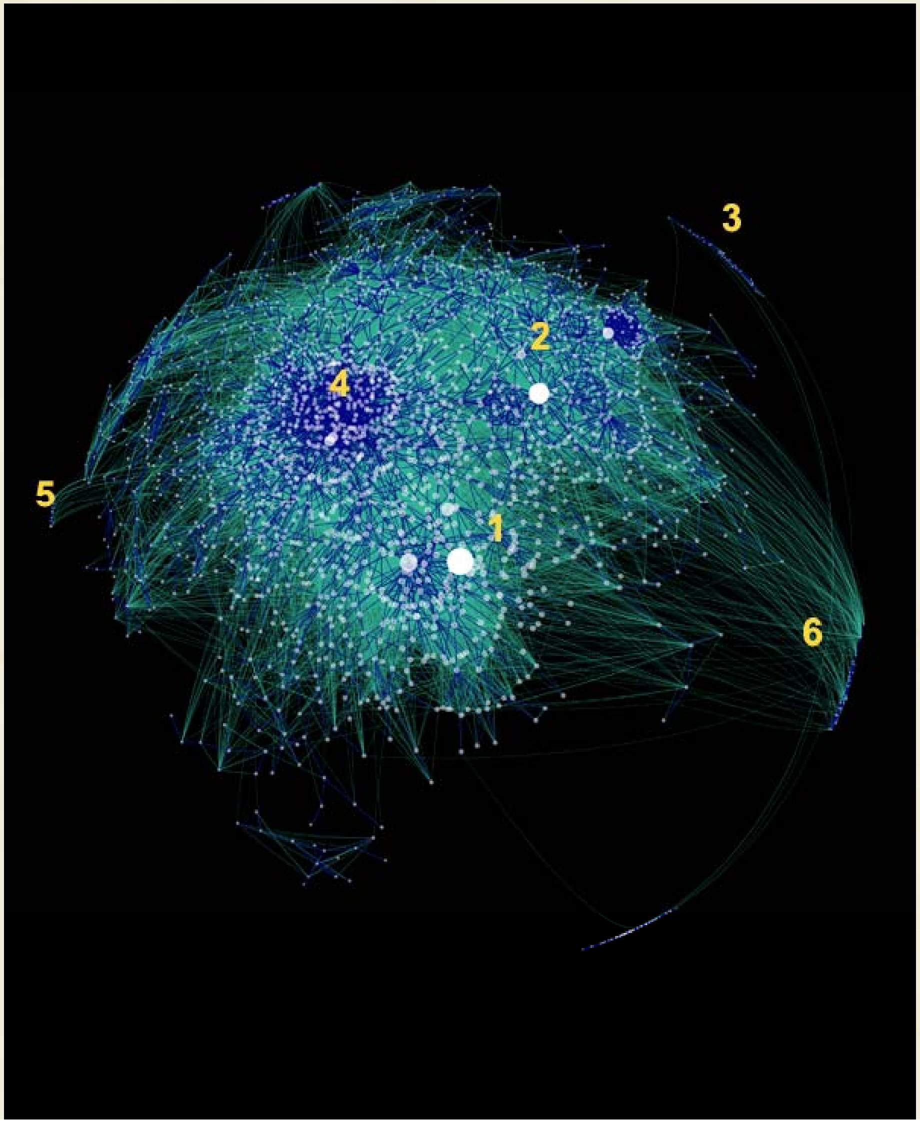
# Evolutionary Clustering

Presenter: Lei Tang

# Evolutionary Clustering

- Processing time stamped data to produce a sequence of clustering.
- Each clustering should be similar to the history, while accurate to reflect corresponding data.
- Trade-off between long-term concept drift and short-term variation.

# Example I: Blogosphere



# Blogosphere

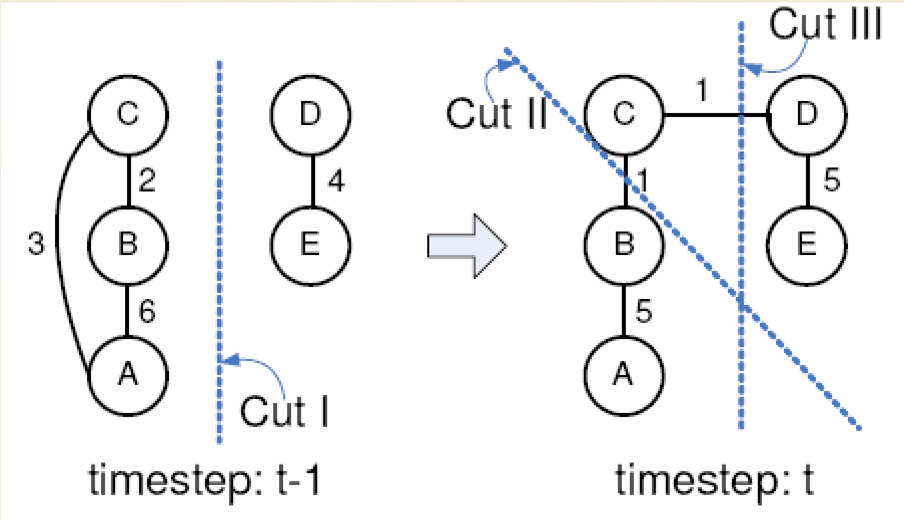
- Community detection
- The overall interest and friendship network is drift slowly.
- Short-term variation is trigged by external event.

# Example II

- Moving objects equipped with GPS sensors are to be clustered (for *traffic jam prediction* or *animal migration analysis*)
- The object follow certain route in the long-term.
- Its estimated coordinate at a given time may vary due to limitations on bandwidth and sensor accuracy.

# The goal

- Current clusters should mainly depend on the current data features.
- Data is expected to change not too quickly. (Temporal Smoothness)



# Related Work

- Online document clustering mainly focusing on novelty detection.
- Clustering data streams: scalability and one-pass-access.
- Incremental clustering: efficiently apply dynamic updates.
- Constrained clustering: must link/can-not link.
- Evolutionary Clustering:
  - The similarity among existing data points varies with time.
  - How cluster evolves smoothly.

# Basic framework

- Snapshot quality: $sq(C_t, M_t)$
- History cost: $hc(C_t, C_{t-1})$
- The total quality of a *cluster sequence*

$$\sum_{t=1}^{T} \mathrm{sq}(C_t, M_t) - \mathrm{cp} \cdot \sum_{t=2}^{T} \mathrm{hc}(C_{t-1}, C_t),$$

- We try to find an optimal cluster sequence greedily without knowing the future.
- Each step, find a cluster that maximize

$$\mathrm{sq}(C_t, M_t) - \mathrm{cp} \cdot \mathrm{hc}(C_{t-1}, C_t).$$

# Construct the similarity matrix

- Local Information Similarity

$$\mathcal{R}(t) = (1-\beta)\cdot\mathcal{B}(t)\mathcal{B}'(t) + \beta\cdot\mathcal{R}(t-1), \quad \text{for } t > 0$$

- Temporal Similarity

$$\mathrm{Corr}(i,j,t_0) = \frac{\sum_{t=1}^{t_0}(x_{i,t}-\mu(i,t))(x_{j,t}-\mu(j,t))}{\sqrt{\mathrm{Var}(i,t)\cdot\mathrm{Var}(j,t)}},$$

- Total Similarity

$$M_t(i,j) = \alpha\cdot S_t(i,j) + (1-\alpha)\cdot\mathrm{Corr}(i,j,t),$$

# Instantiations I: K-means

- Snapshot quality: $\mathrm{sq}(C, M) = \sum_{x \in U} (1 - \min_{c \in C} ||c - x||).$
- History cost: $\mathrm{hc}(C, C') = \min_{f:[k] \to [k]} ||c_i - c'_{f(i)}||,$
- In each k-means iteration, the new centroid between the centroid suggested by non-evolutionary k-means and its closest match from previous time step.

$$
\begin{aligned}
c_j^t \leftarrow \quad & (1-\gamma) \cdot \mathrm{cp} \quad c_{f(j)}^{t-1} \\
& +\gamma \cdot (1 - \mathrm{cp}) \quad \mathop{\mathrm{E}}_{x \in \mathrm{closest}(j)} (x).
\end{aligned}
$$

where

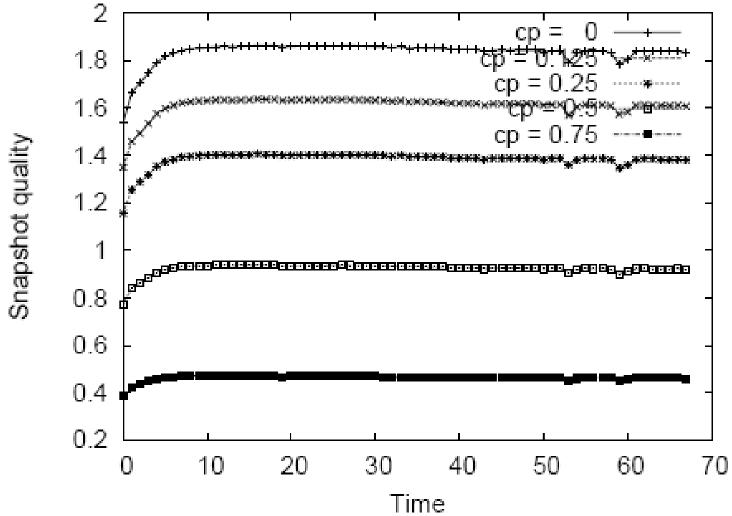
$$\gamma = n_j^t / \left( n_j^t + n_{f(j)}^{t-1} \right)$$

# Agglomerative Clustering

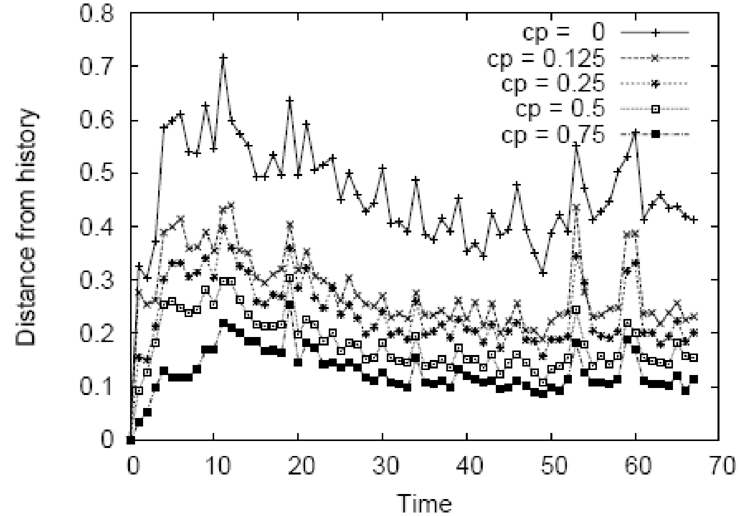
- This is more complicated: need to find out the cluster similarity between two trees (T, T').
- Snapshot quality: the sum of the qualities of all merges performed to create T.
- History cost: $\mathrm{hc}(T', T) = \mathop{\mathrm{E}}_{\substack{i,j \in \mathrm{leaf}(T') \\ i \neq j}} (d_{T',T}(i,j)).$
- 4 greedy heuristics (skipped here):
  - Squared: $\mathrm{sim}_M(m) - \left( \mathrm{cp} \cdot \mathop{\mathrm{E}}_{\substack{i \in \mathrm{leaf}(m_\ell) \\ j \in \mathrm{leaf}(m_r)}} (d_{T',T}(i,j)) \right).$

# Experiment Setup

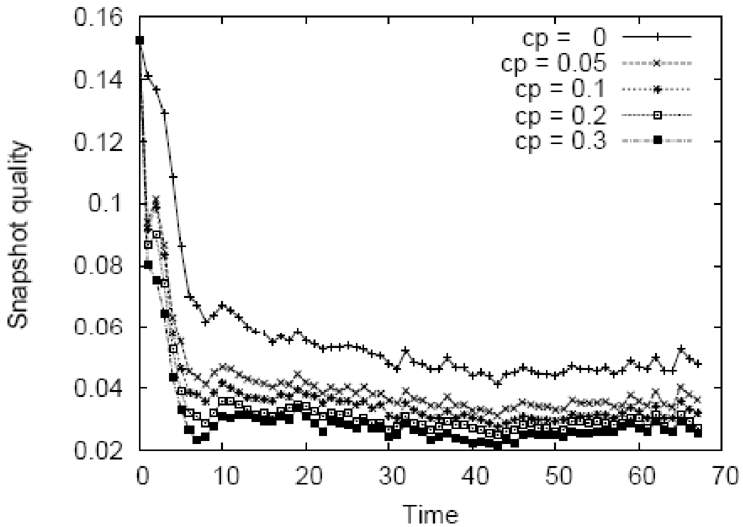
- Data: photo-tag pairs from flickr.com
- Task: Cluster tags
- Two tags are similar if they both occur at the same photo
- However, the experiments in the paper doesn't make much sense for me



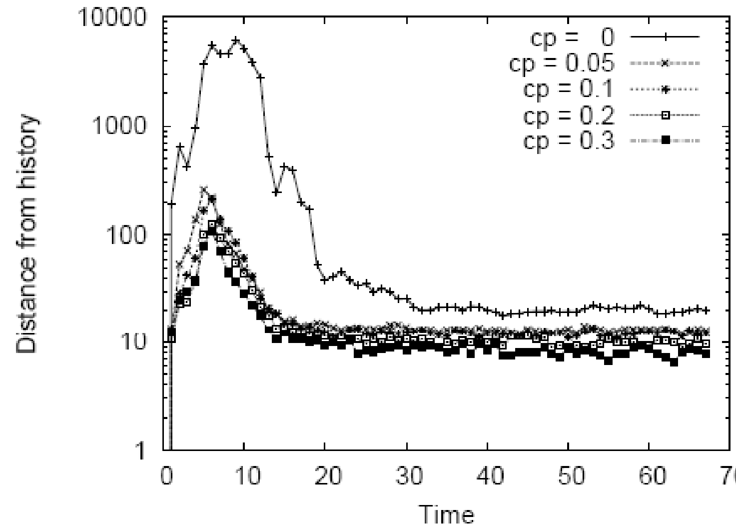

(a) Snapshot quality over time




(b) Distance from history, over time




(a) Linear-Both snapshot quality (log-linear)




(b) Linear-Both distance from history (log-linear)

# Comments

- Pros:
  - New problem
  - Effective heuristics
  - Temporal smoothness is incorporated in both the affinity matrix and the history cost.
- Cons
  - No global solution.
  - Can not handle the change of number of clusters.
  - Experiment seems unreasonable.

# Evolutionary Spectral Clustering

- Idea is almost the same, but here focus on spectral clustering, which preserves nice properties (global solution to a relaxed cut problem, connections to k-means).
- But the idea is presented clearer here.

$$Cost = \alpha \cdot CS + \beta \cdot CT$$

- How to measure the temporal smoothness?
  - Measure the cluster quality on past data
  - Compare the cluster membership

# Spectral Clustering (1)

- K-way average association: $$AA = \sum_{l=1}^{k} \frac{assoc(\mathcal{V}_l, \mathcal{V}_l)}{|\mathcal{V}_l|}$$
- Negated Average Association:

$$NA = Tr(W) - AA = Tr(W) - \sum_{l=1}^{k} \frac{assoc(\mathcal{V}_l, \mathcal{V}_l)}{|\mathcal{V}_l|}$$

- Normalized Cut:

$$NC = \sum_{l=1}^{k} \frac{assoc(\mathcal{V}_l, \mathcal{V} \backslash \mathcal{V}_l)}{assoc(\mathcal{V}_l, \mathcal{V})}$$

- The basic objective is to minimize the normalized cut or negated average association.

# Spectral Clustering (2)

- Typical Procedures
  - Compute eigenvectors X of some variations of the similarity matrix
  - Project all data points into span(X)
  - Applying k-means algorithm to the projected data points to obtain the clustering result.

# K-means Clustering

- Find a partition {v1,v2, … , vk} to minimize the following:

$$KM = \sum_{l=1}^{k} \sum_{i \in \mathcal{V}_l} \|\vec{v}_i - \vec{\mu}_l\|^2$$

# Preserving Cluster Quality

- K-means

$$
\begin{aligned}
Cost_{KM} &= \alpha \cdot CS_{KM} + \beta \cdot CT_{KM} \\
&= \alpha \cdot KM_t\big|_{Z_t} + \beta \cdot KM_{t-1}\big|_{Z_t} \\
&= \alpha \cdot \sum_{l=1}^{k} \sum_{i \in \mathcal{V}_{l,t}} \|\vec{v}_{i,t} - \vec{\mu}_{l,t}\|^2 \\
&\quad + \beta \cdot \sum_{l=1}^{k} \sum_{i \in \mathcal{V}_{l,t}} \|\vec{v}_{i,t-1} - \vec{\mu}_{l,t-1}\|^2
\end{aligned}
$$

Check whether current cluster fits previous cluster.

- A hidden problem, still needs to find the cluster mapping.

# Negated Average Association(1)

- Similar to K-means strategy:

$$\begin{aligned} Cost_{NA} &= \alpha \cdot CS_{NA} + \beta \cdot CT_{NA} \\ &= \alpha \cdot NA_t\big|_{Z_t} + \beta \cdot NA_{t-1}\big|_{Z_t} \end{aligned}$$

- As we know, $NA = Tr(W) - Tr(\tilde{Z}^T W \tilde{Z})$

where $Z^T Z = I_k$.,

$$\begin{aligned} Cost_{NA} &= \alpha \cdot [Tr(W_t) - Tr(\tilde{Z}_t^T W_t \tilde{Z}_t)] \\ &\quad + \beta \cdot [Tr(W_{t-1}) - Tr(\tilde{Z}_t^T W_{t-1} \tilde{Z}_t)] \\ &= Tr(\alpha W_t + \beta W_{t-1}) - Tr\left[\tilde{Z}_t^T (\alpha W_t + \beta W_{t-1}) \tilde{Z}_t\right] \end{aligned} \quad (9)$$

So we just need to maximize the 2nd term.

# Negated Average Association(2)

- The solution to $Tr\left[\tilde{Z}_t^T(\alpha W_t + \beta W_{t-1})\tilde{Z}_t\right]$ are actually the largest k eigenvectors of the matrix.
- Notice that the solution is optimal in terms of a relaxed problem.
- Connection to k-means.
- It is shown that k-means can be reformulated as

$$KM = Tr(A^T A) - Tr(\tilde{Z}^T A^T A \tilde{Z})$$

So k-means is actually a special case of negated average association with a specific similarity definition.

# Normalized Cut

- Normalized cut can be represented as

$$NC = k - Tr\left[Y^T\left(D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\right)Y\right]$$

  with certain constraints.

- Since

$$\begin{aligned} Cost_{NC} &= \alpha \cdot CS_{NC} + \beta \cdot CT_{NC} \\ &= \alpha \cdot NC_t\big|_{Z_t} + \beta \cdot NC_{t-1}\big|_{Z_t} \end{aligned}$$

- We have

$$\begin{aligned} Cost_{NC} &\approx \alpha \cdot k - \alpha \cdot Tr\left[X_t^T\left(D_t^{-\frac{1}{2}}W_tD_t^{-\frac{1}{2}}\right)X_t\right] \qquad (13) \\ &+ \beta \cdot k - \beta \cdot Tr\left[X_t^T\left(D_{t-1}^{-\frac{1}{2}}W_{t-1}D_{t-1}^{-\frac{1}{2}}\right)X_t\right] \\ &= k - Tr\left[X_t^T\left(\alpha D_t^{-\frac{1}{2}}W_tD_t^{-\frac{1}{2}} + \beta D_{t-1}^{-\frac{1}{2}}W_{t-1}D_{t-1}^{-\frac{1}{2}}\right)X_t\right] \end{aligned}$$

Again a trace maximization problem.

# Discussion on PCQ framework

- Very intuitive
- The historic similarity matrix is scaled and combined with current similarity matrix.

# Preserving Cluster Membership

- Temporal cost is measured as the difference between current partition and historical partition.
- Use chi-square statistics to represent the distance:

$$\chi^2(Z_t, Z_{t-1}) = n\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{|\mathcal{V}_{ij}|^2}{|\mathcal{V}_{i,t}|\cdot|\mathcal{V}_{j,t-1}|} - 1\right)$$

So for K-means

$$Cost_{KM} = \alpha \cdot CS_{KM} + \beta \cdot CT_{KM} \quad (15)$$
$$= \alpha \cdot \sum_{l=1}^{k}\sum_{i\in\mathcal{V}_{l,t}} \|\vec{v}_{i,t} - \vec{\mu}_{l,t}\|^2 - \beta \cdot \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{|\mathcal{V}_{ij}|^2}{|\mathcal{V}_{i,t}|\cdot|\mathcal{V}_{j,t-1}|}$$

# Negated Average Association(1)

- Distance: $dist(X_t, X_{t-1}) = \frac{1}{2}\|X_t X_t^T - X_{t-1}X_{t-1}^T\|^2$
- So

$$Cost_{NA} = \alpha \cdot CS_{NA} + \beta \cdot CT_{NA} \quad (17)$$

$$= \alpha \cdot \left[Tr(W_t) - Tr(X_t^T W_t X_t)\right] + \frac{\beta}{2} \cdot \|X_t X_t^T - X_{t-1}X_{t-1}^T\|^2$$

$$= \alpha \cdot \left[Tr(W_t) - Tr(X_t^T W_t X_t)\right] +$$

$$\frac{\beta}{2} Tr\left(X_t X_t^T - X_{t-1}X_{t-1}^T\right)^T \left(X_t X_t^T - X_{t-1}X_{t-1}^T\right)$$

$$= \alpha \cdot \left[Tr(W_t) - Tr(X_t^T W_t X_t)\right] +$$

$$\frac{\beta}{2} Tr(X_t X_t^T X_t X_t^T - 2X_t X_t^T X_{t-1}X_{t-1}^T + X_{t-1}X_{t-1}^T X_{t-1}X_{t-1}^T)$$

$$= \alpha \cdot \left[Tr(W_t) - Tr(X_t^T W_t X_t)\right] + \beta k - \beta Tr\left(X_t^T X_{t-1}X_{t-1}^T X_t\right)$$

$$= \alpha \cdot Tr(W_t) + \beta \cdot k - Tr\left[X_t^T(\alpha W_t + \beta X_{t-1}X_{t-1}^T)X_t\right]$$

# Negated Average Association(2)

- It can be shown that the unrelaxed partition:

$$\frac{1}{2}\|\tilde{Z}_t\tilde{Z}_t^T - \tilde{Z}_{t-1}\tilde{Z}_{t-1}^T\|^2 = k - \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{|\mathcal{V}_{ij}|^2}{|\mathcal{V}_{i,t}| \cdot |\mathcal{V}_{j,t-1}|} \quad (18)$$

- So negated average association can be applied to solve the original evolutionary k-means

# Normalized Cut

- Straight forward

$$
\begin{aligned}
Cost_{NC} &= \alpha \cdot CS_{NC} + \beta \cdot CT_{NC} \qquad (19) \\
&= \alpha \cdot k - \alpha \cdot Tr\left[X_t^T \left(D_t^{-\frac{1}{2}} W_t D_t^{-\frac{1}{2}}\right) X_t\right] \\
&\quad + \frac{\beta}{2} \cdot \| X_t X_t^T - X_{t-1} X_{t-1}^T \|^2 \\
&= k - Tr\left[X_t^T \left(\alpha D_t^{-\frac{1}{2}} W_t D_t^{-\frac{1}{2}} + \beta X_{t-1} X_{t-1}^T\right) X_t\right]
\end{aligned}
$$

# Comparing PQC & PCM

- As for the temporal cost,
  - In PCQ, we need to maximize

$$Tr(X_t^T W_{t-1} X_t)$$

  - In PCM, we need to maximize

$$Tr(X_t^T X_{t-1} X_{t-1}^T X_t)$$

- Connection:

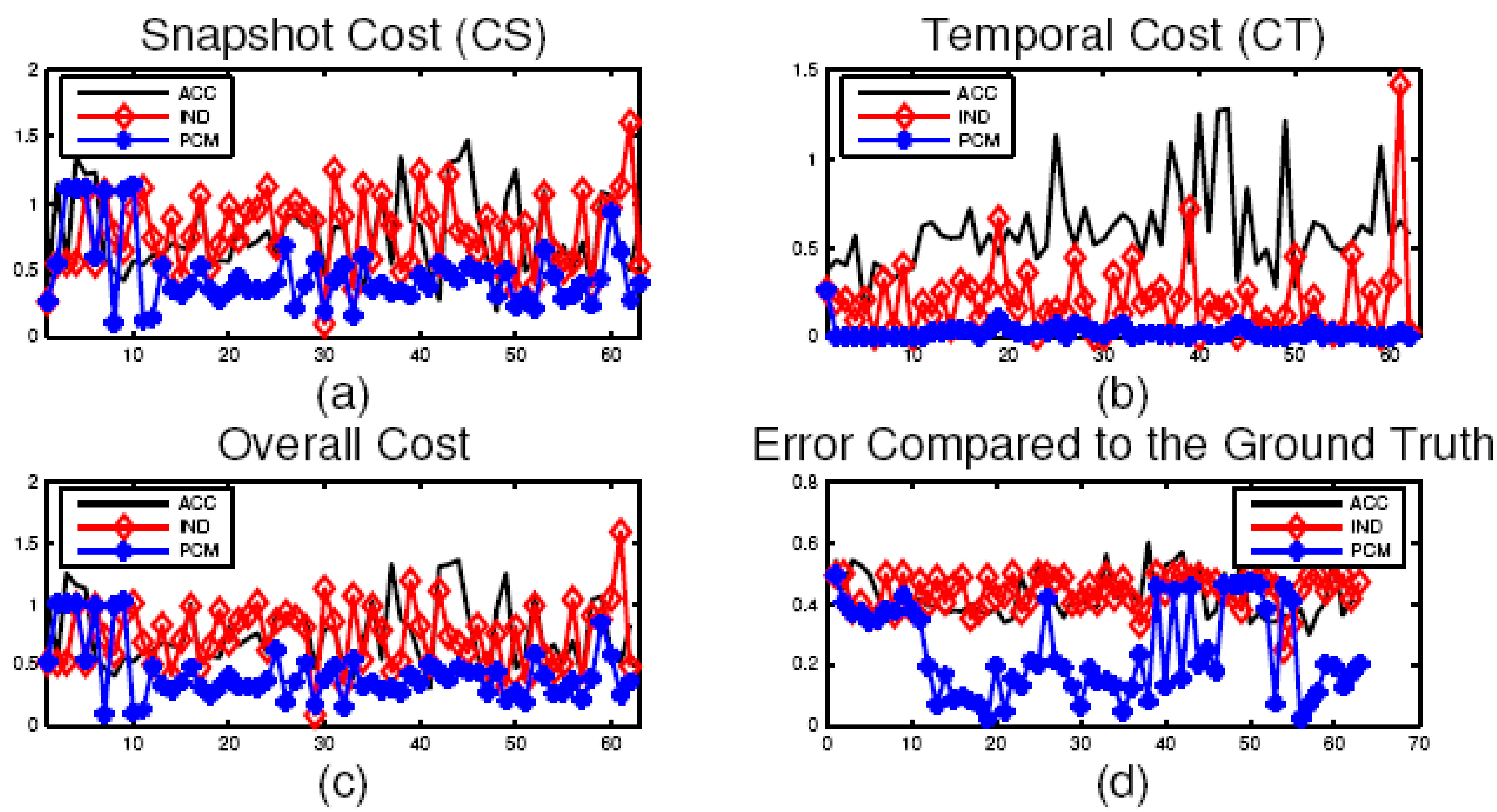
$$X_t^T W_{t-1} X_t = X_t^T (X_{t-1}, X_{t-1}^{\perp}) \Lambda_{t-1} (X_{t-1}, X_{t-1}^{\perp})^T X_t$$

- In PCQ, all the eigen vectors are considered and penalized according to the eigen values.

# Real Blog Data

- 407 blogs during 63 consecutive weeks.
- 148,681 links.
- Two communities (ground truth, labeled manually based on contents)
- Affinity matrix is constructed based on links

# Experiment Result



# Comments

- Nice formulation which has a global solution for the relaxed version.
- Strong connection between k-means and negated average association.
- Can handle new objects or change of number of clusters.

# Any Questions?