

Modularity Kernel

A New Kernel for Classification of Networked Entities

Dell Zhang

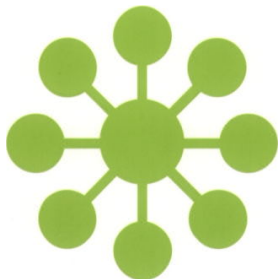
Birkbeck, University of London

dell.z@ieee.org

4 July 2008

Outline

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



Everything is Connected

Introduction

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

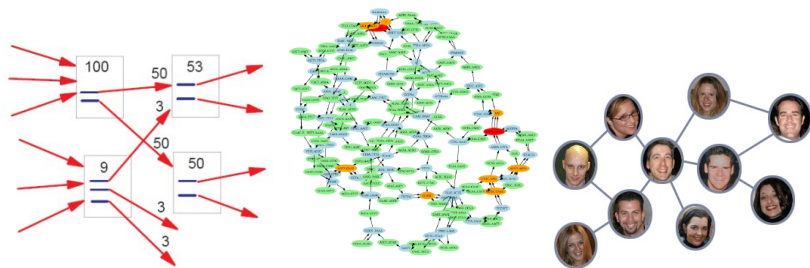


Figure: Examples of complex networks.

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

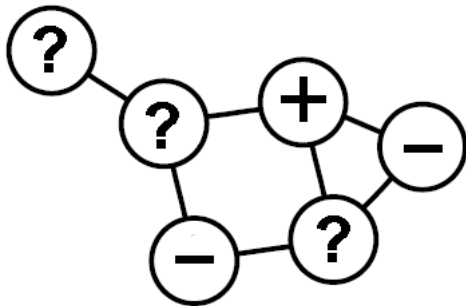


Figure: Classifying networked entities.

Problem

Input:

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

Output:

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

Outline

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

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

Kernel Methods for Semi-Supervised Learning

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

- Kernel $K : X \times X \rightarrow \mathbb{R}$
 - RKHS \mathcal{H}_K of functions $X \rightarrow \mathbb{R}$ with norm $\|\cdot\|_K$
- 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 $X_l \cup X_u$

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 $\{\alpha_i\}_{i=1}^{l+u}$

Community Detection

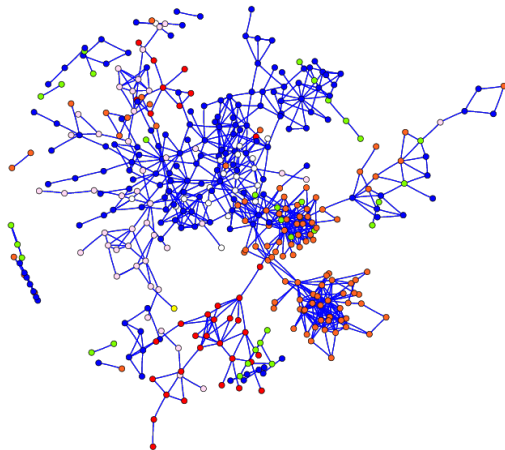


Figure: The community structure of the Cora-HA network.

A Division of the Network into Communities

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

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

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 $\mathbf{L} = \mathbf{D} - \mathbf{A}$
 - $\mathbf{D} = \text{diag}(k_1, \dots, k_n)$

Community Detection

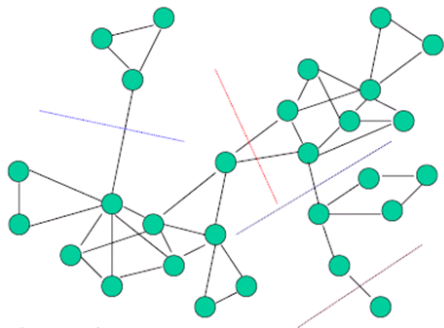


Figure: Spectral graph partitioning via minimising cut-size.

Community Detection

- \mathbf{L} is positive semi-definite
 - eigenvalues: $0 = \lambda_1 = \dots = \lambda_z < \lambda_{z+1} \leq \dots \leq \lambda_n$
 - eigenvectors: $\mathbf{u}_1, \dots, \mathbf{u}_z, \mathbf{u}_{z+1}, \dots, \mathbf{u}_n$
- 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

Laplacian Kernel (LapKer)

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

- resistance distance
- commute time

Community Detection

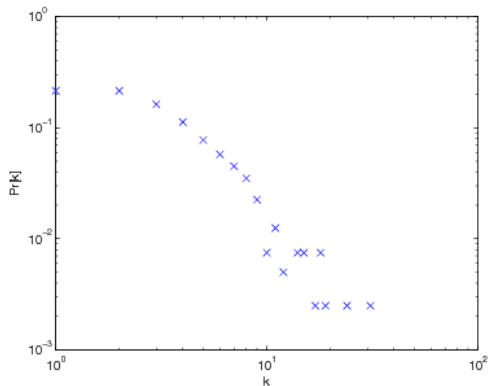


Figure: The degree distribution ($\text{Pr}[k] \sim k$) of the Cora-HA network.

Community Detection

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

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

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 $\mathbf{M} = \mathbf{A} - \mathbf{P}$

Community Detection

- \mathbf{M} is not guaranteed to be positive semi-definite
 - eigenvalues: $\lambda_1 \geq \dots \geq \lambda_n$
 - eigenvectors: $\mathbf{u}_1, \dots, \mathbf{u}_n$
- Optimal division: $\mathbf{f} = \mathbf{u}_1$
 - The number of edges across communities is **smaller than expected**

Community Detection

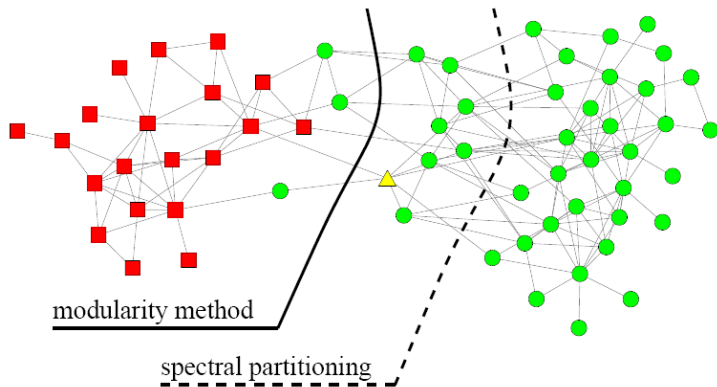


Figure: Modularity vs. Laplacian (the dolphin social network).

Modularity Kernel (ModKer)

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

- $\lambda_1 \geq \dots \geq \lambda_p > 0$ are the p positive eigenvalues of \mathbf{M} and $\mathbf{u}_1, \dots, \mathbf{u}_p$ are the p corresponding eigenvectors
- $\hat{\mathbf{M}}$ is the positive definite matrix that best approximates \mathbf{M}
 - It is a valid kernel function and leads to a convex optimisation problem
- $\|f\|^2 = \langle \mathbf{f}, \mathbf{f} \rangle = \mathbf{f}^T \hat{\mathbf{M}}^{-1} \mathbf{f}$
 - Unsupervised learning backs off to modularity-based community detection

Combining Content and Link

$$0 \leq \mu \leq 1$$

- Regulariser Combination

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

- 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

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

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

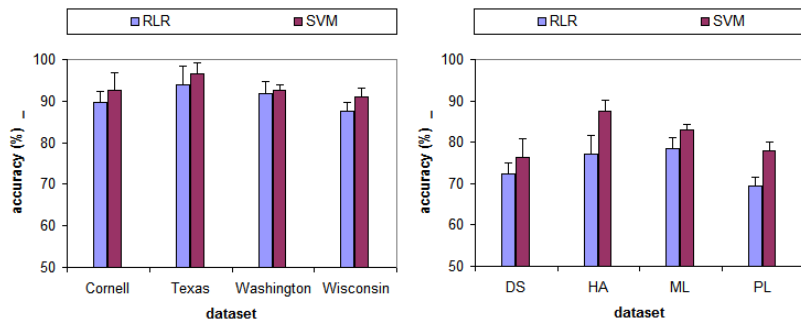


Figure: Comparison of learning algorithms.

Experiments

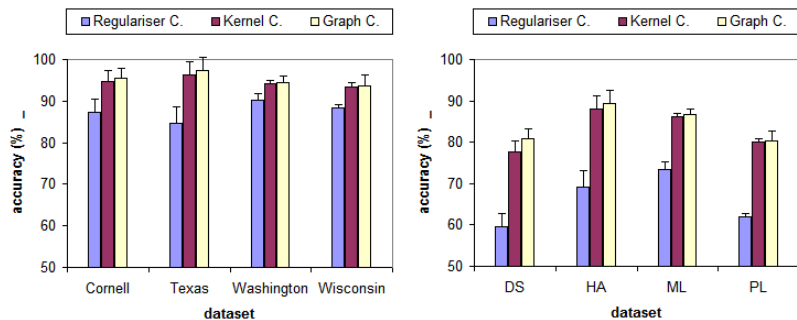


Figure: Comparison of combination methods.

Experiments

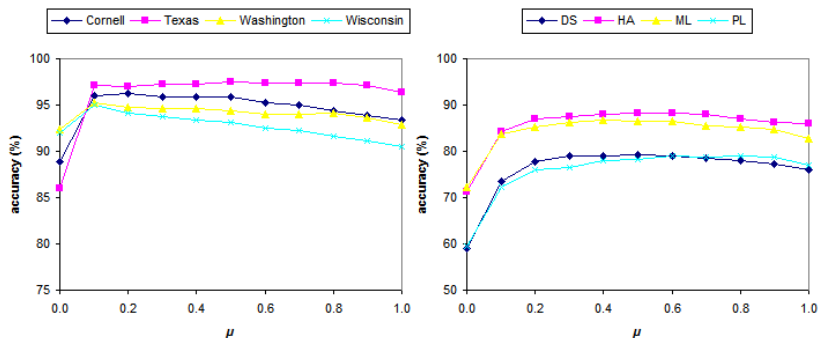


Figure: Graph combination with parameter μ .

Experiments

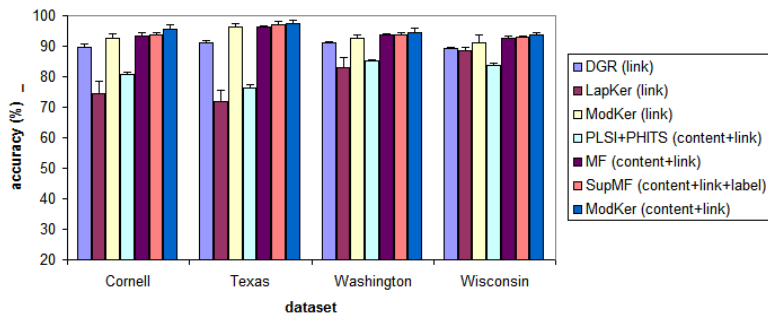


Figure: Comparison with other approaches on the WebKB datasets.

Experiments

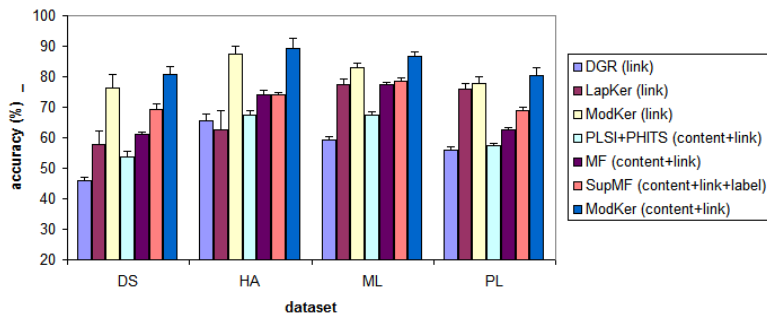


Figure: Comparison with other approaches on the Cora datasets.

Outline

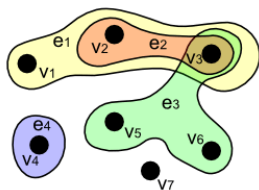
- 1 Introduction
- 2 Problem
- 3 Related Work
- 4 Our Approach
- 5 Experiments
- 6 Conclusions**

Conclusions

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

Future Work

- Undirected Graph \rightarrow Directed Graphs
- Graph \rightarrow Hypergraph and Matrix \rightarrow Tensor



Array RGB

Page 1	Page 2	Page 3
0.112 0.986 0.234 0.432 ...	0.342 0.647 0.515 0.816 ...	0.689 0.709 0.118 0.884 ...
0.765 0.128 0.863 0.521 ...	0.111 0.300 0.205 0.526 ...	0.535 0.532 0.653 0.905 ...
1.000 0.985 0.761 0.698 ...	0.529 0.428 0.712 0.955 ...	0.314 0.288 0.159 0.261 ...
0.455 0.783 0.224 0.395 ...	0.214 0.604 0.918 0.344 ...	0.553 0.633 0.528 0.493 ...
0.021 0.600 0.311 0.129 ...	0.100 0.121 0.173 0.126 ...	0.441 0.465 0.812 0.512 ...
1.000 1.000 0.887 0.951 ...		0.908 0.444 0.481 0.398 ...
1.000 0.945 0.998 0.893 ...		0.912 0.713 ...
0.990 0.941 1.000 0.876 ...		0.219 0.328 ...
0.902 0.887 0.834 0.798 ...		0.128 0.133 ...
...		

- Clustering, Ranking, etc.

- Mining and Learning with Graphs (MLG)
- Graph-based Algorithms for NLP (TextGraphs)
- Algorithms and Models for the Web-Graph (WAW)
- Mining Graphs and Complex Structures (MGCS)
- Graph-Theoretic Concepts in Computer Science (WG)

Laplacian \implies Modularity

Complex Networks + Machine Learning

Thanks :)