Classifying Networked Entities with Modularity Kernels

Dell Zhang

Joint Work with Robert Mao (Microsoft Corp.)

Birkbeck, University of London

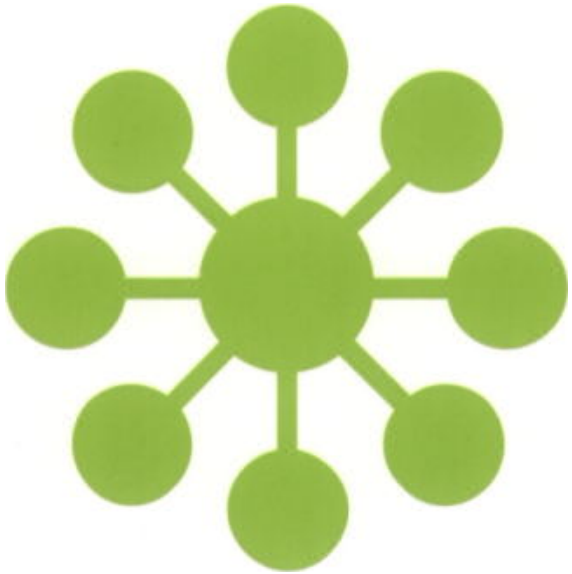
dell.z@ieee.org

CIKM 2008

Outline

1. Introduction
2. Problem
3. Related Work
4. Our Approach
5. Experiments
6. Conclusions

Introduction



Everything is Connected

Introduction

Large. Sparse. Small-World. Scale-Free.

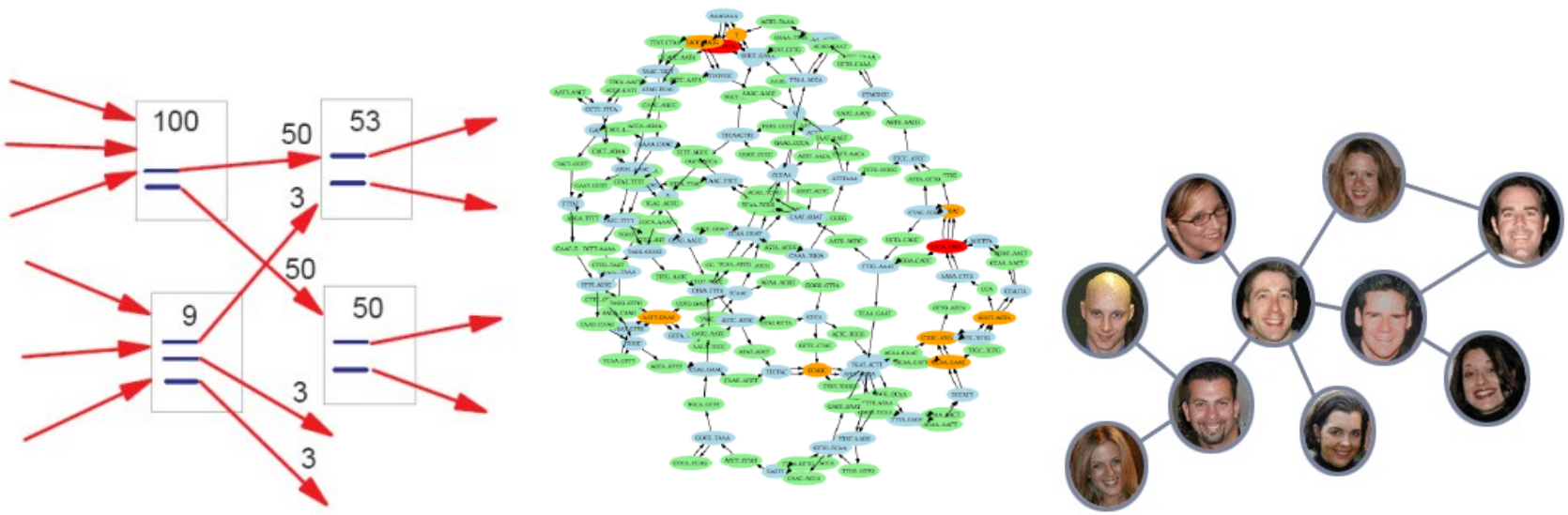


Figure: Examples of complex networks.

Introduction

10 Challenging Problems in Data Mining Research — ICDM'05

- Mining Complex Knowledge from Complex Data
  - Data that are not i.i.d.
- Mining in a Network Setting
- ...

Outline

1. Introduction
2. Problem
3. Related Work
4. Our Approach
5. Experiments
6. Conclusions

Problem

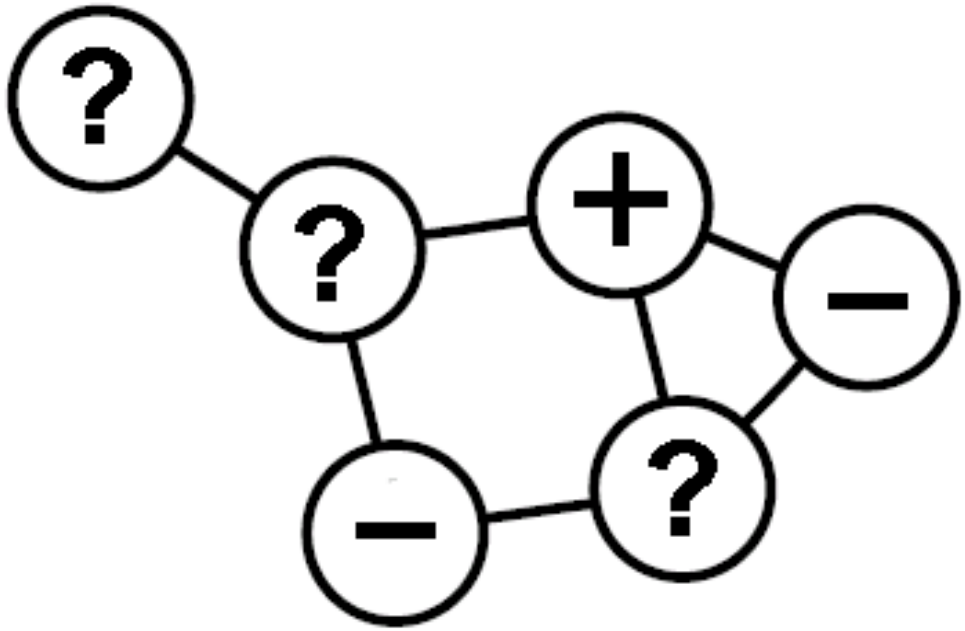


Figure: Classifying networked entities.

Problem

Input:

- $X = \{\mathbf{x}_i\}_{i=1}^{n} \left[\begin{array}{ll} X_l := \{\mathbf{x}_i\}_{i=1}^{l} & Y_l := \{y_i\}_{i=1}^{l} \\ X_u := \{\mathbf{x}_j\}_{j=l+1}^{l+u} & Y_u := \emptyset \end{array} \right.$
- $\mathbf{A} = (A_{ij})_{n \times n}$ adjacency matrix
  - sparse and symmetric
  - $k_i = \sum_j A_{ij}$

Output:

- $f(\mathbf{x})$: classification function

Outline

1. Introduction
2. Problem
3. Related Work
4. Our Approach
5. Experiments
6. Conclusions

Related Work

Feature Engineering or Dimensionality Reduction

- Probabilistic HITS (PHITS) + Probabilistic LSI (PLSI)
- Matrix Factorisation (MF) / Supervised Matrix Factorisation (SupMF)
- ...

Collective Inference or Relational Learning

- Markov Random Fields (MRF)
- ...

Graph-based Semi-Supervised Learning

- Directed Graph Regularisation (DGR)
- Laplacian Kernel (LapKer)
- ...

Outline

1. Introduction
2. Problem
3. Related Work
4. Our Approach
5. Experiments
6. Conclusions

Learning Framework

Kernel Methods for Semi-Supervised Learning

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + C \|f\|_K^2,$$

- V is a loss function defined on X_l
  - RLR: logistic loss $\ln(1 + \exp(-y_i f(\mathbf{x}_i)))$
  - SVM: hinge loss $(1 - y_i f(\mathbf{x}_i))_+ = \max(0, 1 - y_i f(\mathbf{x}_i))$
- $\|f\|_K^2$ is a regulariser defined on $X_l \bigcup X_u$
  - Kernel $K : X \times X \to \mathbb{R}$
  - RKHS $\mathcal{H}_K$ of functions $X \to \mathbb{R}$ with norm $\| \|_K$

Learning Framework

Kernel Methods for Semi-Supervised Learning

- An extension of the Representer Theorem

$$f^*(\mathbf{x}) = \sum_{i=1}^{l+u} \alpha_i K(\mathbf{x}_i, \mathbf{x})$$

- Convex optimisation over $\{\alpha_i\}_{i=1}^{l+u}$

Learning Framework

What regularisers or kernels are good for classification of networked entities?

Community Discovery

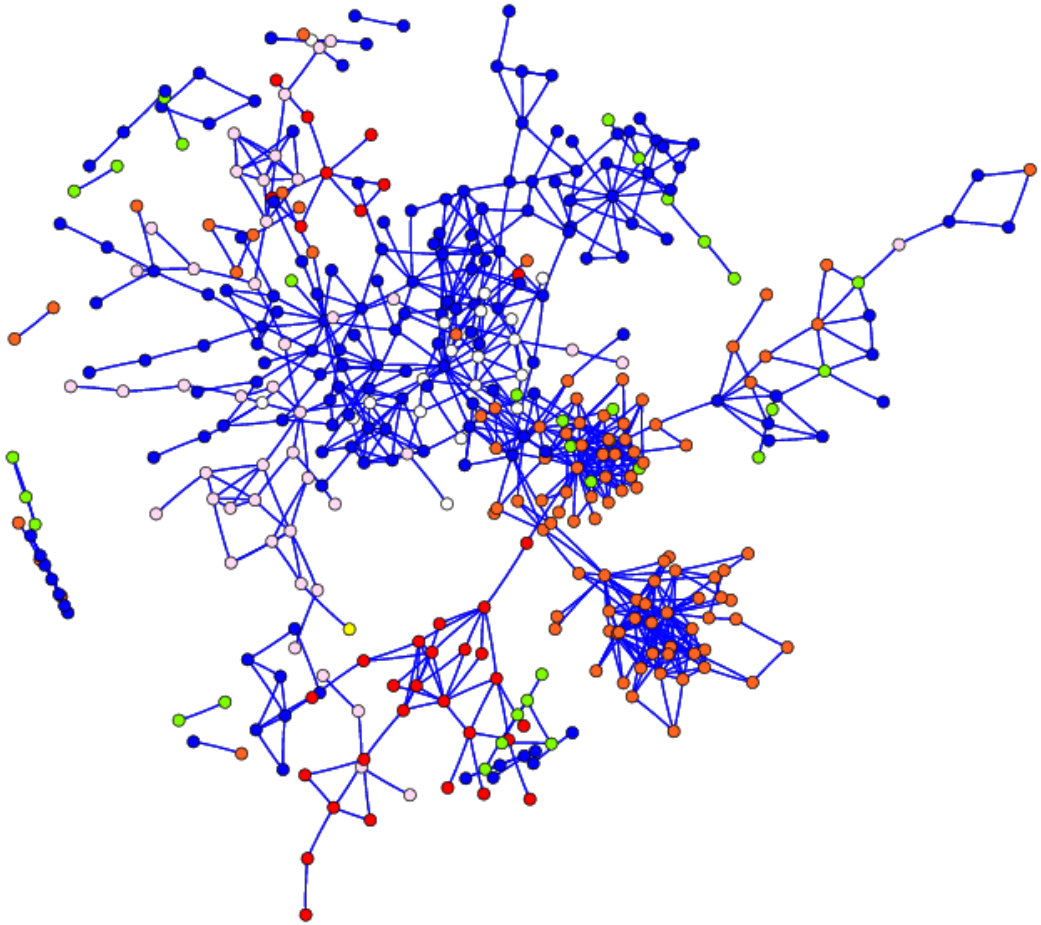


Figure: The community structure of the Cora-HA network.

Community Discovery

A Division of the Network into Communities

$$\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$$

- $f(\mathbf{x}_i) \in \{-1, +1\}$, $\mathbf{f}^T\mathbf{f} = n$
- Real relaxation: $f(\mathbf{x}_i) \in \mathbb{R}$

Community Discovery

Minimising Cut-Size (Spectral Graph Partitioning)

$$S = \frac{1}{2}\sum_{i,j} A_{ij}(1 - \delta(g_i, g_j)) = \frac{1}{4}\mathbf{f}^T\mathbf{L}\mathbf{f}$$

- Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$
  - $\mathbf{D} = \mathrm{diag}(k_1, \ldots, k_n)$

Community Discovery

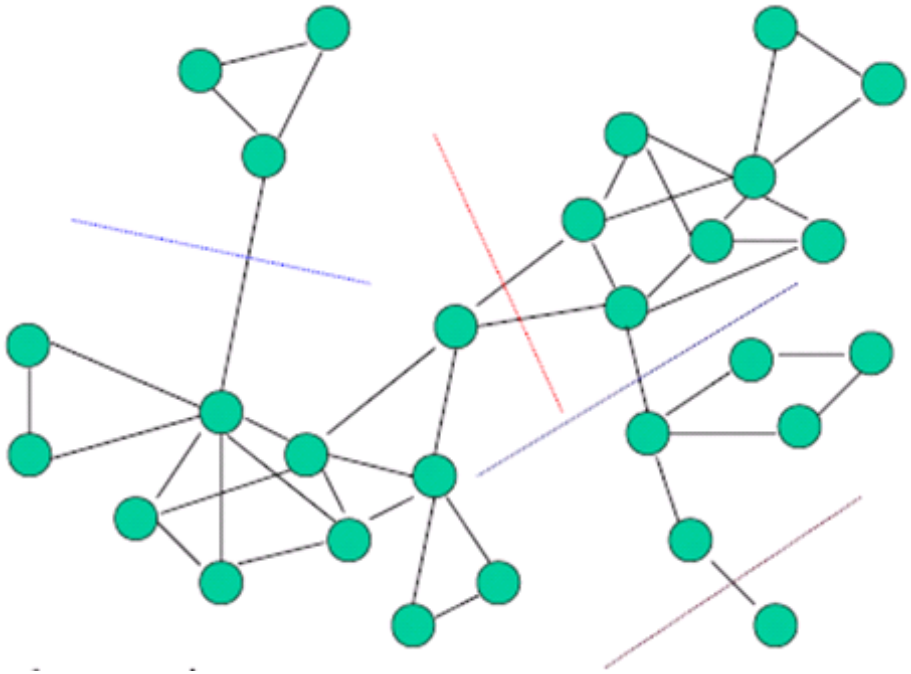


Figure: Spectral graph partitioning via minimising cut-size.

Community Discovery

- $\mathbf{L}$ is positive semi-definite
  - eigenvalues: $0 = \lambda_1 = \ldots = \lambda_z < \lambda_{z+1} \leq \ldots \leq \lambda_n$
  - eigenvectors: $\mathbf{u}_1, \ldots, \mathbf{u}_z, \mathbf{u}_{z+1}, \ldots, \mathbf{u}_n$
- Optimal non-trivial division: $\mathbf{f} = \mathbf{u}_{z+1}$
  - The number of edges across communities is small
- Normalized Laplacian $\tilde{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$
  - Normalized Cut

Kernel

Laplacian Kernel (LapKer)

$$\mathbf{K} = \mathbf{L}^{+}$$

- $\|f\|_K^2 = \mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{k=z+1}^{n} \frac{1}{\lambda_k}\mathbf{u}_k\mathbf{u}_k^T$
- Physical Interpretation
  - Resistance Distance
  - Commute Time

Community Discovery

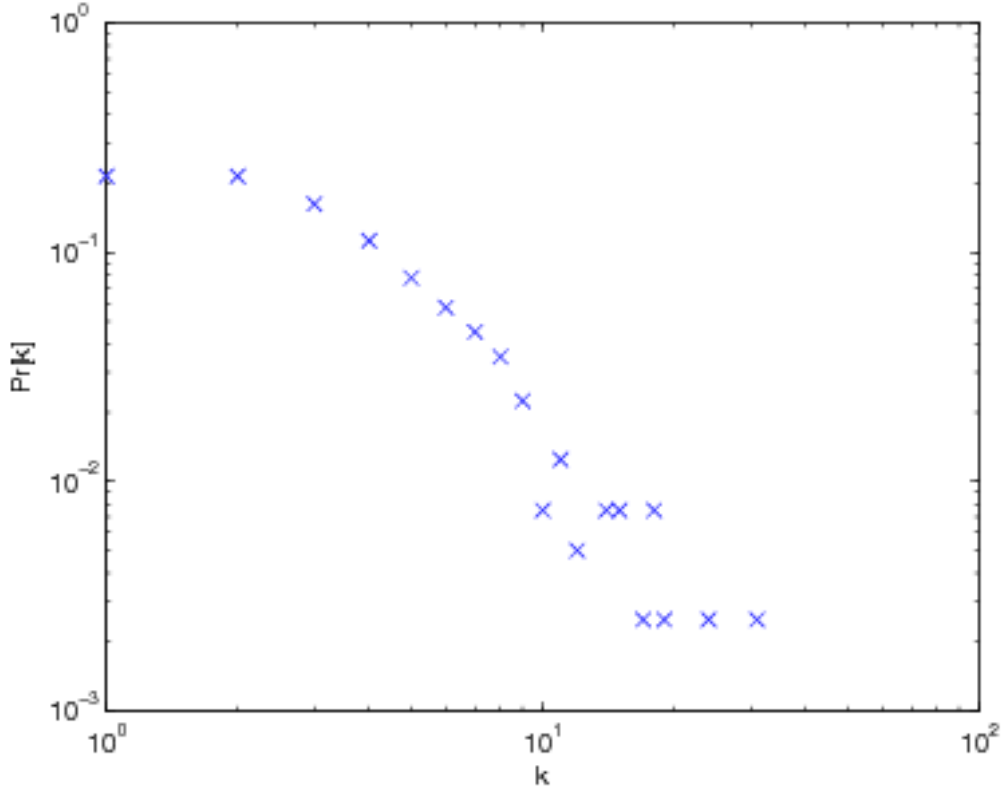


Figure: The degree distribution ($\Pr[k] \sim k$) of the Cora-HA network.

Community Discovery

Intuitive sense: absolute edge weight → relative edge weight

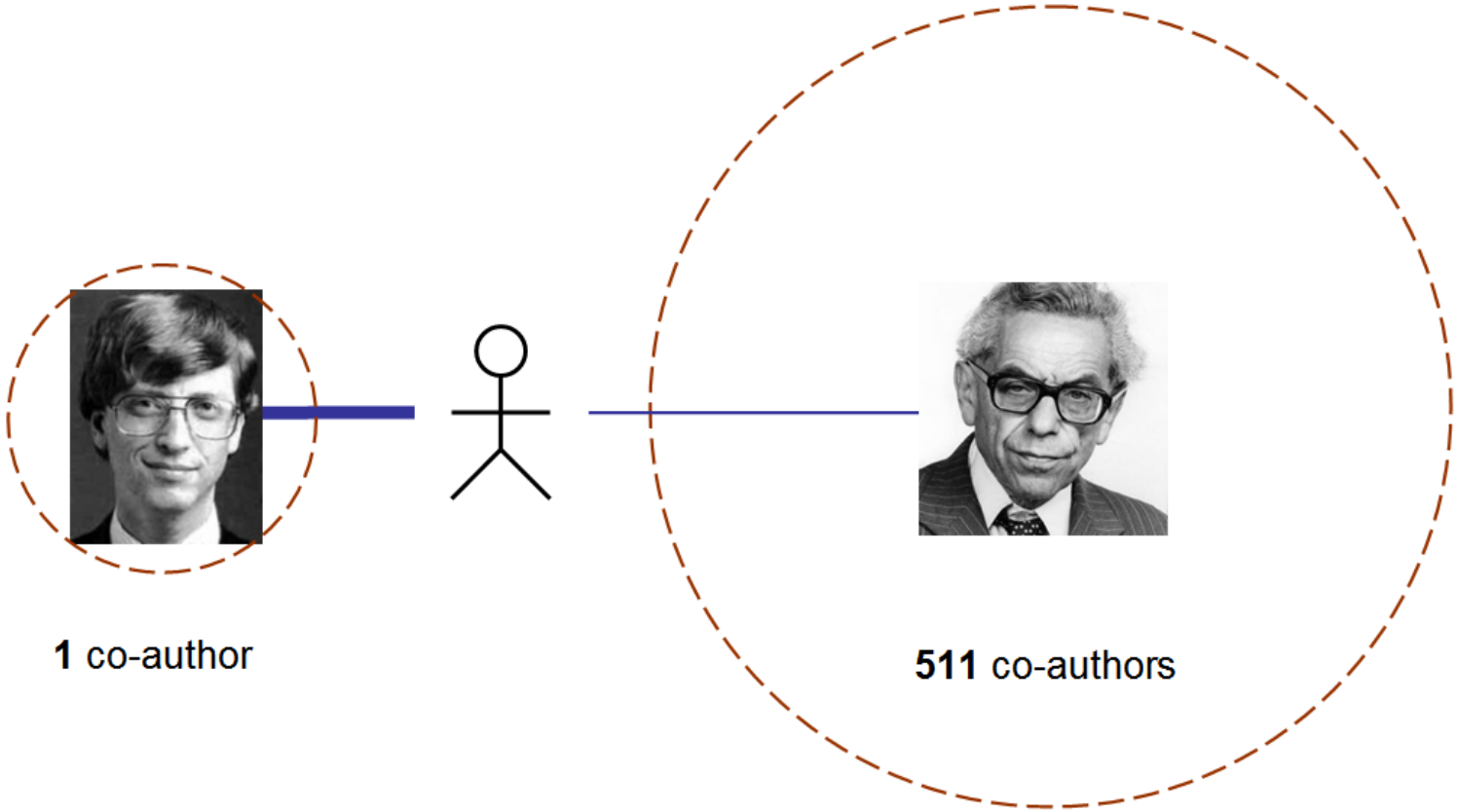


Figure: Links are not equal.

Community Discovery

The **null model**: a *random* graph with the same degree distribution as the given network

- The *expected* number of edges between node $\mathbf{x}_i$ and node $\mathbf{x}_j$ is

$$P_{ij} = (k_i k_j)/(2m)$$

Community Discovery

Maximising Modularity

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - P_{ij}]\delta(g_i, g_j) = \frac{1}{4m}\mathbf{f}^T\mathbf{M}\mathbf{f}$$

- Modularity matrix $\mathbf{M} = \mathbf{A} - \mathbf{P}$
  - Generalisation: $\mathbf{M} = \mathbf{A} - \eta\mathbf{P}$ with $\eta \geq 0$

Community Discovery

- $\mathbf{M}$ is not guaranteed to be positive semi-definite
  - eigenvalues: $\lambda_1 \geq \ldots \geq \lambda_p > 0 \geq \lambda_{p+1} \geq \ldots \geq \lambda_n$
  - eigenvectors: $\mathbf{u}_1, \ldots, \mathbf{u}_p, \mathbf{u}_{p+1}, \ldots, \mathbf{u}_n$
- Optimal division: $\mathbf{f} = \mathbf{u}_1$
  - The number of edges across communities is smaller than expected

Community Discovery

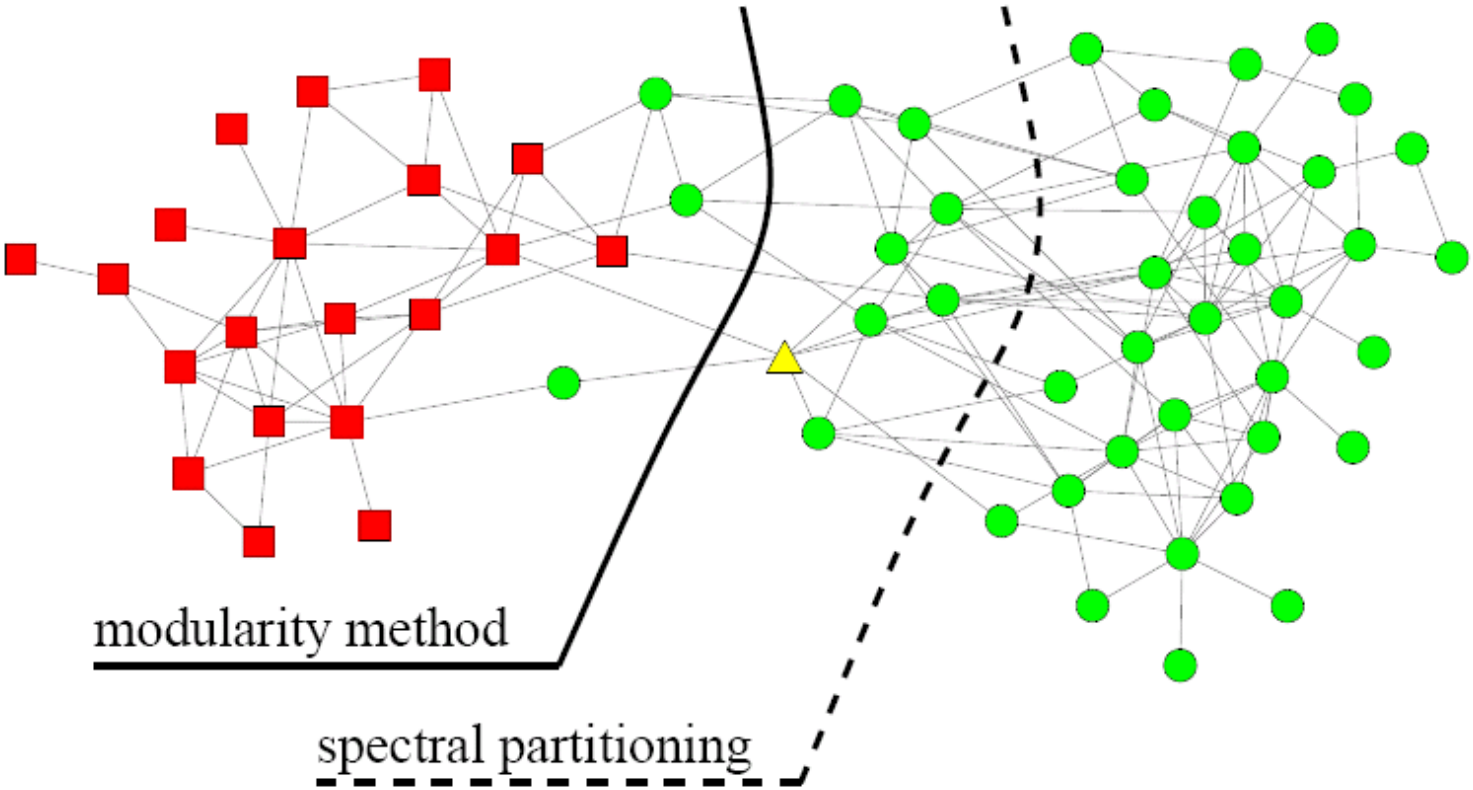


Figure: Modularity vs. Laplacian (the dolphin social network).

Kernel

Modularity Kernel (ModKer)

$$\mathbf{K} = \widehat{\mathbf{M}} = \sum_{k=1}^{p} \lambda_k \mathbf{u}_k \mathbf{u}_k^T$$

- $\|f\|_K{}^2 = \langle \mathbf{f}, \mathbf{f} \rangle = \mathbf{f}^T \widehat{\mathbf{M}}^{-1} \mathbf{f}$
  - Unsupervised learning backs off to modularity-based community discovery
- $\widehat{\mathbf{M}}$ is the positive definite matrix that best approximates $\mathbf{M}$ in terms of least squared errors
  - It is a valid kernel function and leads to a convex optimisation problem

Combining Content and Link

Linear Combination with Parameter $\leq \mu \leq 1$

- Regulariser Combination

$$\|f\|^2 = (1-\mu)\|f\|^2_{content} + \mu\|f\|^2_{link}$$

- Kernel Combination

$$\mathbf{K} = (1-\mu)\mathbf{K}_{content} + \mu\mathbf{K}_{link}$$

- Graph Combination

$$\mathbf{A} = (1-\mu)\mathbf{A}_{content} + \mu\mathbf{A}_{link}$$

Outline

1. Introduction
2. Problem
3. Related Work
4. Our Approach
5. Experiments
6. Conclusions

Experiments

Table: Characteristics of the WebKB datasets.

dataset	Cornell	Texas	Washington	Wisconsin
the number of classes	7	7	7	6
the number of entities (nodes)	827	814	1166	1210
the number of terms	4134	4029	4165	4189
the number of edges	49560	59620	80564	91244
the minimum degree of a node	0	0	0	0
the maximum degree of a node	478	533	912	843
the median degree of a node	14	17	15	21
the average degree of a node	59.93	73.24	69.09	75.41

Co-citation Graph

Experiments

Table: Characteristics of the Cora datasets.

dataset	DS	HA	ML	PL
the number of classes	9	7	7	9
the number of entities (nodes)	751	400	1617	1575
the number of terms	6234	3989	8329	7949
the number of edges	2566	1586	8092	9836
the minimum degree of a node	1	1	1	1
the maximum degree of a node	32	31	55	76
the median degree of a node	2	3	4	4
the average degree of a node	3.42	3.97	5.00	6.25

Undirected Graph

Experiments

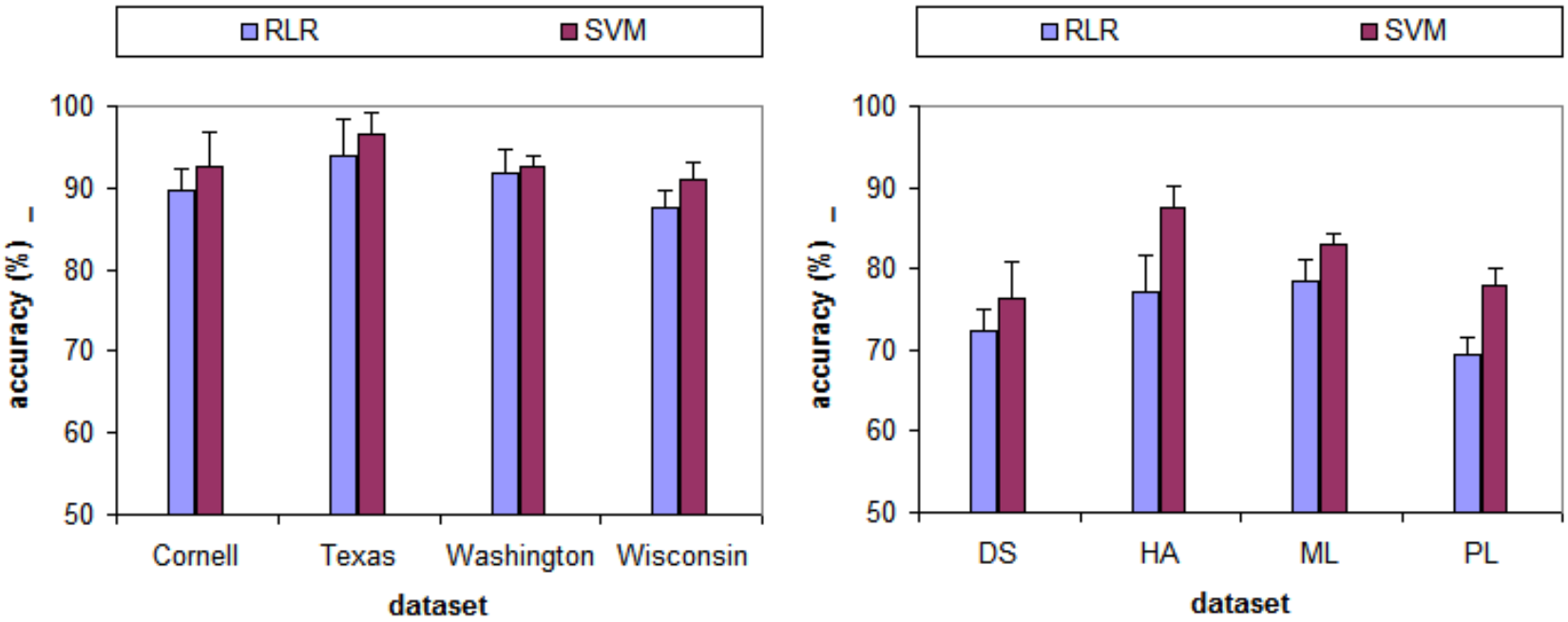


Figure: Comparison of learning algorithms.

Experiments

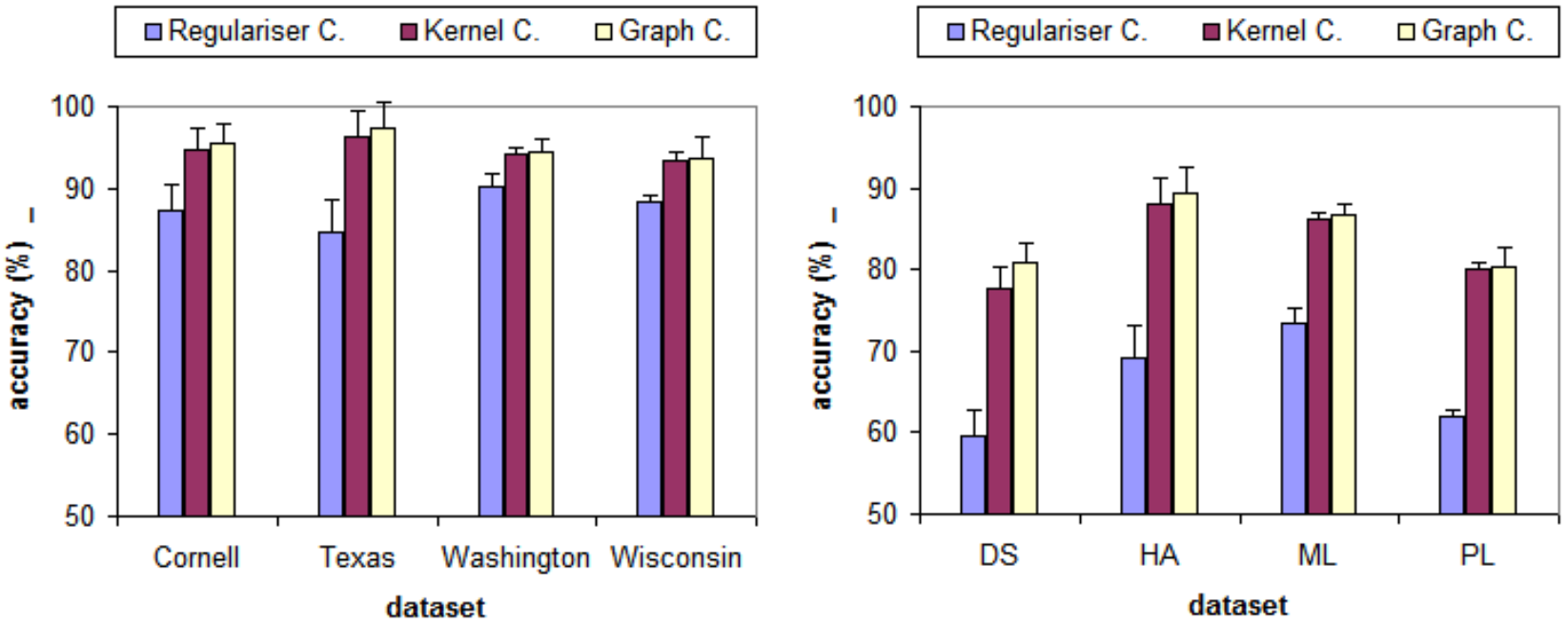


Figure: Comparison of combination methods.

Experiments

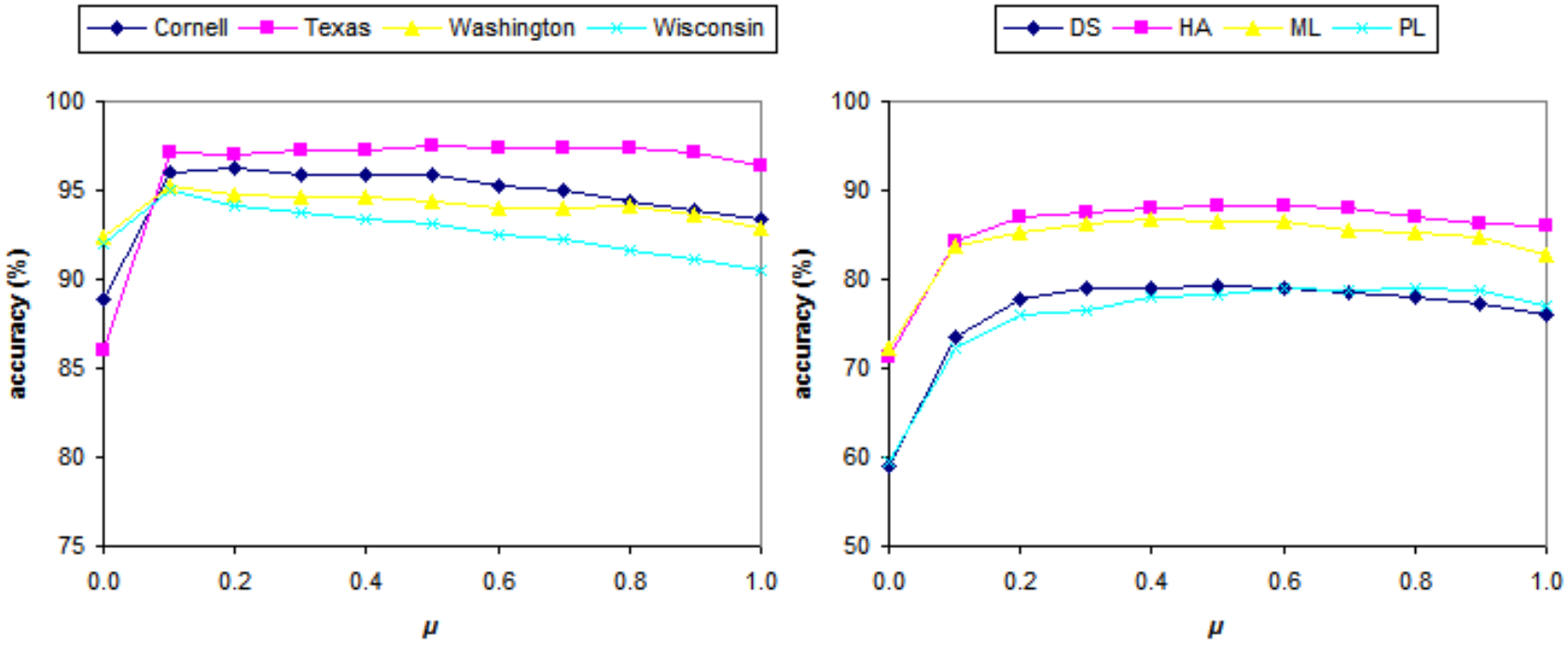


Figure: Graph combination with parameter μ.

Experiments

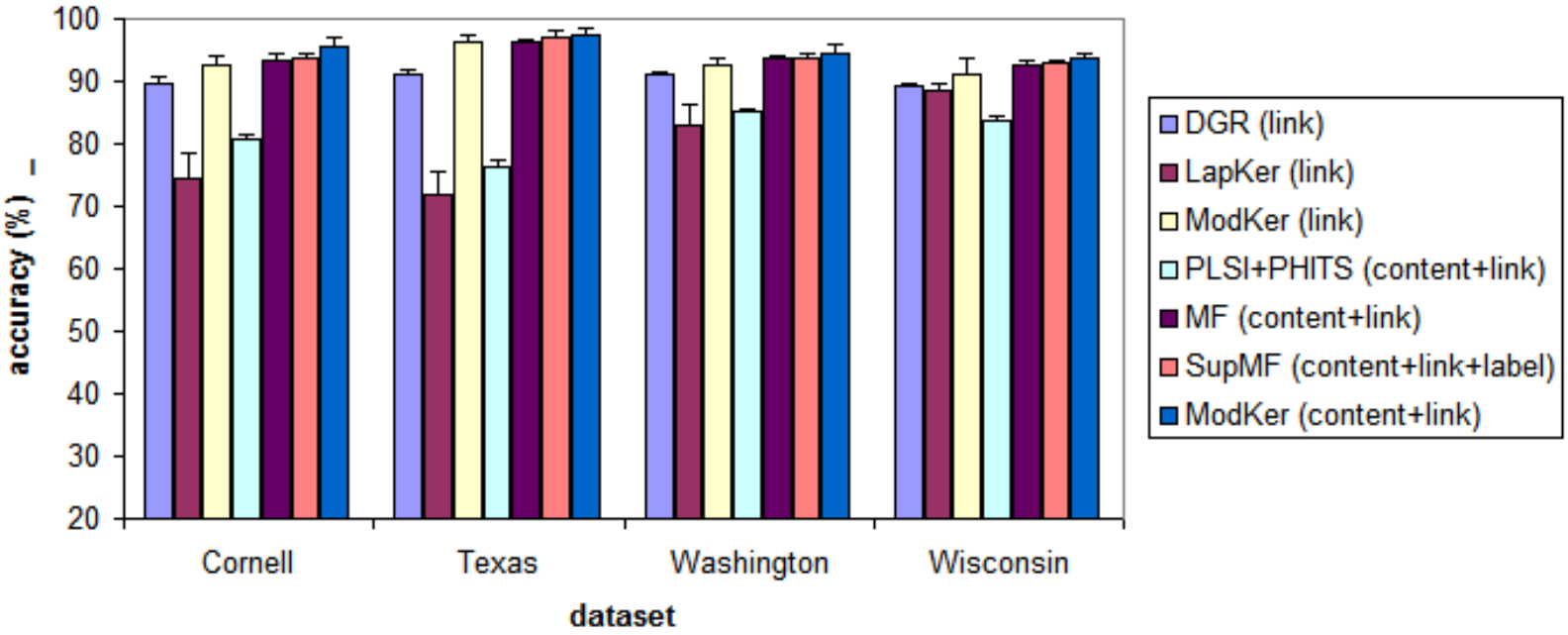


Figure: Comparison with other approaches on the WebKB datasets.

Experiments

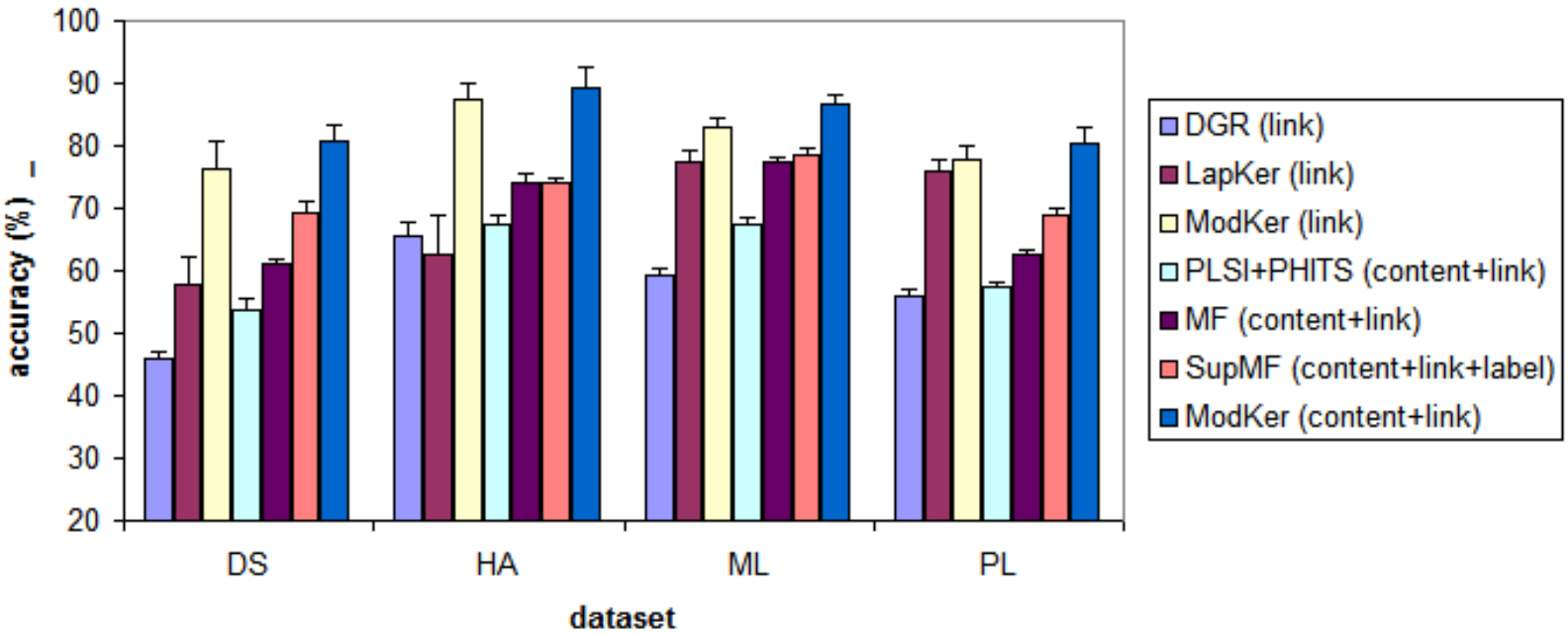


Figure: Comparison with other approaches on the Cora datasets.

Outline

1. Introduction
2. Problem
3. Related Work
4. Our Approach
5. Experiments
6. Conclusions

Conclusions

- Classifying Networked Entities
- Semi-Supervised Learning
- Kernel Methods
- Community Discovery
- Graph Combination

Take Home Message

$$\text{Laplacian} \Longrightarrow \text{Modularity}$$

Q&A

Thanks :)