Learning from Labeled and Unlabeled Data on a Directed Graph

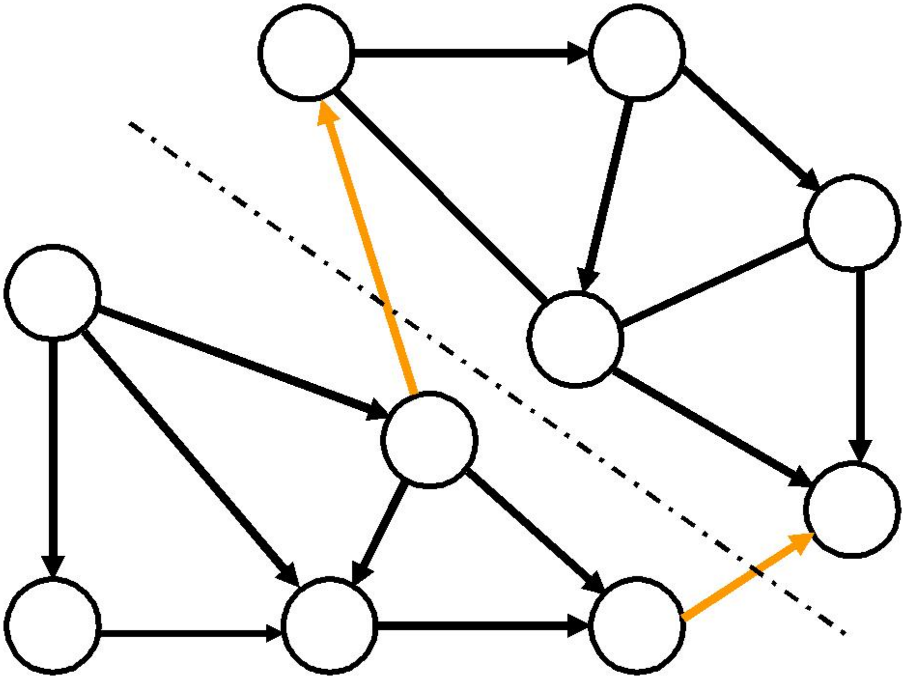
Dengyong Zhou [†] **Jiayuan Huang** [‡†]
Bernhard Schölkopf [†]
[†]Department of Empirical Inference
Max Planck Institute for Biological Cybernetics, Germany
[‡] School of Computer Science
University of Waterloo, Canada

Why should we study learning from directed graphs?

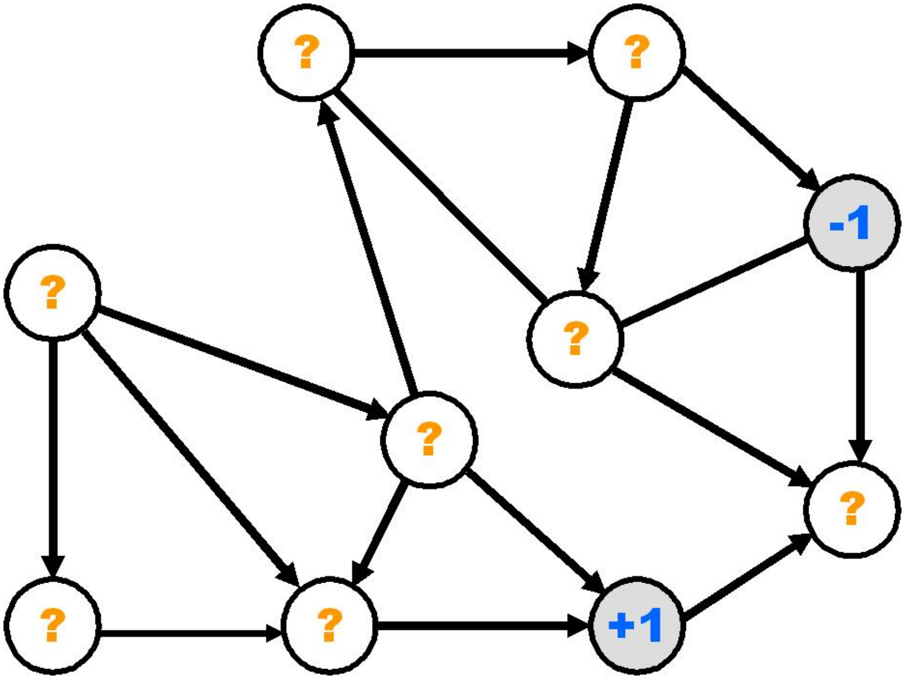
- In typical machine learning approaches, e.g., kernel methods, the pairwise relationships among data are assumed to be symmetric.
- However, in many real-world applications, the pairwise relationships are asymmetric. A typical example is the World Wide Web.
- Transferring asymmetric relationships into symmetric ones leads to loss of information (the directionality).

We analyze the asymmetric relationships directly without the need of transferring.

Learning from directed graphs: clustering



Learning from directed graphs: classification



Some notes

- Shi and Malik (1997) proposed the spectral clustering approach for undirected graphs, which has a nice random walk interpretation (Meilă and Shi, 2001).
- Kleinberg (1997) suggested to use the eigenvectors of $W^T W$ (W denotes the adjacency matrix) for directed graph clustering in his famous paper on the HITS algorithm.
- How to generalize the Shi and Malik's algorithm to the context of directed graphs has been listed as one of six algorithmic challenges in web search engines (Henzinger 2003).

Directed spectral clustering: cut criterion (I)

Our solution

- Defining a random walk over the directed graph $G = (V, E)$ with a transition probability matrix P such that it has a unique stationary distribution π, such as the teleporting random walk used by Google (note: any other random walk can be considered as well, for instance, the two-step random walk).

Directed spectral clustering: cut criterion (II)

- Looking for a cut $V = S \cup S^c$ $(S \cap S^c = \emptyset)$ such that, under the stationary distribution, the probability of transition from one cluster to another $P(S \to S^c) = \sum_{u \in S, v \in S^c} \pi(u) p(u, v)$ is as small as possible, while the probabilities of remaining in the same clusters $P(S) = \sum_{v \in S} \pi(v)$, $P(S^c) = \sum_{v \in S^c} \pi(v)$ are as large as possible. Formally,

$$\min_{S \neq \emptyset \in V} P(S \to S^c) \left(\frac{1}{P(S)} + \frac{1}{P(S^c)} \right).$$

Directed spectral clustering: real-valued relaxation

- The combinatorial optimization can be relaxed into

$$\operatorname*{argmin}_{f \in \mathbb{R}^{|V|}} \Omega(f) = \frac{1}{2} \sum_{[u,v] \in E} \pi(u) p(u,v) \left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}} \right)^2$$

subject to $\quad \| f \| = 1, \ \langle f, \sqrt{\pi} \rangle = 0.$

- Define $\Theta = (\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2})/2$ and $\Delta = I - \Theta$. We can show that $\Omega(f) = \langle f, \Delta f \rangle$.

Summarizing our directed spectral clustering algorithm

It can be implemented with only several lines of Matlab code.

1. Define a random walk over graph $G = (V, E)$ with a transition probability matrix P such that it has a unique stationary distribution.
2. Let Π denote the diagonal matrix with its diagonal elements being the stationary distribution of the random walk. Form the matrix $\Theta = (\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2})/2$.
3. Compute the eigenvector Φ of Θ corresponding to the second largest eigenvalue, and then partition the vertex set V of G into $S = \{v \in V | \Phi(v) \geq 0\}$ and $S^c = \{v \in V | \Phi(v) < 0\}$.

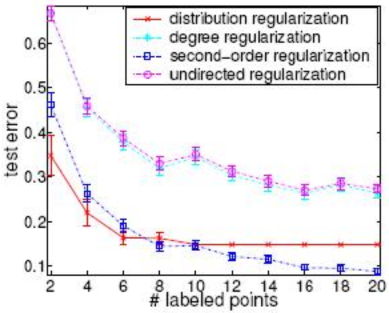
Transductive inference (semi-supervised learning)

It is straightforward from spectral clustering to transductive inference.

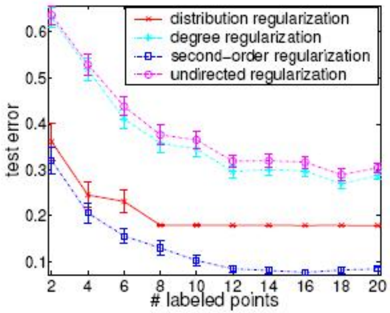
- Given a directed graph $G = (V, E)$, some vertices are labeled. Define a function y on V with $y(v) = 1$ or -1 if vertex v is labeled as 1 or -1, and 0 if v is unlabeled. Then the remaining unlabeled vertices may be classified by using the function

$$f^* = \underset{f \in \mathbb{R}^{|V|}}{\operatorname{argmin}} \left\{ \Omega(f) + \mu \|f - y\|^2 \right\}$$

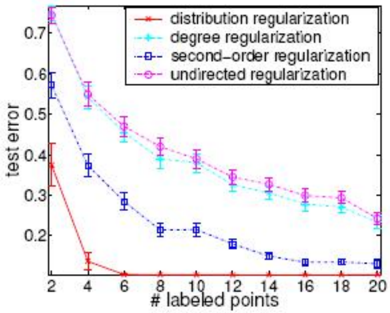
$$\Longrightarrow f^* = \mu(\mu I + \Delta)^{-1} y.$$



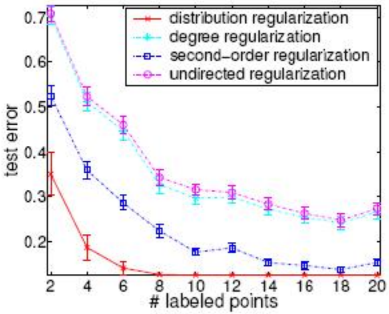
(a) Cornell (student)



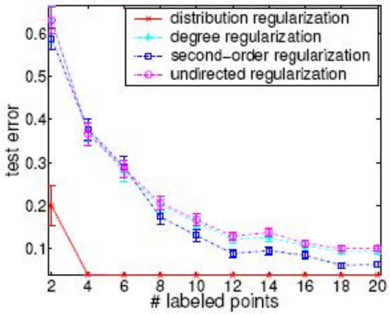
(b) Texas (student)



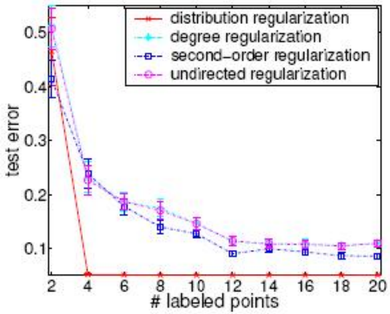
(c) Washington (student)



(d) Wisconsin (student)



(e) Cornell (faculty)



(f) Cornell (course)

Discrete analysis and regularization (I)

We develop discrete analysis for directed graphs to construct a discrete analogue of classical regularization theory.

- Given a directed graph $G = (V, E)$, the functions defined on V can be endowed with the standard inner product in $\mathbb{R}^{|V|}$ as

$$\langle f, g \rangle_{\mathcal{H}(V)} = \sum_{v \in V} f(v) g(v)$$

to form a space denoted by $\mathcal{H}(V)$. Similarly define $\mathcal{H}(E)$.

Discrete analysis and regularization (II)

- We define the graph gradient to be an operator $\nabla : \mathcal{H}(V) \to \mathcal{H}(E)$ which satisfies

$$(\nabla f)([u,v]) := \sqrt{\pi(u)p(u,v)} \left(\frac{f(v)}{\sqrt{\pi(v)}} - \frac{f(u)}{\sqrt{\pi(u)}} \right).$$

- We define the graph divergence to be an operator $\operatorname{div} : \mathcal{H}(E) \to \mathcal{H}(V)$ which satisfies

$$\langle \nabla f, g \rangle_{\mathcal{H}(E)} = \langle f, -\operatorname{div} g \rangle_{\mathcal{H}(V)}.$$

Discrete analysis and regularization (III)

- We define the (directed) graph Laplacian to be an operator $\Delta : \mathcal{H}(V) \to \mathcal{H}(V)$ which satisfies

$$\Delta f := -\frac{1}{2}\operatorname{div}(\nabla f).$$

It can be shown that $\Delta = I - \Theta$ (with Θ as defined earlier).

- We define a general operator $\Delta_p : \mathcal{H}(V) \to \mathcal{H}(V)$ which satisfies

$$\Delta_p f := -\frac{1}{2}\operatorname{div}(\|\nabla f\|^{p-2}\nabla f).$$

Clearly, $\Delta_2 = \Delta$, and $\Delta_p (p \neq 2)$ is nonlinear.

Discrete analysis and regularization (IV)

We can show that the solution f^* of the general optimization problem

$$\underset{f \in \mathcal{H}(V)}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{v \in V} \|\nabla_v f\|^p + \mu \|f - y\|^2 \right\}$$

satisfies

$$p \Delta_p f^* + 2\mu(f^* - y) = 0.$$

(Note that the previous optimization problem is the case of $p = 2$.)

Conclusion

A solid mathematical framework for the web IR

- Generalized the spectral clustering approach to the context of directed graphs;
- Proposed a transductive inference algorithm for directed graphs built on the directed spectral clustering approach;
- Developed discrete analysis for directed graphs and consequently a discrete analogue of classical regularization theory.