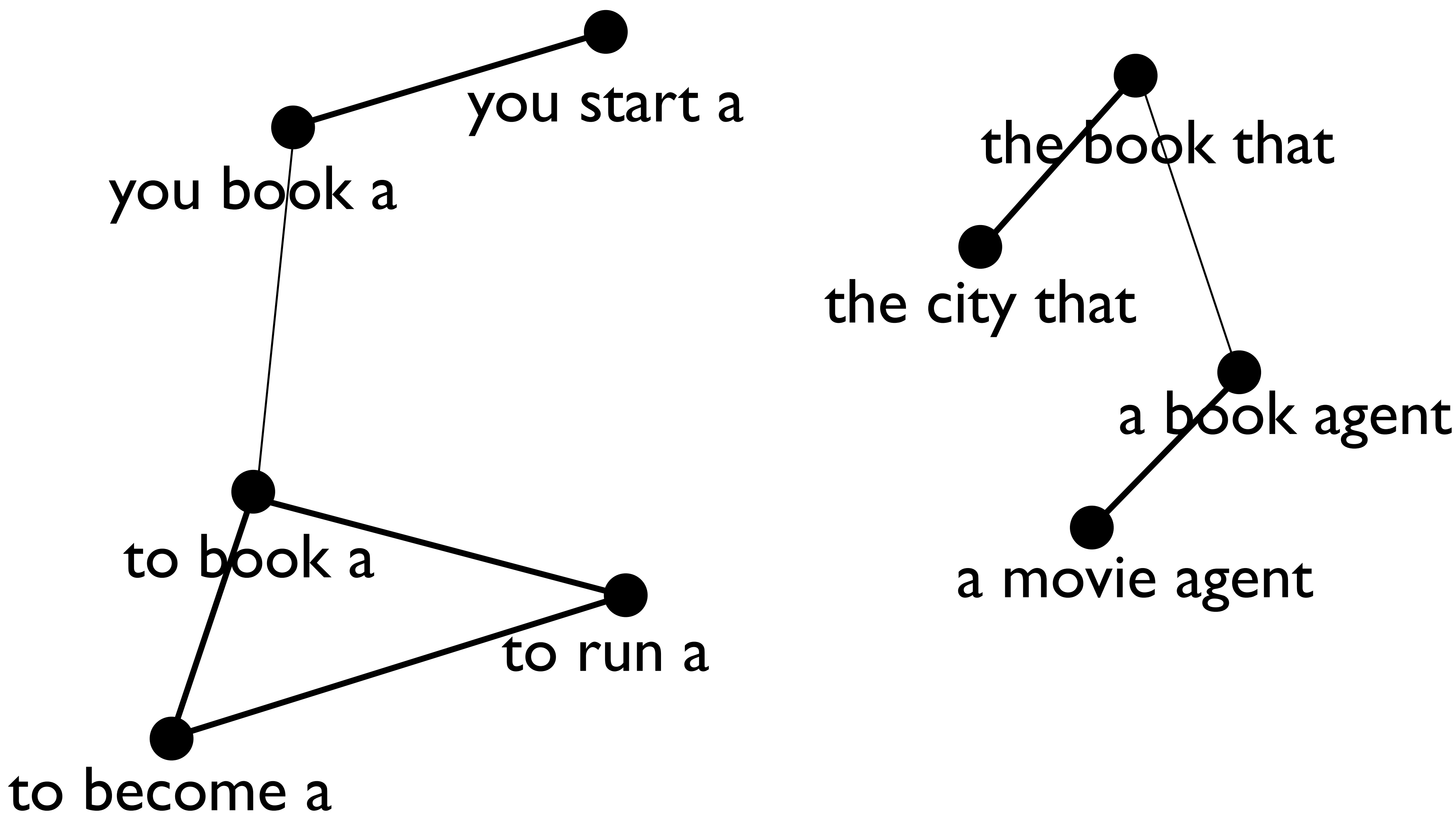
Graph Construction (II)



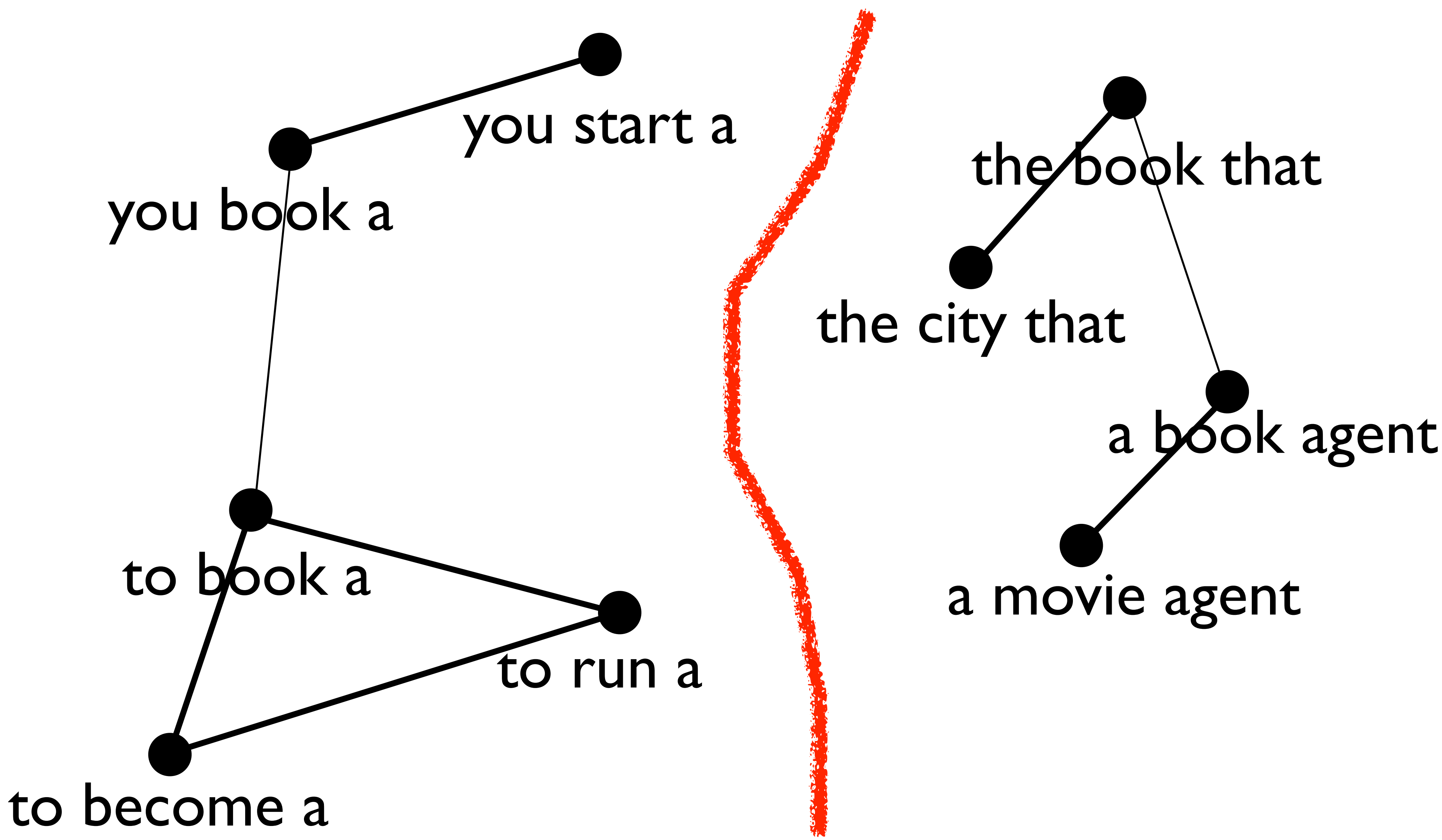
Graph Construction (II)



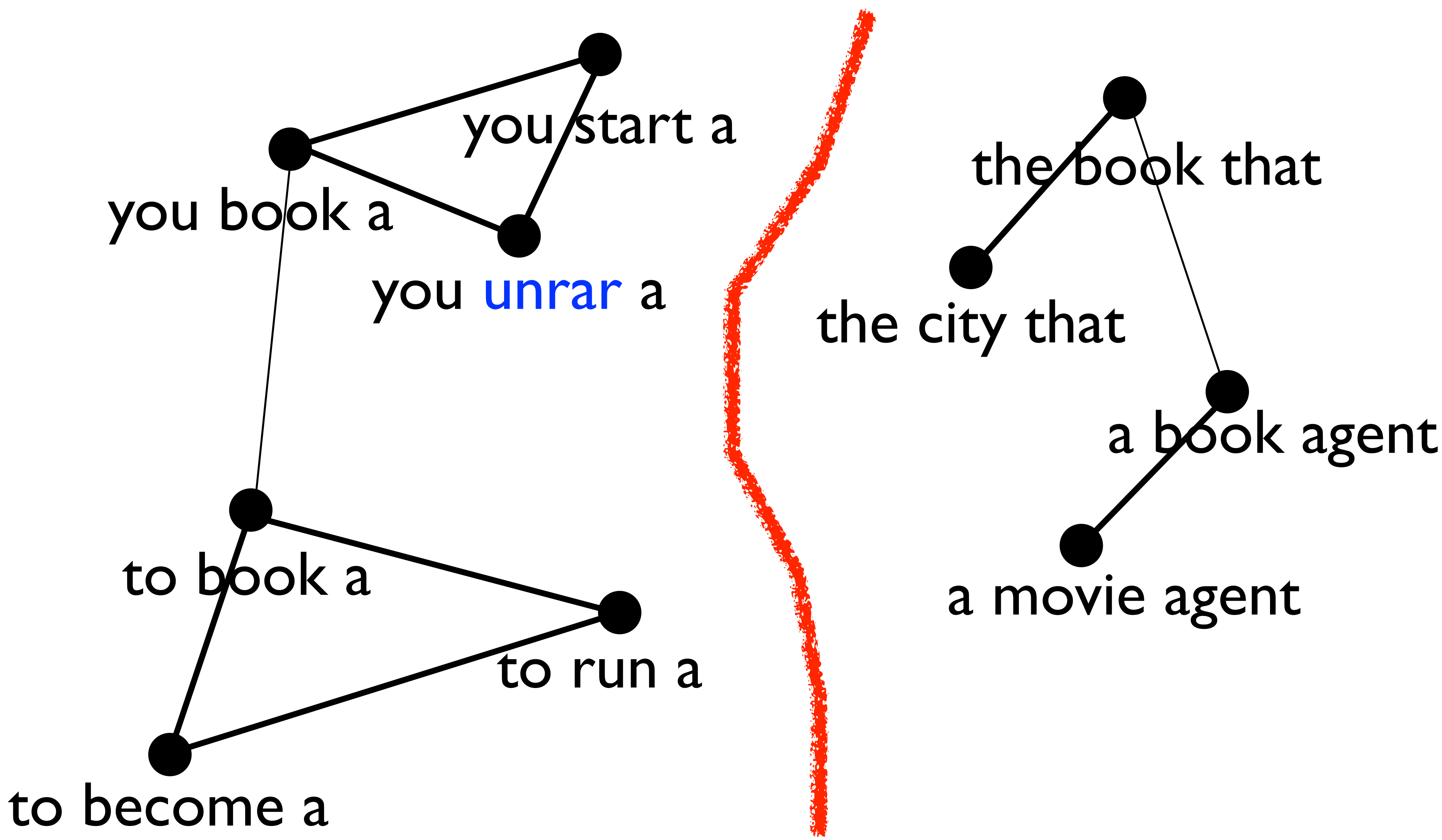
Graph Construction (III)



Graph Construction (III)



Graph Construction (III)



Graph Construction - Features

how much does it cost to book a band ?

Graph Construction - Features

how much does it cost to book a band ?

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to
Right Context	a band

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to
Right Context	a band
Center Word	book

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to
Right Context	a band
Center Word	book
Trigram - Center Word	to ____ a
Left Word + Right Context	to ____ a band
Left Context + Right Word	cost to ____ a
Suffix	none

Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

Graph Construction - Features

how much to book a flight to paris?

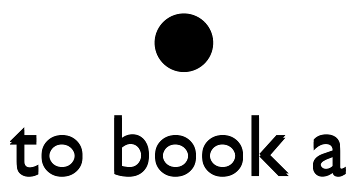
how much does it cost to book a band ?

Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

Graph Construction - Features

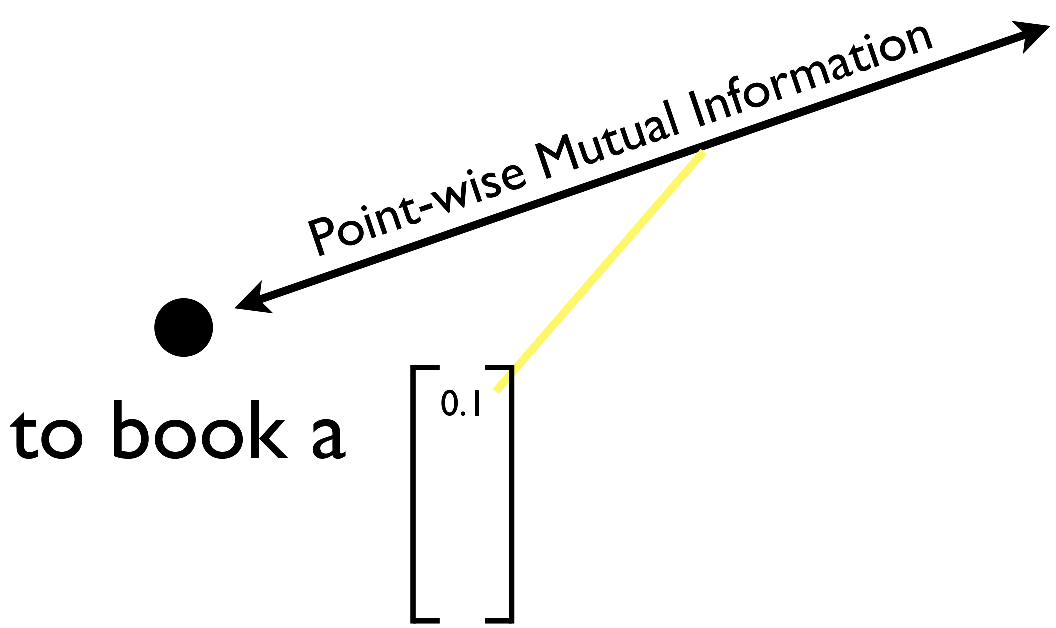


Graph Construction - Features

to book a [ ]

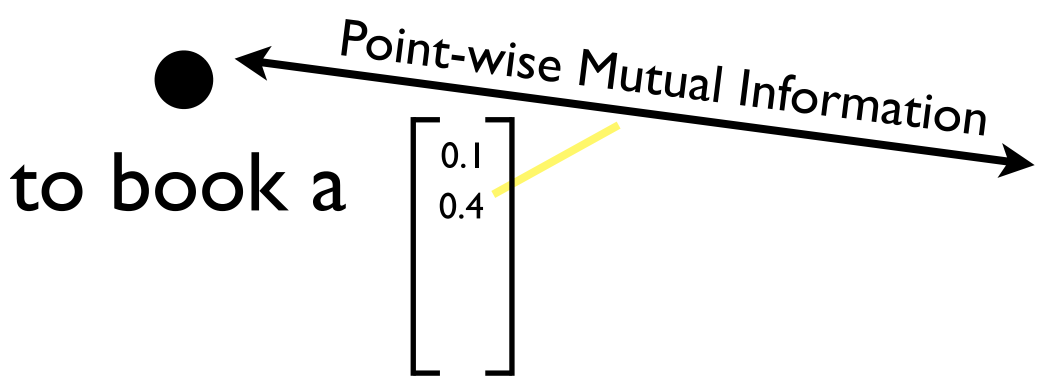
Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Graph Construction - Features



Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Graph Construction - Features



Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Graph Construction - Features

to book a $\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}$

Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Similarity Function

to book a $\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}$

Similarity Function

you unrar a

to book a $\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}$

Similarity Function

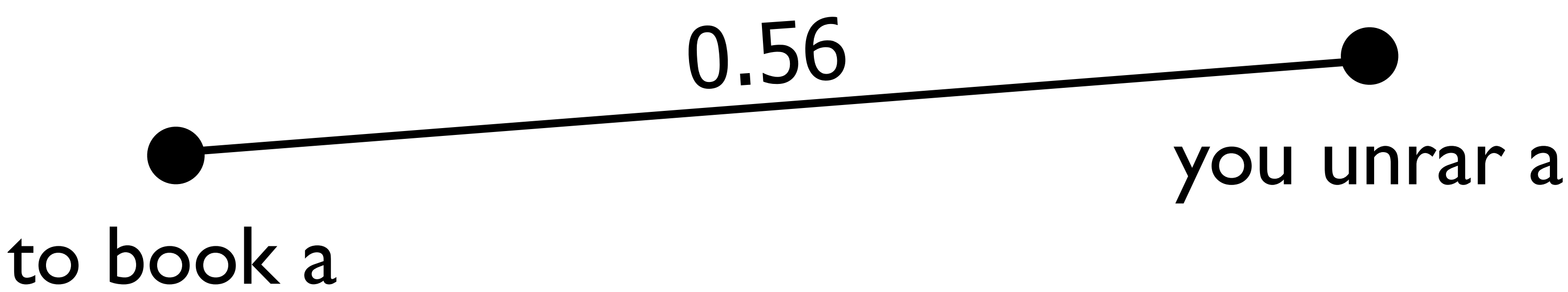
Similarity Function

to book a

you unrar a

$$1 - \cos\left(\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}, \begin{bmatrix} 0.2 \\ 0.3 \\ \vdots \end{bmatrix}\right)$$

Similarity Function



$$1 - \mathbf{cos}\left(\begin{bmatrix}0.1\\0.4\\\vdots\end{bmatrix}, \begin{bmatrix}0.2\\0.3\\\vdots\end{bmatrix}\right) = 0.56$$

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

can you book a day room at hilton hawaiian village ?

how to unrar a zipped file ?

how to get a book agent ?

how do you book a flight to multiple cities ?

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

CRF

can you book a day room at hilton hawaiian village ?

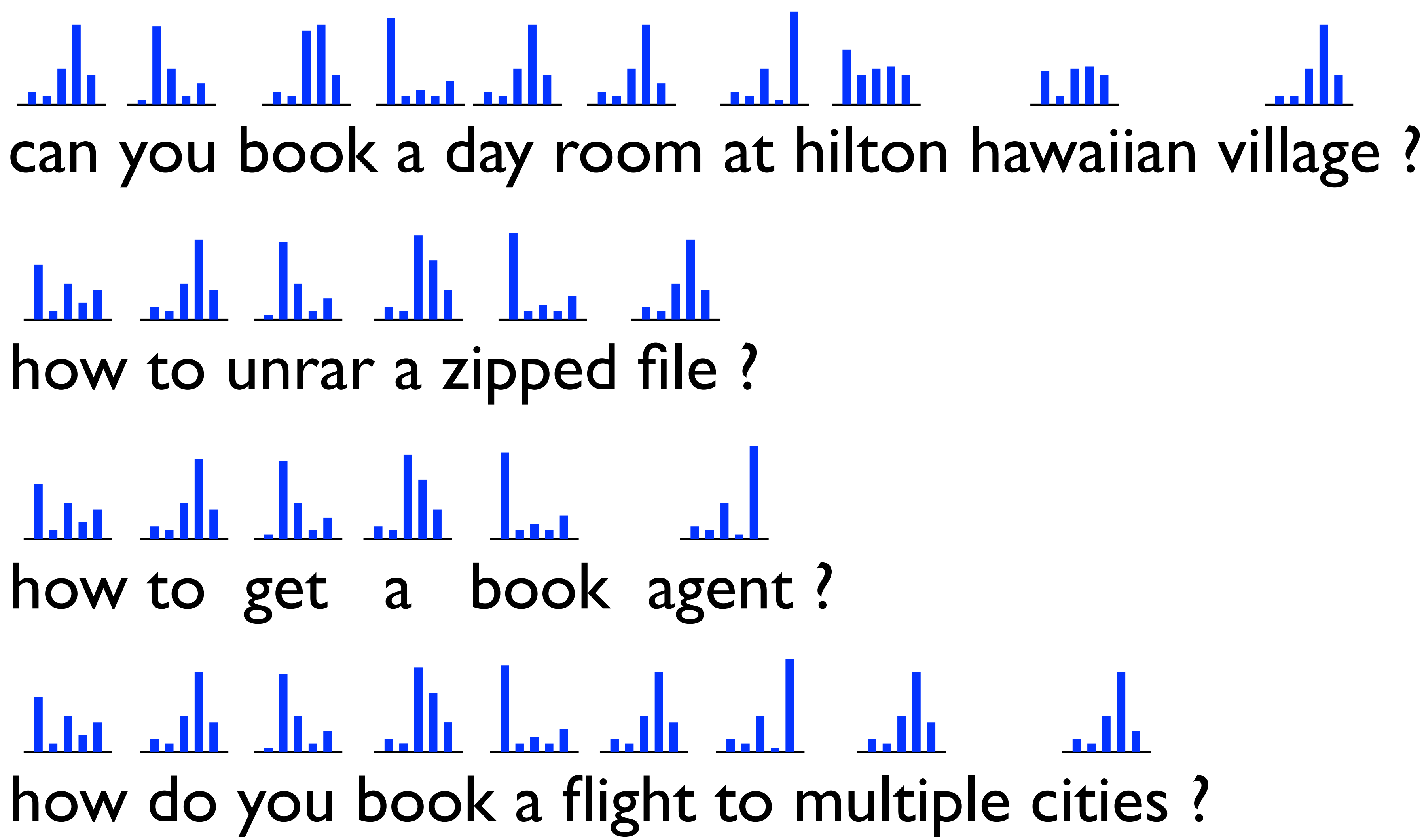
how to unrar a zipped file ?

how to get a book agent ?

how do you book a flight to multiple cities ?

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

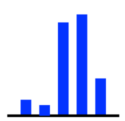


Approach (II)

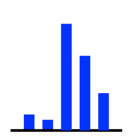
1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)

Approach (II)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)



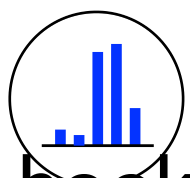
can you book a day room at hilton hawaiian village ?



how do you book a flight to multiple cities ?

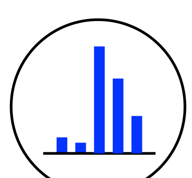
Approach (II)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)



can you book a day room at hilton hawaiian village ?

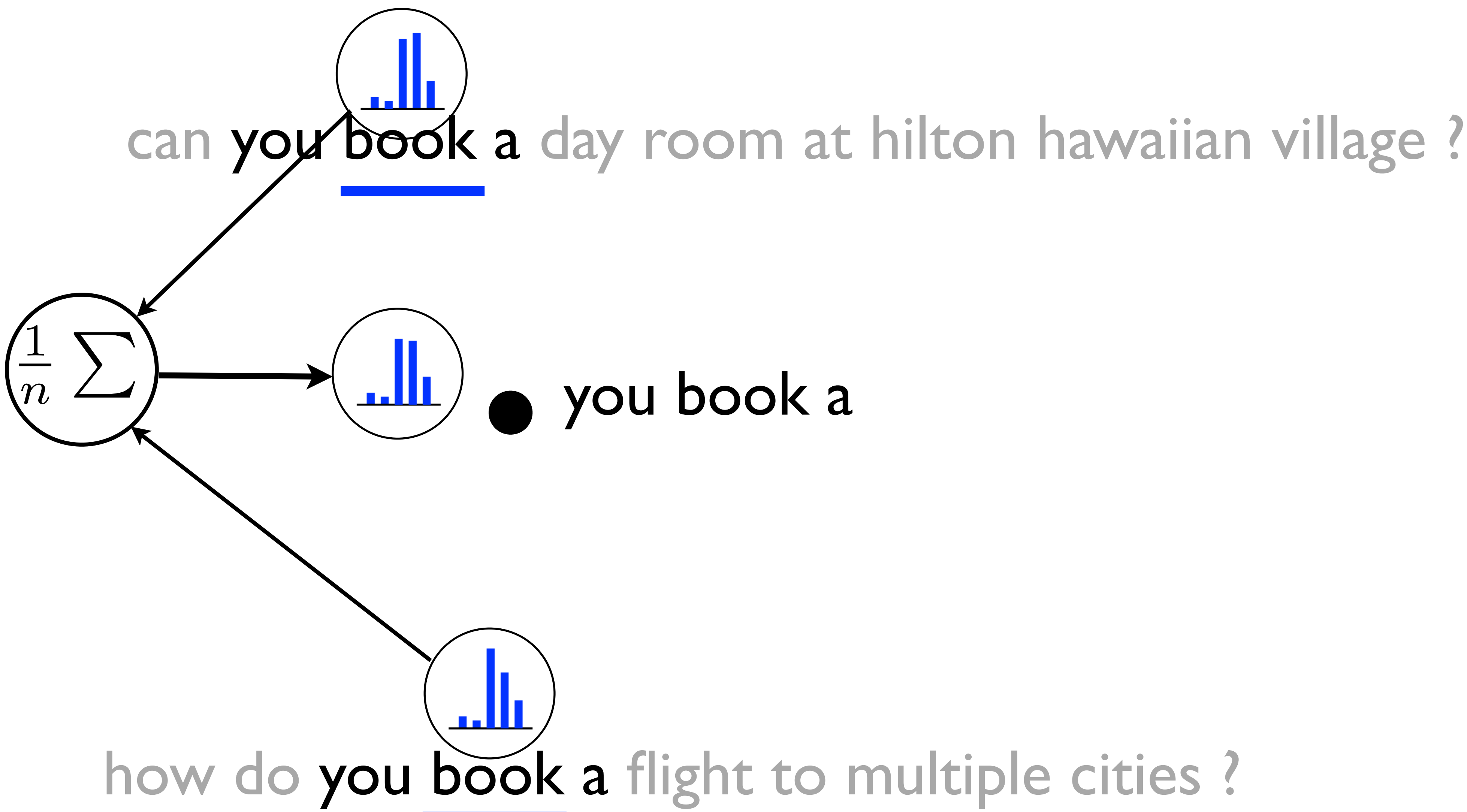
- you book a



how do you book a flight to multiple cities ?

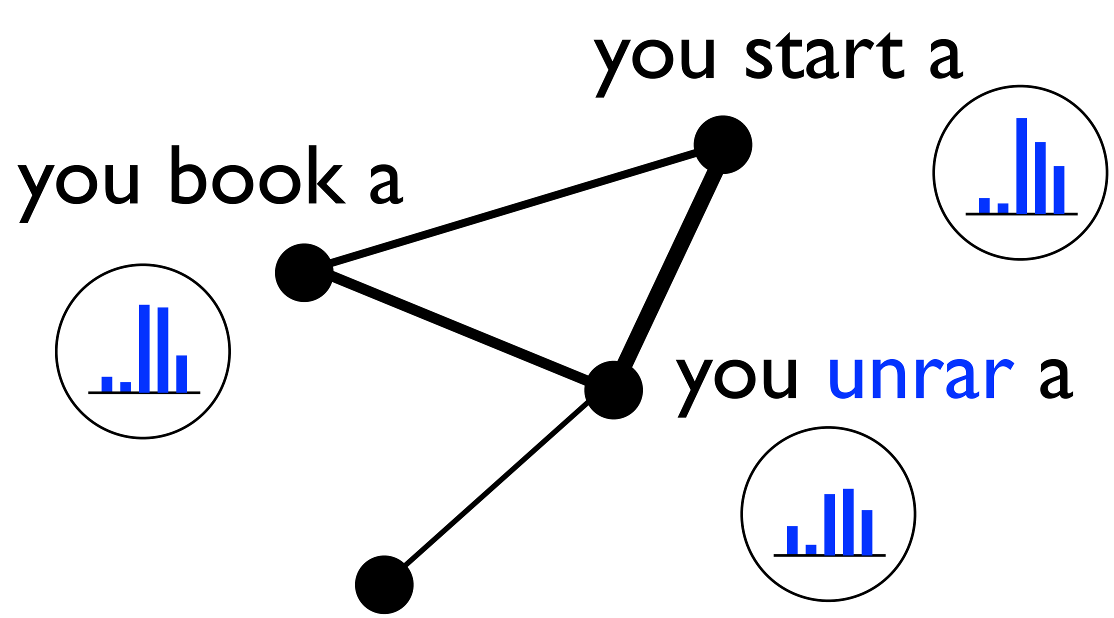
Approach (II)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)



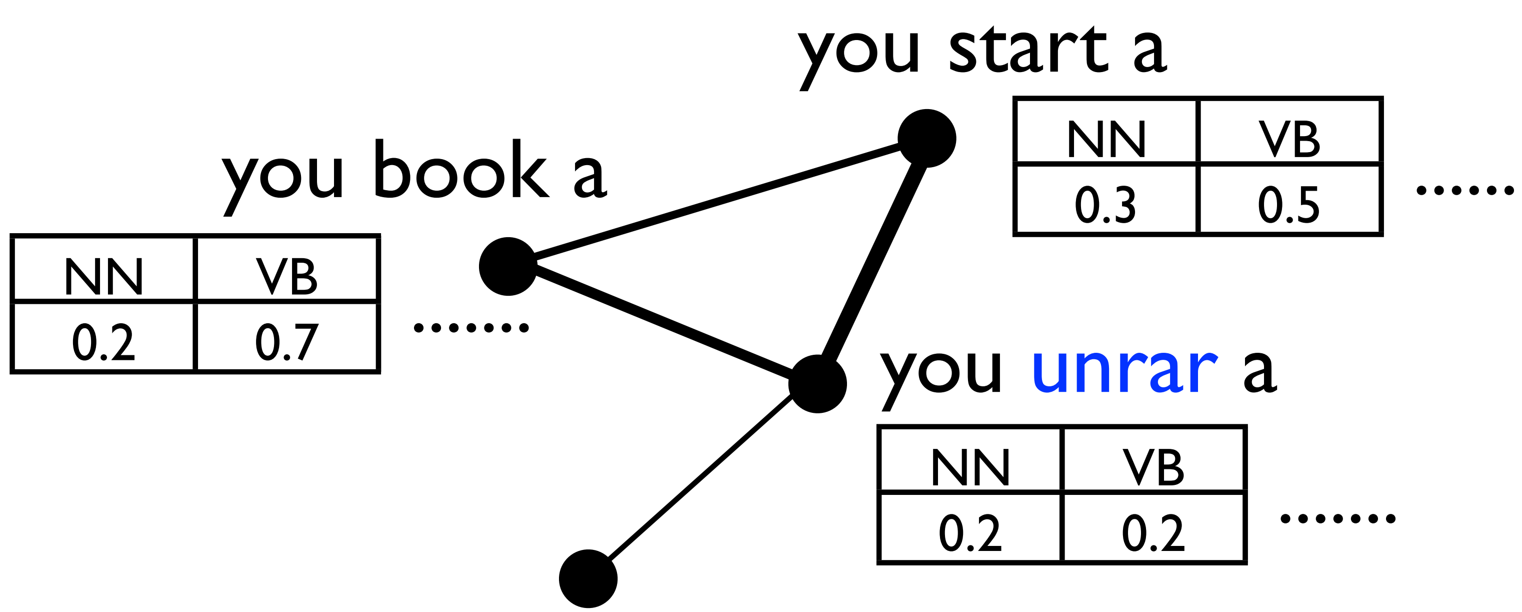
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)’
   2.3. Graph propagation



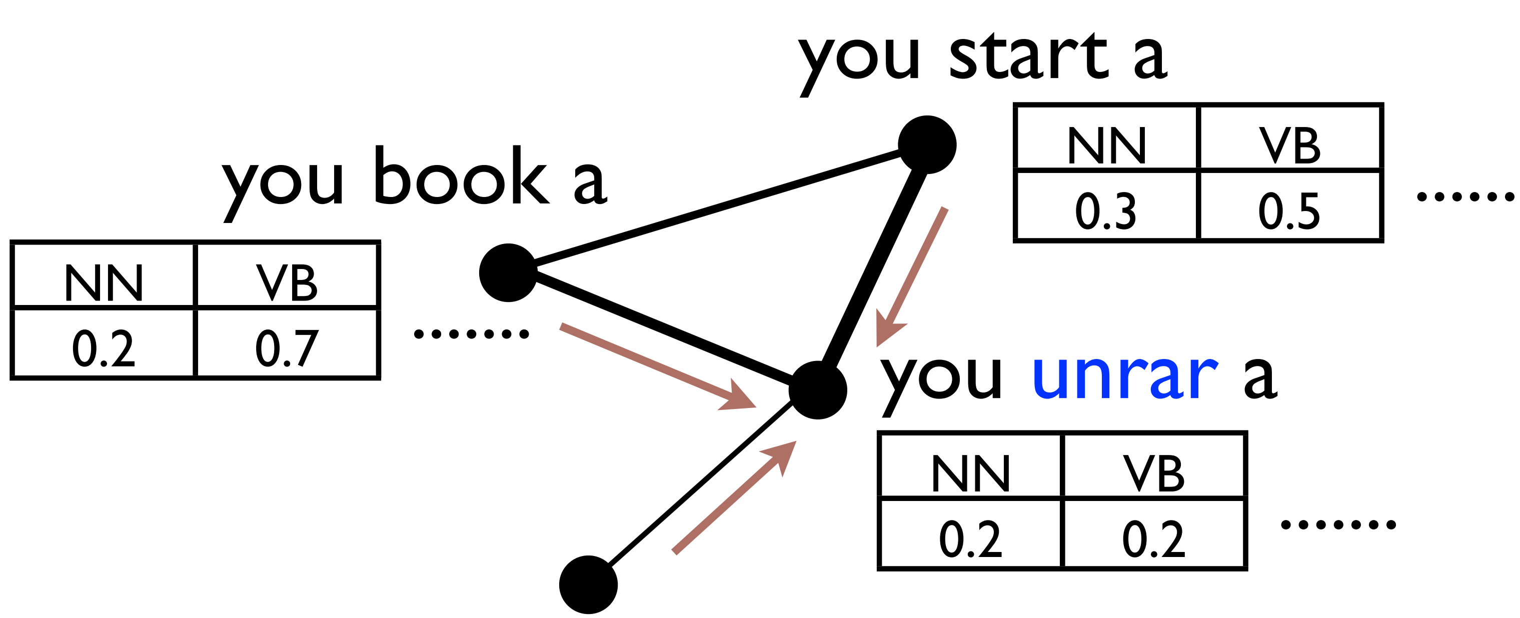
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



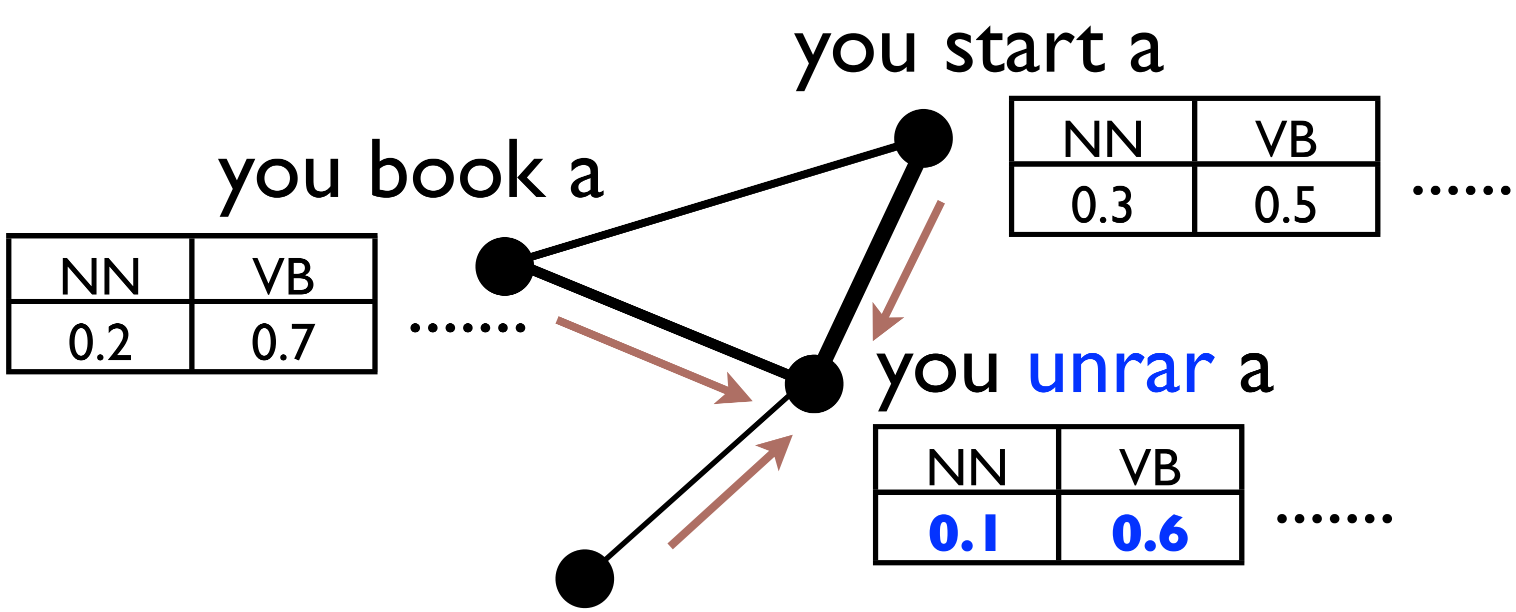
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



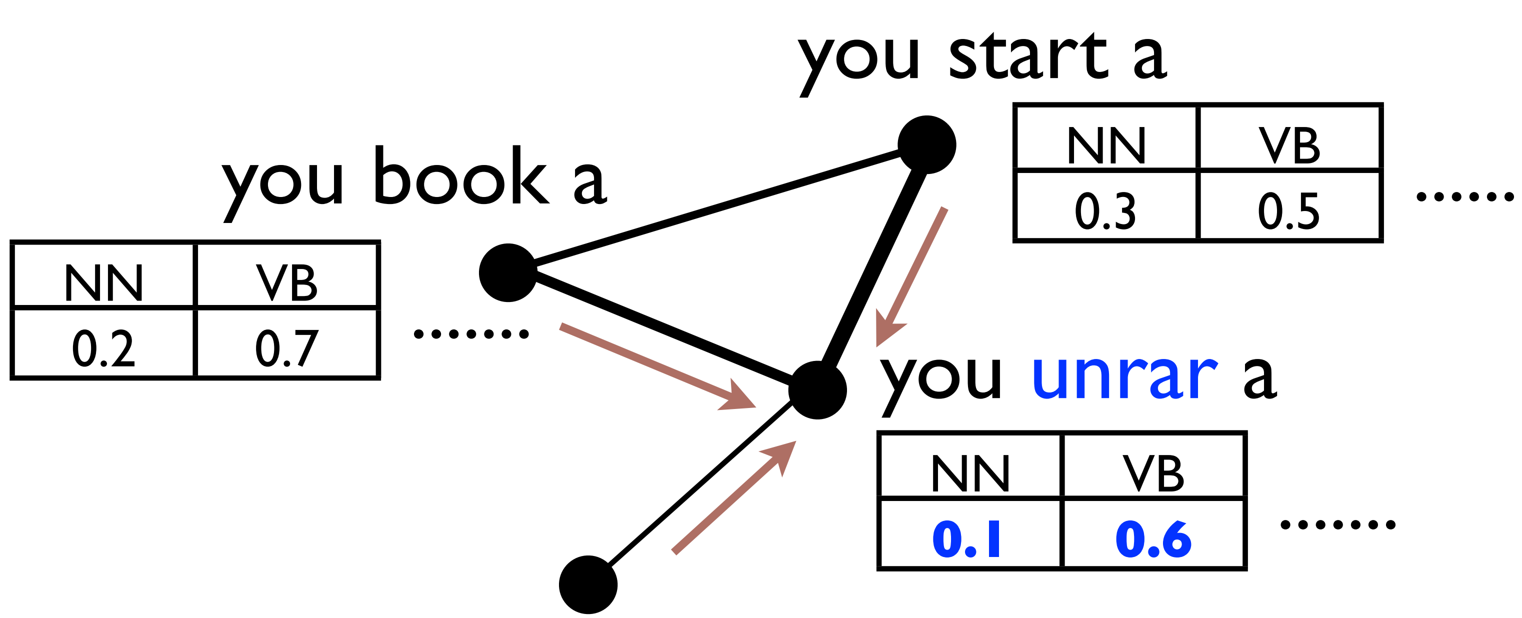
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



If two n-grams are similar according to the graph then their output distributions should be similar

Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Can you unrar a zipped file?

Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Can you unrar a zipped file?

Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Posterior Decode

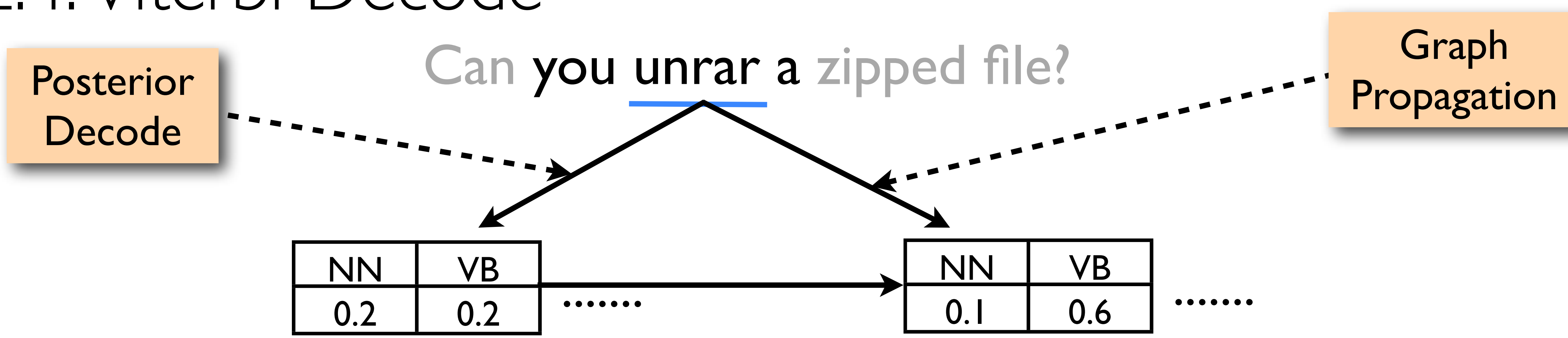
Can you unrar a zipped file?

NN	VB
0.2	0.2

....

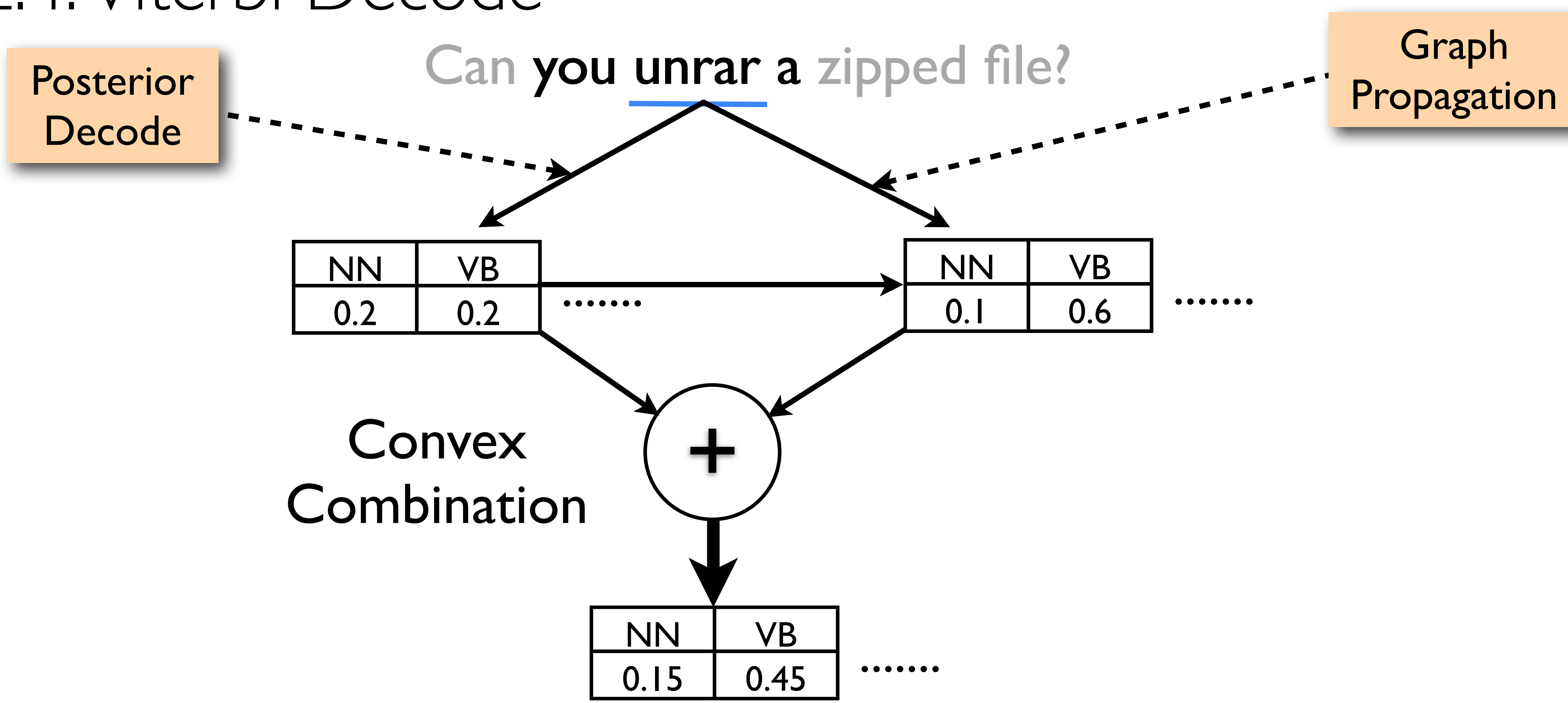
Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode



Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode



Approach (V)

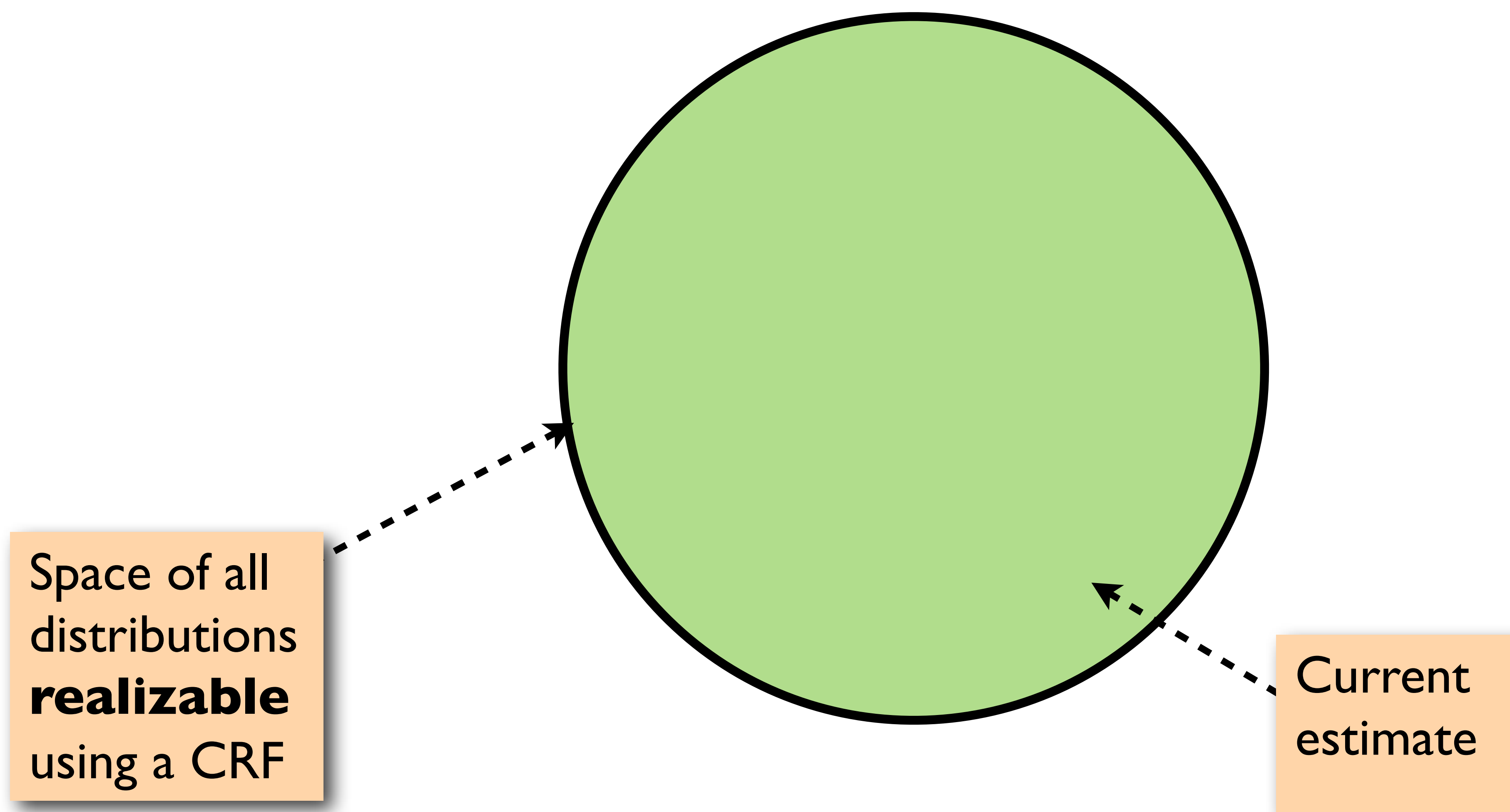
1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)’
   2.3. Graph propagation
   2.4. Viterbi Decode
   2.5. Retrain CRF on labeled & **automatically labeled** unlabeled data

Viterbi Decoding : Intuition

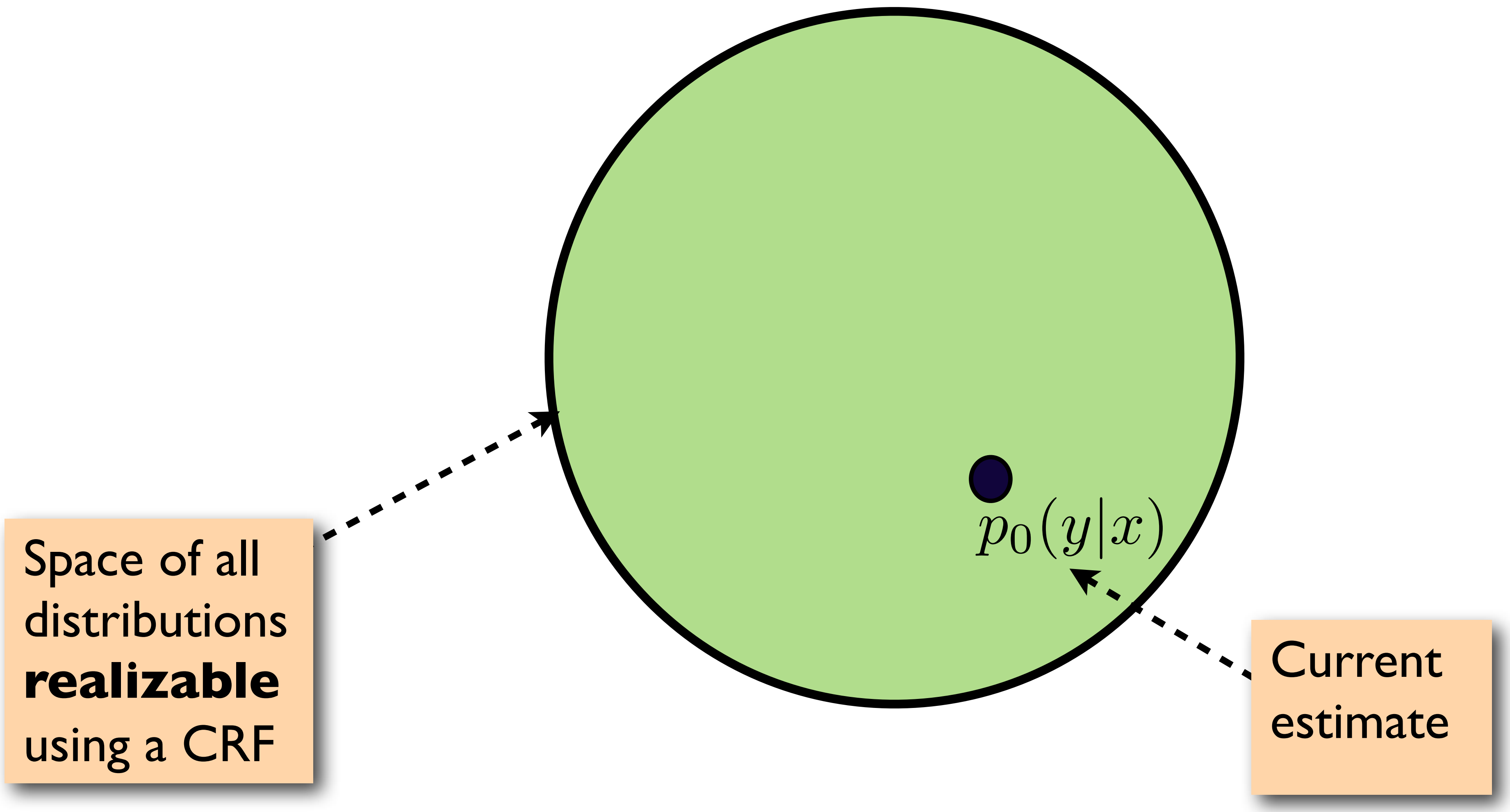
Space of all distributions **realizable** using a CRF

Current estimate

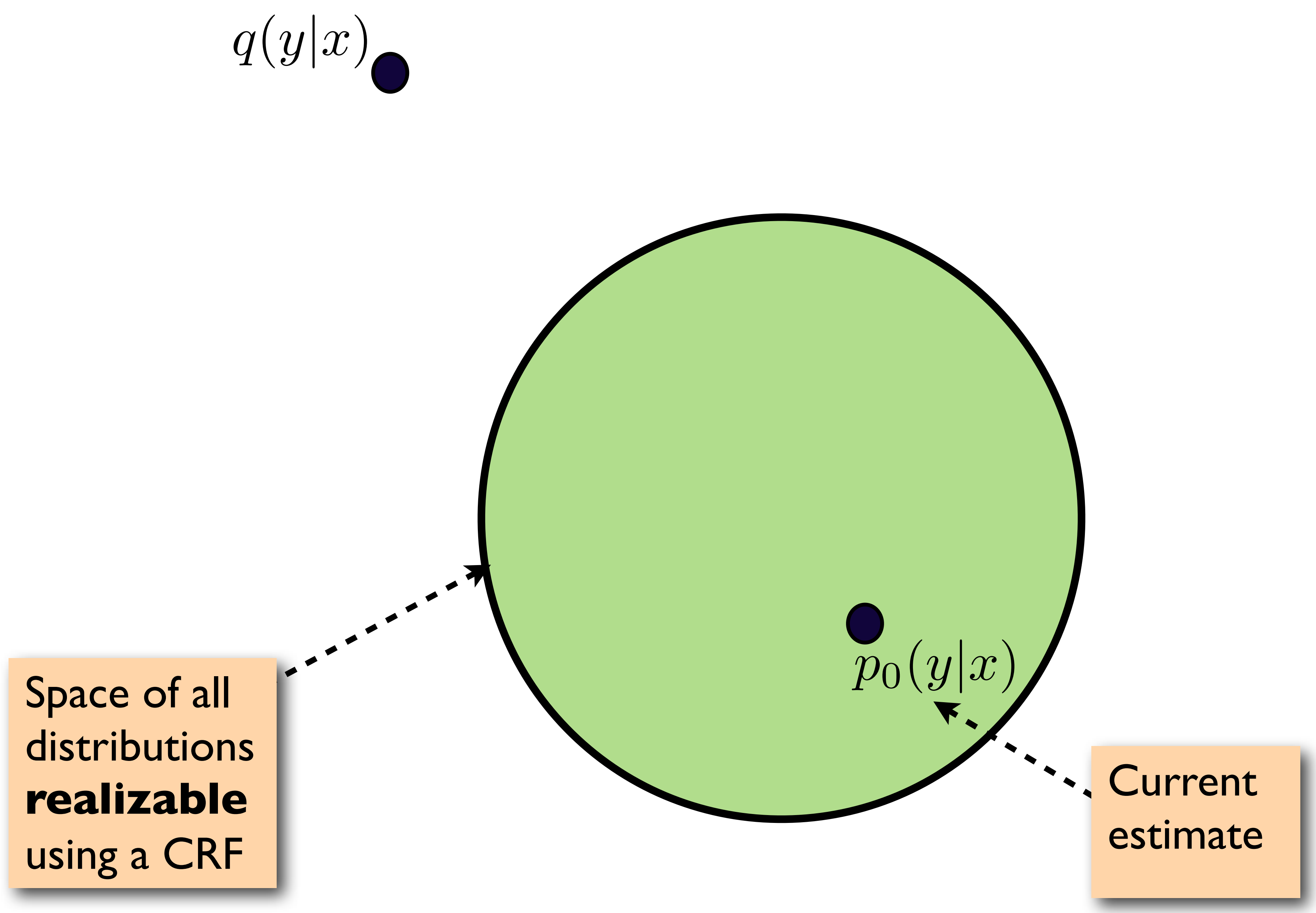
Viterbi Decoding : Intuition



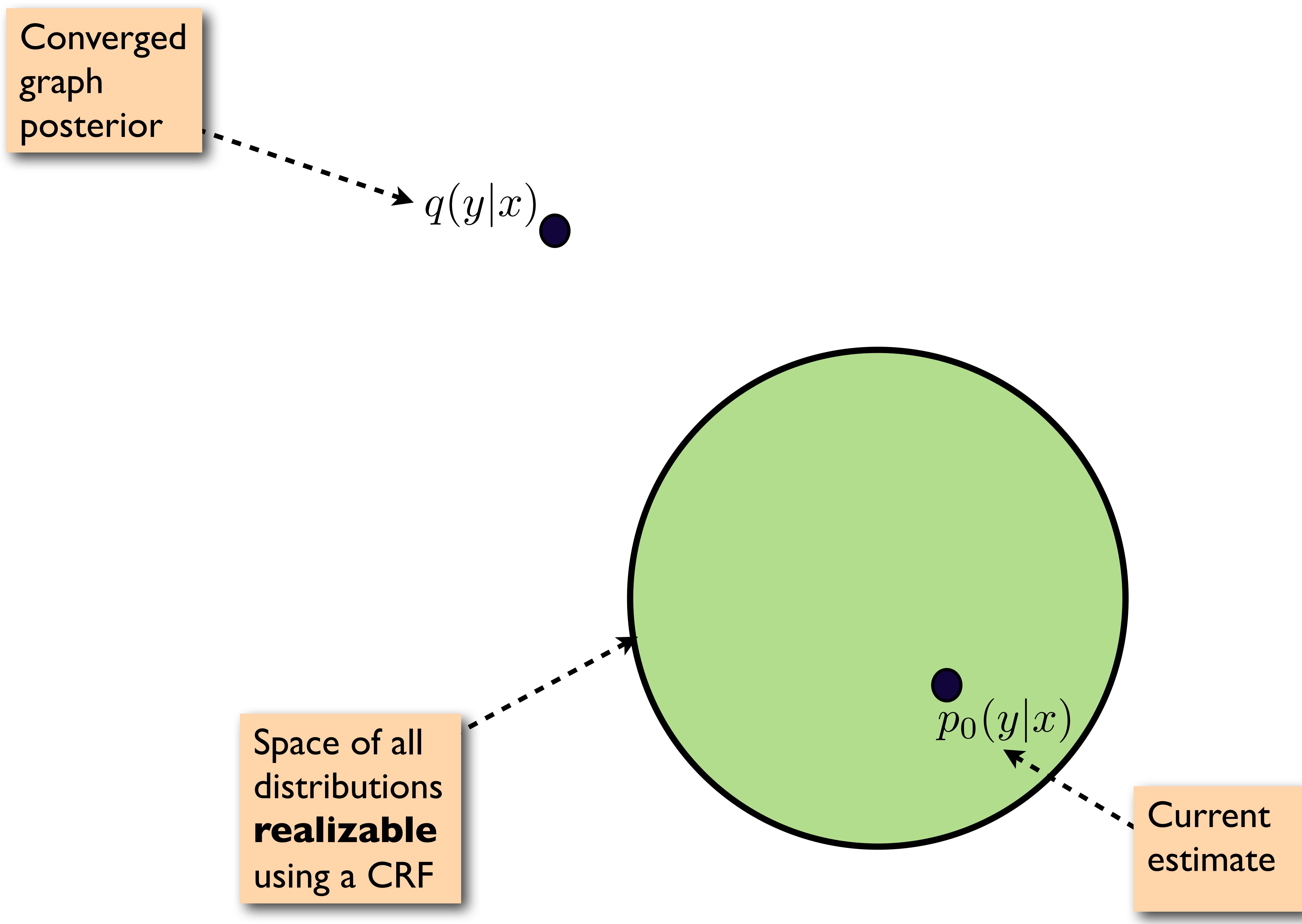
Viterbi Decoding : Intuition



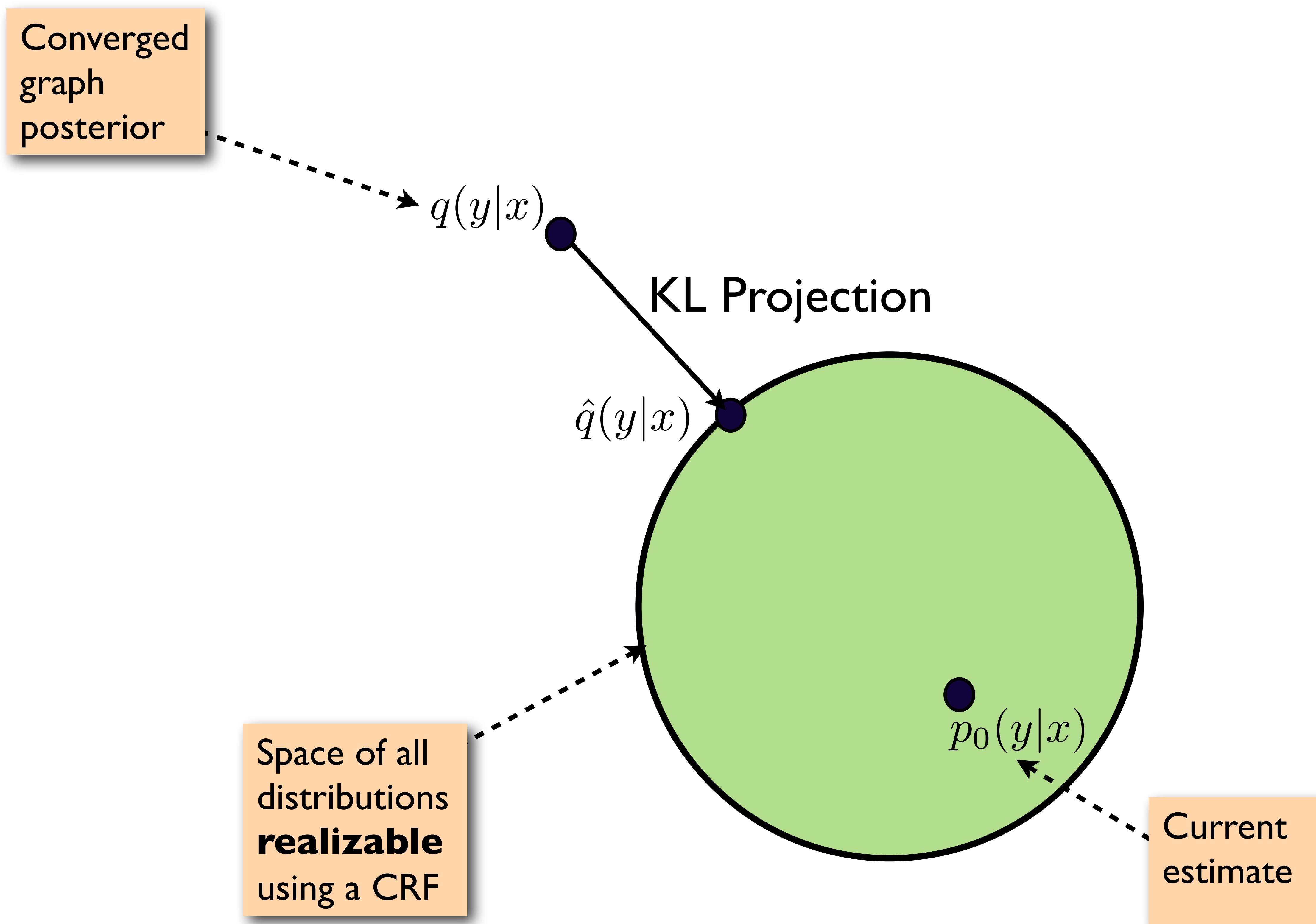
Viterbi Decoding : Intuition



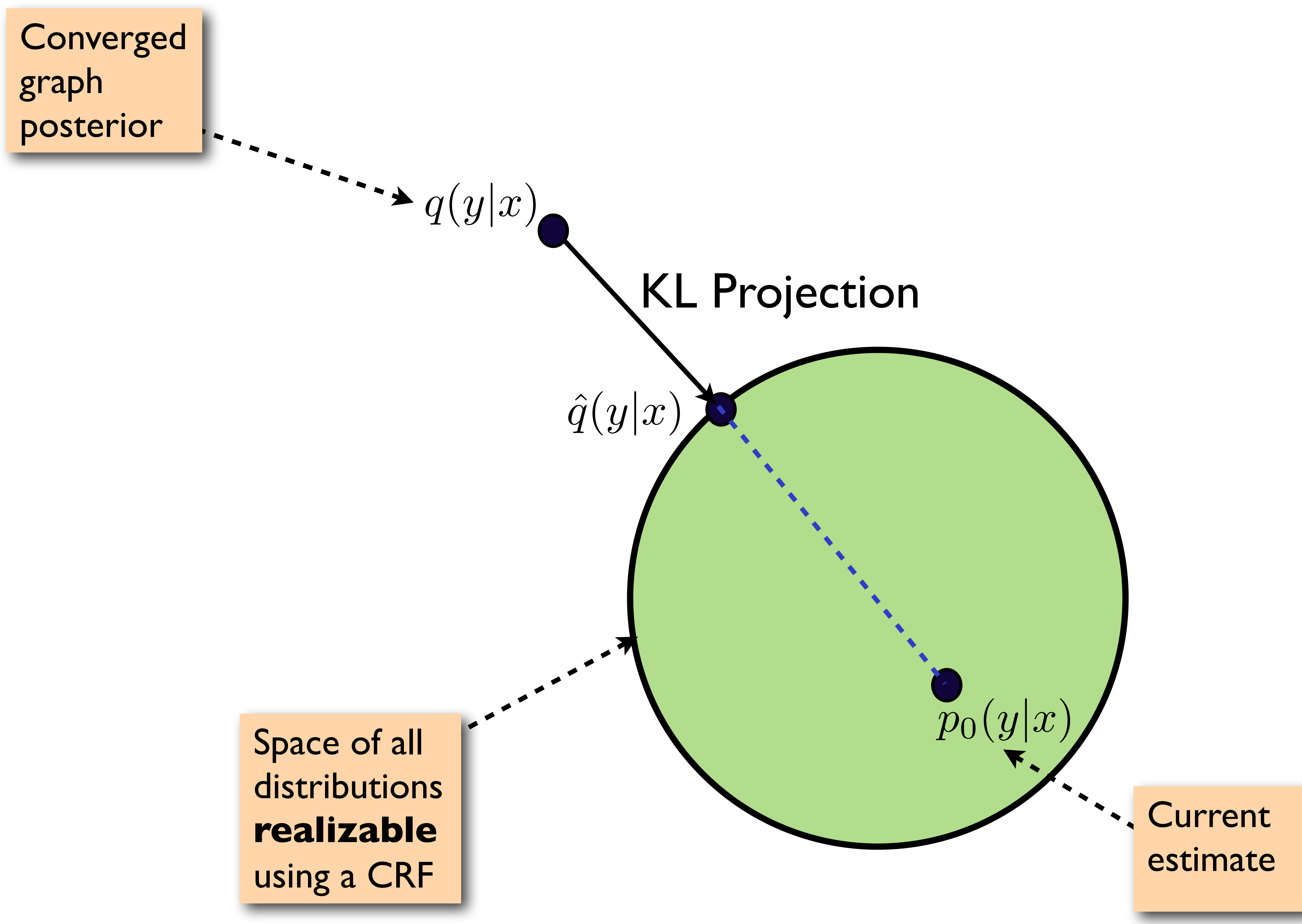
Viterbi Decoding : Intuition



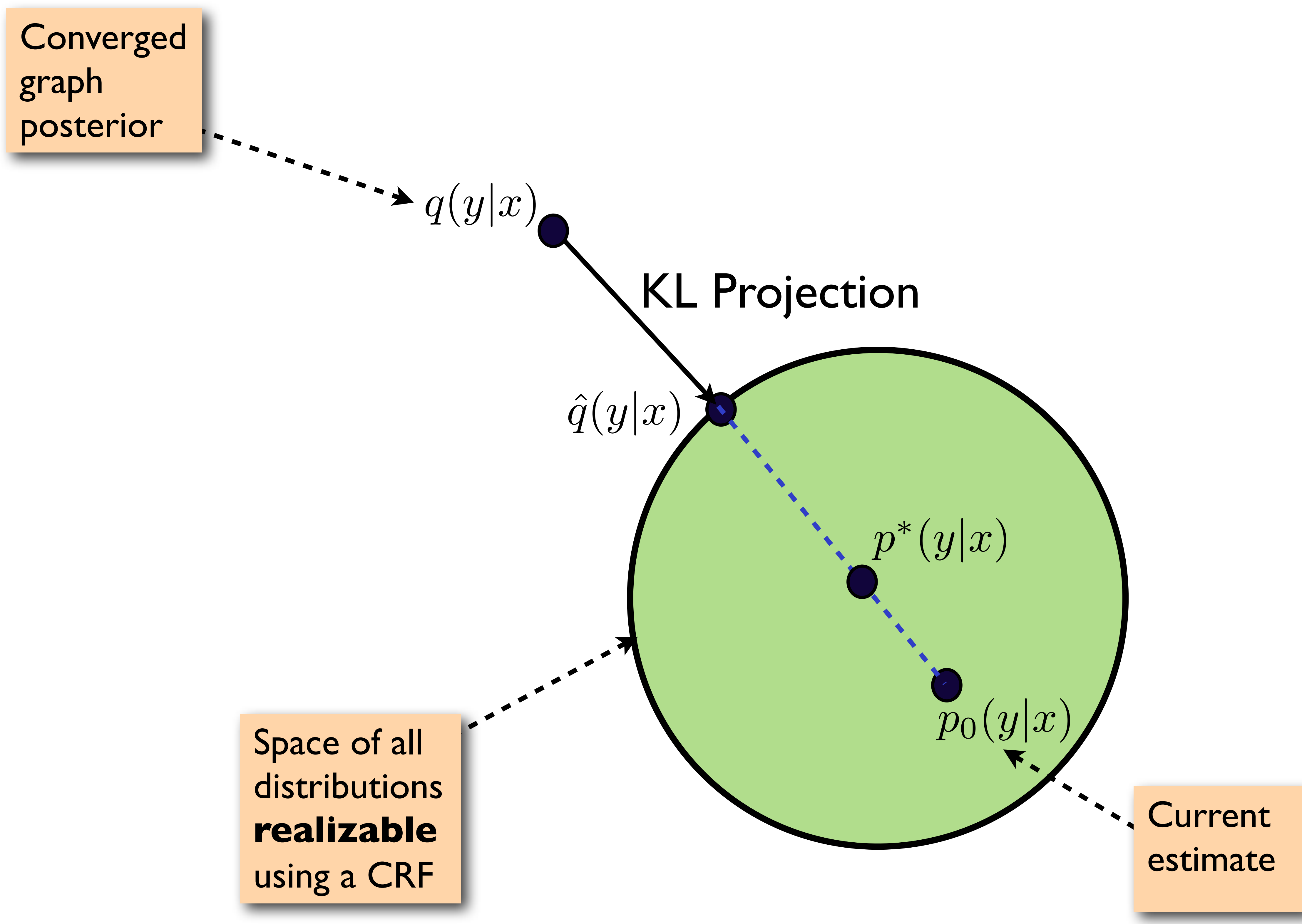
Viterbi Decoding : Intuition



Viterbi Decoding : Intuition



Viterbi Decoding : Intuition

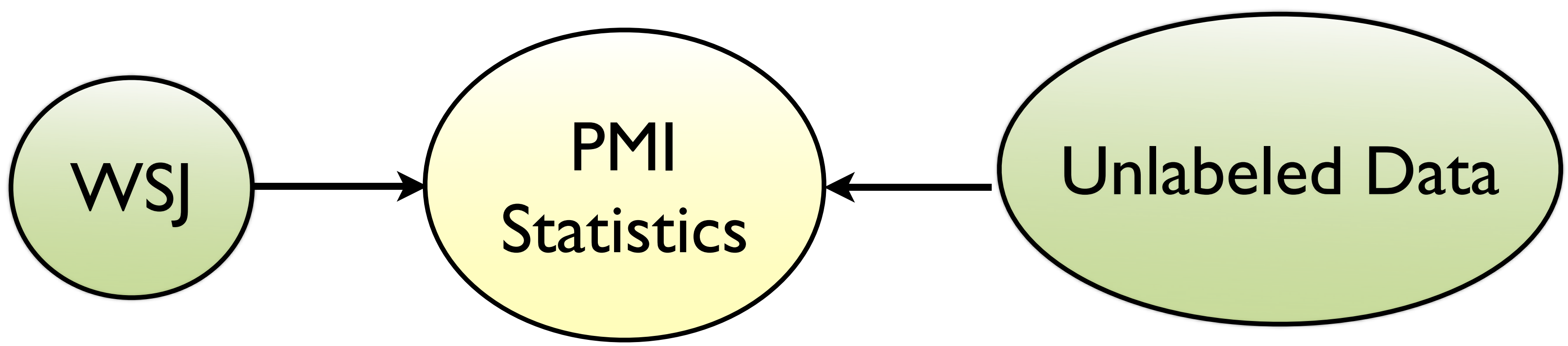


Corpora

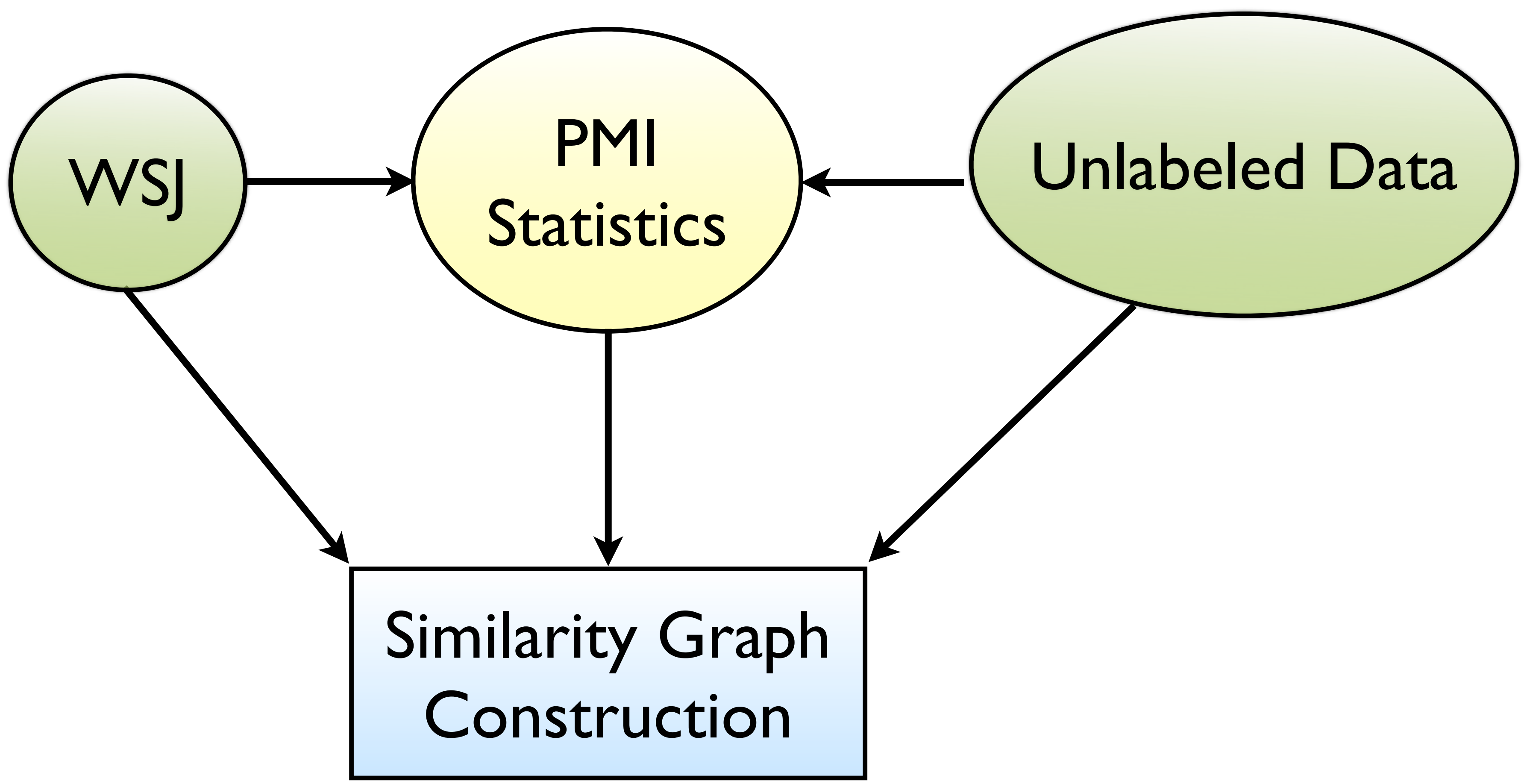
- Source Domain (labeled): Wall Street Journal (WSJ) section of the Penn Treebank.
- Target Domain:
  - QuestionBank: 4000 labeled sentences
  - Penn BioTreebank: 1061 labeled sentences

Graph Construction: Question Bank

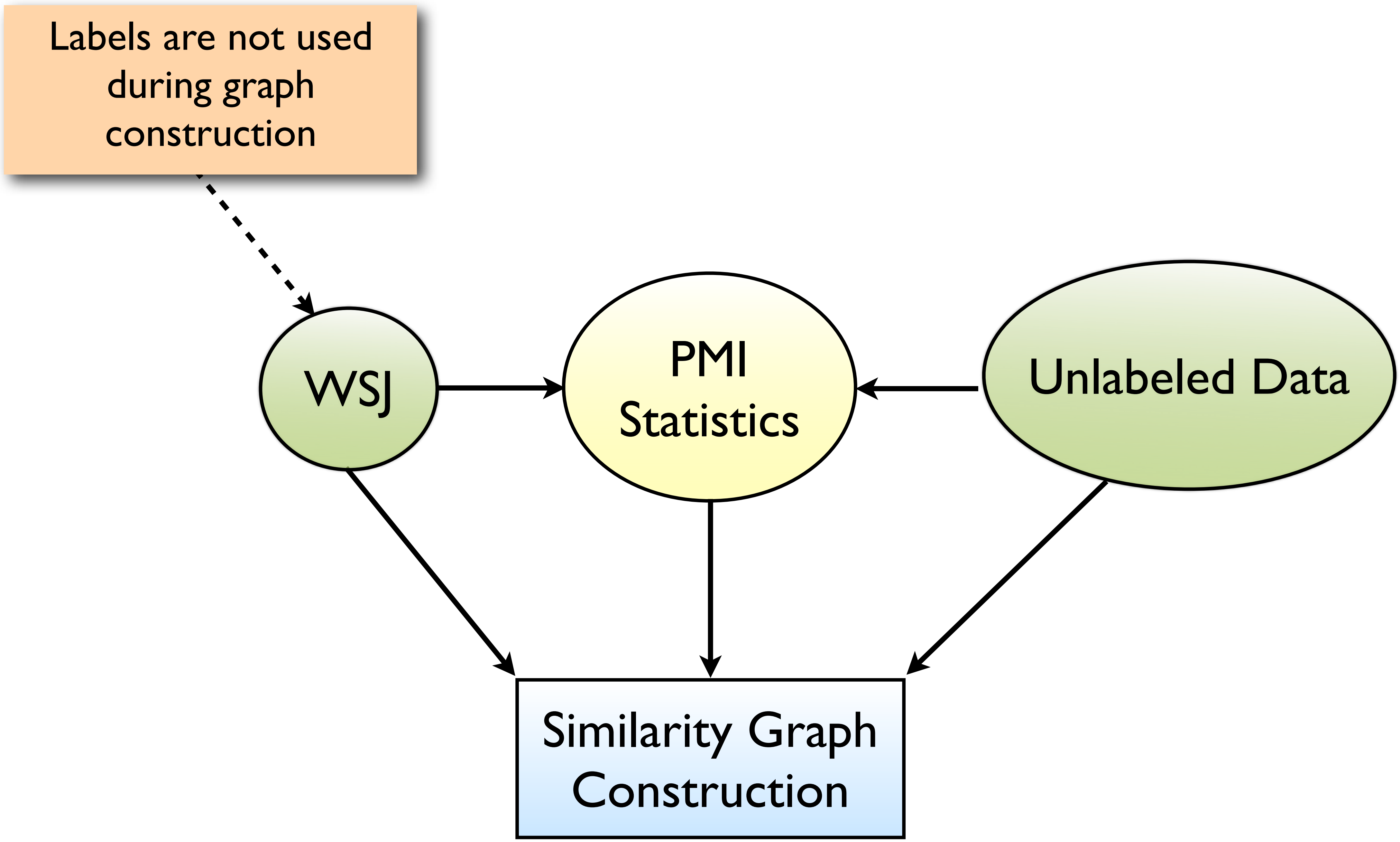
Graph Construction: Question Bank



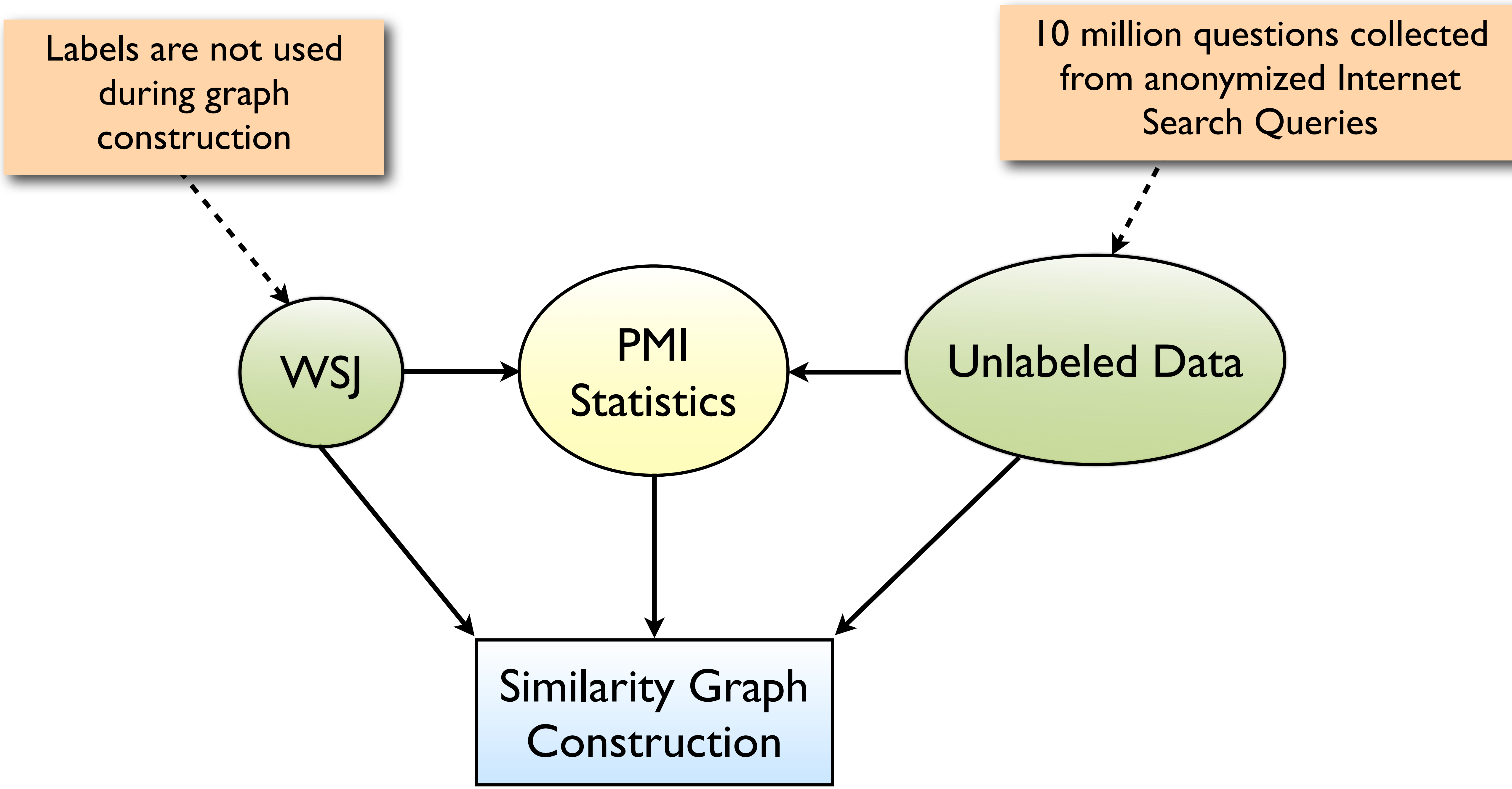
Graph Construction: Question Bank



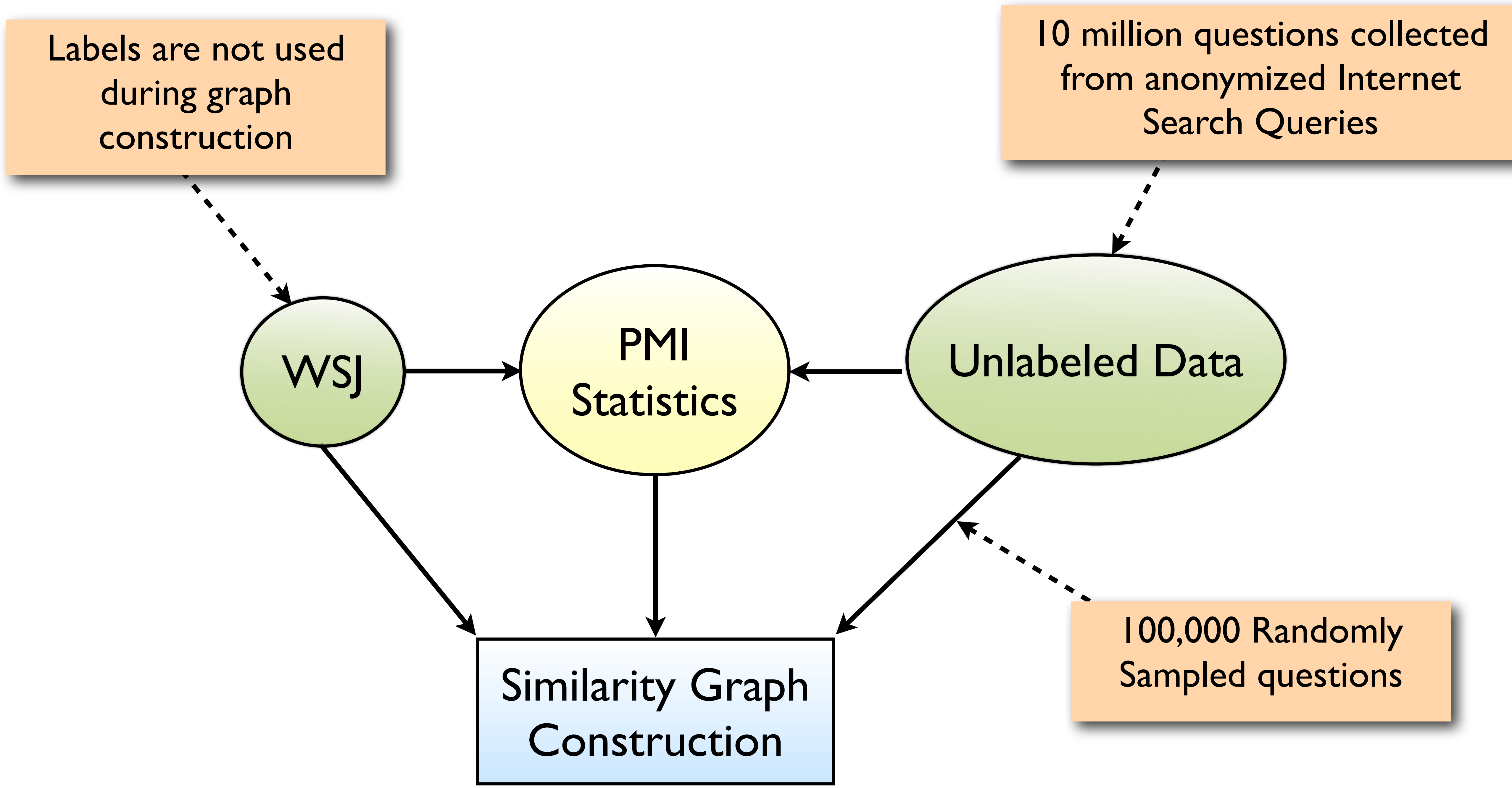
Graph Construction: Question Bank



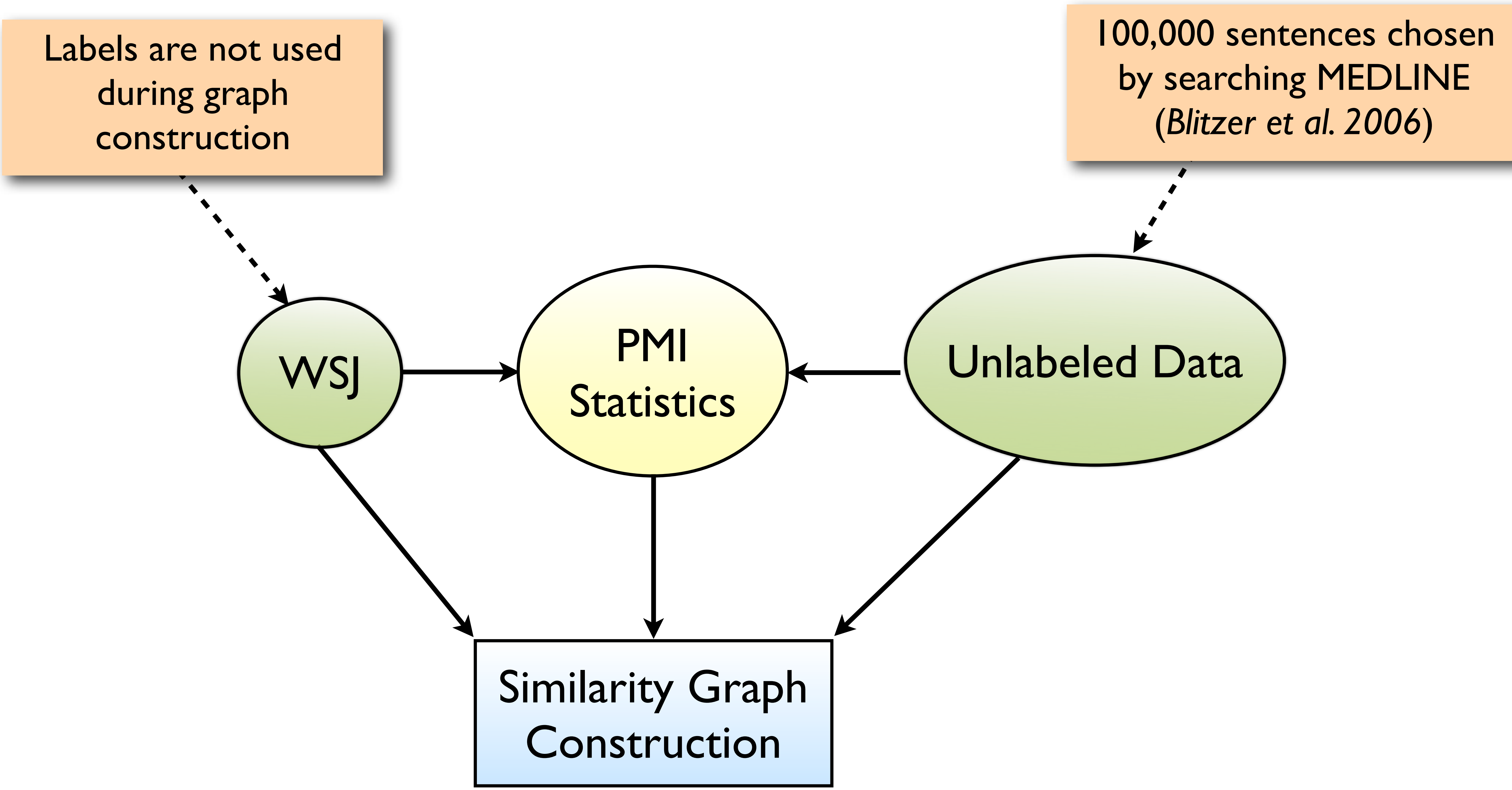
Graph Construction: Question Bank



Graph Construction: Question Bank



Graph Construction: Bio



Baseline (Supervised)

Not the same as features used using graph construction

- Features: word identity, suffixes, prefixes & special character detectors (dashes, digits, etc.).
- Achieves 97.17% accuracy on WSJ development set.

Results

	Questions	Bio
Baseline	83.8	86.2
Self-training	84.0	87.1
Semi-supervised CRF	**86.8**	**87.6**

Analysis

	Questions	Bio
percentage of unlabeled trigrams not connected to and any labeled trigram	12.4	46.8
average path length between an unlabeled trigram and its nearest labeled trigram	9.4	22.4

Analysis

Sparse Graph

	Questions	Bio
percentage of unlabeled trigrams not connected to and any labeled trigram	12.4	46.8
average path length between an unlabeled trigram and its nearest labeled trigram	9.4	22.4

Analysis

- Pros
  - Inductive
  - Produces a CRF (standard CRF inference infrastructure may be used)
- Issues
  - Graph construction
  - Graph is not integrated with CRF training

Outline

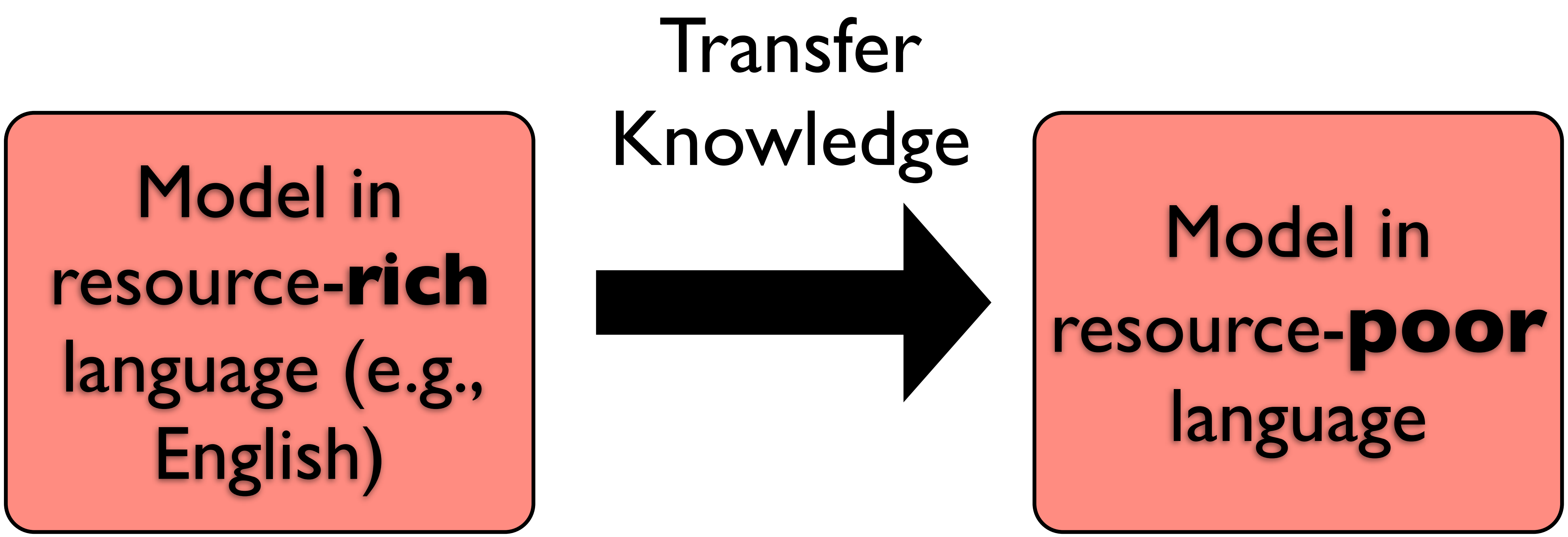
- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Text Categorization
  - Sentiment Analysis
  - Class Instance Acquisition
  - POS Tagging
  - MultiLingual POS Tagging [Das & Petrov, ACL 2011]
  - Semantic Parsing
- Conclusion & Future Work

Motivation

- Supervised POS taggers for English have accuracies in the high 90's for most domains
- By comparison taggers in other languages are not as accurate
  - Performance ranges from between 60 - 80%

Motivation

- Supervised POS taggers for English have accuracies in the high 90's for most domains
- By comparison taggers in other languages are not as accurate
  - Performance ranges from between 60 - 80%



Cross-Lingual Projection

The food at Google is good .

Cross-Lingual Projection

96% Accuracy

DET	NOUN	ADP	NOUN	VERB	ADJ	.
The	food	at	Google	is	good	.

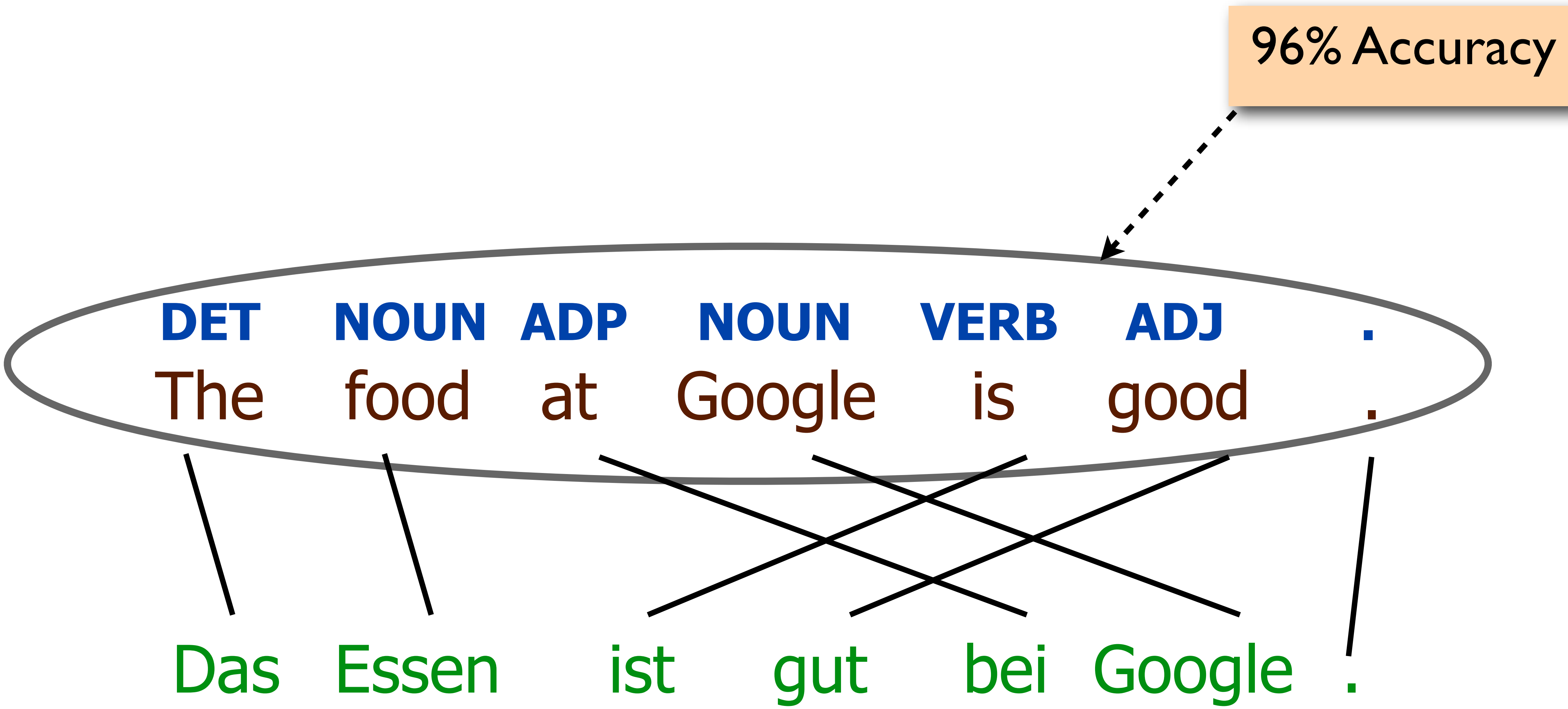
Cross-Lingual Projection

96% Accuracy

DET	NOUN	ADP	NOUN	VERB	ADJ	.
The	food	at	Google	is	good	.

Das Essen ist gut bei Google .

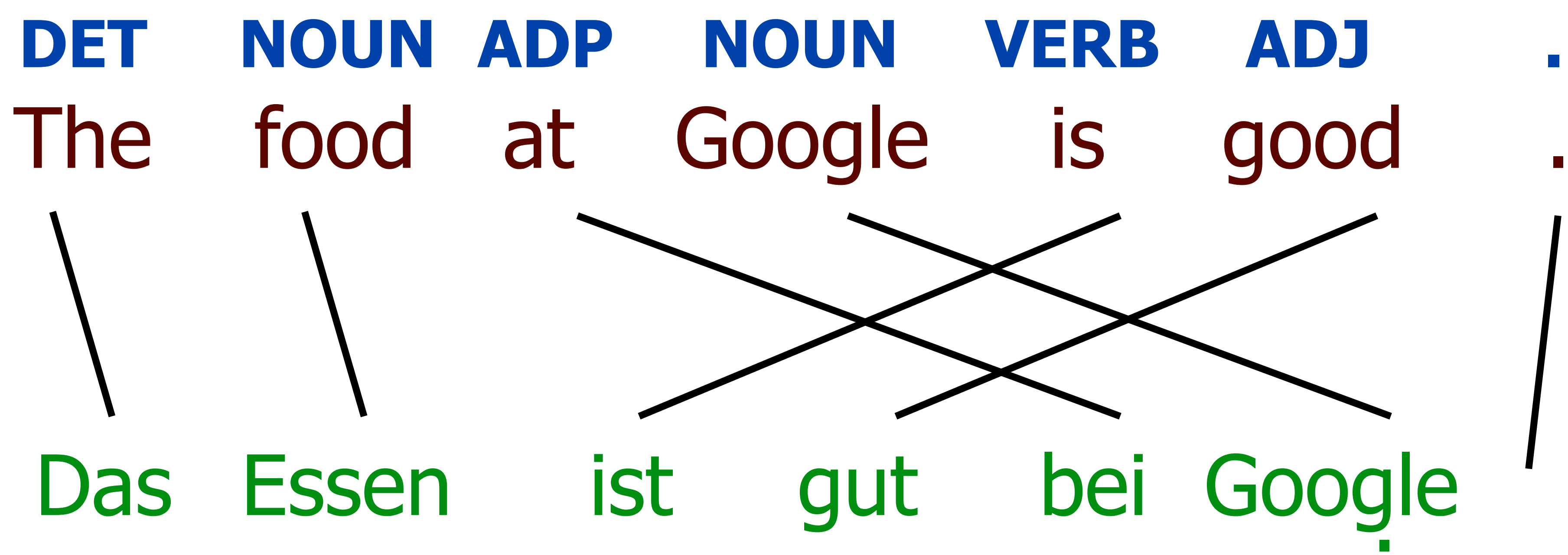
Cross-Lingual Projection



Automatic alignments from translation data

(available for more than 50 languages)

Cross-Lingual Projection



Cross-Lingual Projection

NOUN
food

DET
The

Essen

Das

VERB
is

ADJ
good

ist

gut

.

bei

Google

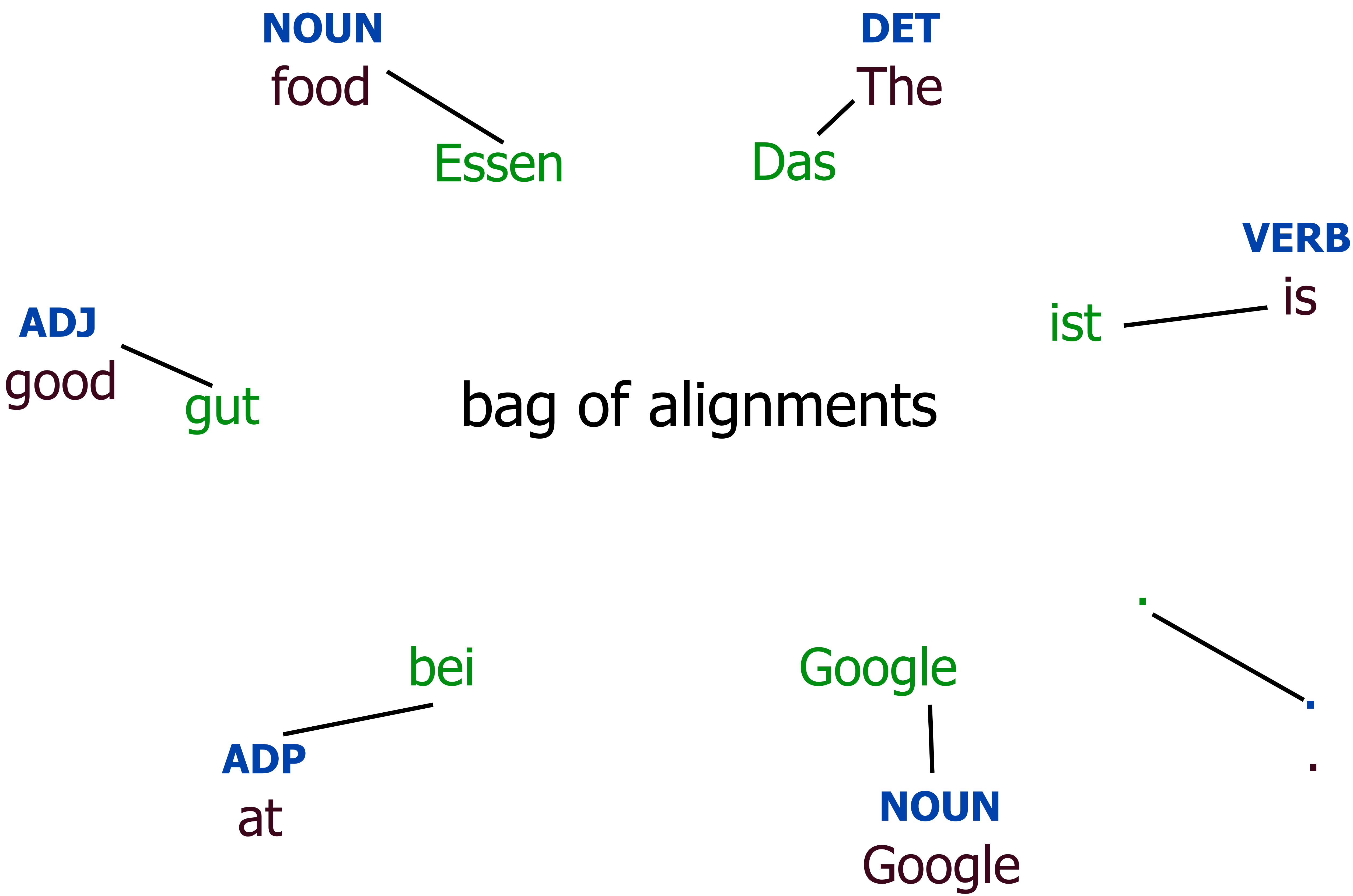
.

ADP
at

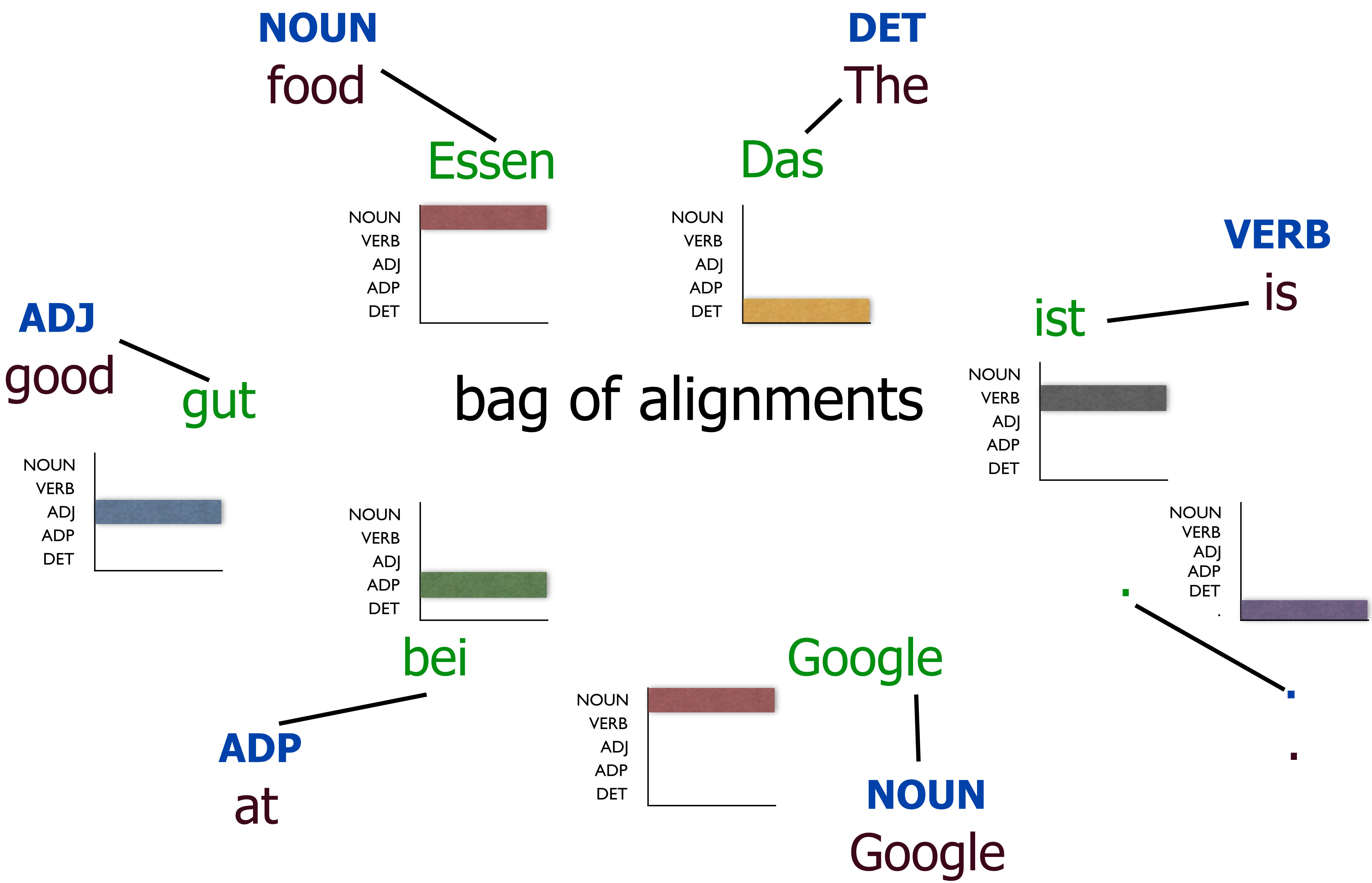
.

NOUN
Google

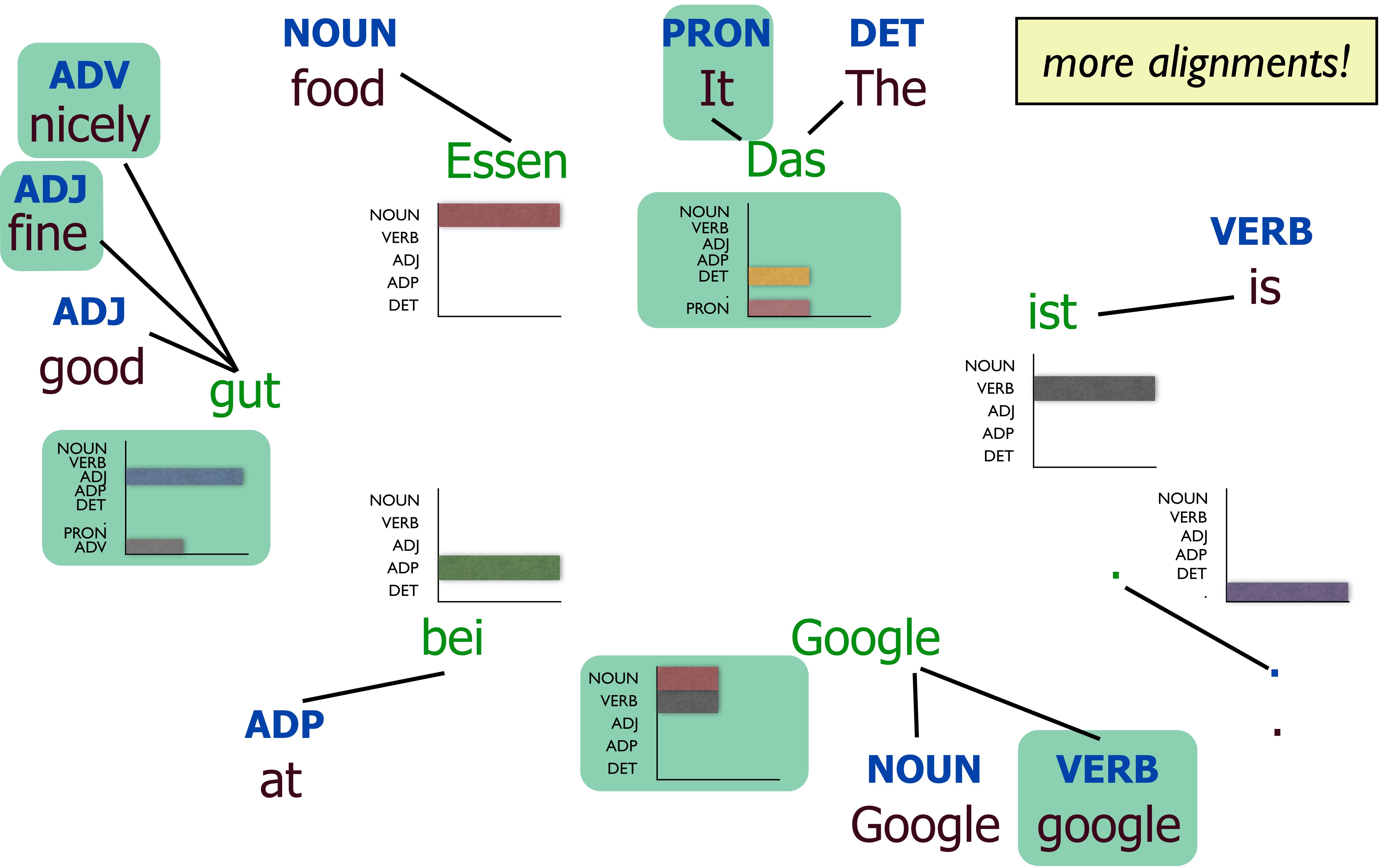
Cross-Lingual Projection



Cross-Lingual Projection



Cross-Lingual Projection



Cross-Lingual Projection Results

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0

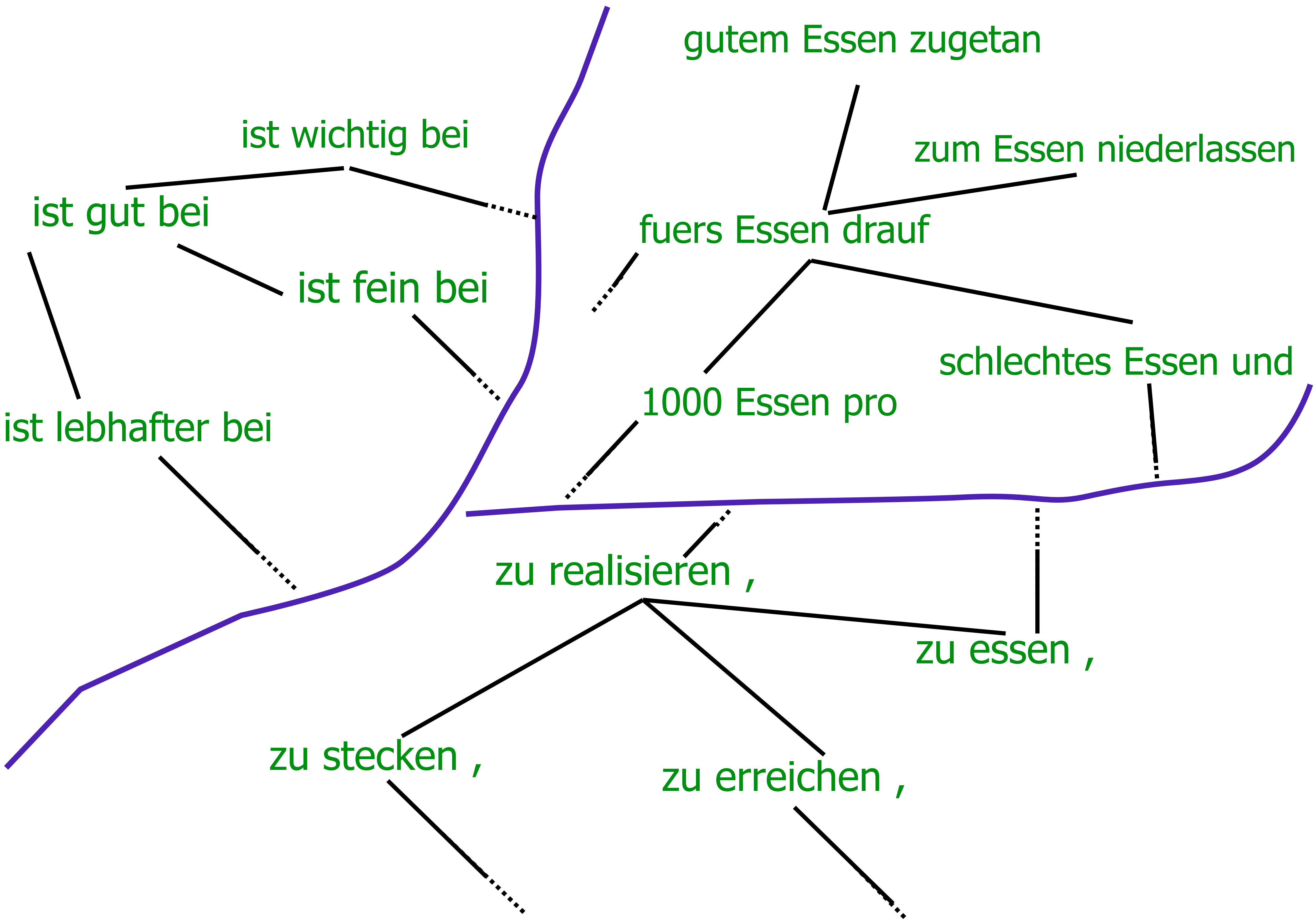
Feature-HMM [Berg-Kirkpatrick, NAACL 2010]

Cross-Lingual Projection Results

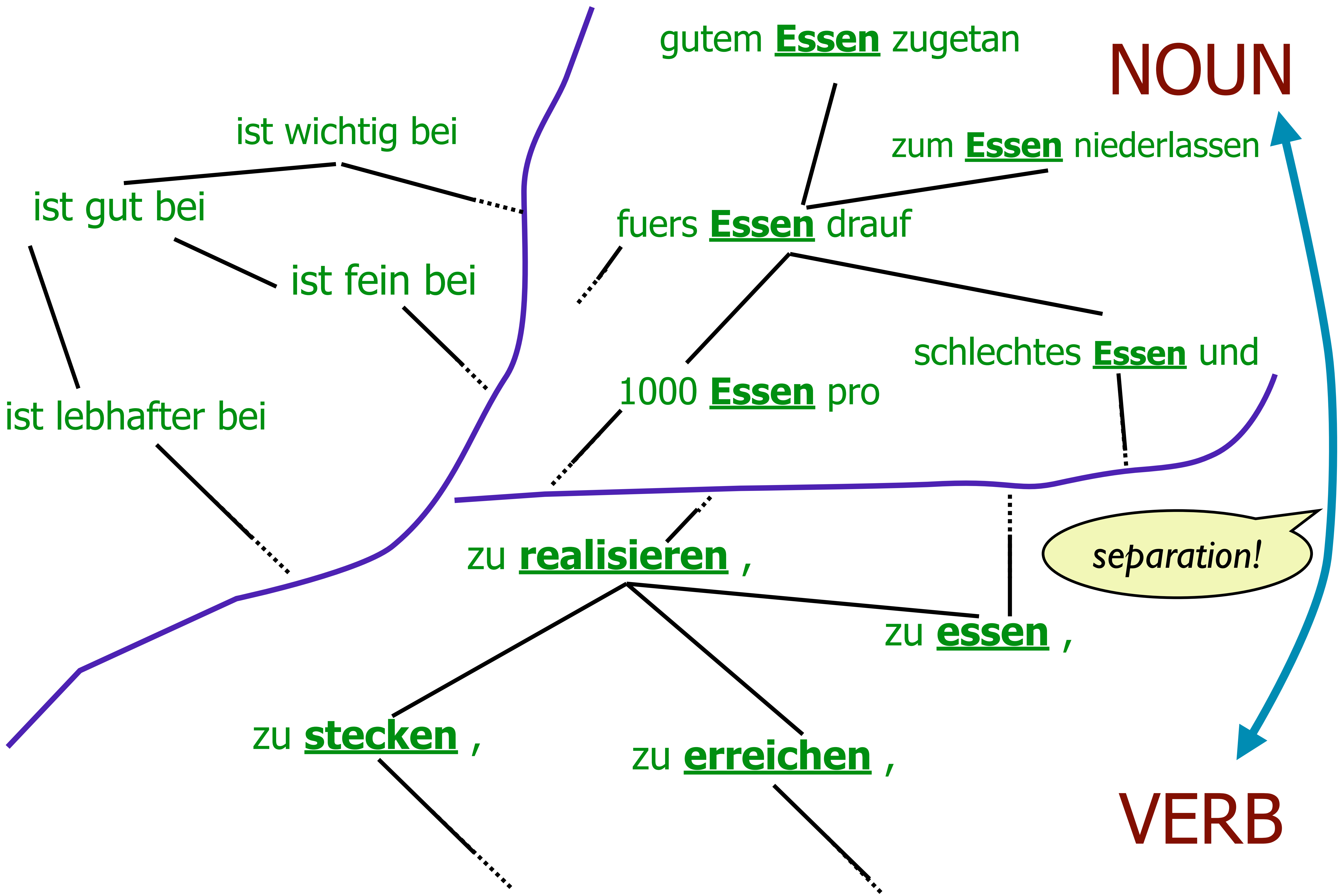
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	**73.6**	**77.0**	**83.2**	**79.3**	**79.7**	**82.6**	**80.1**	**74.7**	**78.8**

Feature-HMM [Berg-Kirkpatrick, NAACL 2010]

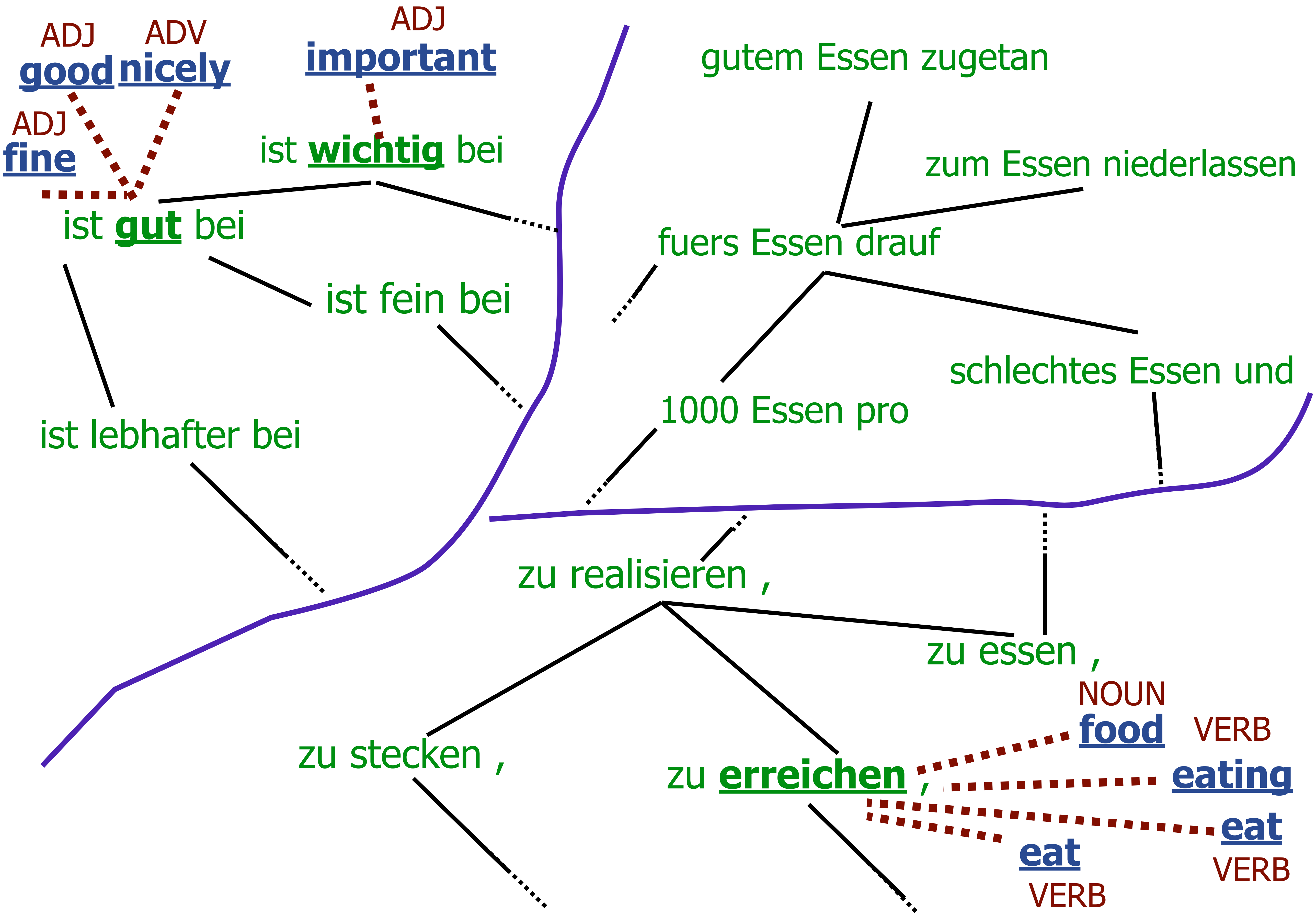
Graph Regularization



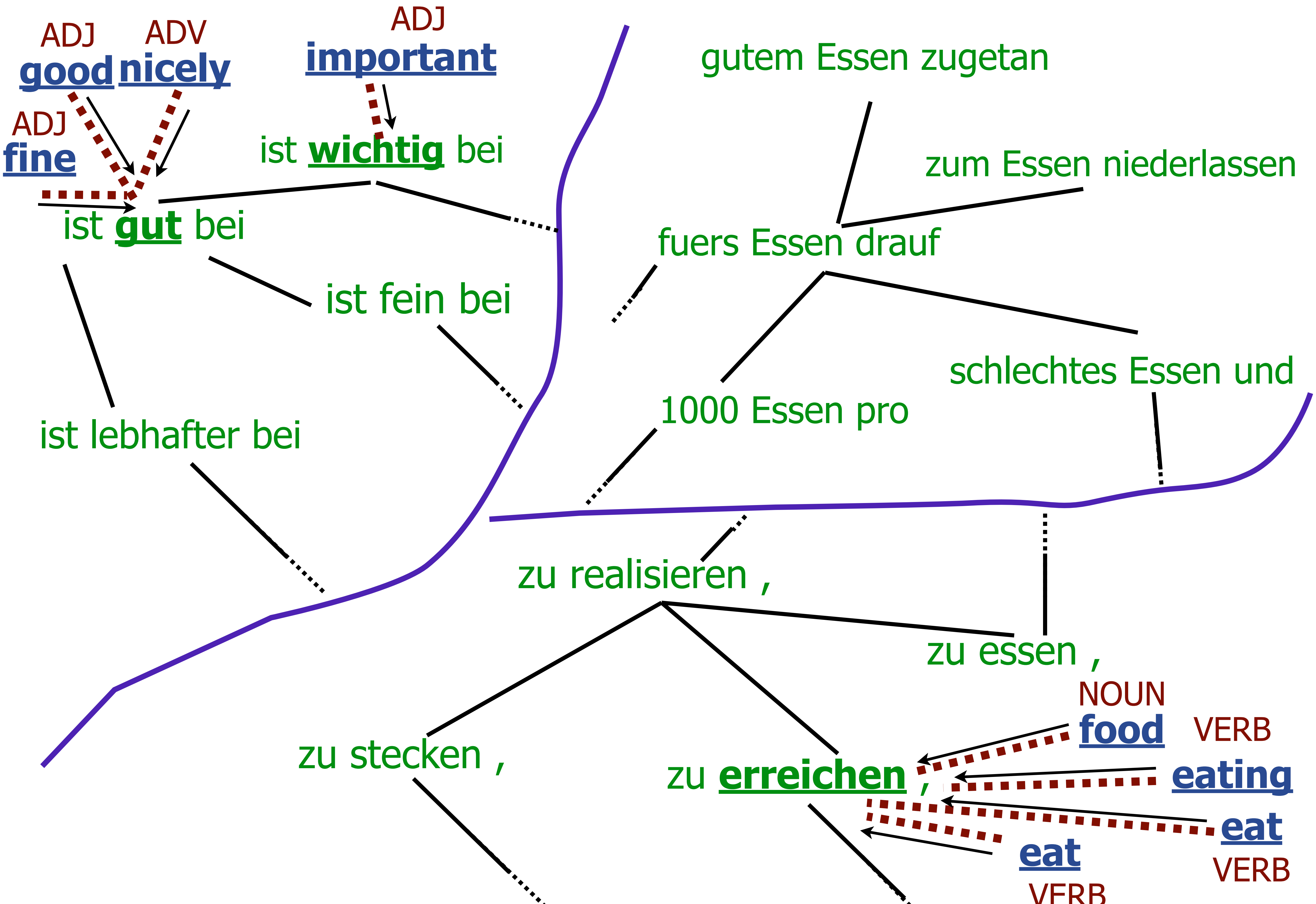
Graph Regularization



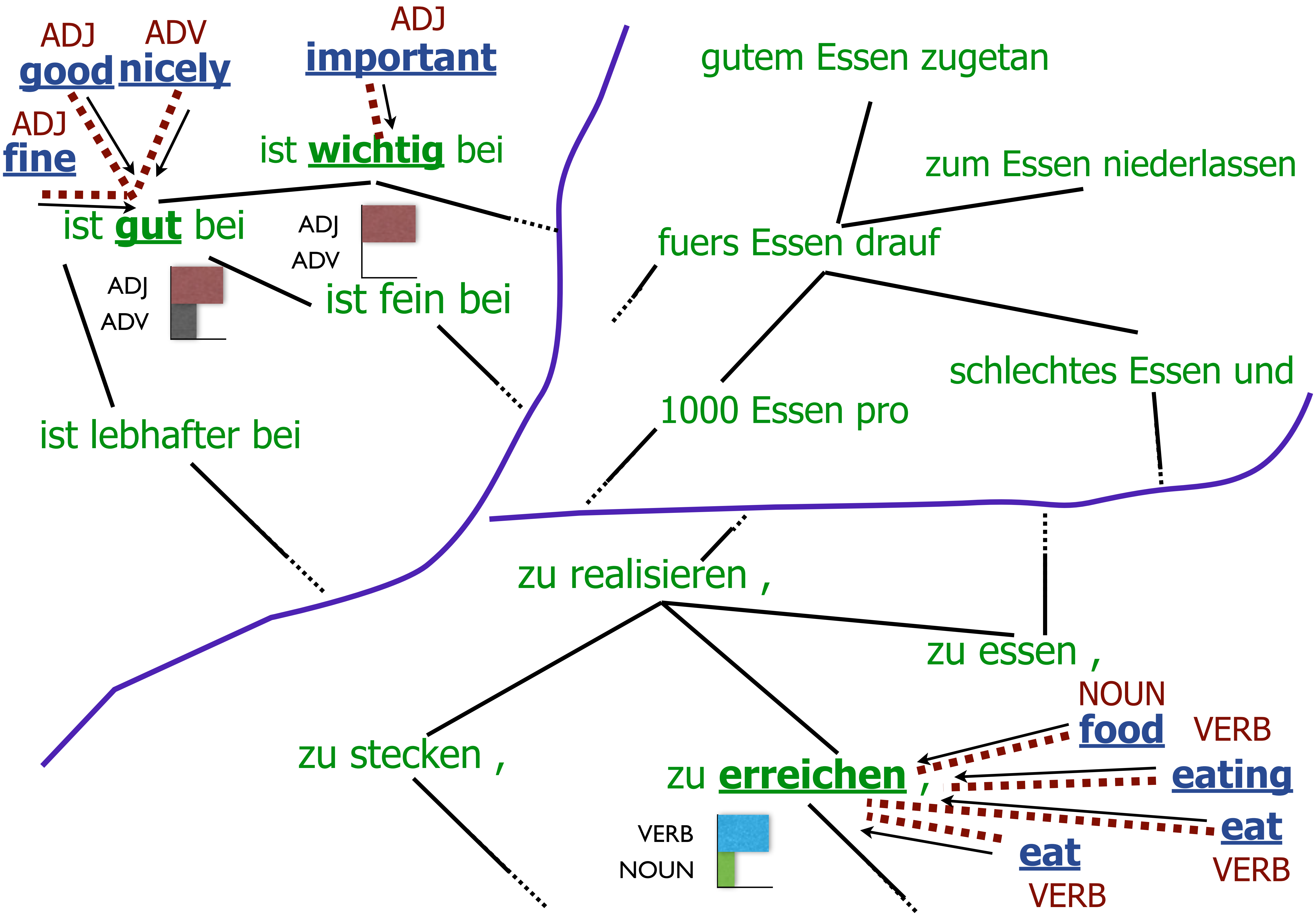
Graph Regularization



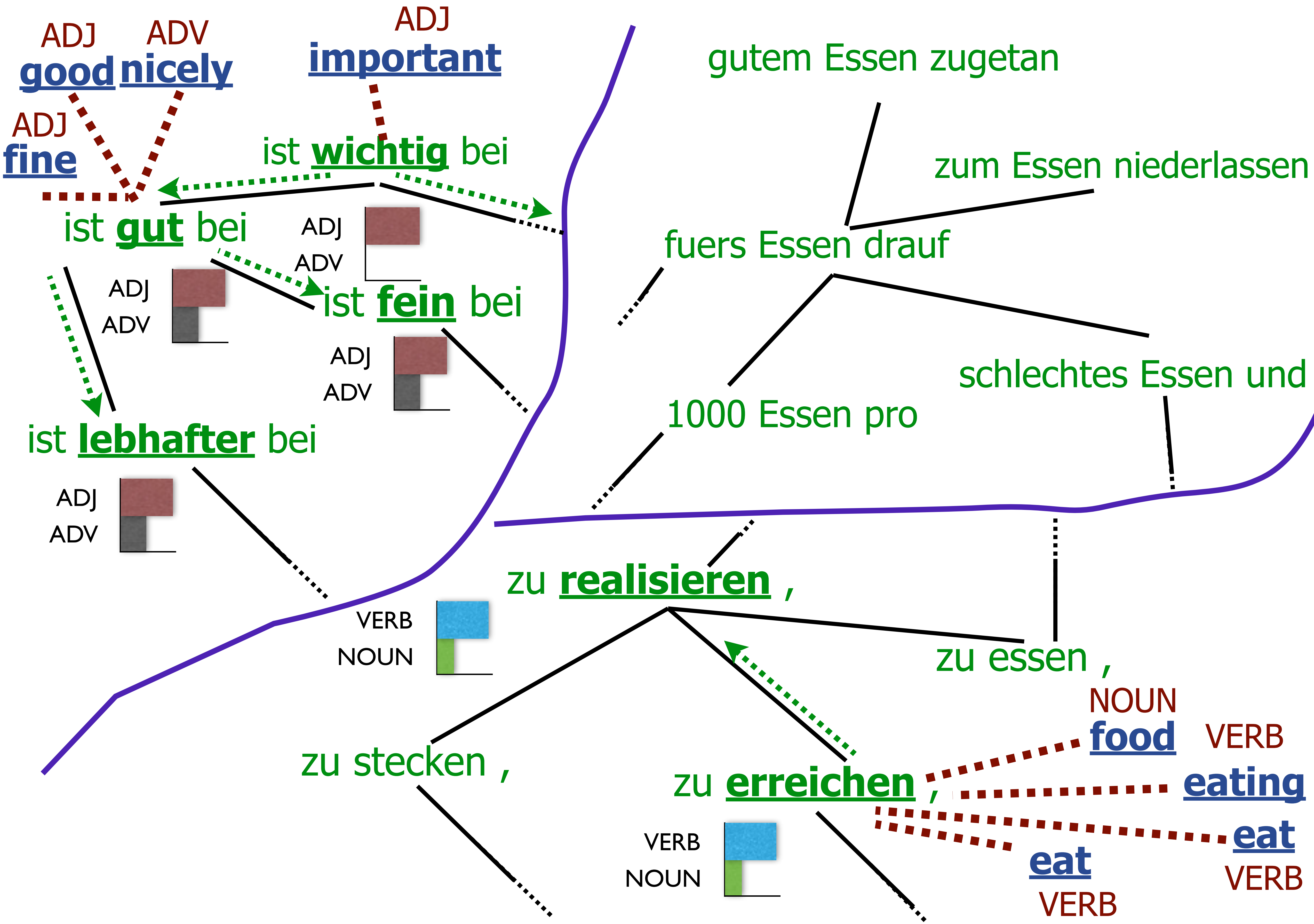
Graph Regularization



Graph Regularization

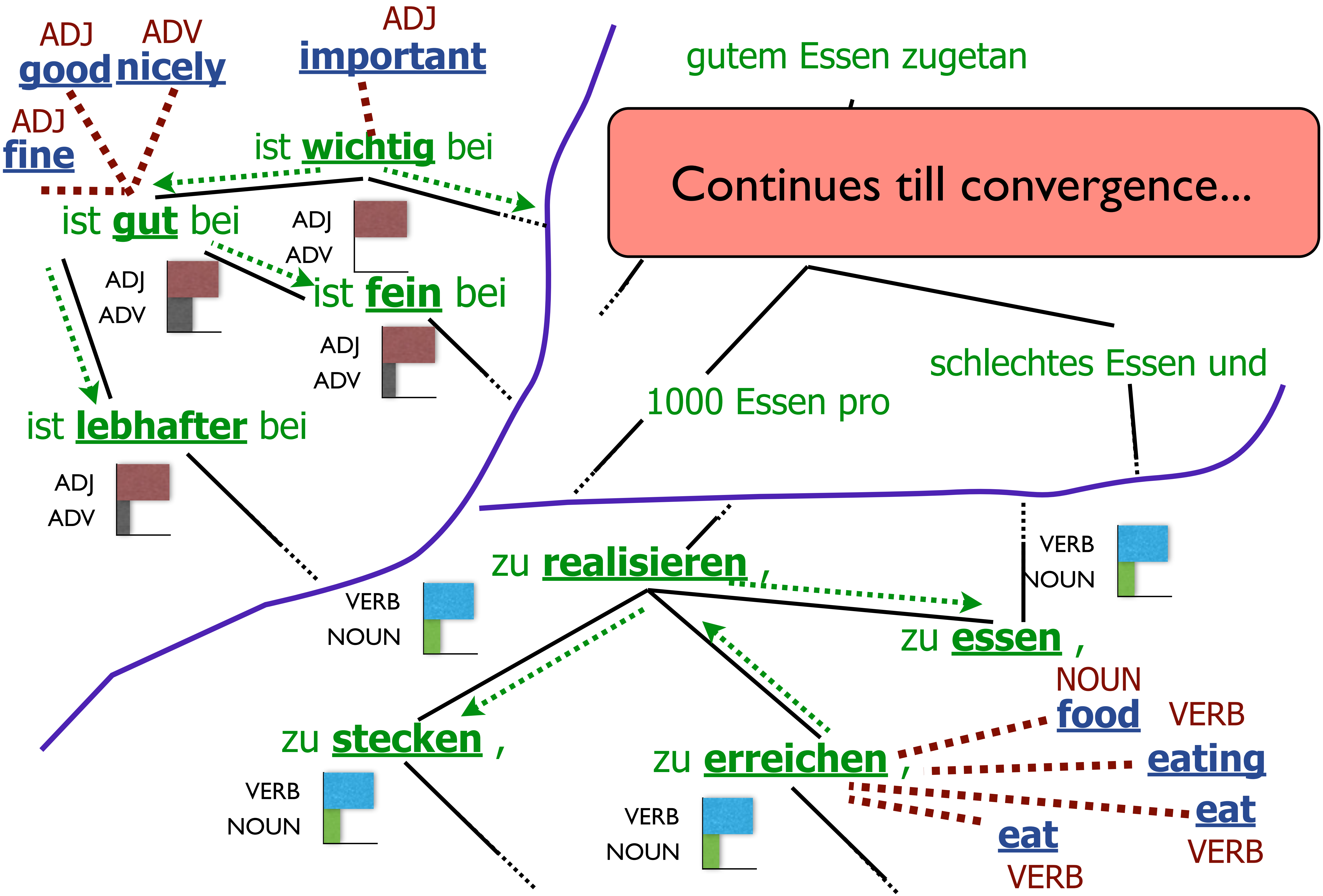


Graph Regularization



Graph Regularization

Graph Regularization



Results

	Danish	Dutch	German	Greek	Italian	Portugese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8

Results

	Danish	Dutch	German	Greek	Italian	Portugese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8
Graph-based Projection	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	**83.4**

Results

	Danish	Dutch	German	Greek	Italian	Portugese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8
Graph-based Projection	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	**83.4**
Oracle (Supervised)	96.9	94.9	98.2	97.8	95.8	97.2	96.8	94.8	96.6

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Text Categorization
  - Sentiment Analysis
  - Class Instance Acquisition
  - POS Tagging
  - MultiLingual POS Tagging
  - Semantic Parsing
    [Das & Smith, ACL 2011]
- Conclusion & Future Work

Problem Description

- Extract shallow semantic structure: **Frames** and **Roles**

I want to go to Jeju Island on Sunday

Problem Description

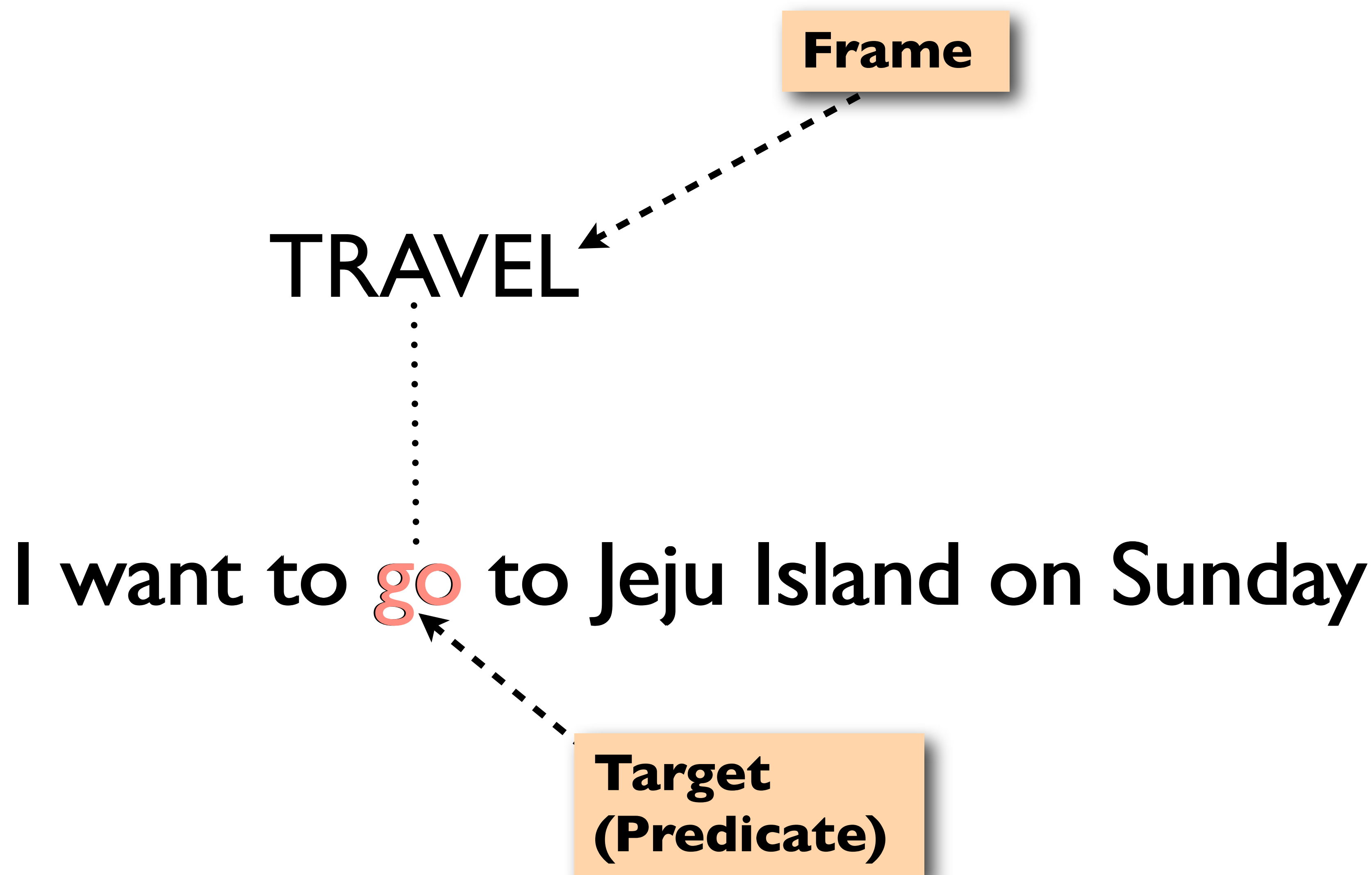
- Extract shallow semantic structure: **Frames** and **Roles**

I want to go to Jeju Island on Sunday

Target (Predicate)

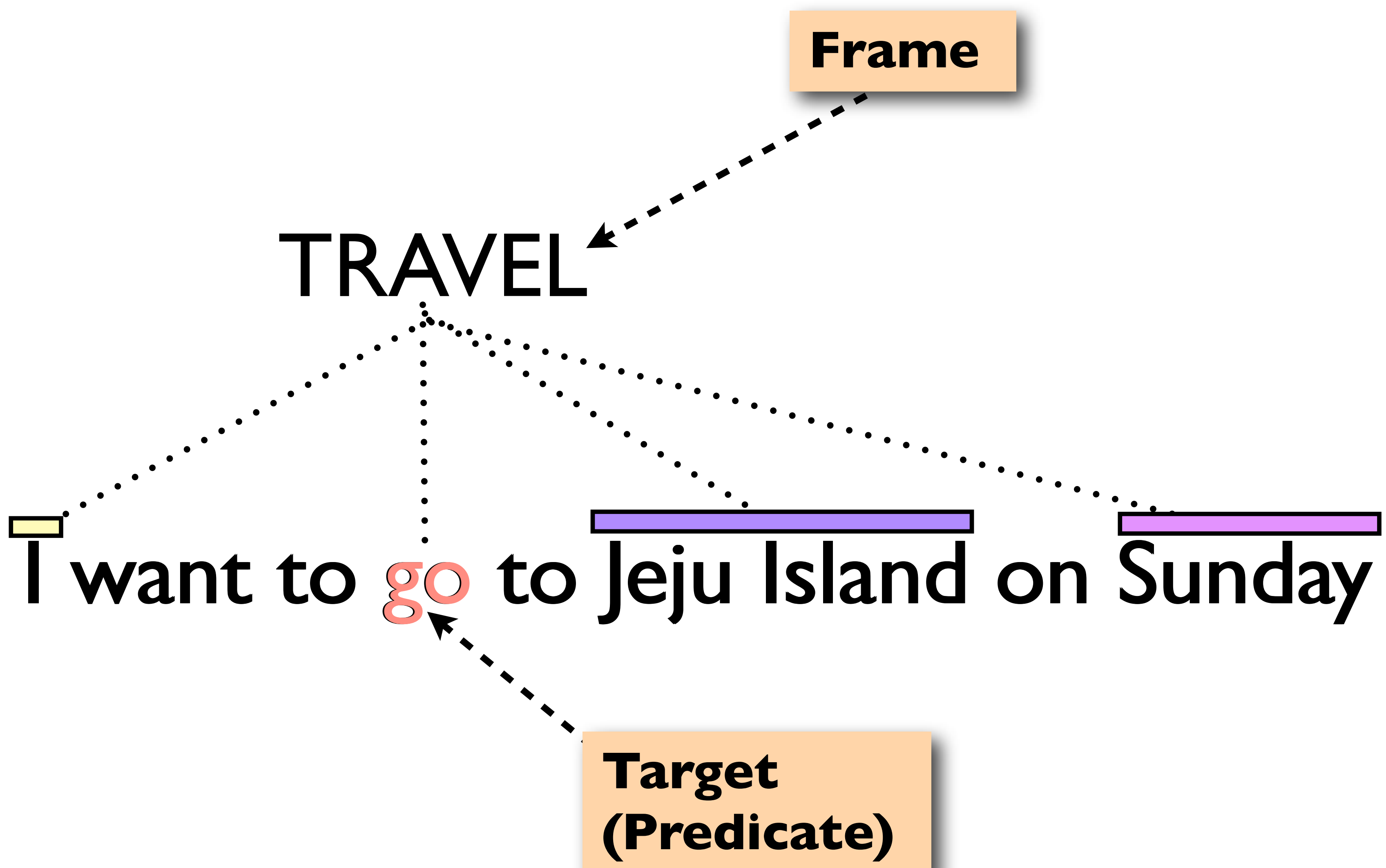
Problem Description

- Extract shallow semantic structure: **Frames** and **Roles**



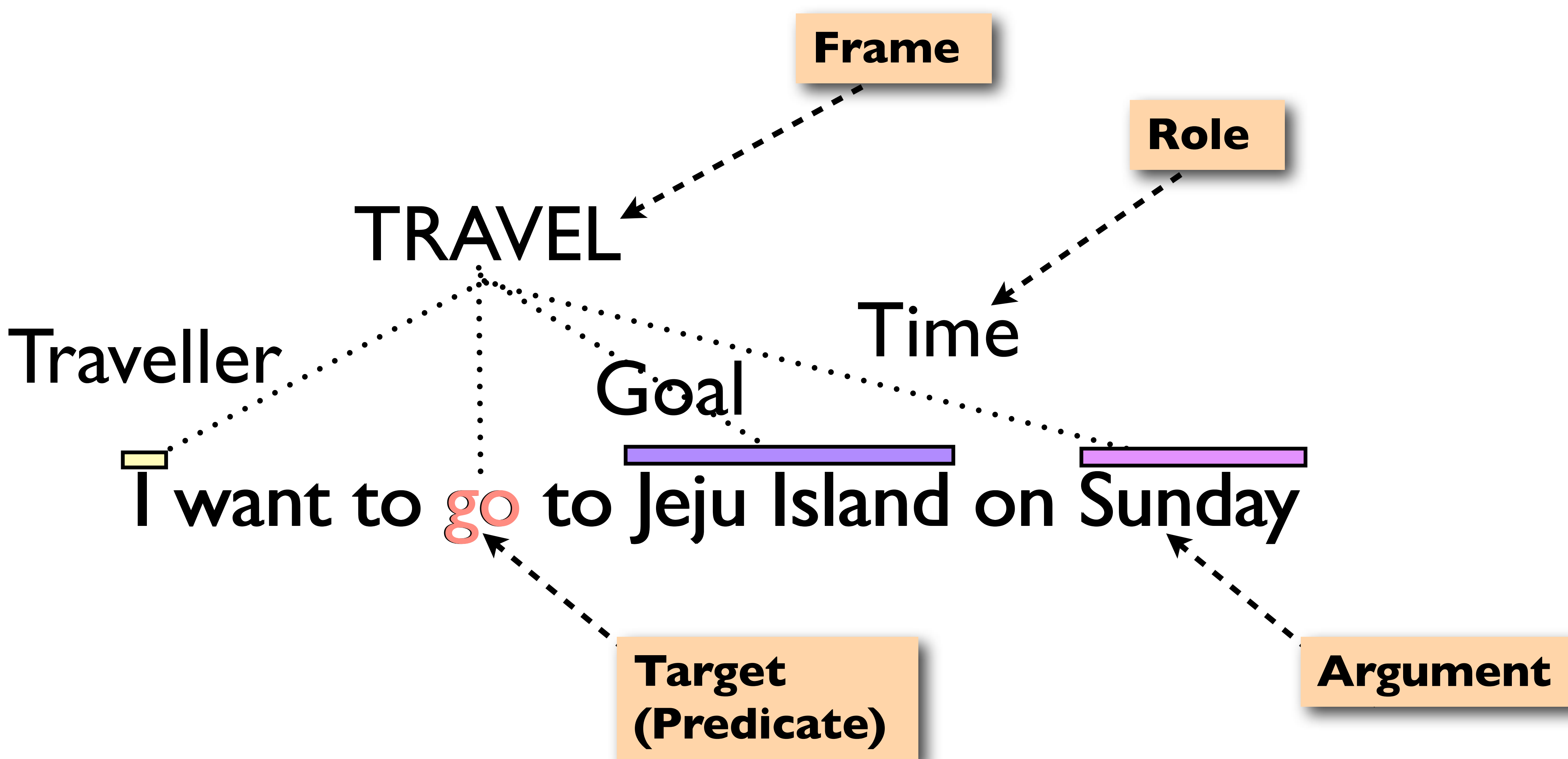
Problem Description

- Extract shallow semantic structure: **Frames** and **Roles**



Problem Description

- Extract shallow semantic structure: **Frames** and **Roles**

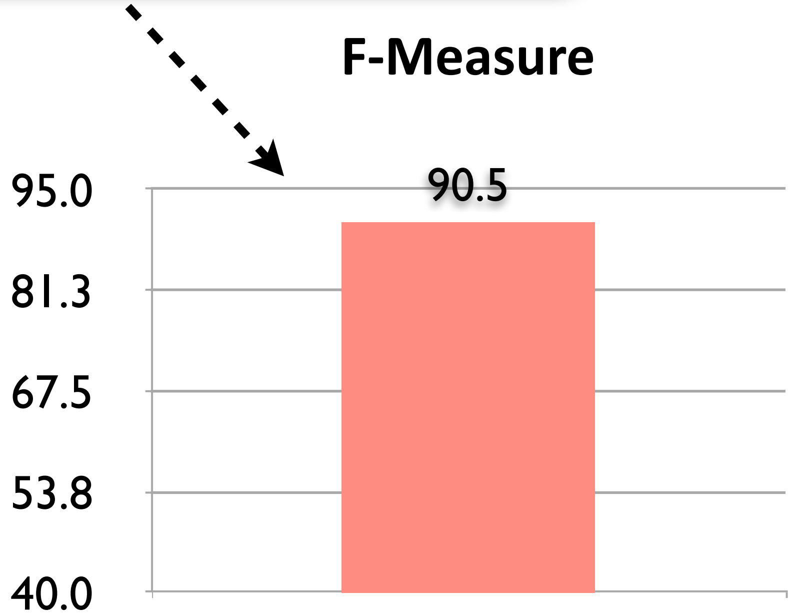


Problem Description

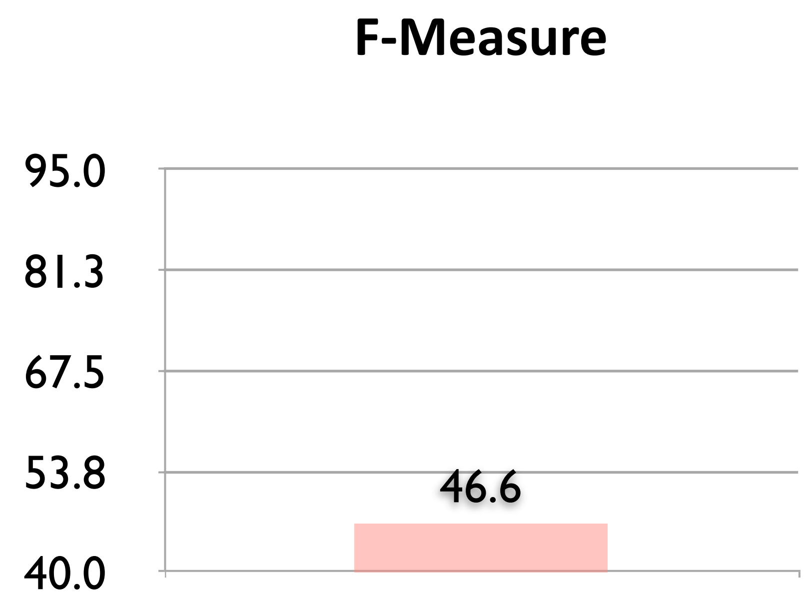
- Target identification
  - Most approaches assume this is given
- **Frame identification**
- Argument identification

Motivation

Frame Identification

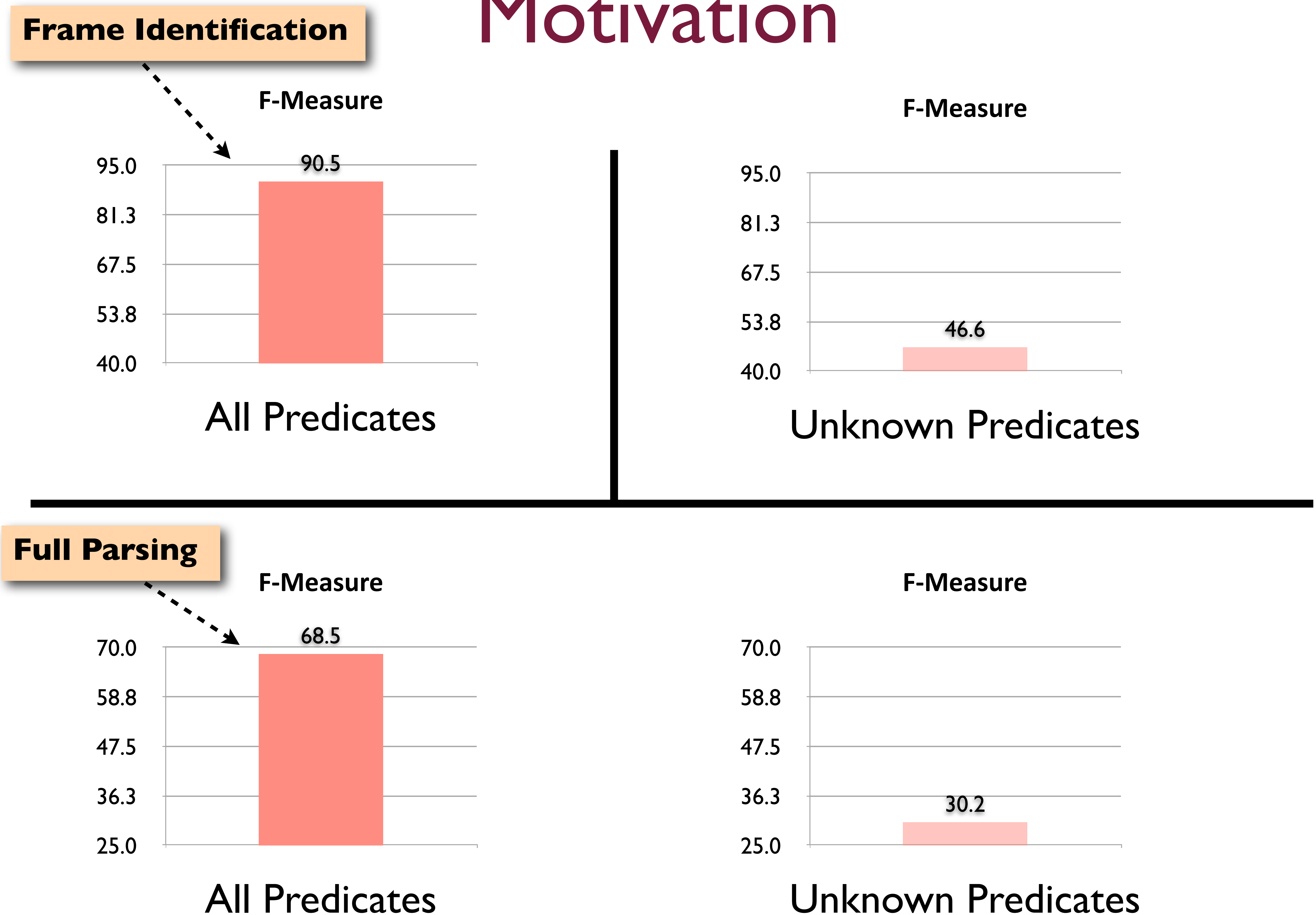


All Predicates



Unknown Predicates

Motivation



Frame Identification
F-Measure
95.0
81.3
67.5
53.8
40.0
90.5
All Predicates
F-Measure
95.0
81.3
67.5
53.8
40.0
46.6
Unknown Predicates
Full Parsing
F-Measure
70.0
58.8
47.5
36.3
25.0
68.5
All Predicates
F-Measure
70.0
58.8
47.5
36.3
25.0
30.2
Unknown Predicates

Sparse label data

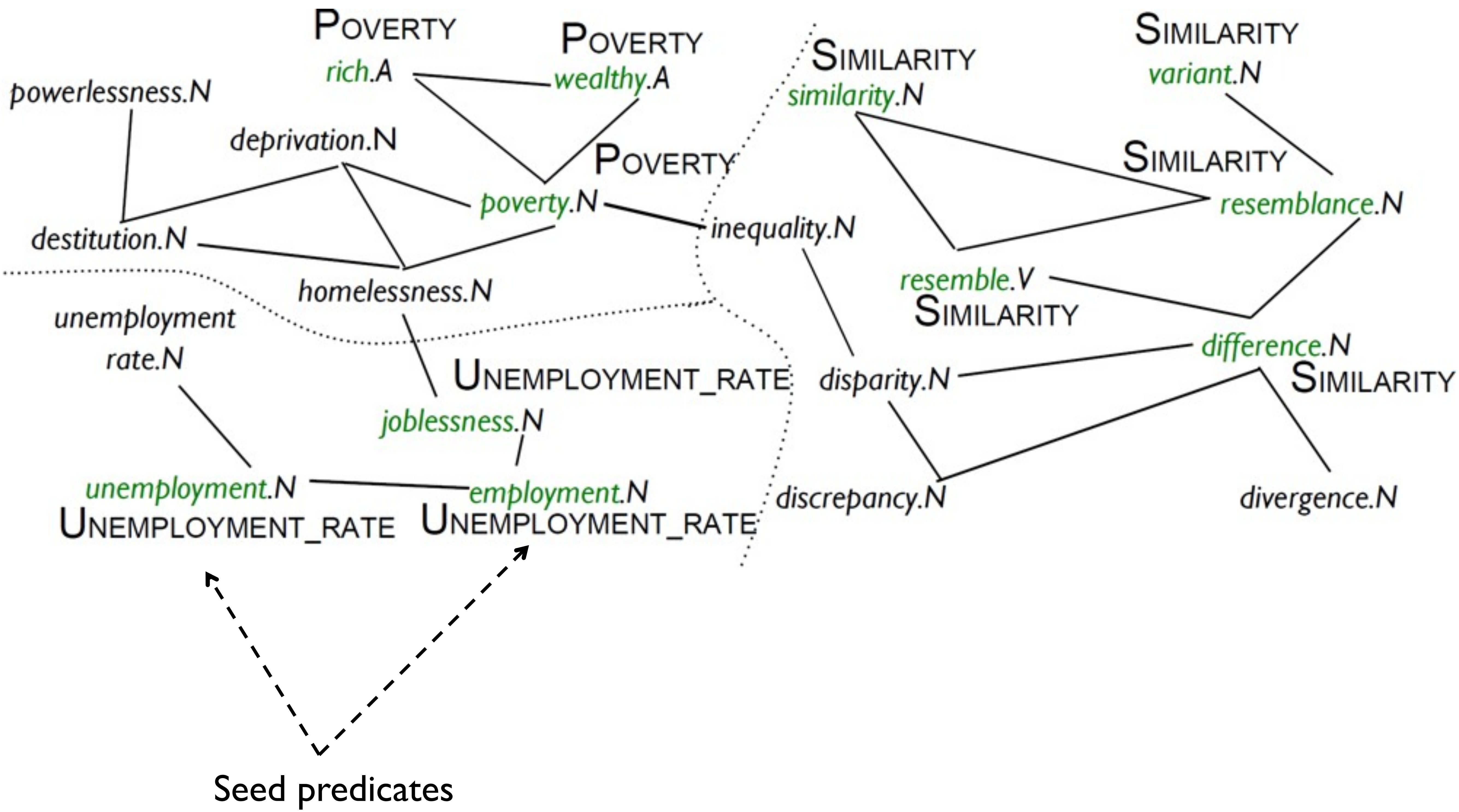
- Labeled data has only about 9,263 labeled predicates (targets)
    - English on the other hand has a lot more potential predicates (~65,000 in newswire)

Sparse label data

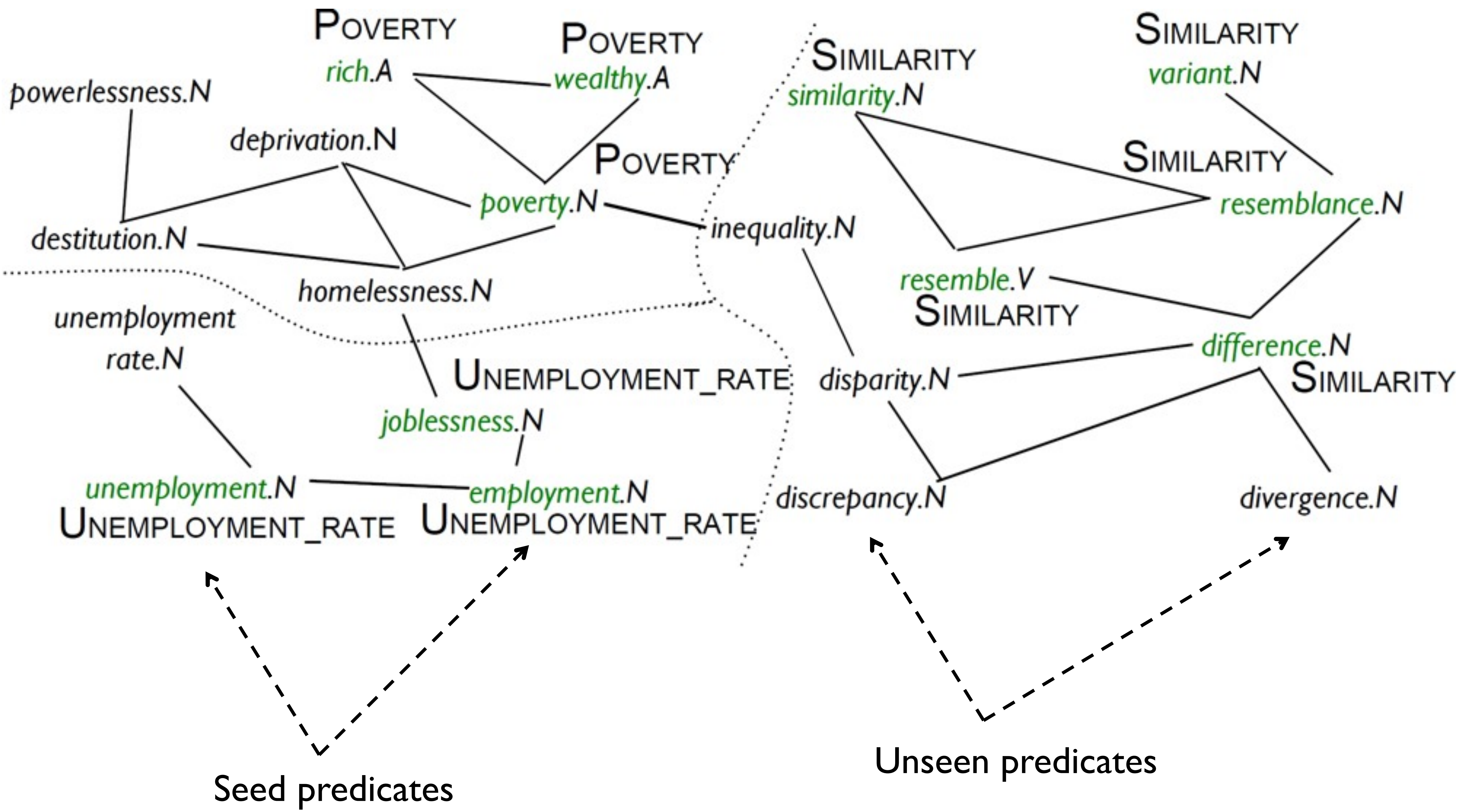
- Labeled data has only about 9,263 labeled predicates (targets)
  - English on the other hand has a lot more potential predicates (~65,000 in newswire)

- Construct a graph with potential predicates as vertices
- Expand the lexicon by using graph-based SSL

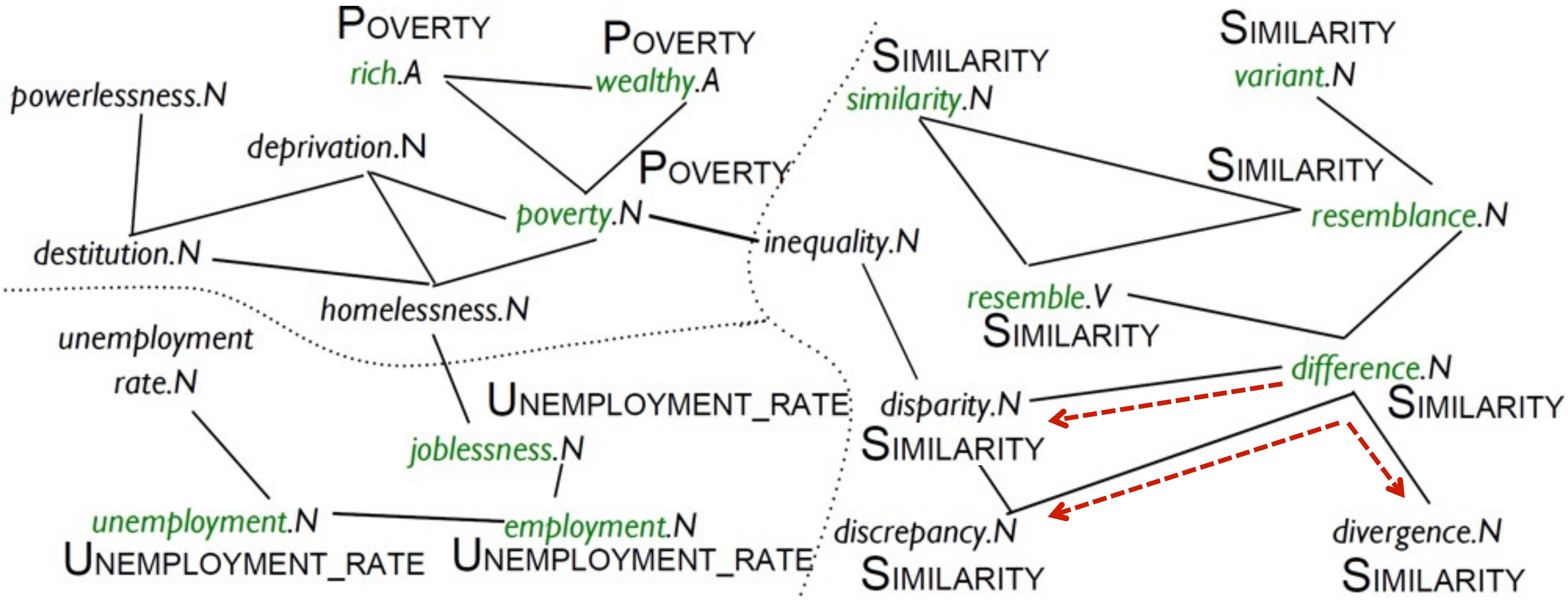
Graph Propagation (I)



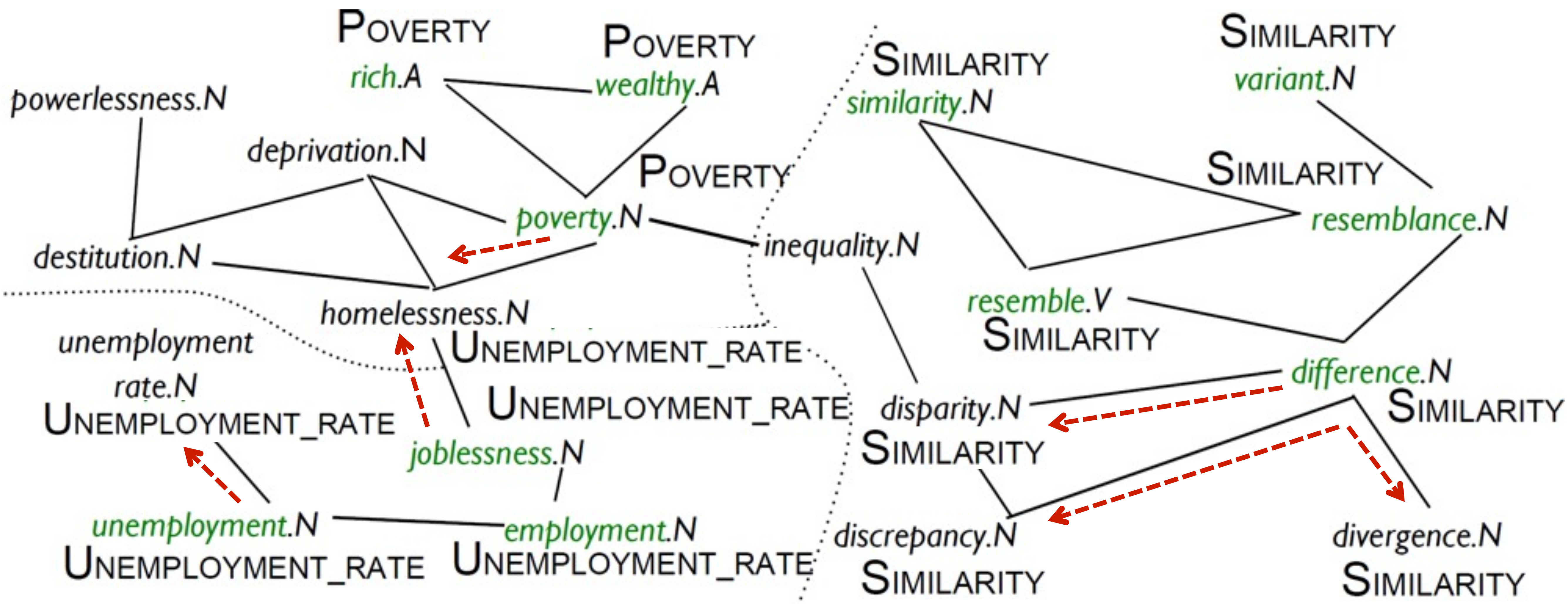
Graph Propagation (II)



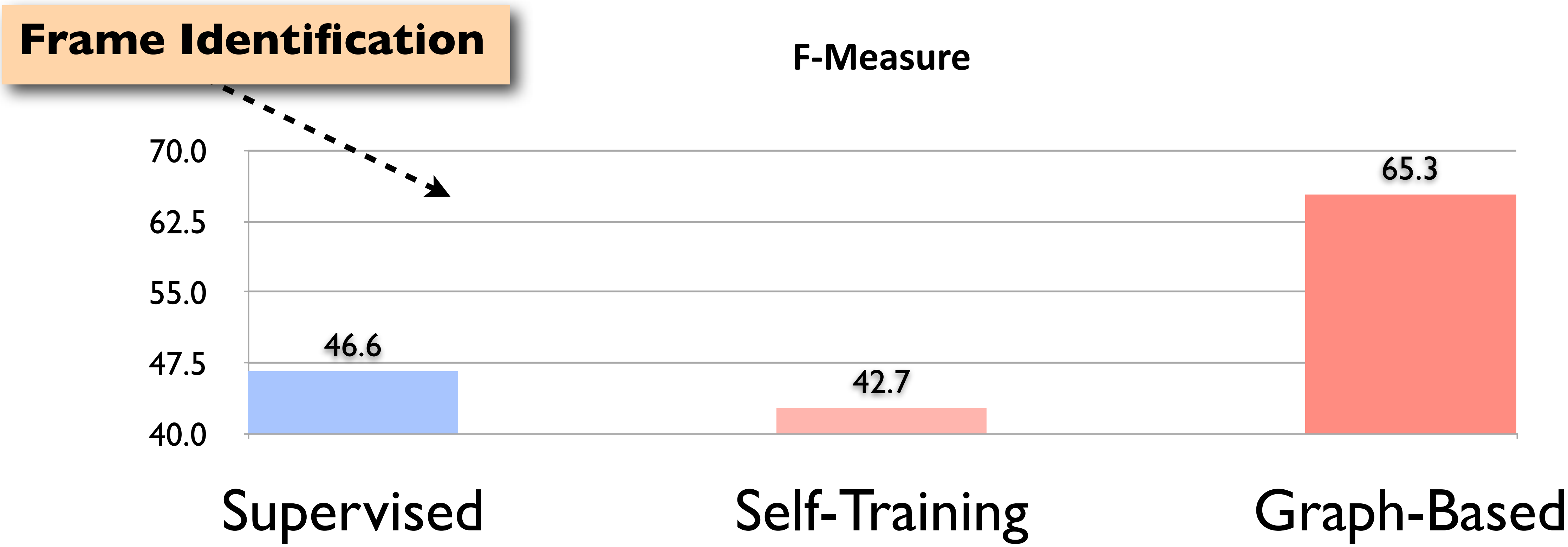
Graph Propagation (III)



Graph Propagation (IV)

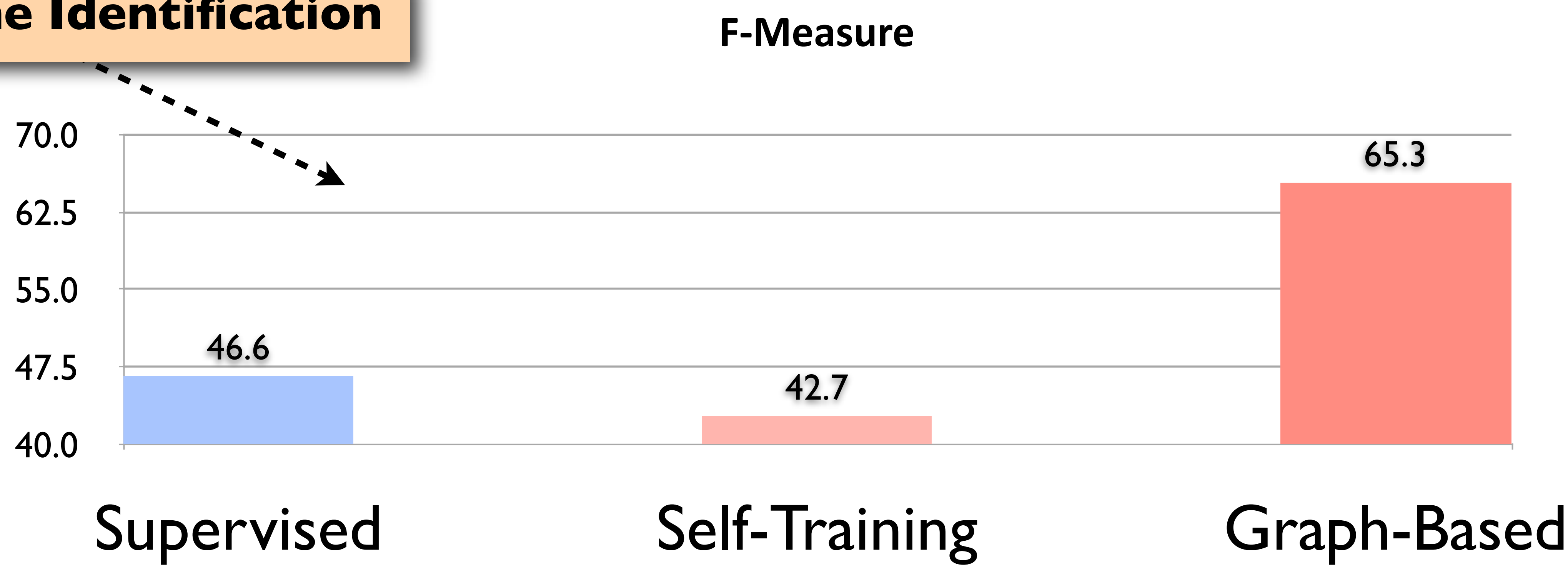


Results on Unknown Predicates

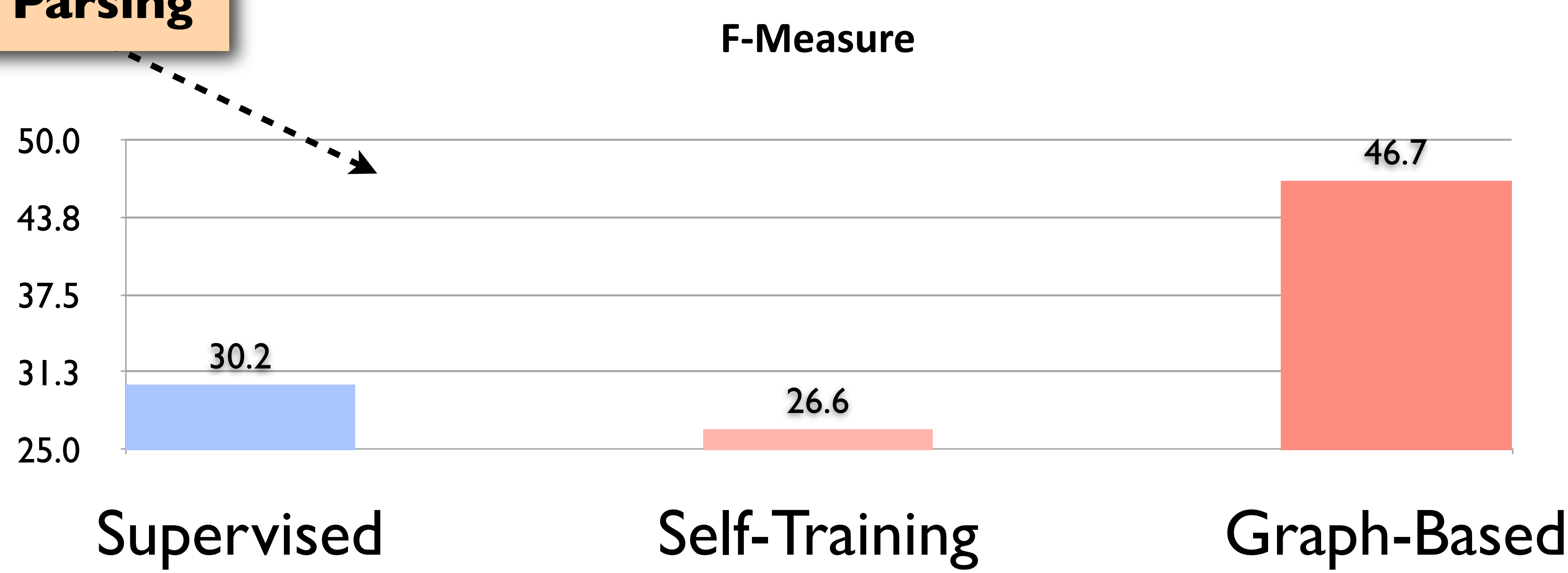


Results on Unknown Predicates

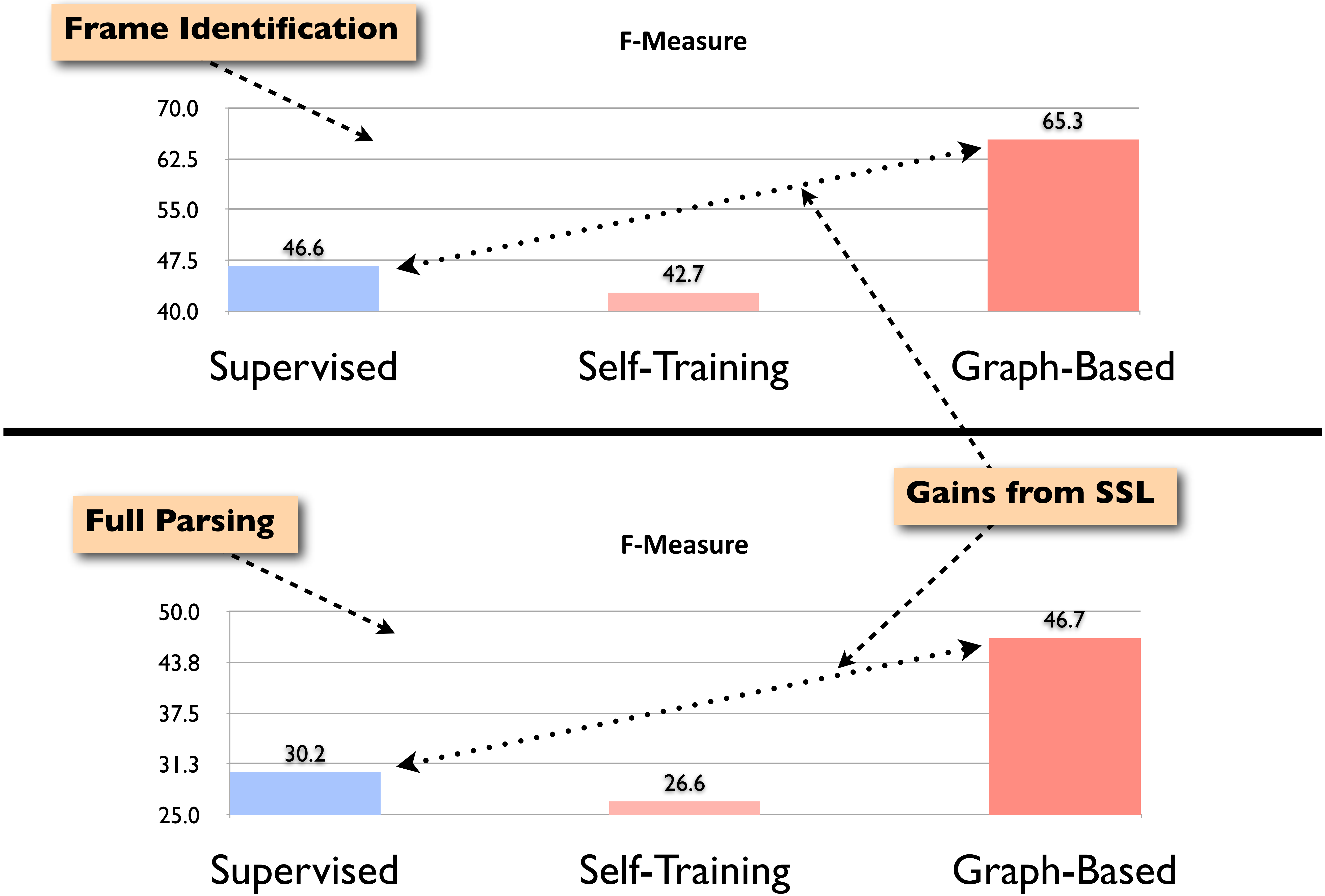
Frame Identification



Full Parsing



Results on Unknown Predicates



Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

When to use Graph-based SSL and which method?

- When input data itself is a graph
  - or, when the data is expected to lie on a manifold
- Measure Propagation (MP)
  - for probabilistic interpretation
- Quadratic Criteria (QC), MAD, MADDL
  - when labels are not mutually exclusive
- Manifold Regularization
  - for generalization to unseen data (induction)

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data
- Graph construction can be key
- Scalable: Node Reordering and MapReduce
- Can handle labeled as well as unlabeled data
- Can handle multi class, multi label settings
- Effective in practice

Open Challenges

- Use in structured prediction problems
  - Constituency and dependency parsing
- Combining Inductive and Graph-based methods
  - Joint optimization and parallel training [Altun et al., NIPS 2006]
- Scalable graph construction, especially with multi-modal data
- Extensions with other loss functions, sparsity, etc.
- Using side information

Acknowledgments

- National Science Foundation (NSF) IIS-0447972
- DARPA HRO1107-1-0029, FA8750-09-C-0179
- Google Research Award
- Dipanjan Das (Google), Ryan McDonald (Google), Fernando Pereira (Google), Slav Petrov (Google), Noah Smith (CMU)

References (I)

[1] A. Alexandrescu and K. Kirchhoff. Data-driven graph construction for semi- supervised graph-based learning in NLP. In NAACL HLT, 2007.
[2] Y. Altun, D. McAllester, and M. Belkin. Maximum margin semi-supervised learn- ing for structured variables. NIPS, 2006.
[3] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. J. Mach. Learn. Res., 3:1183–1208, 2003.
[4] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7:2399–2434, 2006.
[5] Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. Semi-supervised learning, 2006.
[6] T. Berg-Kirkpatrick, A. Bouchard-Coˆt́ e, J. DeNero, and D. Klein. Painless unsupervised learning with features. In HLT-NAACL, 2010.
[7] J. Bilmes and A. Subramanya. Scaling up Machine Learning: Parallel and Distributed Approaches, chapter Parallel Graph-Based Semi-Supervised Learning. 2011.
[8] S. Blair-goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar. Building a sentiment summarizer for local service reviews. In In NLP in the Information Explosion Era, 2008.
[9] M. Cafarella, A. Halevy, D. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. VLDB, 2008.
[10] O. Chapelle, B. Scho ̈lkopf, A. Zien, et al. Semi-supervised learning. MIT press Cambridge, MA:, 2006.
[11] Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear program- ming for domain specific sentiment classification. In EMNLP, 2009.
[12] S. Daitch, J. Kelner, and D. Spielman. Fitting a graph to vector data. In ICML, 2009.
[13] D. Das and S. Petrov. Unsupervised part-of-speech tagging with bilingual graph- based projections. In ACL, 2011.
[14] D. Das, N. Schneider, D. Chen, and N. A. Smith. Probabilistic frame-semantic parsing. In NAACL-HLT, 2010.
[15] D. Das and N. Smith. Graph-based lexicon expansion with sparsity-inducing penalties. NAACL-HLT, 2012.
[16] D. Das and N. A. Smith. Semi-supervised frame-semantic parsing for unknown predicates. In ACL, 2011.
[17] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In ICML, 2007.
[18] O. Delalleau, Y. Bengio, and N. L. Roux. Efficient non-parametric function induction in semi-supervised learning. In AISTATS, 2005.
[19] P. Dhillon, P. Talukdar, and K. Crammer. Inference-driven metric learning for graph construction. Technical report, MS-CIS-10-18, University of Pennsylvania, 2010.
[20] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In CIKM, 1998.

References (II)

[21] J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. ACM Transaction on Mathematical Software, 3, 1977.
[22] J. Garcke and M. Griebel. Data mining with sparse grids using simplicial basis functions. In KDD, 2001.
[23] A. Goldberg and X. Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, 2006.
[24] A. Goldberg, X. Zhu, and S. Wright. Dissimilarity in graph-based semi-supervised classification. AISTATS, 2007.
[25] M. Hu and B. Liu. Mining and summarizing customer reviews. In KDD, 2004.
[26] T. Jebara, J. Wang, and S. Chang. Graph construction and b-matching for semi-supervised learning. In ICML, 2009.
[27] T. Joachims. Transductive inference for text classification using support vector machines. In ICML, 1999.
[28] T. Joachims. Transductive learning via spectral graph partitioning. In ICML, 2003.
[29] M. Karlen, J. Weston, A. Erkan, and R. Collobert. Large scale manifold transduction. In ICML, 2008.
[30] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In Proceedings of the 20th International conference on Computational Linguistics, 2004.
[31] F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. Information Theory, IEEE Transactions on, 47(2):498–519, 2001
[32] K. Lerman, S. Blair-Goldensohn, and R. McDonald. Sentiment summarization: evaluating and learning user preferences. In EACL, 2009.
[33] D.Lewisetal.Reuters-21578.http://www.daviddlewis.com/resources/testcollections/reuters21578, 1987.
[34] J. Malkin, A. Subramanya, and J. Bilmes. On the semi-supervised learning of multi-layered perceptrons. In InterSpeech, 2009.
[35] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In EMNLP, 2002.
[36] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In EACL, 2009.
[37] A. Subramanya and J. Bilmes. Soft-supervised learning for text classification. In EMNLP, 2008.
[38] A. Subramanya and J. Bilmes. Entropic graph regularization in non-parametric semi-supervised classification. NIPS, 2009.
[39] A. Subramanya and J. Bilmes. Semi-supervised learning with measure propagation. The Journal of Machine Learning Research, 2011.
[40] A. Subramanya, S. Petrov, and F. Pereira. Efficient graph-based semi-supervised learning of structured tagging models. In EMNLP, 2010.

References (III)

[41] P. Talukdar. Topics in graph construction for semi-supervised learning. Technical report, MS-CIS-09-13, University of Pennsylvania, 2009.
[42] P. Talukdar and K. Crammer. New regularized algorithms for transductive learning. ECML, 2009.
[43] P. Talukdar and F. Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In ACL, 2010.
[44] P. Talukdar, J. Reisinger, M. Pa¸ sca, D. Ravichandran, R. Bhagat, and F. Pereira. Weakly-supervised acquisition of labeled class instances using graph random walks. In EMNLP, 2008.
[45] B. Van Durme and M. Pasca. Finding cars, goddesses and enzymes: Parametrizable acquisition of labeled instances for open-domain information extraction. In AAAI, 2008.
[46] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald. The viability of web-derived polarity lexicons. In HLT-NAACL, 2010.
[47] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In ICML, 2006.
[48] J. Wang, T. Jebara, and S. Chang. Graph transduction via alternating minimization. In ICML, 2008.
[49] R. Wang and W. Cohen. Language-independent set expansion of named entities using the web. In ICDM, 2007.
[50] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research, 10:207–244, 2009.
[51] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase- level sentiment analysis. In HLT-EMNLP, 2005.
[52] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scho¨ lkopf. Learning with local and global consistency. NIPS, 2004.
[53] D. Zhou, J. Huang, and B. Scho¨ lkopf. Learning from labeled and un- labeled data on a directed graph. In ICML, 2005.
[54] D. Zhou, B. Scho¨ lkopf, and T. Hofmann. Semi-supervised learning on directed graphs. In NIPS, 2005.
[55] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University, 2002.
[56] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University, 2002.
[57] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In ICML, 2003.
[58] X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph- based methods for inductive and scalable semi-supervised learning. In ICML, 2005.

Thanks!

Web: http://graph-ssl.wikidot.com/