Semi-supervised learning on graphs

via stationary empirical correlations

Ya Xu

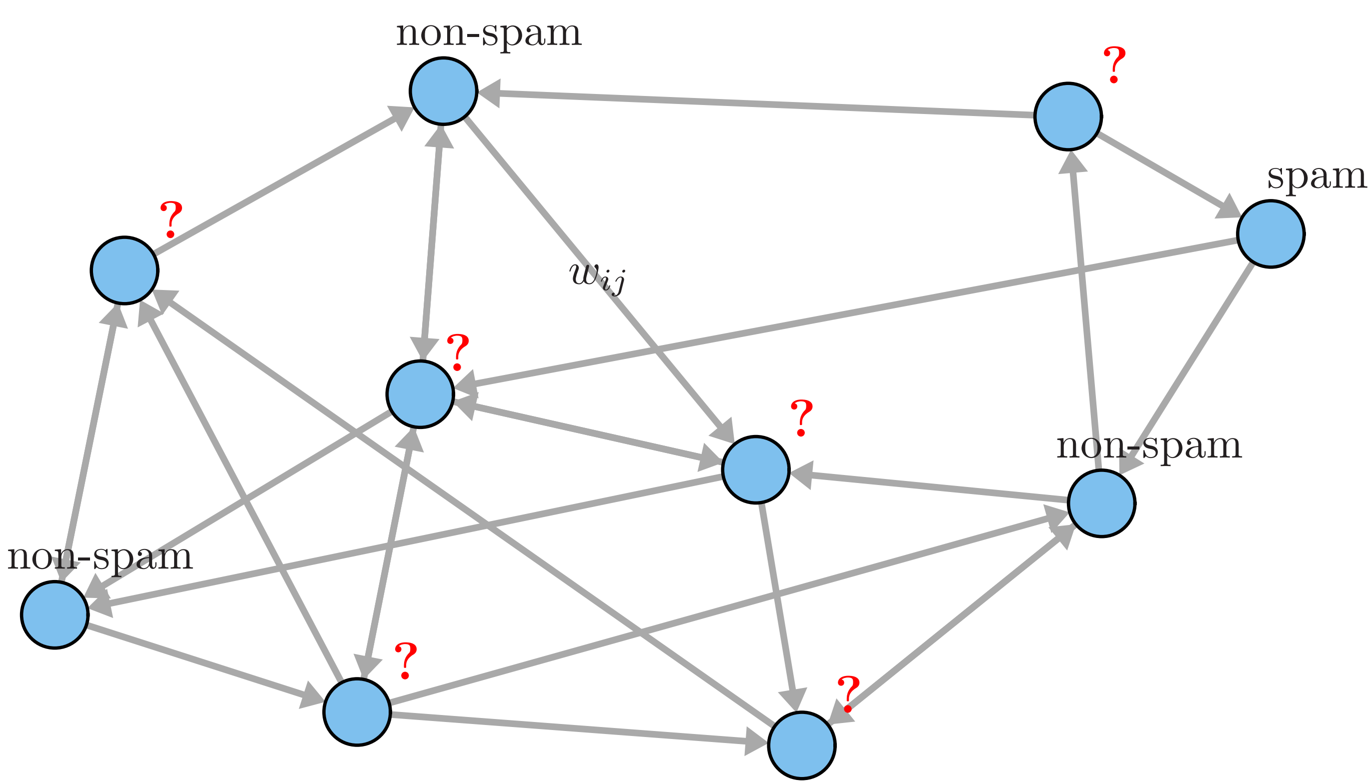
Justin S. Dyer

Art B. Owen

Department of Statistics

Stanford University

Prediction on graphs



Some nodes are labeled, some not.

We want to predict the unlabeled using labels and graph structure.

Operative assumption: nearby nodes are similar.

The story in one slide

1) Many graph-based predictions are linear in the observed responses.
2) So there's a "Gaussian model" story.
3) We find the implied correlations,
4) and replace them with empirical ones.
5) So far, it makes a big improvement.
6) We did small examples, but with scaling in mind

Why it improves

The semi-supervised learning methods we found had preconceived notions of how correlation varies with graph distance.

We estimate the correlation vs distance pattern from the data.

Graph Notation

G	The graph
w_{ij}	Edge weight from i to j
W	Adjacency matrix
$w_{i+} = \sum_j w_{ij}$	In-degree of i
$w_{+j} = \sum_i w_{ij}$	Out-degree of j
$w_{++} = \sum_{ij} w_{ij}$	Graph volume
Y_i	Response value at node i
$Y^{(0)}$	Measured responses $i = 1, \ldots, r$
$Y^{(1)}$	Unknown responses $i = r+1, \ldots, n$

Graph random walk

Transition probability $\quad P_{ij} = \dfrac{w_{ij}}{w_{i+}}$

Stationary distribution $\quad \pi_i \quad$ e.g. PageRank

The associated random walk leaves node i for node j with probability proportional to w_{ij}.

We assume it is aperiodic and irreducible. (If necessary add teleportation.)

$\therefore$ it has a stationary distribution π

Graph Laplacian

$$\Delta_{ij} = \begin{cases} w_{i+} - w_{ii} & i = j \\ -w_{ij} & i \neq j \end{cases}$$

needed later

Zhou, Huang, Schölkoff (2005)

Node similarity:

$$s_{ij} \equiv \pi_i P_{ij} + \pi_j P_{ji}$$

Variation functional:

$$\Omega(Z) = \frac{1}{2} \sum_{i,j} s_{ij} \Big(\frac{Z_i}{\sqrt{\pi_i}} - \frac{Z_j}{\sqrt{\pi_j}} \Big)^2$$

Criterion:

$$\widehat{Z} = \arg \min_{Z \in \mathbb{R}^n} \Omega(Z) + \lambda \|Z - Y^*\|^2$$

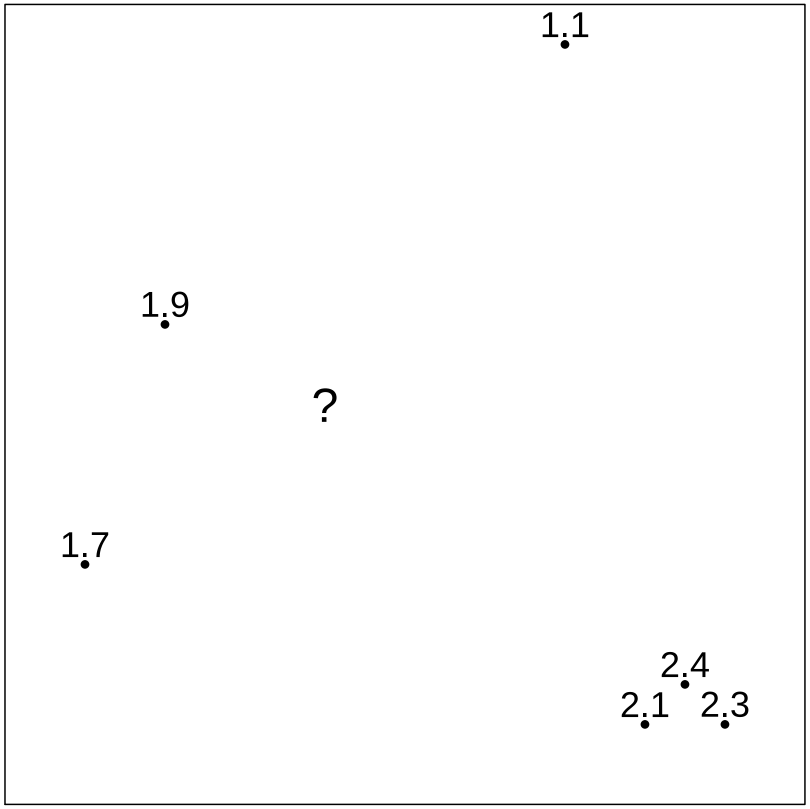
$$Y_i^* = \begin{cases} Y_i & \text{observed} \\ \mu_i & \text{(default, e.g. } 0\text{) otherwise} \end{cases}$$

ZHS trade off fit to observations vs graph smoothness via λ.

Result is a linear function of Y^*

There must be an equivalent Gaussian process story

Kriging



1) Predict at '?' by weighting the obs
2) 1.9 gets more weight than 1.7 because it is closer
3) the $\doteq$ 2s get more weight than the 1.1, because there are 3 of them
4) but not triple the weight, because they're somewhat redundant
5) the tradeoffs come from a Gaussian covariance model

The model originated in geostatistics

Kriging model

obs	$Y = \nu\beta + S + \varepsilon \in \mathbb{R}^n$	
coefficients	$\beta \in \mathbb{R}^k$	we'll have $k = 1$
predictors	$\nu \in \mathbb{R}^{n\times k}$	e.g. $\nu = \sqrt{\pi}$ or $\mathbf{1}_n$
correlated part	$S \sim \mathrm{N}(0, \Sigma)$	
noise	$\varepsilon \sim \mathrm{N}(0, \Gamma)$	Γ is diagonal

Predictions

Now $Y = Z + \varepsilon$, for **signal** $Z = \nu\beta + S$

Taking ν fixed and $\beta \sim \mathrm{N}(\mu, \delta^{-1})$

makes $Z \sim \mathrm{N}(\mu\nu, \Psi)$, $\Psi = \nu\nu'\delta^{-1} + \Sigma$

Predict by $\widehat{Z} = \mathbb{E}(Z \mid Y^{(0)})$

Kriging some more

Z is signal $Y^{(0)}$ has observed responses

Partition Ψ

$$\Psi = \text{Cov}\begin{pmatrix} Z^{(0)} \\ Z^{(1)} \end{pmatrix} = \begin{pmatrix} \Psi_{00} & \Psi_{01} \\ \Psi_{10} & \Psi_{11} \end{pmatrix} = \begin{pmatrix} \Psi_{\bullet 0} & \Psi_{\bullet 1} \end{pmatrix}.$$

Joint distribution of signal (everywhere) and observations

$$\begin{pmatrix} Z \\ Y^{(0)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu\nu \\ \mu\nu_0 \end{pmatrix}, \begin{pmatrix} \Psi & \Psi_{\bullet 0} \\ \Psi_{0\bullet} & \Psi_{00} + \Gamma_{00} \end{pmatrix}\right)$$

$\cdots$ yields expression for $\mathbb{E}(Z \mid Y^{(0)})$

ZHS method as kriging

Let $\Pi = \textbf{diag}(\pi_i)$ and define

$$\widetilde{\Delta}_{ij} = \begin{cases} s_{i+} - s_{ii} & i = j \\ -s_{ij} & i \neq j \end{cases}$$

The Laplacian after replacing w_{ij} by $s_{ij} = \pi_i P_{ij} + \pi_j P_{ji}$

Choose

noise variance $\Gamma = \lambda^{-1} I_n$

signal variance $\Sigma = \Pi^{1/2} \widetilde{\Delta}^{+} \Pi^{1/2}$ ($+$ for generalized inverse)

predictors $\nu = \textbf{diag}(\sqrt{\pi_i})'$

defaults $\mu_i = \mu \nu_i, \quad r + 1 \leq i \leq n$

Then

$\lim_{\delta \to 0^+}$ Kriging$(\Gamma, \Sigma, \nu, \delta) =$ ZHS method

Interpretation

The ZHS method is a kind of kriging

The correlation matrix depends on the graph but not on the nature of the response

This seems strange: shouldn't some variables correlate strongly with their neighbors, others weakly and still others negatively?

It also anticipates $Z \propto \sqrt{\pi}$ (for every response variable)

Belkin, Matveeva, Niyogi (2004)

Graph Tikhonov regularization

$$Z'\Delta Z + \lambda_0 \|Z^{(0)} - Y^{(0)}\|^2$$

Δ is the graph Laplacian, penalty is only on observed responses

As kriging

noise variance $\Gamma = \textbf{diag}(\lambda_0^{-1} I_r, \lambda_1 I_{n-r})$

signal variance $\Sigma = \Delta^+$ (no $\Pi^{1/2}$)

predictors $\nu = \mathbf{1}_n$ (no $\sqrt{\pi_i}$)

let $\delta \to 0^+$ and then let $\lambda_1 \to 0^+$

Zhou et al (2004)

Undirected graph precursor to ZHS, using $D_{ii} = w_{i+} = w_{+i}$:

$$\frac{1}{2}\sum_{i,j} w_{ij}\Big(\frac{Z_i}{\sqrt{D_{ii}}} - \frac{Z_j}{\sqrt{D_{jj}}}\Big)^2 + \lambda\|Z - Y^*\|^2$$

As kriging

noise variance $\Gamma = \lambda^{-1} I$

signal variance $\Sigma = D^{1/2}\Delta^{+}D^{1/2}$

predictors $\nu = \textbf{diag}(\sqrt{D_{ii}})$

with $\delta \rightarrow 0^{+}$

More examples

Zhou, Schölkopf, Hofmann (2005)

They define a hub walk and an authority walk. Each has a transition matrix, stationary distribution, similarity matrix and similarity-Laplacian. They replace $\Omega(Z)$ by the convex combination

$$\gamma\Omega_H(Z) + (1-\gamma)\Omega_A(Z), \qquad 0 < \gamma < 1.$$

The resulting signal variance is the corresponding convex combination of hub and authority signal variance matrices.

Belkin, Niyogi, Sindhwani (2006) Manifold regularization. Get covariance $(K + \gamma\Delta)^{-1}$ when their Mercer kernel is linear with matrix K.

Kondor and Lafferty (2002) and Smola and Kondor (2003) and Zhu, Ghahramani and Lafferty (2003) use spectral criterion $Z'LZ$ where $L = \sum_i f(d_i)u_i u_i'$ where (d_i, u_i) are eigen-val/vects of Λ. Kriging covariance is $\Sigma = \sum_i f(d_i)^{-1} u_i u_i'$.

Empirical stationary correlations

In Random walk smoothing ZHS

$$Y \sim \mathrm{N}\Big(\mu\sqrt{\pi},\ \Pi^{1/2}(\widetilde{\Delta}^{+} + \mathbf{1}\mathbf{1}'\delta^{-1})\Pi^{1/2} + \lambda^{-1}I\Big)$$

In Tikhonov smoothing BMN

$$Y \sim \mathrm{N}\Big(\mu\mathbf{1},\ I(\Delta^{+} + \mathbf{1}\mathbf{1}'\delta^{-1})I + \lambda^{-1}I\Big)$$

Our proposal XDO

$$Y \sim \mathrm{N}\Big(\mu\nu,\ V^{1/2}(\sigma^2 R)V^{1/2} + \lambda^{-1}I\Big)$$

where $\nu \in \mathbb{R}^n$ and $V = \mathbf{diag}(v_i)$ are given,

R is a correlation matrix we choose, via $R_{ij} = \rho(s_{ij})$

for a smooth function $\rho(\cdot)$ of similarity s_{ij}

(eg $s_{ij} = \pi_i P_{ij} + \pi_j P_{ji}$)   We also choose $\sigma > 0$.

Stationary because ρ depends only on s,

Empirical because we get ρ from data

NB: $\mathbb{E}(Y)$ and Var(Y) not necessarily stationary

Variogram estimator

$$\Phi_{ij} \equiv \frac{1}{2}\mathbb{E}\Big(\big((Y_i - \mu\nu_i) - (Y_j - \mu\nu_j)\big)^2\Big)$$

$$= \frac{1}{\lambda} + \frac{1}{2}\sigma^2(\nu_i^2 + \nu_j^2 - 2\nu_i\nu_j R_{ij}) \qquad \text{(by model)}$$

$$\hat{\Phi}_{ij} \equiv \frac{1}{2}\big((y_i - \mu\nu_i) - (y_j - \mu\nu_j)\big)^2 \qquad 1 \le i < j \le r$$

1) $\hat{\Phi}_{ij}$ is a naive estimator of Φ_{ij}.
2) We plug it in to solve for a naive $\hat{R}_{ij}$.
3) Then fit a spline curve to $(\log(1+s_{ij}), \widehat{R}_{ij})$ pairs: $\widetilde{R}_{ij} \doteq \hat{\rho}(s_{ij})$.
4) Put $\hat{\Sigma} = \sigma^2 V \widetilde{R} V$, and make positive definite: $\hat{\Sigma}_+$

4′) (Variant) Use low rank approx to $\widehat{\Sigma}$ (might scale better for large n)

Then we use kriging with the estimated correlation matrix.

UK web link dataset

- Nodes are 107 UK universities
- Edges are web links
- Weights w_{ij}: # links from i to j
- Y_i: research score measuring quality of Uni i's research

We will try to predict the university research scores from the graph structure and some of the scores.

Data features

- RAE scores in $[0.4, 6.5]$ with mean ~ 3 and standard deviation ~ 1.9.
- 15% of weights w_{ij} are 0, 50% are below 7, max is 2130

Experiment

1) Randomly hold out some universities (ranging from ~ 10% to ~ 90%)
2) Predict held out scores
3) Find mean square error
4) Repeat 50 times

Methods:

Random walk smoothing,

Tikhonov smoothing

and empirical correlation versions of both

Tuning

Empirical correlation has two tuning parameters: λ and σ

The other methods have just one

The comparison is fair because we use hold outs

For RW & Tikhonov methods we eventually just took their best parameter value and it still did not beat cross-validated empirical correlations

Implementation notes

Tikhonov

This method is defined for undirected graphs

So we use $\widetilde{W} = W + W'$

. . . in both original and empirical stationary versions

Choosing μ for which $\beta \sim \mathrm{N}(\mu, \delta^{-1})$

For RW: use $\mu = 0$ for binary responses, but for UNI data take

$$\hat{\mu} = \frac{1}{r} \sum_{i=1}^{r} \frac{y_i}{\nu_i}$$

on 'held in' nodes

For Tikhonov: μ disappears from equations in $\delta \to 0$ limit, so we don't need it

Random walk ZHS for Uni data

Recall the criterion

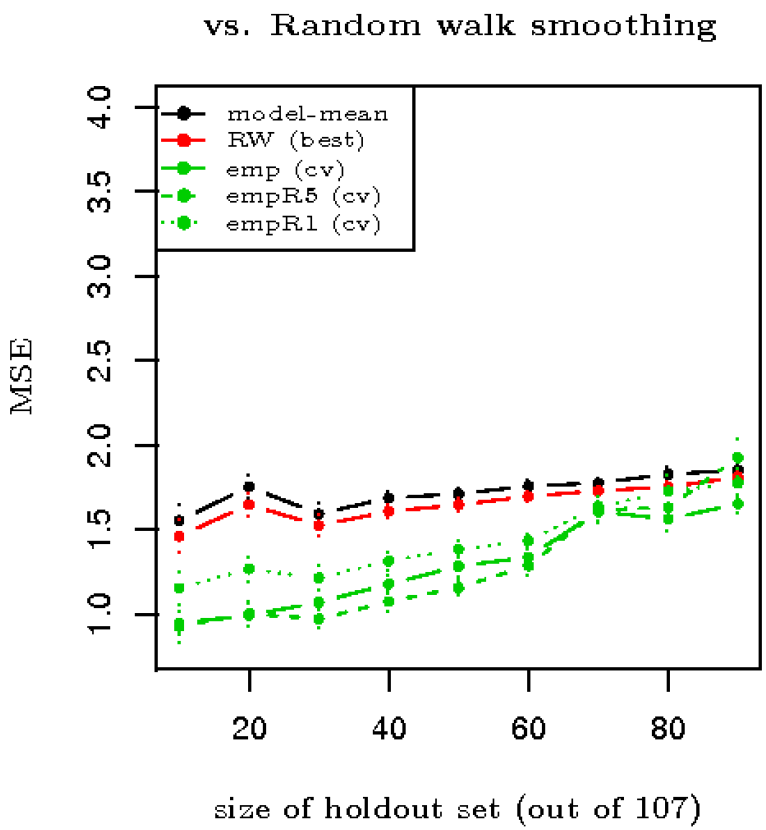
$$\frac{1}{2}\sum_{i,j} s_{ij}\Big(\frac{Z_i}{\sqrt{\pi_i}} - \frac{Z_j}{\sqrt{\pi_j}}\Big)^2 + \lambda\|Z - Y^*\|^2$$

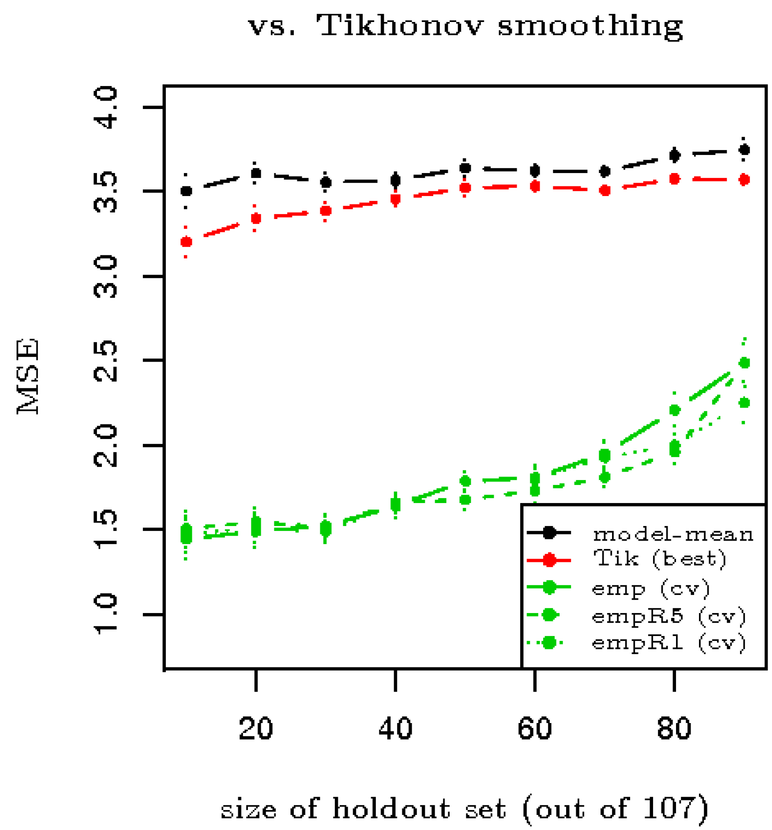
We find (empirically) that the estimate $\widehat{Z}_i$ is nearly $\propto \sqrt{\pi_i}$

Nodes with comparable PageRank π_i get similar predictions

The similarity s_{ij} is virtually ignored

Results for University data





Notes

- RW has $\widehat{Z}$ nearly $\propto \nu = \sqrt{\pi}$
- Tikhonov ignores direction of links
- Empirical correlation performance not sensitive to rank reduction

Numerical summary

Improvement over baseline

	Random walk	Tikhonov
Baseline MSE	1.71	3.64
Random walk	3.8%	-
Tikhonov	-	3.2%
Empirical	25.0%	50.9%
Empirical R5	32.4%	53.9%
Empirical R1	19.1%	50.9%

Mean square prediction errors when 50 of 107 university scores are held out.

Baseline is plain regression on ν with no other graphical input.

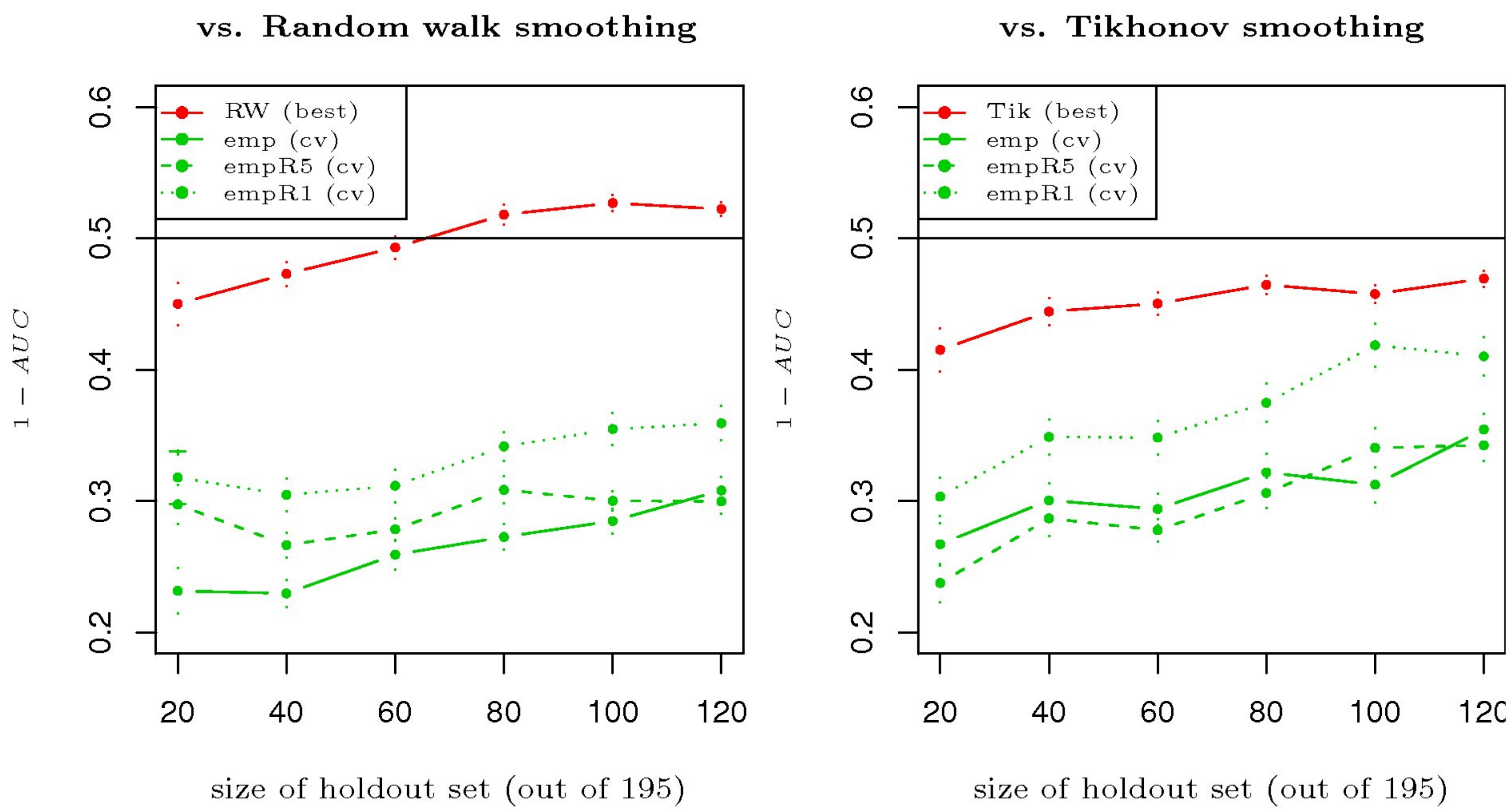
Web KB data

We used the data for Cornell, omitting 'other'.

$$Y = \begin{cases} 1 & \text{student web page} \\ -1 & \text{faculty, staff, dept, course, project} \end{cases}$$

$$W_{ij} = \begin{cases} 1 & i \text{ links to } j \\ 0 & \text{else.} \end{cases}$$

Results for Web KB data



Notes

- Now $\nu = 1$ so $\propto \nu$ is not helpful; solid line is coin toss
- Tikhonov ignores direction of links, but now it helps!
- Empirical correlation performance not sensitive to rank reduction

Numerical results for webKB

Improvement over baseline

	Random walk	Tikhonov
Baseline ($1-$AUC)	0.5	0.5
Random walk	-5.4%	-
Tikhonov	-	8.5%
Empirical	43.0%	37.5%
Empirical R5	40.0%	31.9%
Empirical R1	29.0%	16.3%

Baseline is a coin toss, AUC$= 0.5$

Next steps

1) more examples
2) scaling issues
3) more similarity measures