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) Sometimes it makes a big improvement.
6) We did small examples, but with scaling in mind
7) Now · · · early results for Wikipedia graph

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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>The graph</td>
</tr>
<tr>
<td>$w_{ij}$</td>
<td>Edge weight from $i$ to $j$</td>
</tr>
<tr>
<td>$W$</td>
<td>Adjacency matrix</td>
</tr>
<tr>
<td>$w_{i+} = \sum_j w_{ij}$</td>
<td>Out-degree of $i$</td>
</tr>
<tr>
<td>$w_{+j} = \sum_i w_{ij}$</td>
<td>In-degree of $j$</td>
</tr>
<tr>
<td>$w_{++} = \sum_{ij} w_{ij}$</td>
<td>Graph volume</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>Response value at node $i$</td>
</tr>
<tr>
<td>$Y^{(0)}$</td>
<td>Measured responses $i = 1, \ldots, r$</td>
</tr>
<tr>
<td>$Y^{(1)}$</td>
<td>Unknown responses $i = r + 1, \ldots, n$</td>
</tr>
</tbody>
</table>
Graph random walk

Transition probability \( P_{ij} = \frac{w_{ij}}{w_{i+}} \)

Stationary distribution \( \pi_i \) 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.)

\[ \therefore \text{it has a stationary distribution} \pi \]

Graph Laplacian

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

needed later
What is a graph Laplacian (anyway)?

For a function \( f(i) \in \mathbb{R} \), let

\[
  f = \begin{pmatrix}
    f(1) \\
    f(2) \\
    \vdots \\
    f(n)
  \end{pmatrix} \in \mathbb{R}^n
\]

\[
  (\Delta f)(i) = \sum_j w_{ij} (f(i) - f(j))
\]

In 1 dimension with \( i \) and \( i + 1 \) neighbors:

\[
  (f(i) - f(i - 1)) + (f(i) - f(i + 1)) = -\left(f(i + 1) - 2f(i) + f(i - 1)\right)
\]

So \( \Delta \) is like negative curvature with respect to the graph
Zhou, Huang, Schölkoff (2005)

Node similarity:

\[ s_{ij} \equiv \pi_i P_{ij} + \pi_j P_{ji} \]

Variation functional:

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

Criterion:

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

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

ZHS trade off fit to observations vs graph smoothness via \( \lambda \).

Result is a linear function of \( Y^* \)

There must be an equivalent Gaussian process story
1) Predict at ‘?’ by weighting the obs
2) 1.9 gets more weight than 1.7 because it is closer
3) the 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 are automatic ... given covariance in a Gaussian model

Kriging originated in geostatistics
Kriging model

obs $Y = X\beta + S + \varepsilon \in \mathbb{R}^n$

coefficients $\beta \in \mathbb{R}^k$

predictors $X \in \mathbb{R}^{n \times k}$

correlated part $S \sim N(0, \Sigma)$

noise $\varepsilon \sim N(0, \Gamma)$

we’ll have $k = 1$

e.g. $X = \sqrt{\pi}$ or $1_n$

$\Gamma$ is diagonal

Predictions

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

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

makes $Z \sim N(X\mu, \Psi)$, $\Psi = XX^T\delta^{-1} + \Sigma$

Predict by $\hat{Z} = \mathbb{E}(Z \mid Y^{(0)})$
Kriging some more

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

**Partition $\Psi$**

$$\Psi = \text{Cov} \begin{pmatrix} Z^{(0)} \\ Z^{(1)} \end{pmatrix} = \begin{pmatrix} \Psi_{00} & \Psi_{01} \\ \Psi_{10} & \Psi_{11} \end{pmatrix} = \begin{pmatrix} \Psi_0 & \Psi_1 \end{pmatrix}.$$  

Joint distribution of signal (everywhere) and observations

$$\begin{pmatrix} Z \\ Y^{(0)} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} X \mathbb{E}(\beta) \\ X_0 \mathbb{E}(\beta) \end{pmatrix}, \begin{pmatrix} \Psi & \Psi_0 \\ \Psi_0 & \Psi_{00} + \Gamma_{00} \end{pmatrix} \right)$$

\[\cdots\] yields expression for $\mathbb{E}(Z \mid Y^{(0)})$
ZHS method as kriging

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

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

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 $\Sigma = \Pi^{1/2} \tilde{\Delta}^+ \Pi^{1/2}$ (+ for generalized inverse)
- predictors $X = \text{diag}(\sqrt{\pi_i})^T$
- defaults $\mu_i = \mu X_i$, $r + 1 \leq i \leq n$

Then

$$\lim_{\delta \to 0^+} \text{Kriging}(\Gamma, \Sigma, X, \delta) = \text{ZHS method}$$
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, Matveeava, Niyogi (2004)

Graph Tikhonov regularization

\[ Z^T \Delta Z + \lambda_0 \| Z^{(0)} - Y^{(0)} \|^2 \]

\(\Delta\) is the graph Laplacian, penalty is only on observed responses

As kriging

noise variance \( \Gamma = \begin{pmatrix} \lambda_0^{-1} I_r \\ \lambda_1^{-1} I_{n-r} \end{pmatrix} \)

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

predictors \( X = 1_n \quad (\text{no } \sqrt{\pi_i}) \)

let \( \delta \to 0^+ \) and then let \( \lambda_1 \to 0^+ \)

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

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

As kriging

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

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

predictors $X = \text{diag}(\sqrt{D_{ii}})$

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

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

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 $Z^T L Z$ where $L = \sum_i f(d_i) u_i u_i^T$ where $(d_i, u_i)$ are eigen-val/vects of $\Lambda$. Kriging covariance is $\Sigma = \sum_i f(d_i)^{-1} u_i u_i^T$. 

Kriging on graphs
Empirical stationary correlations

In Random walk smoothing ZHS

\[ Y \sim N\left(\mu \sqrt{\pi}, \Pi^{1/2}(\tilde{\Delta}^+ + 11^T\delta^{-1})\Pi^{1/2} + \lambda^{-1}I\right) \]

In Tikhonov smoothing BMN

\[ Y \sim N\left(\mu 1, I(\Delta^+ + 11^T\delta^{-1})I + \lambda^{-1}I\right) \]

Our proposal XDO

\[ Y \sim N\left(\mu X, V^{1/2}(\sigma^2 R)V^{1/2} + \lambda^{-1}I\right) \]

where \( X \in \mathbb{R}^n \) and \( V = \text{diag}(v_i) \) 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 \( \rho \) depends only on \( s \),
Empirical because we get \( \rho \) from data
NB: \( \mathbb{E}(Y) \) and \( \text{Var}(Y) \) not necessarily stationary
What we’ll do

Estimate $\rho(s_{ij})$ at every pair of nodes $i$ and $j$ using (essentially)

$$\widehat{\mathbb{E}}(Y_i - Y_j)^2 \doteq (Y_i - Y_j)^2, \quad \forall 1 \leq i < j \leq r$$

Extreme overfitting.

Then smooth them.
The variogram estimator

\[ \Phi_{ij} \equiv \frac{1}{2} \mathbb{E} \left( \left( (Y_i - \mu X_i) - (Y_j - \mu X_j) \right)^2 \right) \]

\[ = \frac{1}{\lambda} + \frac{1}{2} \sigma^2 \left( X_i^2 - 2X_iX_j R_{ij} + X_j^2 \right) \quad \text{(by model)} \]

\[ \hat{\Phi}_{ij} \equiv \frac{1}{2} \left( (Y_i - \mu X_i) - (Y_j - \mu X_j) \right)^2 \quad 1 \leq i < j \leq r \]

Set \( \Phi_{ij} = \hat{\Phi}_{ij} \) and solve for \( R_{ij} \)

a naive correlation for signals \( Z_i \) and \( Z_j \)
Using the variogram estimator

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

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

Then we use kriging with the estimated correlation matrix.
Smoothing to get $\rho$
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 $\sim 3$ and standard deviation $\sim 1.9$.
- 15% of weights $w_{ij}$ are 0, 50% are below 7, max is 2130
Experiment

1) Randomly hold out some universities (ranging from $\sim 10\%$ to $\sim 90\%$)
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: $\lambda$ and $\sigma$
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 \( \tilde{W} = W + W^T \)

... in both original and empirical stationary versions

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

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

\[
\hat{\mu} = \frac{1}{r} \sum_{i=1}^{r} \frac{Y_i}{X_i}
\]

on 'held in' nodes

For Tikhonov: \( \mu \) disappears from equations in \( \delta \to 0 \) limit, so we don’t need it
Random walk ZHS for Uni data

Recall the criterion

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

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

Nodes with comparable PageRank $\pi_i$ get similar predictions

The similarity $s_{ij}$ is virtually ignored
Results for University data

Notes

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

Kriging on graphs
## Numerical summary

### Improvement over baseline

<table>
<thead>
<tr>
<th></th>
<th>Random walk</th>
<th>Tikhonov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline MSE</td>
<td>1.71</td>
<td>3.64</td>
</tr>
<tr>
<td>Random walk</td>
<td>3.8%</td>
<td>-</td>
</tr>
<tr>
<td>Tikhonov</td>
<td>-</td>
<td>3.2%</td>
</tr>
<tr>
<td>Empirical</td>
<td>25.0%</td>
<td>50.9%</td>
</tr>
<tr>
<td>Empirical R5</td>
<td>32.4%</td>
<td>53.9%</td>
</tr>
<tr>
<td>Empirical R1</td>
<td>19.1%</td>
<td>50.9%</td>
</tr>
</tbody>
</table>

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

Baseline is plain regression on $X$ 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 $X = 1$ so $\propto X$ 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

<table>
<thead>
<tr>
<th>Improvement over baseline</th>
<th>Random walk</th>
<th>Tikhonov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (1 − AUC)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Random walk</td>
<td>−5.4%</td>
<td>-</td>
</tr>
<tr>
<td>Tikhonov</td>
<td>-</td>
<td>8.5%</td>
</tr>
<tr>
<td>Empirical</td>
<td>43.0%</td>
<td>37.5%</td>
</tr>
<tr>
<td>Empirical R5</td>
<td>40.0%</td>
<td>31.9%</td>
</tr>
<tr>
<td>Empirical R1</td>
<td>29.0%</td>
<td>16.3%</td>
</tr>
</tbody>
</table>

Baseline is a coin toss, AUC = 0.5
Next steps

1) more examples
   - Biological examples . . . noisy edges
   - Wikipedia preliminary results

2) scaling issues
   - sparsity helps; sparse plus low rank promising
   - Ya Xu has Markov chain algorithms for some similarities

3) more similarity measures
Peek ahead

For Wikipedia graph: can make up responses like 1 iff web page links to a given page such as the one for US, China, France or Bob Dylan.

Have 2.4 million nodes and observe 10%

Empirical stationary correlation got 22.5% (1-AUC) error in rank 1 version vs near 53.8% for a local average.

PROPACK sparse SVD code works for empirical algorithm because it needs eigenvectors for largest eigenvalues but fails for random walk smoothing and Tikhonov smoothing which need smallest eigen.
Randomized labels

Protein  

Webspam  

Wikipedia

Observed roughness $\Omega(Y) = \bullet$ vs histogram of $\Omega(\text{permuted}(Y))$

RW on top, Tikhonov below

Kriging on graphs
Second peek ahead

Work in progress with Justin Dyer

Copula of links between Wikipedia pages
Thanks

- Ying Lu asked me to speak
- San Francisco Bay Area Chapter of the ASA for organizing:
  - Spencer Graves, Chris Barker, Dean Fearn
  - Steve Scott and Google for hosting
- NSF division of mathematics and statistics