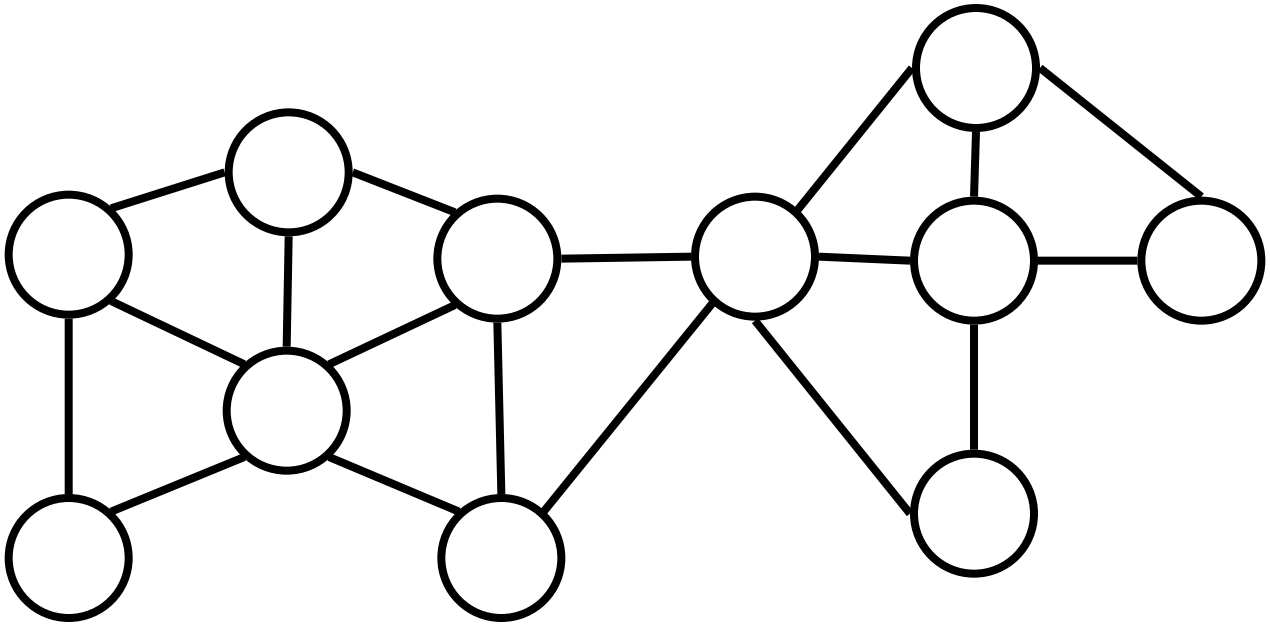
Active Semi-Supervised Learning using Submodular Functions

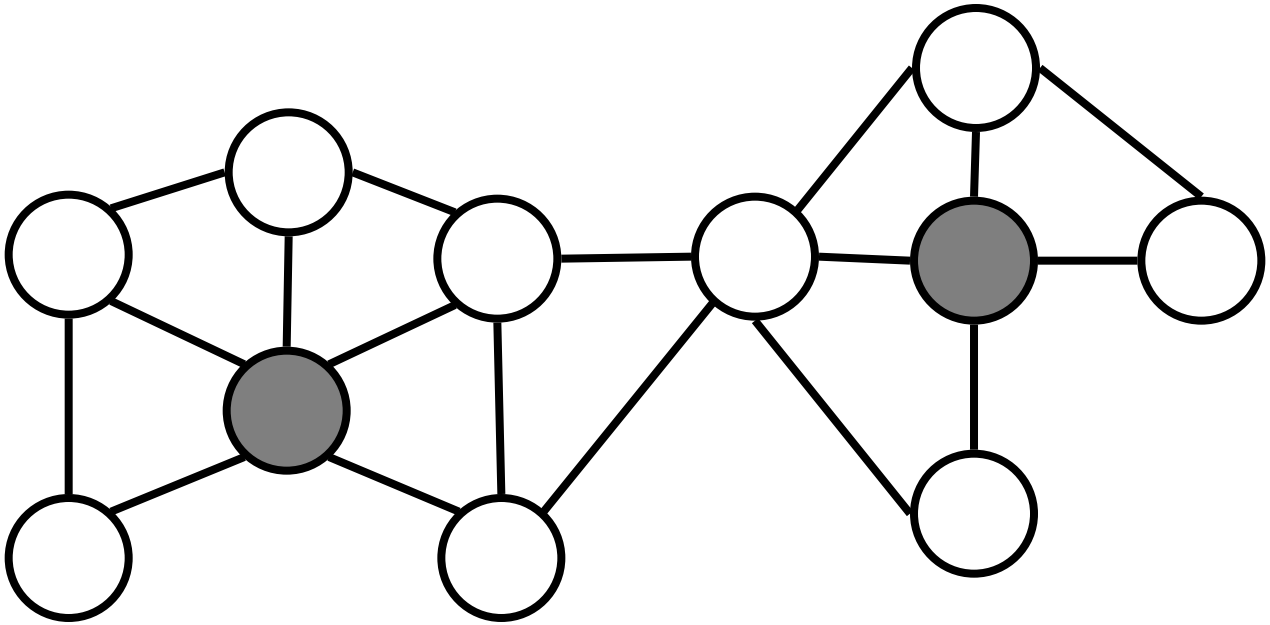
Andrew Guillory, Jeff Bilmes
University of Washington

Given unlabeled data

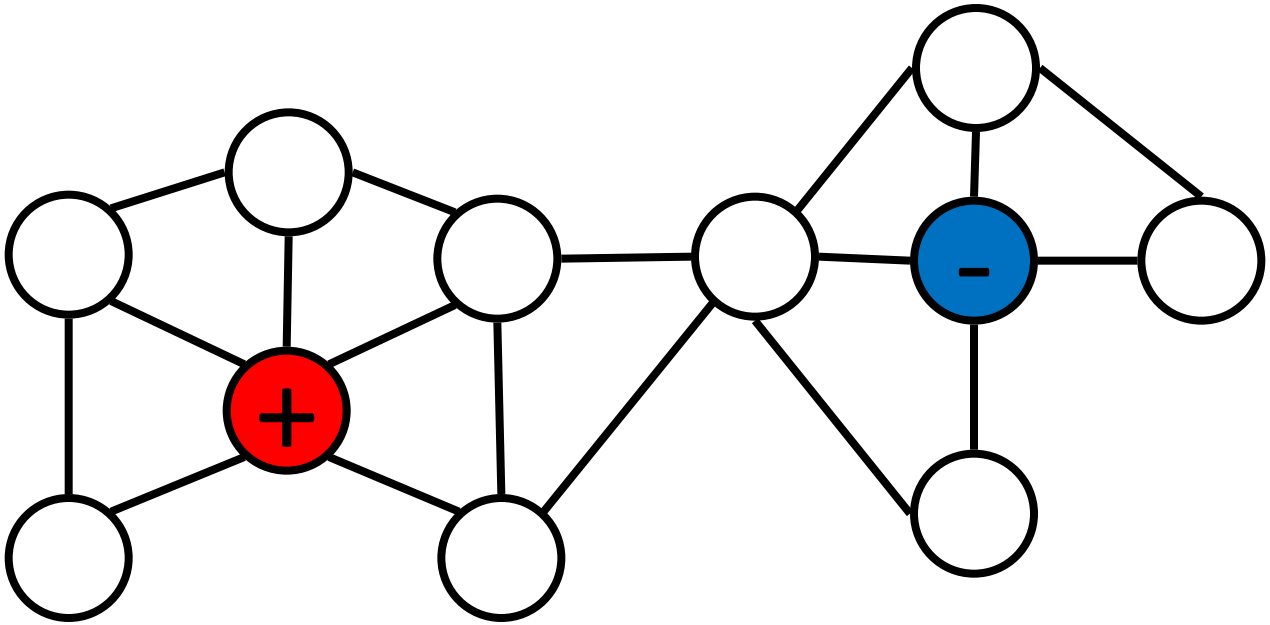


for example, a graph

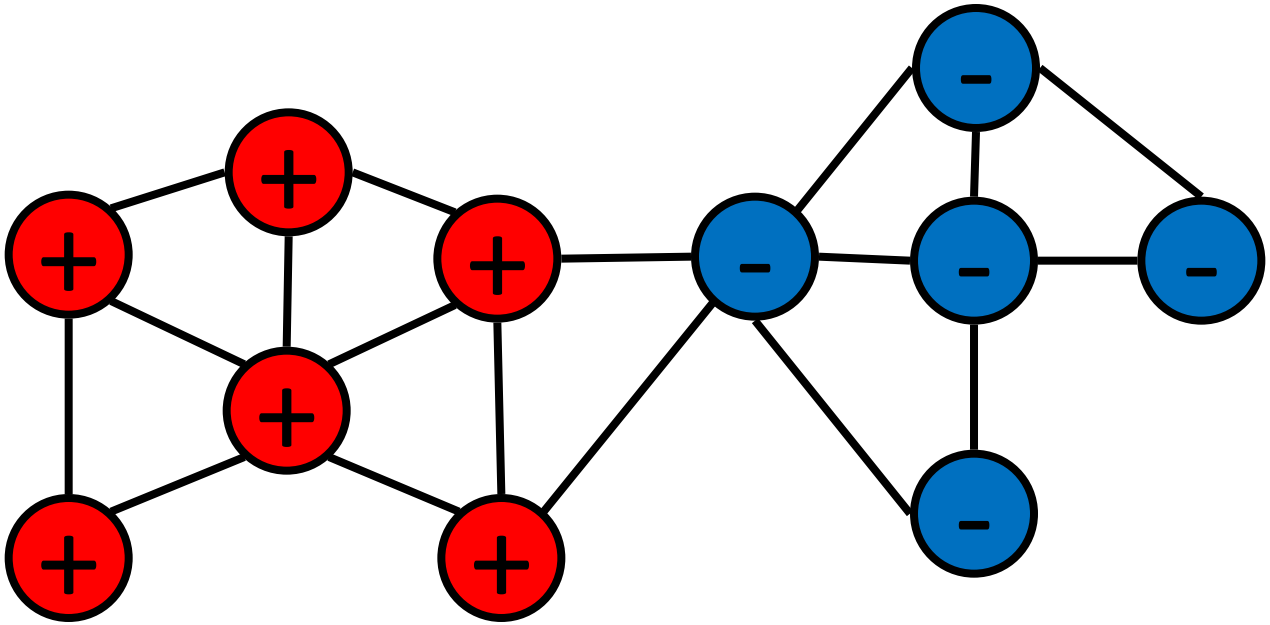
Learner chooses a labeled set $L \subseteq V$



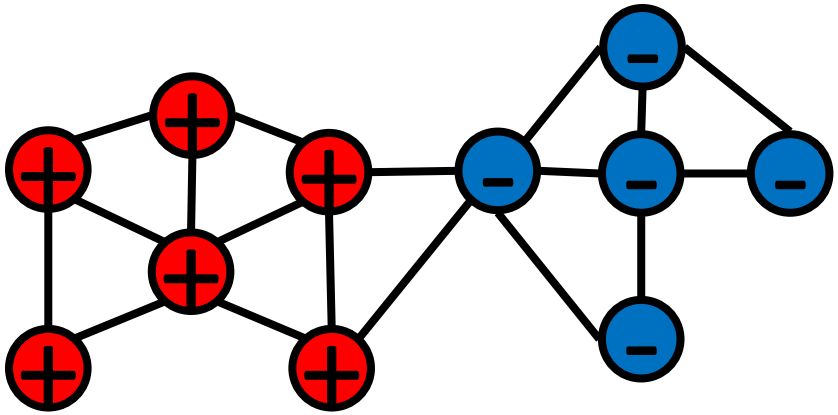
Nature reveals labels $\mathrm{y}_L \in \{0,1\}^{\mathrm{L}}$



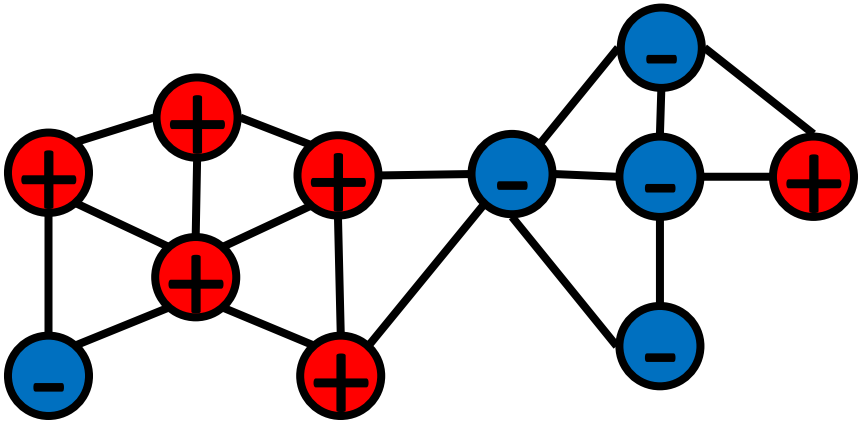
Learner predicts labels $\hat{y} \in \{0,1\}^V$



Learner suffers loss $\|\hat{y} - y\|_1$



Predicted



Actual

$$\|\hat{y} - y\|_1 = 2$$

Basic Questions

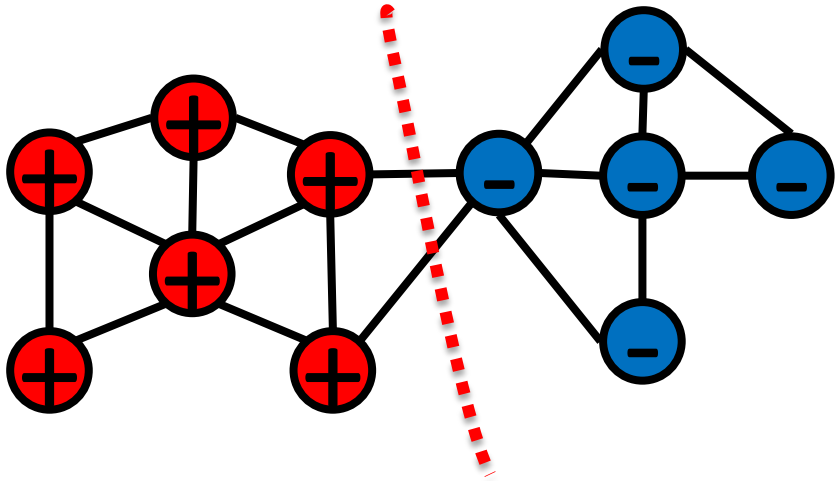
- What should we assume about y?
- How should we predict $\hat{y}$ using y_L?
- How should select L?
- How can we bound error?

Outline

- **Previous work: learning on graphs**
- More general setting using submodular functions
- Experiments

Learning on graphs

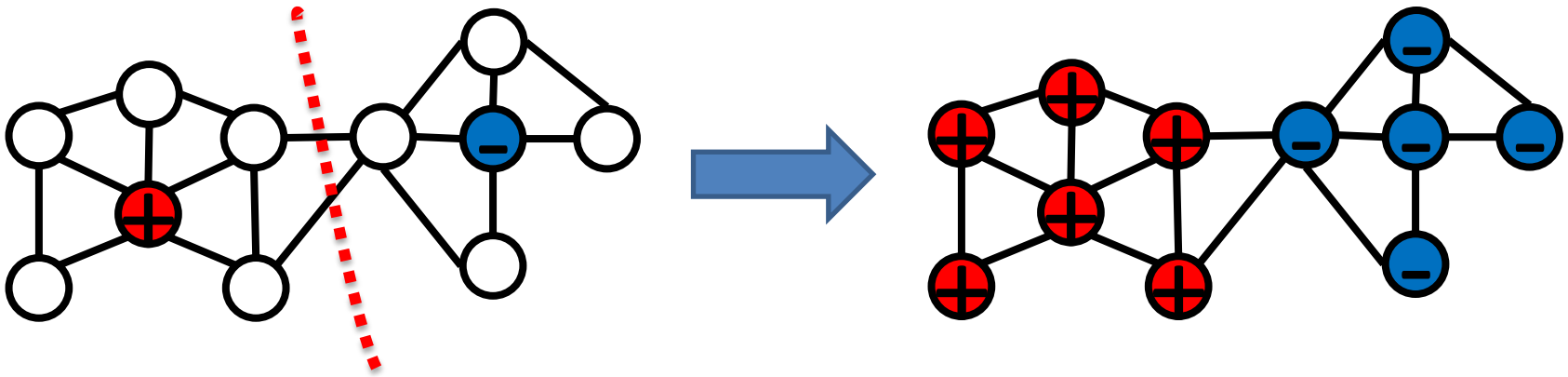
- What should we assume about y?
- Standard assumption: small cut value
- $\Phi(y) = \sum_{i<j} (y_i - y_j)^2 W_{i,j}$
- A "smoothness" assumption



$\Phi(y) = 2$

Prediction on graphs

- How should we predict $\hat{y}$ using y_L?
- Standard approach: min-cut (Blum & Chawla 2001)
- Choose $\hat{y}$ to minimize $\Phi(\hat{y})$ s.t. $\hat{y}_L = y_L$
- Reduces to a standard min-cut computation

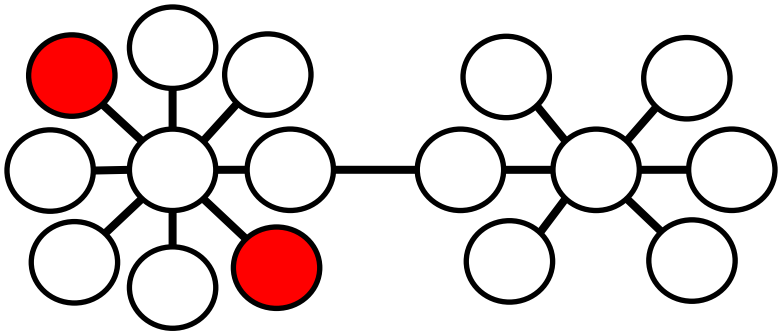


Active learning on graphs

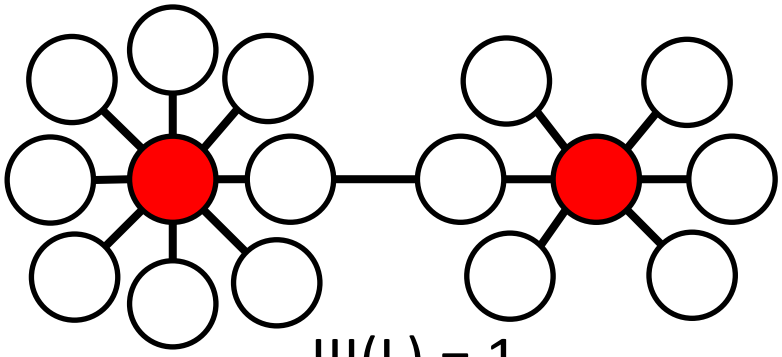
- How should select L?
- In previous work, we propose the following objective

$$\Psi(L) = \min_{T \subseteq V \setminus L : T \neq \emptyset} \frac{\Gamma(T)}{|T|}$$

  where $\Gamma(T)$ is cut value between T and $V \setminus T$
- Small $\Psi(L)$ means an adversary can cut away many points from L without cutting many edges



Ψ(L) = 1/8



Ψ(L) = 1

Error bound for graphs

How can we bound error?

Theorem (Guillory & Bilmes 2009): Assume $\hat{y}$ minimizes $\Phi(\hat{y})$ subject to $\hat{y}_L = y_L$. Then

$$\|\hat{y} - y\|_1 \leq 2\frac{\Phi(y)}{\Psi(L)}$$

- Intuition: $Error \leq \frac{Complexity\ of\ true\ labels}{Quality\ of\ labeled\ set}$
- Note: Deterministic, holds for *adversarial* labels

Drawbacks to previous work

- Restricted to graph based, min-cut learning
- Not clear how to *efficiently* maximize $\Psi(L)$
  - Can compute in polynomial time (Guillory & Bilmes 2009)
  - Only heuristic methods known for maximizing
  - Cesa-Bianchi et al 2010 give an approximation *for trees*
- Not clear if this bound is the right bound

Our Contributions

- A new, more general bound on error parameterized by an arbitrarily chosen submodular function
- An active, semi-supervised learning method for approximately minimizing this bound
- Proof that minimizing this bound exactly is NP-hard
- Theoretical evidence this is the “right” bound

Outline

- Previous work: learning on graphs
- **More general setting using submodular functions**
- Experiments

Submodular functions

- A function $F(S)$ defined over a ground set V is submodular iff for all $A \subseteq B \subseteq (V \setminus \{v\})$

$$F(A + v) - F\ (A) \geq F(B + v) - F(B)$$

- Example:

- Real World Examples: Influence in a social network (Kempe et al. 03), sensor coverage (Krause, Guestrin 09), document summarization (Lin, Bilmes 11)
- $F(S)$ is symmetric if $F(S) = F(V \setminus S)$

Submodular functions for learning

- $\Gamma(T)$ (cut value) is symmetric and submodular
- This makes $\Gamma(T)$ "nice" for learning on graphs
  - Easy to analyze
  - Can minimize exactly in polynomial time
- For other learning settings, other symmetric submodular functions make sense
  - Hypergraph cut is symmetric, submodular
  - Mutual information is symmetric, submodular
  - An arbitrary submodular function F can be symmetrized

$$\Gamma(S) = F(S) + F(V \setminus S) - F(V)$$

Generalized error bound

Theorem: For any symmetric, submodular $\Gamma(S)$, assume $\hat{y}$ minimizes $\Phi(\hat{y})$ subject to $\hat{y}_L = y_L$. Then

$$\|\hat{y} - y\|_1 \leq 2\frac{\Phi(y)}{\Psi(L)}$$

- Φ and Ψ are defined in terms of Γ, not graph cut

$$\Phi(y) = \Gamma(V_{y=1}) \quad \Psi(S) = \min_{T \subseteq V \setminus S : T \neq \emptyset} \frac{\Gamma(T)}{|T|}$$

- Each choice of Γ gives a different error bound
- Minimizing $\Phi(\hat{y})$ s.t. $\hat{y}_L = y_L$ can be done in polynomial time (submodular function minimization)

Can we efficiently maximize Ψ?

- Two related problems:
  1. Maximize $\Psi(L)$ subject to $|L| < k$
  2. Minimize $|L|$ subject to $\Psi(L) \geq \lambda$
- If $\Psi(L)$ were submodular, we could use well known results for greedy algorithm:
  - $\left(1 - \frac{1}{e}\right)$ approximation to (1) (Nemhauser et al. 1978)
  - $1 + \ln F(V)$ approximation for (2) (Wolsey 1981)*
- Unfortunately $\Psi(L)$ is **not** submodular

*Assuming integer valued F

Approximation result

- Define a surrogate objective $F_\lambda(S)$ s.t.
  - $F_\lambda(S)$ is submodular
  - $F_\lambda(\mathrm{S}) \geq 0$ iff $\Psi(S) \geq \lambda$
- In particular we use

$$F_\lambda(S) = \min_{T \subseteq V \setminus S:\, T \neq \emptyset} \Gamma(T) - \lambda|T|$$

- Can then use standard methods for $F_\lambda(S)$

Theorem: For any integer, symmetric, submodular $\Gamma(S)$, integer λ, greedily maximizing $F_\lambda(L)$ gives L with $\Psi(L) \geq \lambda$ and $|L| \leq (1 + \ln \lambda) \min_{L:\Psi(L)\geq\lambda} |L|$

Can we do better?

- Is it possible to maximize $\Psi(L)$ exactly?

Probably not, we show the problem is NP-Complete

  - Holds also if we assume $\Gamma(S)$ is the cut function
  - Reduction from vertex cover on fixed degree graphs
  - Corollary: no PTAS for min-cost version

- Is there a strictly better bound?

Not of the same form, up to the factor 2 in the bound.

  - Holds without factor of 2 for slightly different version
  - No function larger than $\Psi(L)$ for which the bound holds
  - Suggests this is the “right” bound

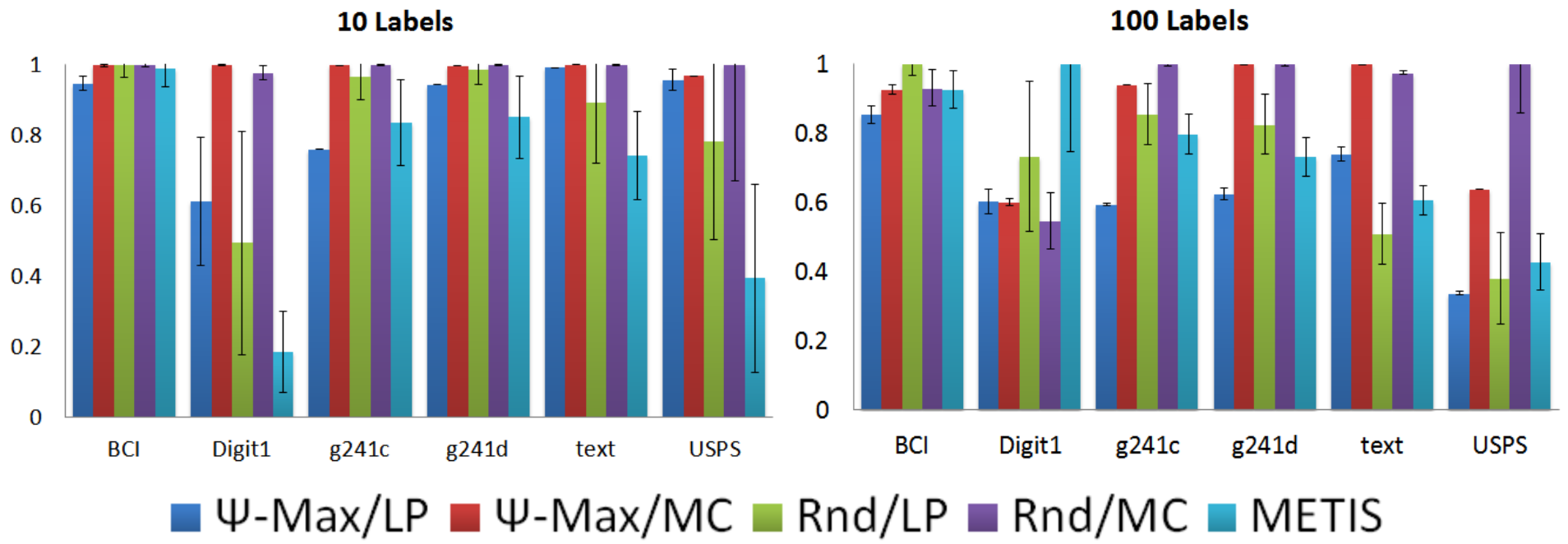
Outline

- Previous work: learning on graphs
- More general setting using submodular functions
- **Experiments**

Experiments: Learning on graphs

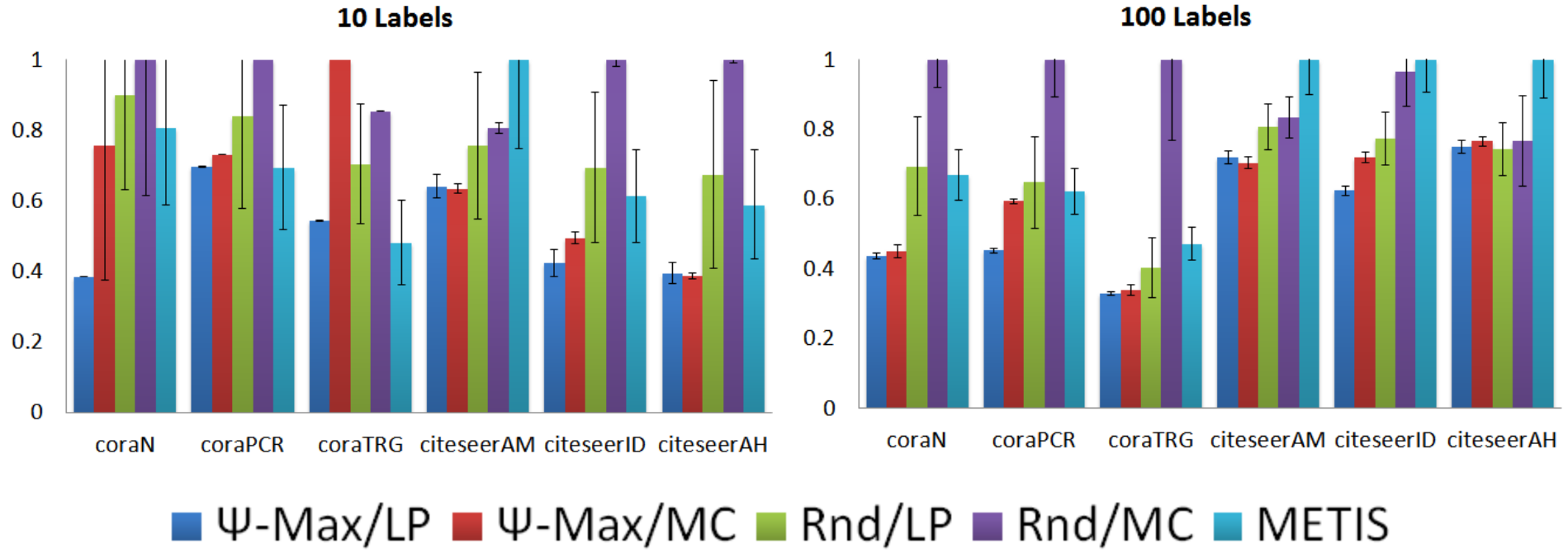
- With $\Gamma(S)$ set to cut, we compared our method to random selection and the METIS heuristic
- We tried min-cut and label propagation prediction
- We used benchmark data sets from *Semi-Supervised Learning,* Chapelle et al. 2006 (using knn neighbors graphs) and two citation graph data sets

Benchmark Data Sets



- Our method + label prop best in 6/12 cases, but not a consistent, significant trend
- Seems cut may not be suited for knn graphs

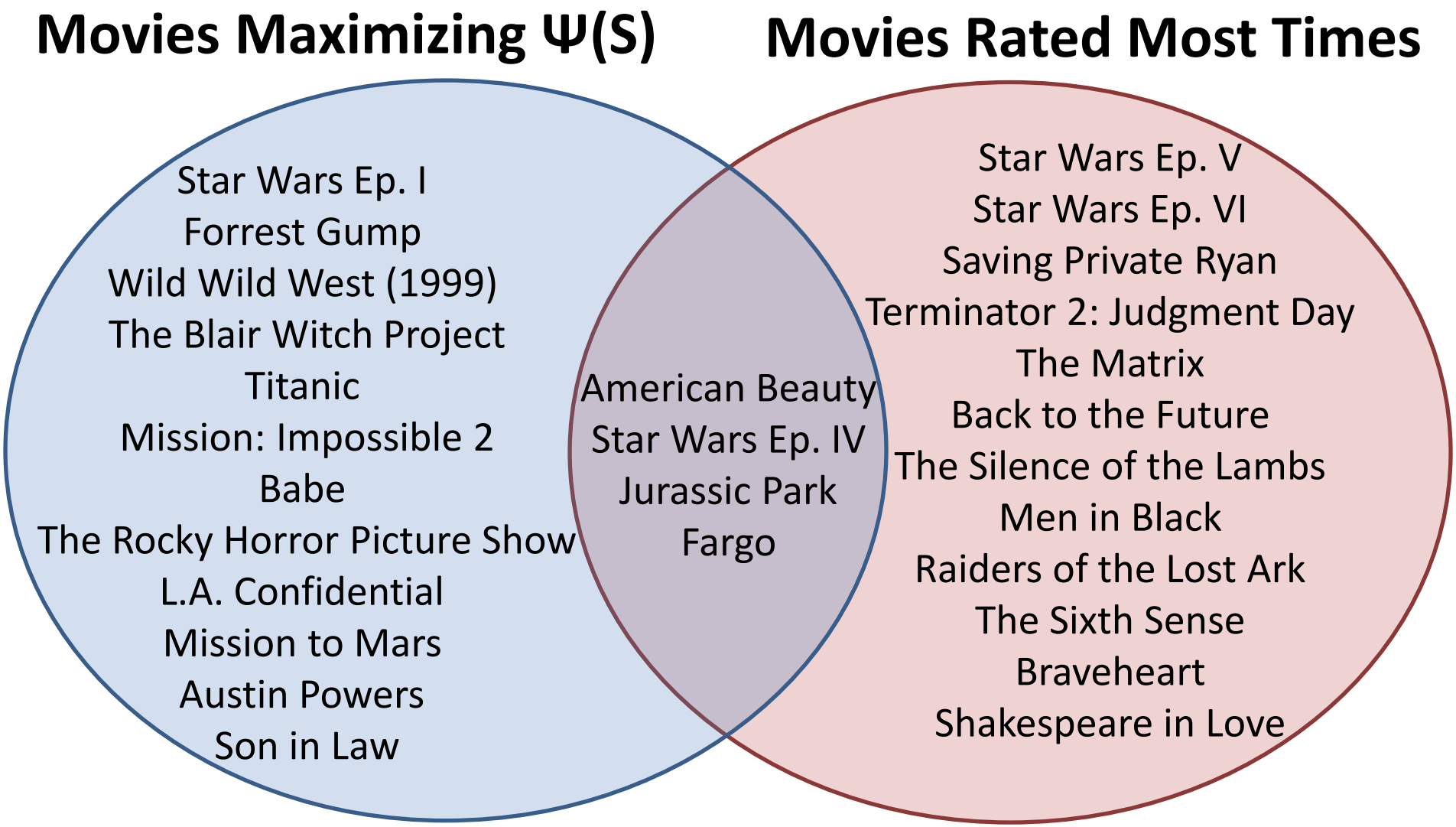
Citation Graph Data Sets



- Our method gives consistent, significant benefit
- On these data sets the graph is not constructed by us (not knn), so we expect more irregular structure.

Experiments: Movie Recommendation

- Which movies should a user rate to get accurate recommendations from collaborative filtering?
- We pose this problem as active learning over a hypergraph encoding user preferences, using $\Gamma(S)$ set to hypergraph cut
- Two hypergraph edges for each user:
  - Hypergraph edge connecting all movies a user likes
  - Hypergraph edge connecting all movies a user dislikes
- Partitions with low hypergraph cut value are consistent (on average) with user preferences



Using Movielens data

Our Contributions

- A new, more general bound on error parameterized by an arbitrarily chosen submodular function
- An active, semi-supervised learning method for approximately minimizing this bound
- Proof that minimizing this bound exactly is NP-hard
- Theoretical evidence this is the “right” bound
- Experimental results