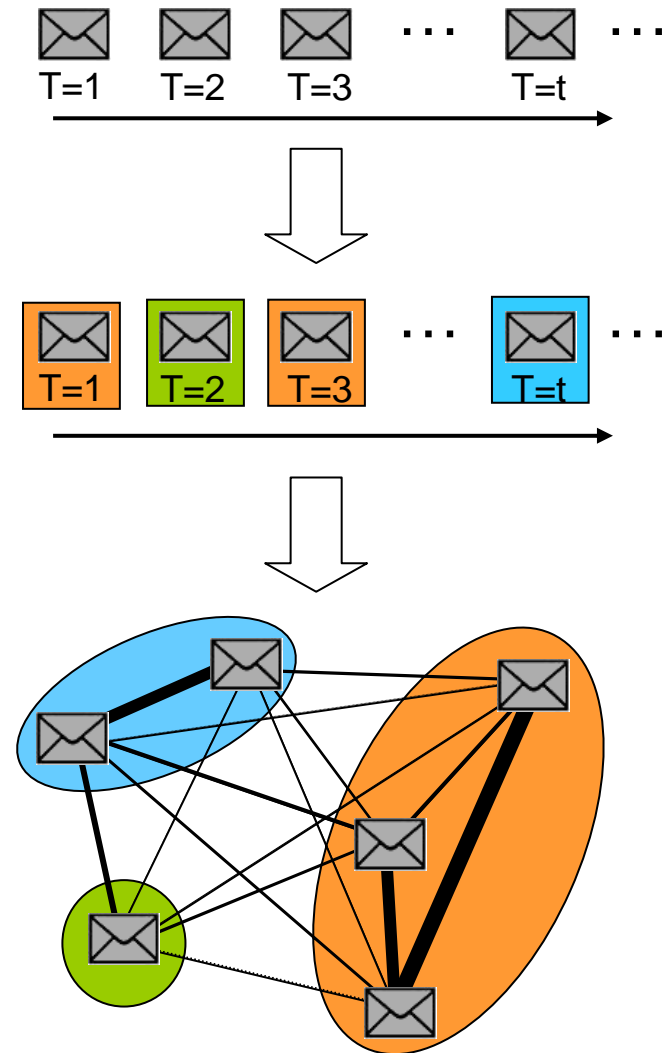


Supervised Clustering of Streaming Data for Email Batch Detection

P. Haider, U. Brefeld, T. Scheffer

- ▶ Problem Setting.
 - ▶ Spam, phishing, virus email are generated according to templates.
 - ▶ Given: A stream of emails.
 - ▶ Task: Detect email batches in the stream.
- ▶ Supervised Clustering task.
 - ▶ Ground-truth exist!
 - ▶ Learn similarity measure with structural SVMs (Finley & Joachims, ICML, 2005).
- ▶ BUT:
 - ▶ Number of constraints in QP is cubic in the number of messages.
 - ▶ Not feasible for real-time applications.



Supervised Clustering of Streaming Data for Email Batch Detection P. Haider, U. Brefeld, T. Scheffer

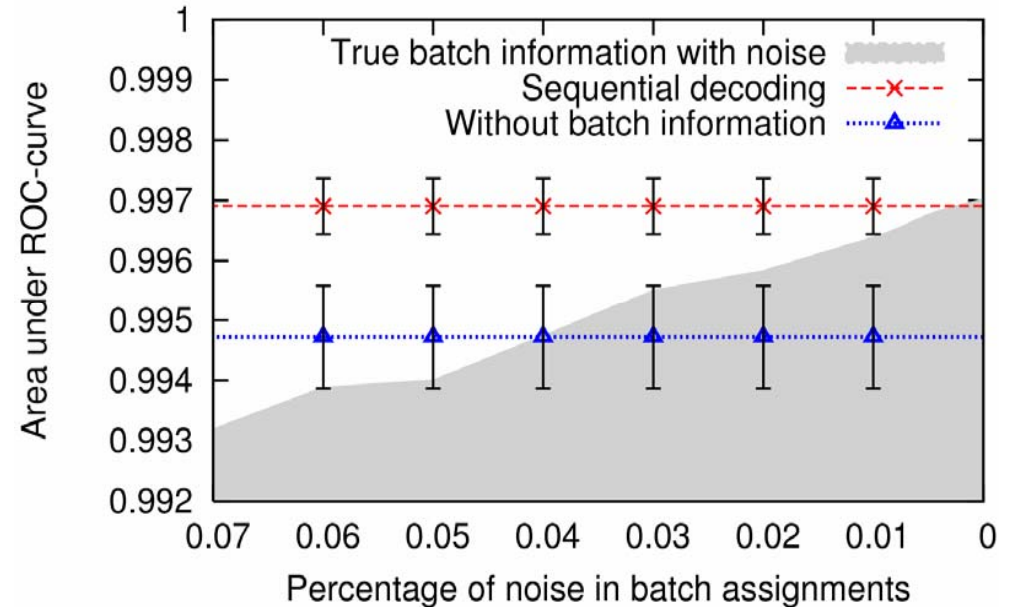
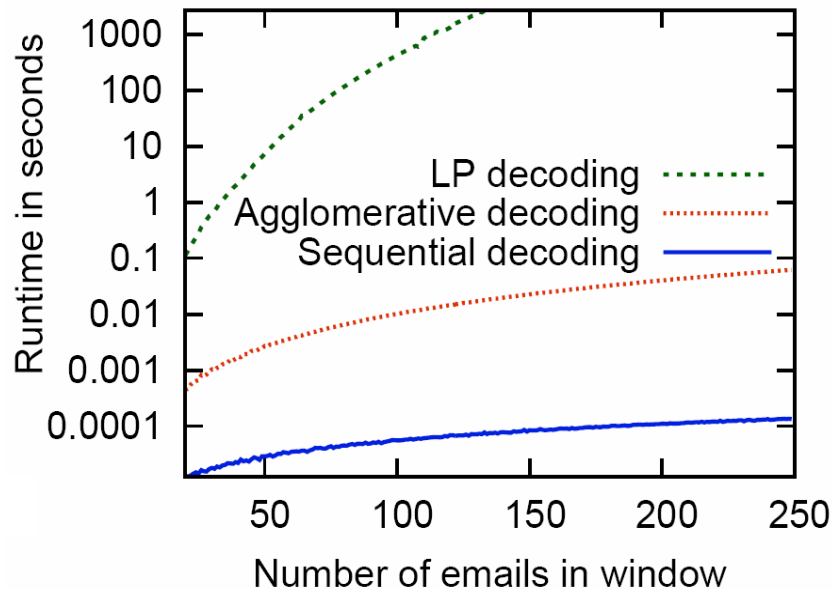
- ▶ Our solution:
 - ▶ Exploit streaming nature of the data.
 - ▶ Cluster assignment cannot be altered once they have been made.
 - ▶ Decompose clustering objective:

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}) &= \sum_{j=1}^T \sum_{k=1}^{j-1} y_{jk} \text{sim}_{\mathbf{w}}(x_j, x_k) \\ &= \underbrace{\sum_{j=1}^{T-1} \sum_{k=1}^{j-1} y_{jk} \text{sim}_{\mathbf{w}}(x_j, x_k)}_{\text{constant}} + \underbrace{\sum_{k=1}^{T-1} y_{Tk} \text{sim}_{\mathbf{w}}(x_T, x_k)}_{\text{objective of sequential update computation in } O(T)} \end{aligned}$$

- ▶ Contributions:
 - ▶ Sequential clustering in linear time in the number of messages.
 - ▶ Sequential large margin approach for parameter estimation.

Supervised Clustering of Streaming Data for Email Batch Detection

P. Haider, U. Brefeld, T. Scheffer



- ▶ Sequential approach is linear in the number of emails.
- ▶ No significant difference in accuracy/error.
- ▶ Reduction of spam misclassification risk by 40%.