Predicting Graph Labels using Perceptron

Shuang Song
shs037@eng.ucsd.edu

Online learning over graphs

M. Herbster, M. Pontil, and L. Wainer, Proc. 22nd Int. Conf. Machine Learning (ICML'05), 2005

Prediction on a graph with a perceptron

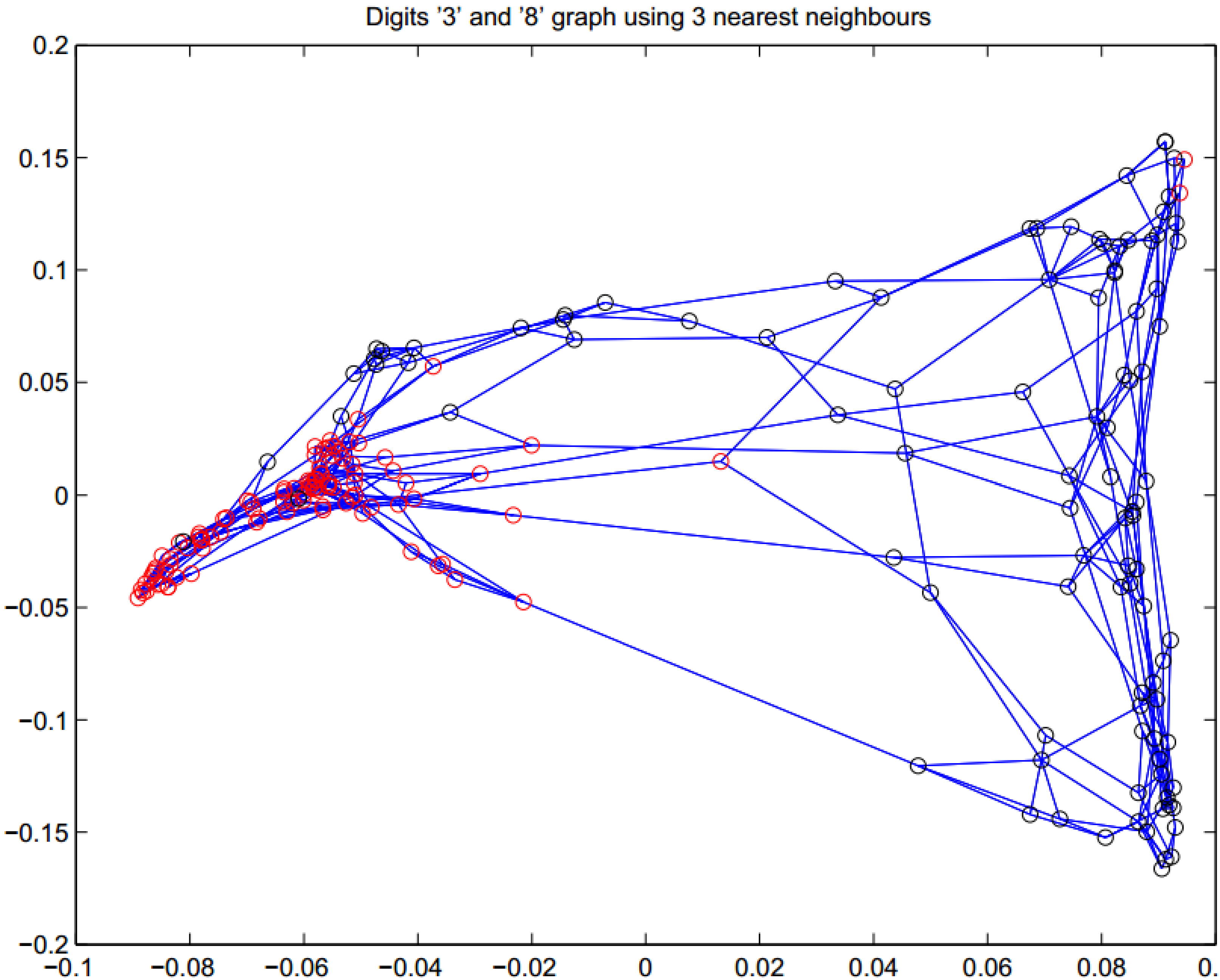
M. Herbster, and M. Pontil, NIPS 20, 2006

Outline

1. Problem Setting
2. Perceptron
3. Properties of Graphs
4. Bound # of mistakes

Problem Setting

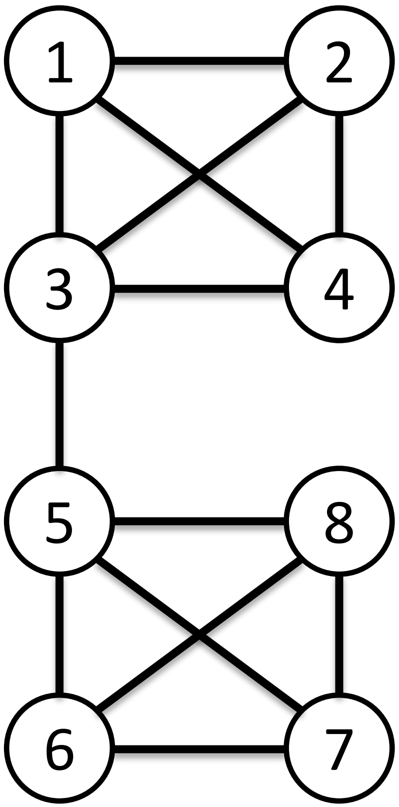
- Known graph, unknown labels on vertices
  - eg. Advertisement service on web page
  - eg. Digit recognition task on USPS (graph is built using NN)



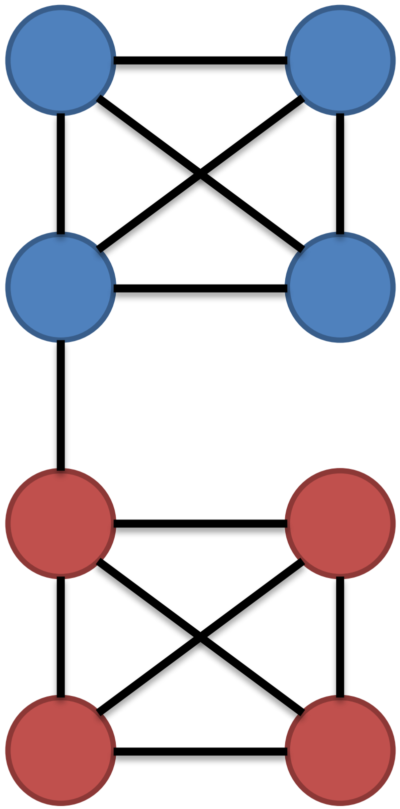
Digits '3' and '8' graph using 3 nearest neighbours
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
−0.1
−0.08
−0.06
−0.04
−0.02
0
0.02
0.04
0.06
0.08
0.1

Problem Setting

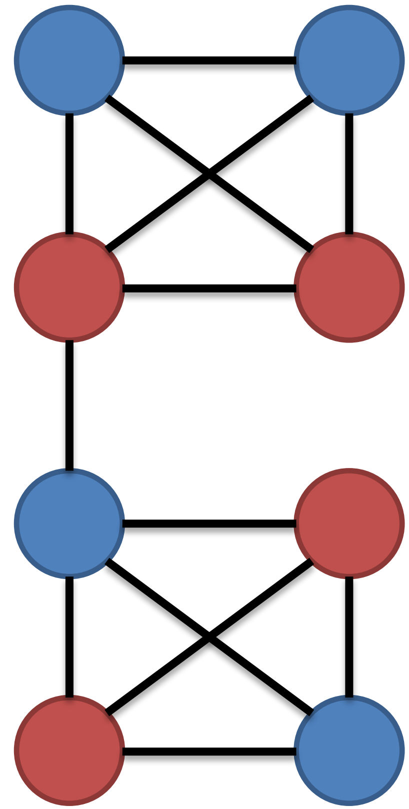
- Given a graph $G = (V, E)$ where $V = \{1, \ldots, n\}$
- for $t = 1, \ldots, l$
  - nature selects $v_t \in V$
  - learner predicts $\hat{y}_t \in \{1, -1\}$
  - nature reveals $y_t \in \{1, -1\}$
  - if $\hat{y}_t \neq y_t, mistakes = mistakes + 1$
- minimize $mistakes$



Node	Predict	Nature	Mistakes
1	1	-1	1
2	-1	-1	1
3	-1	-1	1
4	-1	-1	1
5	-1	1	2
6	1	1	2
7	1	1	2
8	1	1	2



"easy"



"hard"

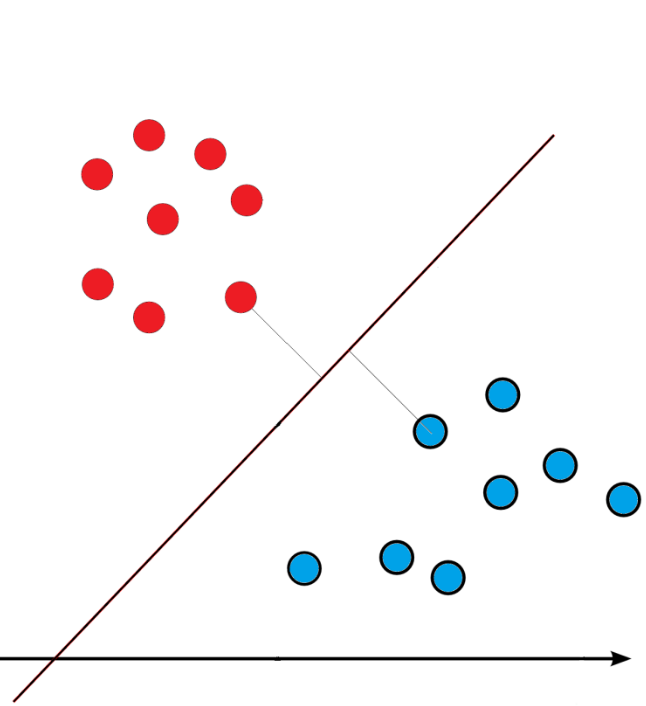
Problem Setting

- Implicit assumption: adjacent nodes have similar labels
- The nature can be adversarial, and the learner can always make mistake; yet if the nature is regular and simple, then it is possible for the learner to make only a few mistake.
- Bound $mistakes$ using complexity of nature's labelling
- Assume graph is connected, unweighted

What algorithm are we going to use?

Perceptron

- simply linear classification
- assume linearly separable with margin 1



Perceptron: algorithm

- data: $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_l, y_l)\} \subset (\mathcal{H} \times \{1, -1\})^l$
- Initial $\boldsymbol{w}_1 = \boldsymbol{0} \in \mathcal{H}$
- For $t = 1, \dots, l$
  - receive $\boldsymbol{x}_t \in \mathcal{H}$
  - predict $\hat{y}_t = \operatorname{sign}(\langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle)$
  - receive $y_t \in \{1, -1\}$
  - if $\hat{y}_t \neq y_t$
    - $mistake = mistake + 1$
    - $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_t \boldsymbol{x}_t$
  - else
    - $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$

Perceptron: mistake bound

- Theorem: given a sequence $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^{l} \in \mathcal{H} \times \{-1,1\}$, and M as the set of trails in which the perceptron predicted incorrectly, then

$$|M| \leq \|\boldsymbol{w}\|^2 \max_{t \in M} \|\boldsymbol{x}_t\|^2$$

for all $\boldsymbol{w} \in \{-1,1\}^n$ st. $w_t = y_t, t = 1, \dots, l$

norm is taken w.r.t. the inner product of $\mathcal{H}$

Perceptron: how to use

- For us, what is the inner product? what is $\boldsymbol{x}_t$?
- We would want a $\boldsymbol{x}_t$ that captures the structure of the whole graph.

Perceptron: algorithm

- data: $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_l, y_l)\} \subset (\mathcal{H} \times \{1, -1\})^l$
- Initial $\boldsymbol{w}_1 = \mathbf{0} \in \mathcal{H}$
- For $t = 1, \dots, l$
  - receive $\boldsymbol{x}_t \in \mathcal{H}$
  - predict $\hat{y}_t = \mathrm{sign}(\langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle)$
  - receive $y_t \in \{1, -1\}$
  - if $\hat{y}_t \neq y_t$
    - $mistake = mistake + 1$
    - $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_t \boldsymbol{x}_t$
  - else
    - $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$

For us, what is $\boldsymbol{x}_t$? What is the inner product?

We would want a $\boldsymbol{x}_t$ that captures the structure of the whole graph.

Properties of Graphs

- Graph Laplacian $L = D - A$, where A is adjacency matrix and $D = \mathrm{diag}(d_1, \ldots, d_n)$

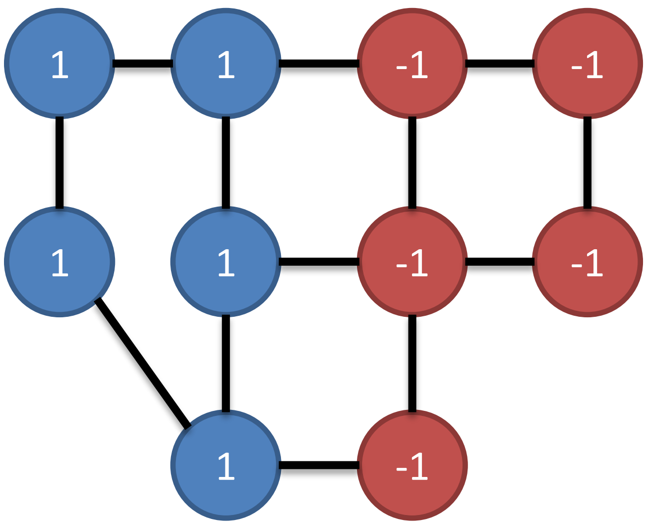
- Inner product: $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \boldsymbol{f}^T L \boldsymbol{g}, \forall \boldsymbol{f}, \boldsymbol{g} \in \mathbb{R}^n$

- Semi-norm:

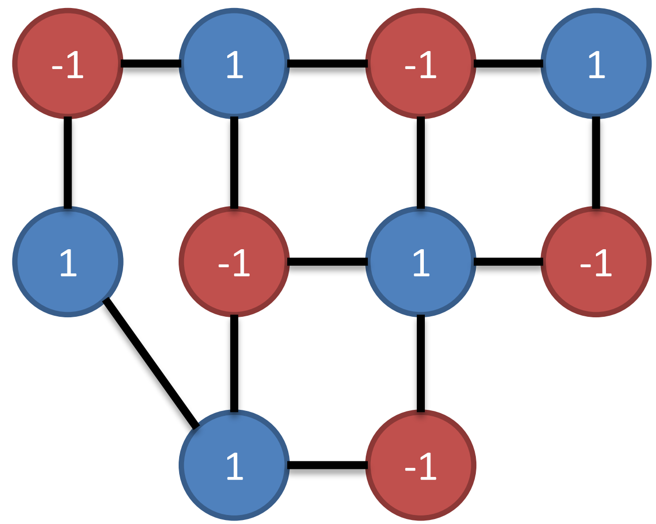
$$\|\boldsymbol{f}\|^2 = \langle \boldsymbol{f}, \boldsymbol{f} \rangle = \sum_{(i,j)\in E} \left(f_i - f_j\right)^2$$

Properties of Graphs

Norm measures "smoothness" or "complexity" of a labelling $\boldsymbol{g}$:

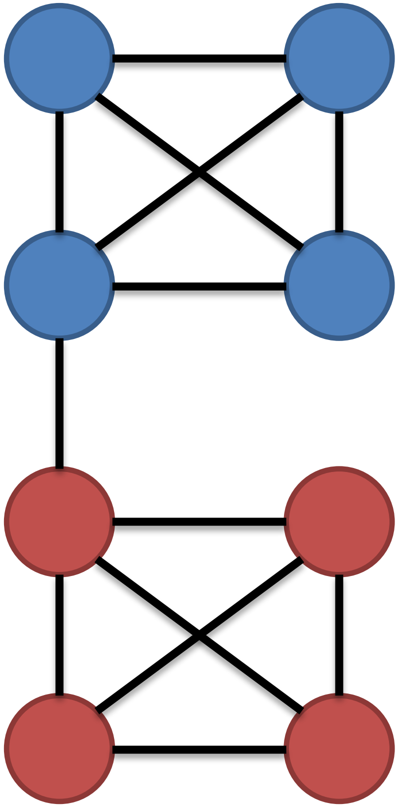


$$\|\boldsymbol{g}\|^2 = 3 \times 4$$

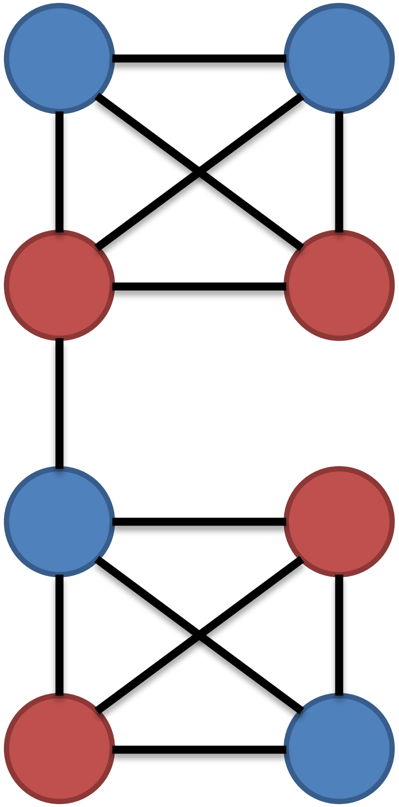


$$\|\boldsymbol{g}\|^2 = 12 \times 4$$

Properties of Graphs



$$\|g\|^2 = 1 \times 4$$



$$\|g\|^2 = 9 \times 4$$

Properties of Graphs

- Eigenvalue λ_i and eigenvector $\boldsymbol{u}_i$ of L:
  - Connected → $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_n$ with $\boldsymbol{u}_1$ as constant vector

- $\mathcal{H} = \mathrm{span}\{\boldsymbol{u}_{\mathbf{2}}, \dots, \boldsymbol{u}_n\} = \{\boldsymbol{g}: \boldsymbol{g}^T \boldsymbol{u}_1 = 0\}$
- $= \{\boldsymbol{g}: \sum_{i=1}^{n} g_i = 0\}$
  - Semi-norm becomes norm

Properties of Graphs

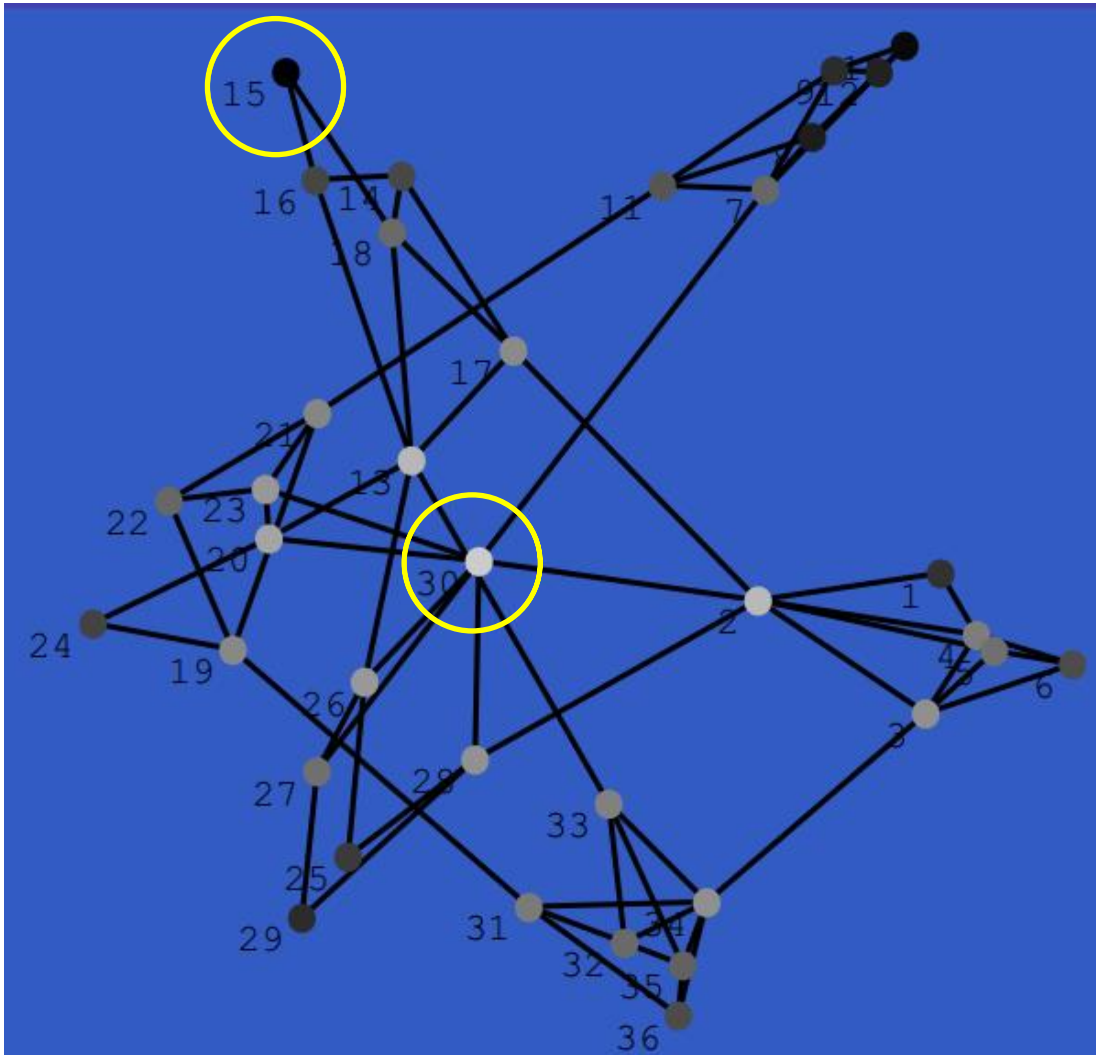
- Pseudoinverse

$$K = L^{+} = \sum_{i=2}^{n} \lambda_i^{-1} \boldsymbol{u}_i \boldsymbol{u}_i^T$$

- It is the reproducing kernel of $\mathcal{H}$: $\forall\ \boldsymbol{g} \in \mathcal{H},\ K_i = K(:, i)$

$$\langle K_i, \boldsymbol{g} \rangle = g_i$$

- $\|K_t\|^2 = K_{tt}$ measures the "remoteness" of vertex t and it decreases with connectivity



Grey-scaled $\mathbf{K}_{tt}$: $\mathbf{K}_{30,30} = .21$ (min), $\mathbf{K}_{15,15} = .94$ (max)

- This will be our “feature” of a vertex, i.e.,

$$\boldsymbol{x}_t = K_t$$

Properties of Graphs

- For $p \in V$, define
  - Distance (of two vertices): $d(p,q) = \min|P(p,q)|$ where P is a path from p to q
  - Eccentricity (of a vertex): $\rho_p = \max_{q \in V} d(p,q)$
  - Diameter (of a graph): $D_G = \max_{p} \rho_p$

Bound

- $|M| \leq \|\boldsymbol{w}\|^2 \max_{t \in M} \|\boldsymbol{x}_t\|^2$
- We want to know what $\|\boldsymbol{w}\|^2$ and $\max_{t \in M} \|\boldsymbol{x}_t\|^2$ is with the properties of graph

Bound # of mistakes： $\|\boldsymbol{w}\|^2$

- Firstly we look at $\boldsymbol{w}$:
- $\boldsymbol{w} \in \{-1, +1\}^n$
- $\|\boldsymbol{w}\|^2 = \sum_{(i,j)\in E}\left(w_i - w_j\right)^2$
- “smoothness” or “complexity”
- $\|\boldsymbol{w}\|^2$
  $= 4 \times (\text{\# edges spanning different labels})$

Bound # of mistakes: $\|\boldsymbol{x}_t\|^2$

- Then we look at $\boldsymbol{x}_t = K_t$:
- $\|K_t\|^2 = K_{tt}$. So we want to bound K_{tt}
- Theorem: For a connected graph G with Laplacian kernel K,

$$K_{tt} \le \min\left(\frac{1}{\lambda_2}, \rho_t\right), t \in V$$

2nd smallest eigenvalue

eccentricity:
$\rho_t = \max \min_{q \in V} |P(p, q)|$

Bound # of mistakes : $\|\boldsymbol{x}_t\|^2$

- Proof:
- $K_{tt} \leq \frac{1}{\lambda_2}$
  - $g^T L g \geq \lambda_2 g^T g\ , \forall g \in \mathcal{H}$
  - Taking $g = K_t$, $K_{tt} \geq \lambda_2 \sum {g_p}^2 \geq \lambda_2 {K_{tt}}^2$
- $K_{tt} \leq \rho_t$
  - If $g_t > 0$, then $\exists s$, s.t. $g_s < 0$
  - $\exists$path P from t to s, s.t. $|E(P)| \leq \rho_t$

Bound # of mistakes : $\|\boldsymbol{x}_t\|^2$

- $\sum_{(i,j)\in E(P)} |g_i - g_j| \geq g_t - g_s > g_t$
- By $n \sum_{i=1}^{n} a_i^2 \geq \left(\sum_{i=1}^{n} a_i\right)^2$ for non−negative $\{a_i\}$, we have

$$\sum_{(i,j)\in E(P)} \left(g_i - g_j\right)^2 \geq \frac{\left(\sum_{(i,j)\in E(P)} |g_i - g_j|\right)^2}{|E(P)|}$$

$$\geq \frac{\left(\sum_{(i,j)\in E(P)} |g_i - g_j|\right)^2}{\rho_t} \geq \frac{g_t^2}{\rho_t}$$

Bound # of mistakes : $\|\boldsymbol{x}_t\|^2$

– Taking $g = K_t$, we have $\|K_t\|^2 = \sum_{(i,j)\in E(G)} \left(K_{ti} - K_{tj}\right)^2$ and $\|K_t\|^2 = \langle K_t, K_t \rangle = K_{tt}$

– $K_{tt} = \sum_{(i,j)\in E(G)} \left(K_{ti} - K_{tj}\right)^2 \geq \frac{K_{tt}^2}{\rho_t}$

- $K_{tt} \leq \frac{1}{\lambda_2}$ and $K_{tt} \leq \rho_t$

Bound # of mistakes

- $|M| \leq \|\boldsymbol{w}\|^2 \max_{t \in M} \|\boldsymbol{x}_t\|^2$

$$4(\text{\# edges spanning different labels})$$

$$\times \min\left(\frac{1}{\lambda_2}, \rho_t\right)$$

Further improvement

- Noisy samples:

$$|M| \leq 2|M \cap M_w| + \frac{\|w\|^2 X^2}{2} + \sqrt{2|M \cap M_w| \|w\|^2 X^2 + \frac{\|w\|^4 X^4}{4}}$$

- Bound K_{pp} using resistance

$$K_{pp} \leq \max_{(p,q) \in V} r(p, q)$$