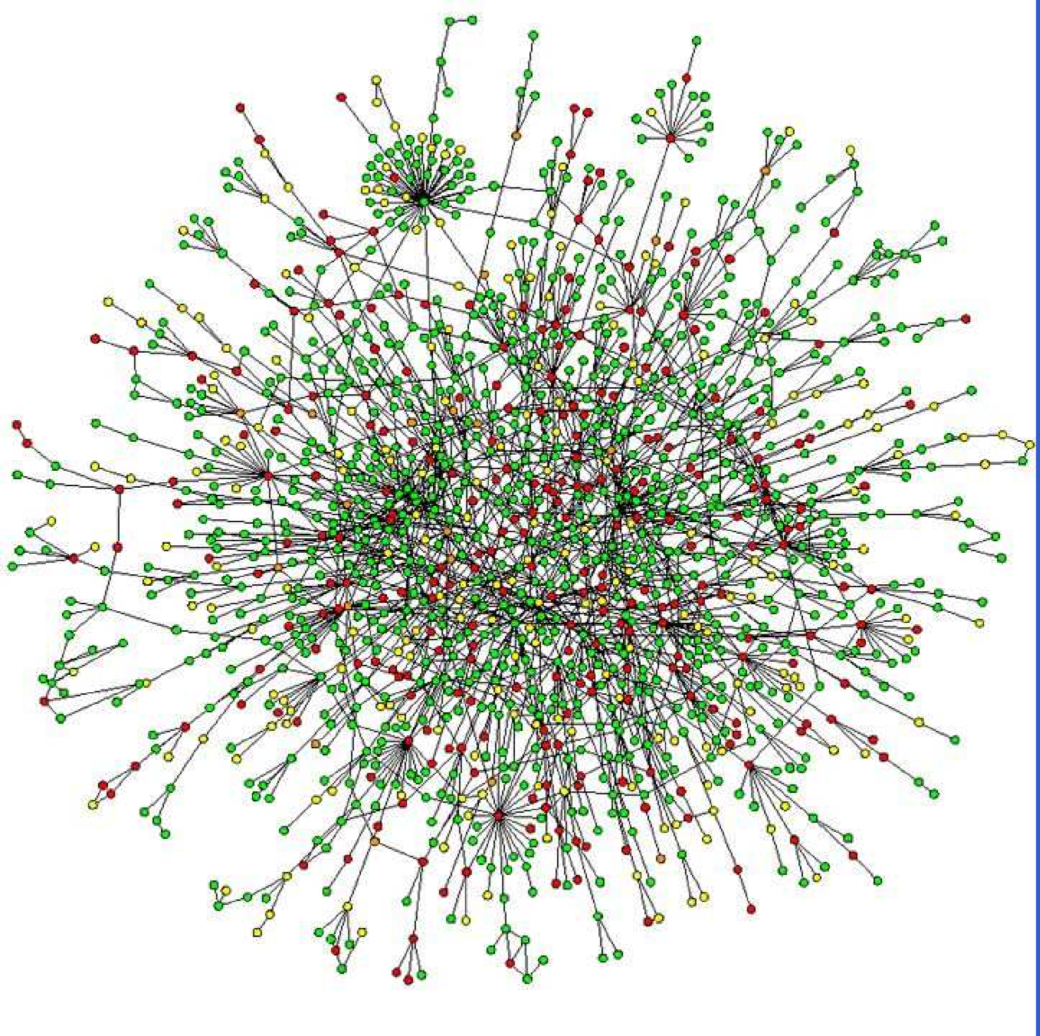
Online Learning over Graphs

Mark Herbster, Massimiliano Pontil, Lisa Wainer

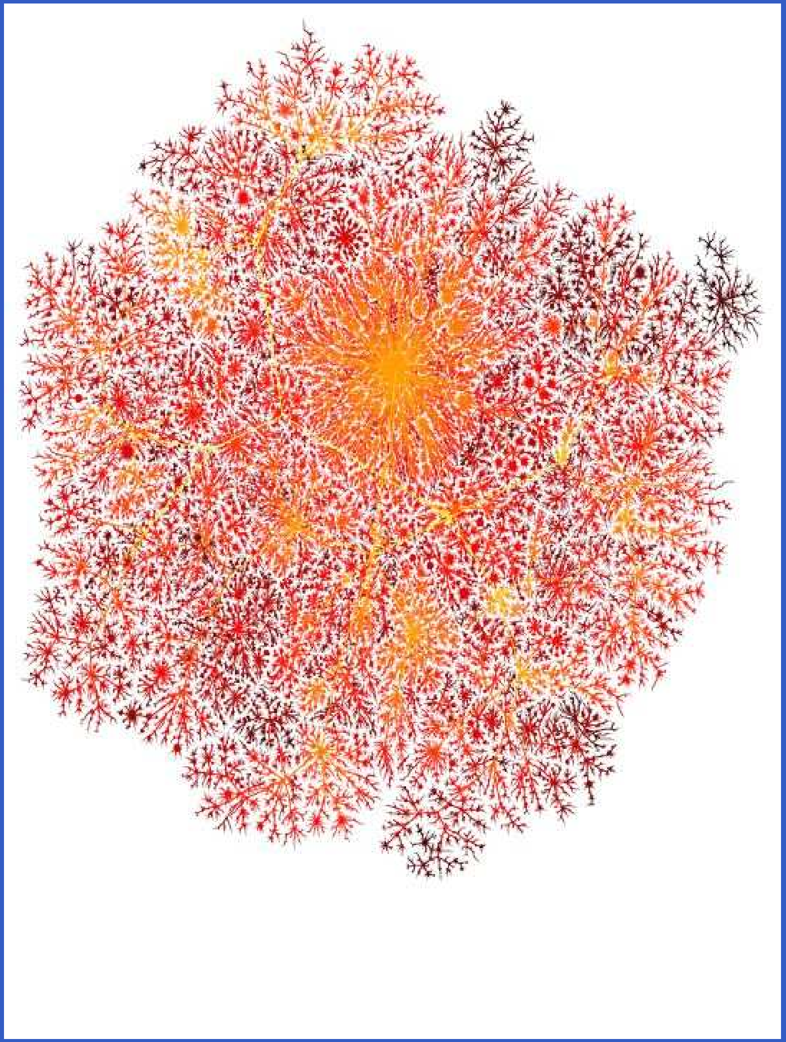
Computer Science Department

University College London

Inspiration – 1

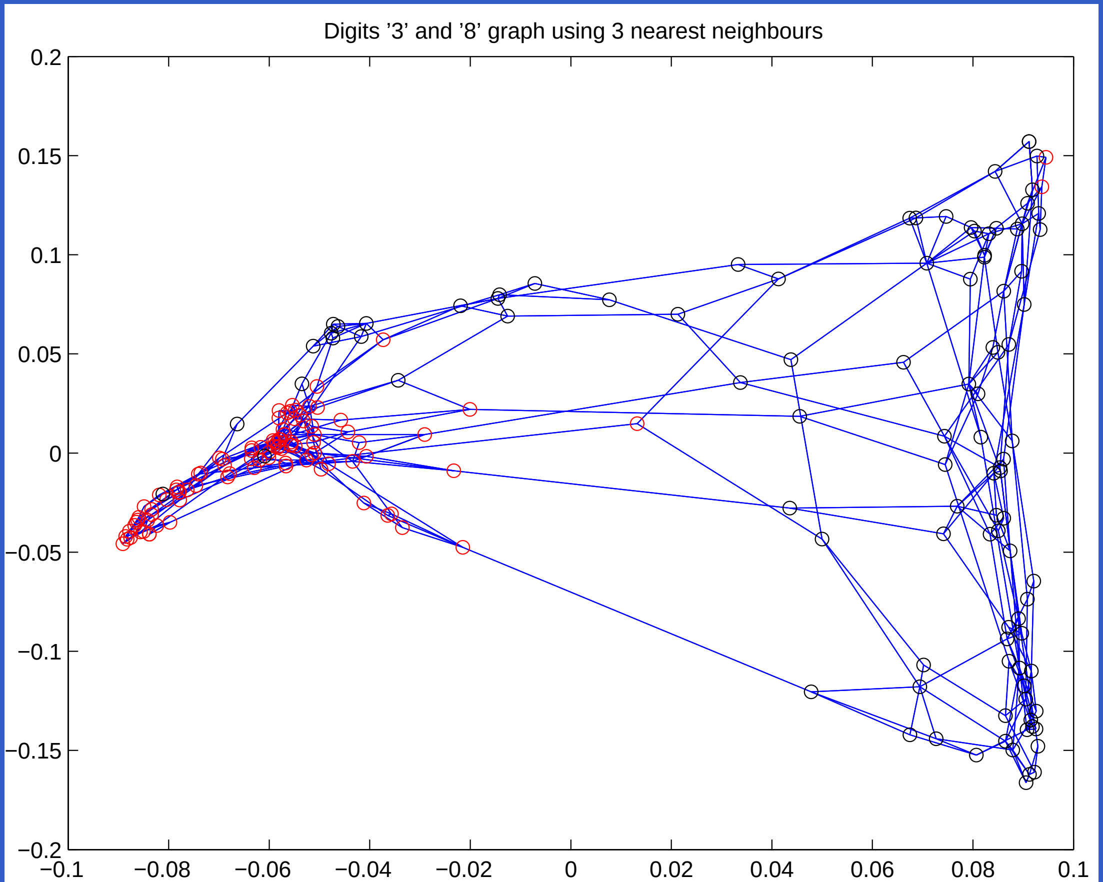


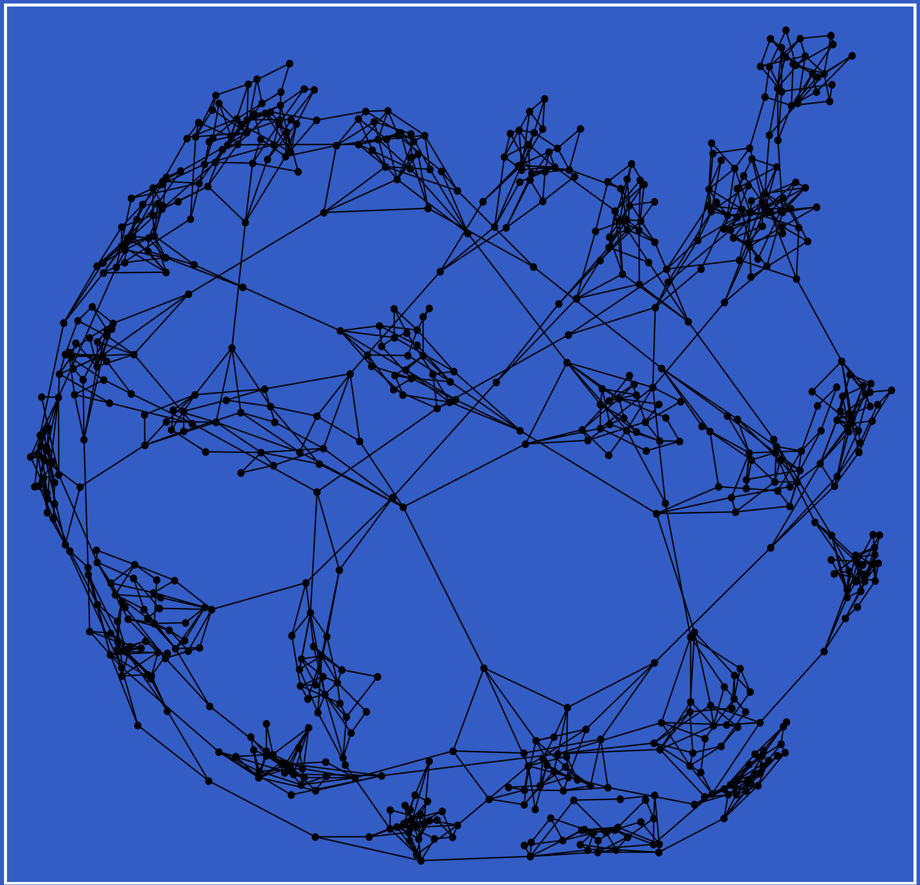
Yeast protein network



Internet hosts

Inspiration – 2





USPS digits 3 and 8

Random graph $G^2_{k-\text{out}}(26;2)$

Online Learning Model

- Aim: learn a function $\mathsf{g} : V \to \{-1, +1\}$ corresponding to a labeling of a graph $G = (V, E)$ and $V = \{1, \ldots, n\}$.
- Learning proceeds in trials

for $t = 1, \ldots, \ell$ **do**

1. Nature selects $v_t \in V$
2. Learner predicts $\hat{y}_t \in \{-1, +1\}$
3. Nature selects $y_t \in \{-1, +1\}$
4. If $\hat{y}_t \neq y_t$ then mistakes $=$ mistakes $+ 1$

- Learner's goal: *minimize* mistakes
- Bound: mistakes $\leq$ *f(complexity*$(\mathsf{g}))$
- What is a natural *complexity* for a graph labeling?

RKHS on a Graph

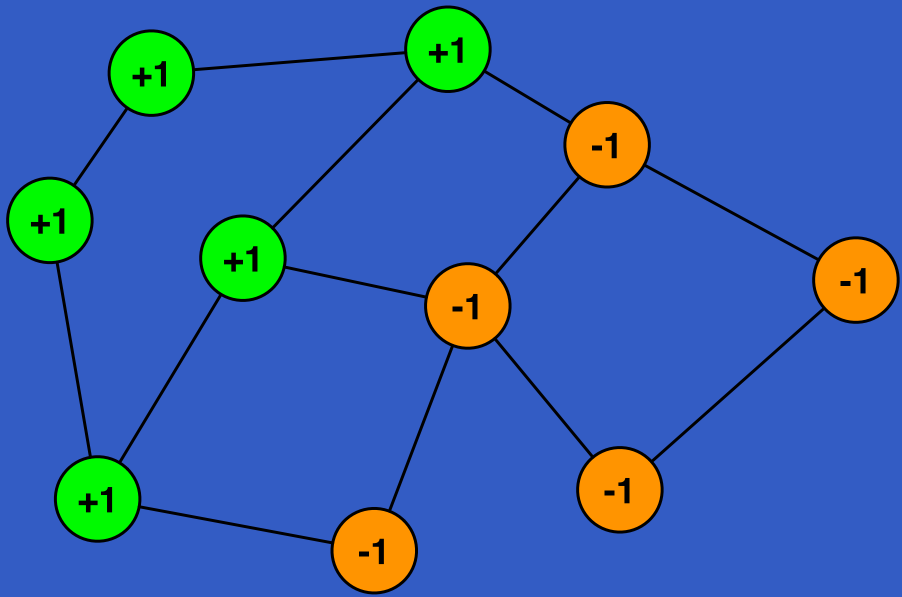
- Graph Laplacian $\mathbf{L} := \mathbf{D} - \mathbf{A}$ where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D} := \mathrm{diag}(d_1, \ldots, d_n)$
- Define $\mathbb{S}_n := \{\mathsf{g} : \sum_{i=1}^n g_i = 0\}$; assume G is connected
- Define

$$\langle \mathbf{f}, \mathsf{g} \rangle := \mathbf{f}^\top \mathbf{L} \mathsf{g} \quad \mathbf{f}, \mathsf{g} \in \mathbb{S}_n$$
$$\|\mathsf{g}\|^2 = \sum_{(i,j)\in E(G)} (g_i - g_j)^2$$

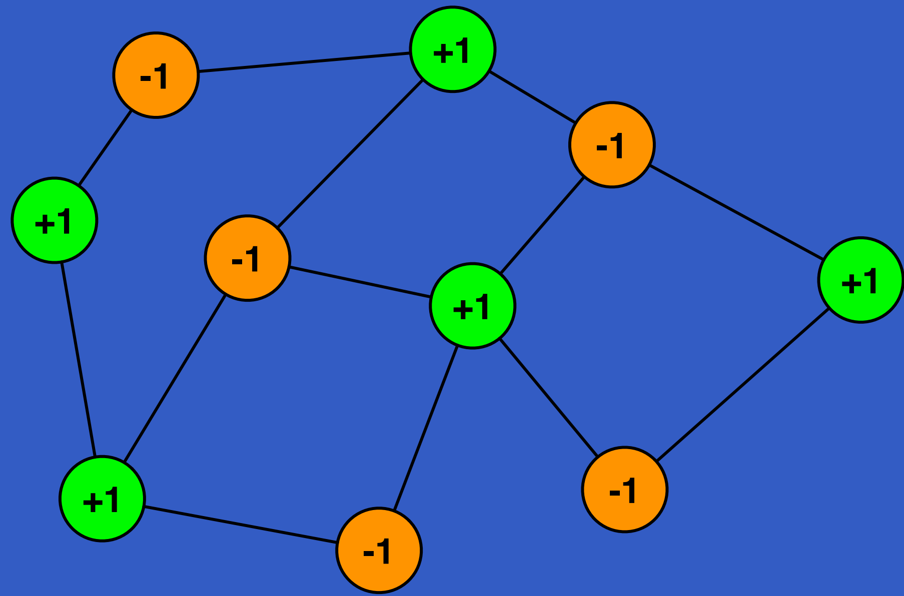
- Graph kernel: $\mathbf{K}_G = \mathbf{L}^+$ (pseudoinverse)
- Reproducing property:

$$g_i = \boldsymbol{e}_i \mathbf{L}^+ \mathbf{L} \mathsf{g} = \mathbf{K}_i \mathbf{L} \mathsf{g} = \langle \mathbf{K}_i, \mathsf{g} \rangle$$

Example



$$\|g\|^2 = 3 \times 4$$



$$\|g\|^2 = 12 \times 4$$

Projection Algorithms

- Perceptron-inspired (but non-conservative)
- Projection: $P(\mathcal{N}; \boldsymbol{w}) := \arg\min_{\boldsymbol{u} \in \mathcal{N}} \|\boldsymbol{u} - \boldsymbol{w}\|$
- Protypical algorithm:

Input: A sequence of closed convex sets $\{\mathcal{U}_t\}_{t=1}^{\ell} \subset \mathcal{H}$
Initialization: $\mathsf{g}_1 = \mathbf{0}$
For $t = 1, \ldots, \ell$ **do**
$\quad \mathsf{g}_{t+1} = P(\mathcal{U}_t; \mathsf{g}_t)$

- Three variants:
  1. `1-Proj` : $\mathcal{U}_t = \{\mathsf{g} : y_t \langle \mathbf{K}_{v_t}, \mathsf{g} \rangle \geq 1\}$
  2. `MNI` : $\mathcal{U}_t = \bigcap_{i=1}^{t} \{\mathsf{g} : y_i \langle \mathbf{K}_{v_i}, \mathsf{g} \rangle = 1\}$
  3. `C-proj` : cycle thru $\{\mathcal{U}_1, \ldots, \mathcal{U}_t\}$ until no “mistakes”
- Prediction: $\hat{y}_t = \mathrm{sign}(\langle \mathbf{K}_{v_t}, \mathsf{g}_t \rangle) = \mathrm{sign}(g_{t,v_t})$

Bounds

Theorem: Given a sequence of vertex label pairs $\{(v_i, y_i)\}_{i=1}^{\ell} \subseteq V \times \{-1, 1\}$ where M is the index set of mistaken trials then the cumulative mistakes of the algorithm is bounded by

$$|M| \leq \|\mathsf{g}^*\|^2 B$$

for all consistent $\mathsf{g}^* \in \mathbb{S}_n$ where

$$B = \text{harmonic-mean}(\{K_{v_i v_i}\}_{i \in M}) \leq \max_{i \in V}(K_{ii})$$

- mistake $\rightarrow$ generalization bounds see [CCG04, etc]

Interpretation of Bounds – 1

- $d_G(p,q)$ length of shortest path p to q for $p,q \in V$.
- Eccentricity: $\rho_p := \max_{q \in V} d_G(p,q)$
- Diameter: $D_G := \max_{p \in V} \rho_p$
- Algebraic connectivity: λ_2 2nd smallest eigenvalue of $\mathbf{L}$

Theorem:

$$\mathbf{K}_{pp} \leq \min\left(\frac{1}{\lambda_2}, \rho_p\right)$$

- Label partition: $(\mathsf{g}^+, \mathsf{g}^-) := (\{i : g_i = 1\}, \{i : g_i = -1\})$ s.t. $|\mathsf{g}^+| + |\mathsf{g}^-| = n$.
- Cut : $\partial(\mathsf{g}^+, \mathsf{g}^-) := \{(i,j) \in E(G) : i \in \mathsf{g}^+, j \in \mathsf{g}^-\}$

Interpretation of Bounds – 2

Theorem: For `1-Proj` and `C-Proj`

$$|M| \leq \underbrace{|\partial(\mathsf{g}^+, \mathsf{g}^-)|}_{\text{cut size}} \underbrace{\left(\frac{n}{\min(|\mathsf{g}^+|, |\mathsf{g}^-|)}\right)^2}_{\text{partition balance}} \underbrace{\min(\frac{1}{\lambda_2}, D_G)}_{\text{structure of } G}$$

for all *consistent* label partitions.

Observations:

- First terms data-dependent, 3rd data-independent
- Bound implies `C-Proj` polynomially many updates
- Compare to cyclic updating with
  - Alternate graph kernel $\mathbf{K} = (L + aI)^{-1}\ ,\ \ 0 < a$.
  - $\mathbb{R}_n$

Online Active Learning – 1

Protocol: A is an index set of active trials. On an active trial t *now* the learner selects v_t.

Theorem: Given a sequence of vertex label pairs $\{(v_i, y_i)\}_{i=1}^{\ell} \subseteq V \times \{-1, 1\}$ where M is the index set of mistaken trials then the cumulative mistakes of the algorithm is bounded by

$$|M \backslash A| \leq \left(\|\mathsf{g}^*\|^2 - Z_A \right) B$$

for all consistent g^* where $Z_A = \underbrace{\sum_{t \in A} \|\mathsf{g}_t - \mathsf{g}_{t+1}\|^2}_{\text{“progress”}}$

Idea: Maximize the “progress”: Z_A

Online Active Learning – 2

Strategy: On trial $t \in A$ choose v_t to max-minimize the per-trial progress $\|\mathsf{g}_t - \mathsf{g}_{t+1}\|^2$ thus choose vertex

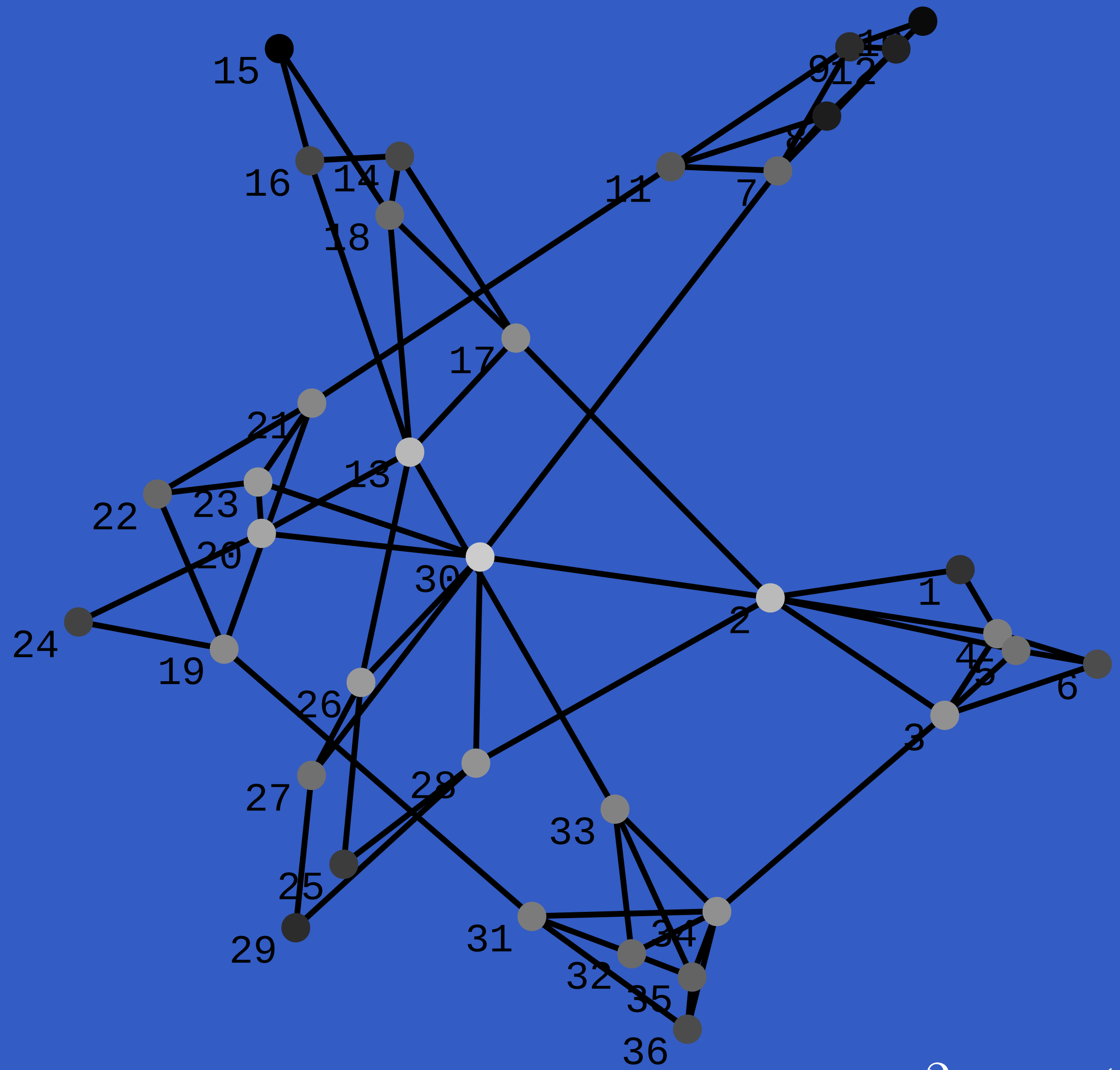
$$
\begin{aligned}
v_t &= \arg\max_{i \in V} \min_{y \in \{-1,1\}} \left\|\mathsf{g}_t - P(\{\mathsf{g} : \langle \mathsf{g}, \mathbf{K}_i \rangle y \geq 1\}; \mathsf{g}_t)\right\|^2 \\
&= \arg\max_{i \in V} \frac{(\min(|g_{t,i}|, 1) - 1)^2}{K_{ii}}
\end{aligned}
$$

Observations:

- The top is the *current* “uncertainty” (margin) of vertex i
- The bottom is a “structural” property of vertex i
- Since $K_{ii} \leq \rho_i$ as an approximation posit that $K_{ii} \sim \rho_i$

Interpretation: the above criteria trades the label uncertainty (“nearness to previous labels”) with “centrality” of vertex

Online Active Learning – 3

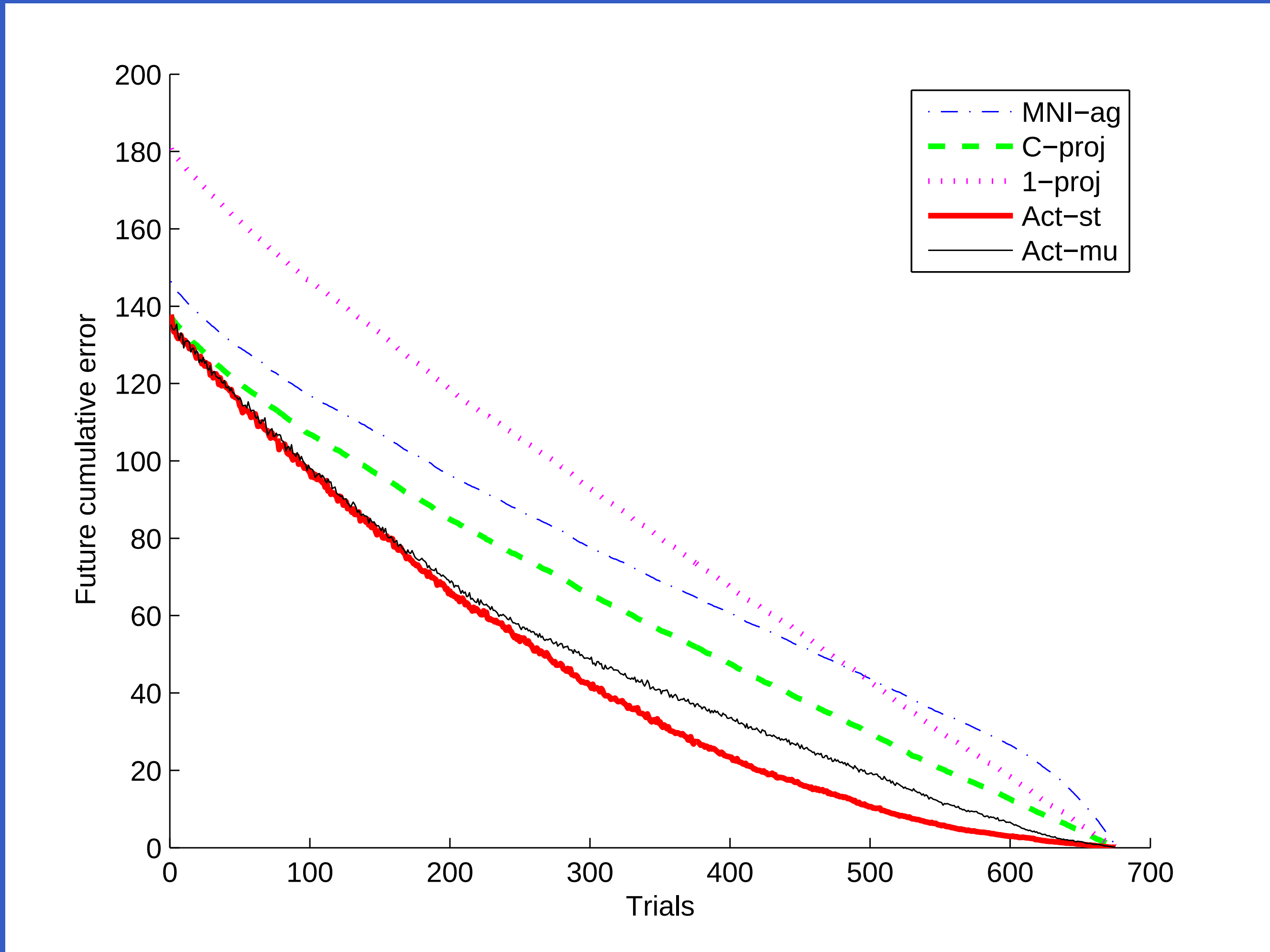


Observe: even if vertex 15 is maximally uncertain ($g_{15} = 0$) vertex 30 is still preferred if the margin $|g_{30}| \leq 0.51$.

Hierarchal random graph $G^2_{k\text{–out}}(6;2)$

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

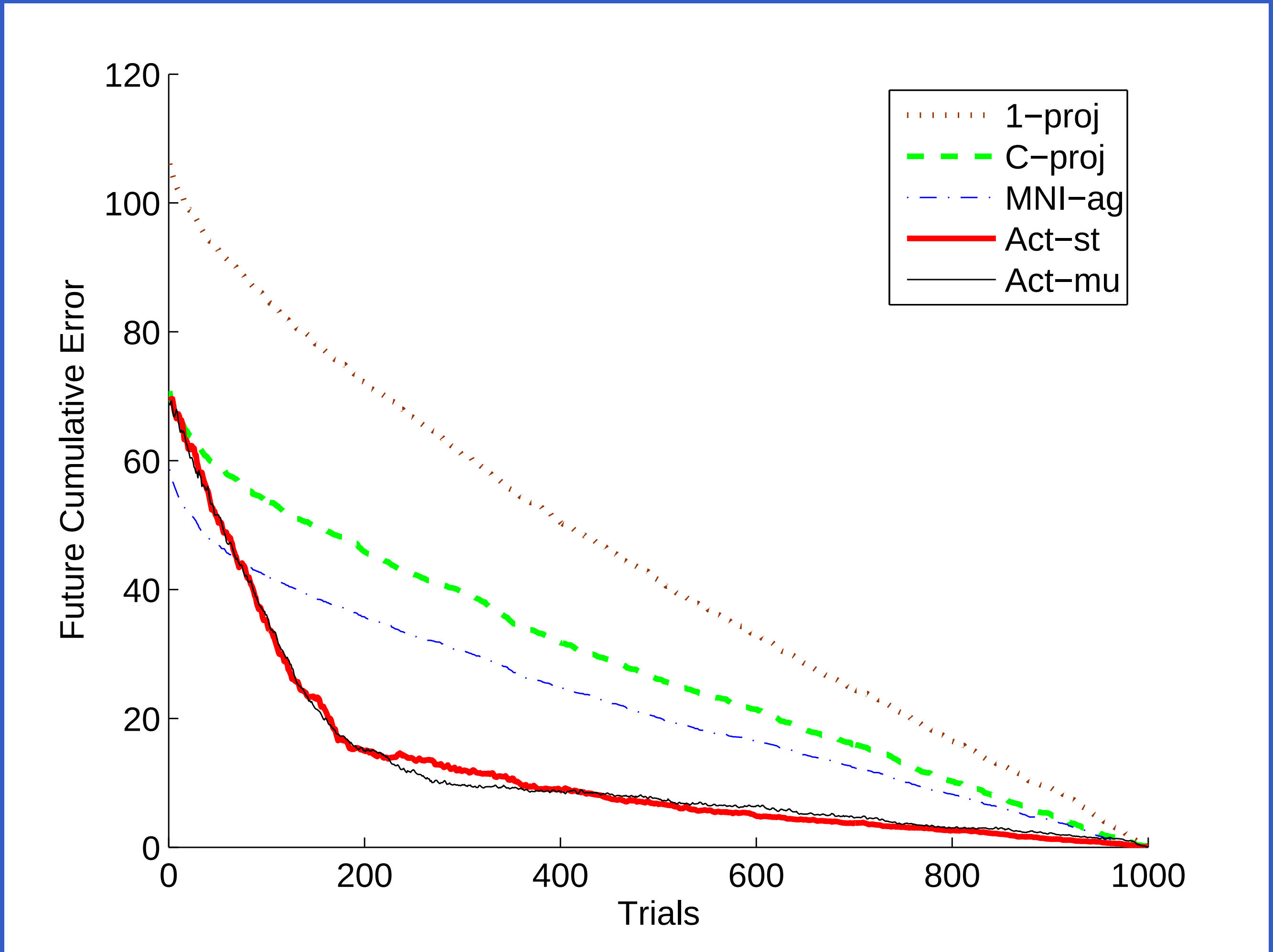
Experiments – 1 (Random Graph)



Hierarchal random graph $G^2_{k-\text{out}}(26; 2)$

726 Nodes labeled by a noisy diffusion process

Experiments – 2 (USPS Even vs Odd)



1000 random samples 100 per digit

Graph built via 3-NN with Euclidean distance

Selected References

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on riemannian manifolds. *Machine Learning*, *56*, 209–239.

Chung, F. R. (1997). *Spectral graph theory*. No. 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Smola, A., & Kondor, R. (2003). Kernels and regularization on graphs. *COLT 2003, Proc.* (pp. 144–158).

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003b). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. *Proc. of the ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in ML and Data Mining* (pp. 58–65).

Active Learning Demo