Similarity Estimation Techniques from Rounding Algorithms

Moses Charikar
Princeton University

Compact sketches for estimating similarity

- Collection of objects, e.g. mathematical representation of documents, images.
- Implicit similarity/distance function.
- Want to estimate similarity without looking at entire objects.
- Compute compact sketches of objects so that similarity/distance can be estimated from them.

Similarity Preserving Hashing

- Similarity function *sim(x,y)*
- Family of hash functions *F* with probability distribution such that

$$\Pr_{h \in F}[h(x) = h(y)] = sim(x, y)$$

Applications

- Compact representation scheme for estimating similarity

$$x \rightarrow (h_1(x), h_2(x), \ldots, h_k(x))$$

$$y \rightarrow (h_1(y), h_2(y), \ldots, h_k(y))$$

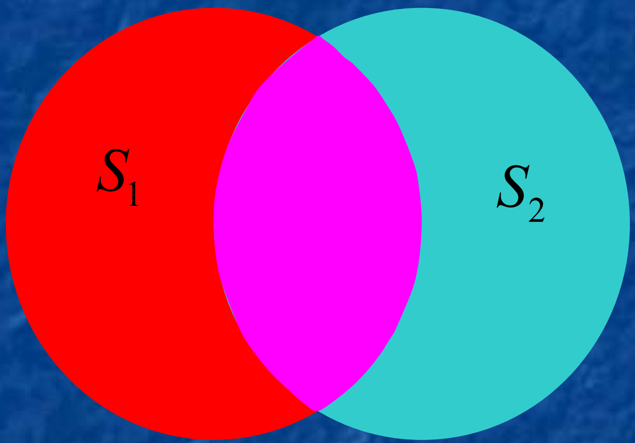
- Approximate nearest neighbor search [Indyk,Motwani] [Kushilevitz,Ostrovsky,Rabani]

Estimating Set Similarity

[Broder,Manasse,Glassman,Zweig]

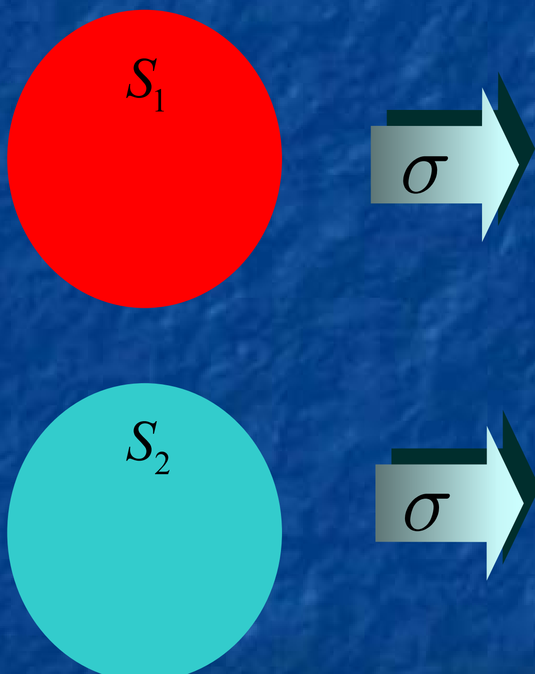
[Broder,C,Frieze,Mitzenmacher]

- Collection of subsets



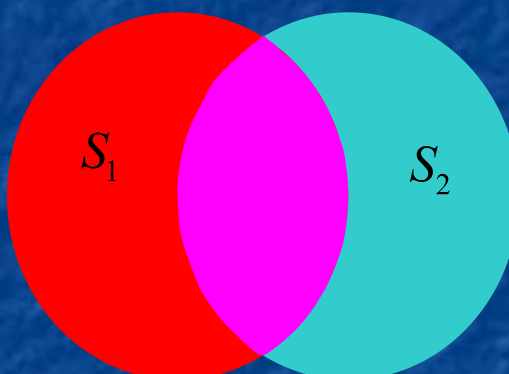
$$\text{similarity} = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Minwise Independent Permutations



$$\min(\sigma(S_1))$$

$$\min(\sigma(S_2))$$



$$\mathrm{prob}(\min(\sigma(S_1) = \min(\sigma(S_2)) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Related Work

- Streaming algorithms
  - Compute f(data) in one pass using small space.
  - Implicitly construct sketch of data seen so far.
- Synopsis data structures [Gibbons,Matias]
- Compact distance oracles, distance labels.
- Hash functions with similar properties:
  [Linial,Sassoon]
  [Indyk,Motwani,Raghavan,Vempala]
  [Feige, Krauthgamer]

Results

- Necessary conditions for existence of similarity preserving hashing (SPH).

- SPH schemes from rounding algorithms
  - Hash function for vectors based on random hyperplane rounding.
  - Hash function for estimating Earth Mover Distance based on rounding schemes for classification with pairwise relationships.

Existence of SPH schemes

- *sim(x,y)* admits an SPH scheme if $\exists$ family of hash functions *F* such that

$$\Pr_{h \in F}[h(x) = h(y)] = sim(x, y)$$

Theorem: If *sim(x,y)* admits an SPH scheme then *1-sim(x,y)* satisfies triangle inequality.

Proof:

$$1 - sim(x,y) = \Pr_{h \in F}(h(x) \neq h(y))$$

$\Delta_h(x,y)$: indicator variable for $h(x) \neq h(y)$

$$\Delta_h(x,y) + \Delta_h(y,z) \geq \Delta_h(x,z)$$

$$1 - sim(x,y) = \mathrm{E}_{h \in F}[\Delta_h(x,y)]$$

Stronger Condition

Theorem: If *sim(x,y)* admits an SPH scheme then *(1+sim(x,y))/2* has an SPH scheme with hash functions mapping objects to {0,1}.

Theorem: If *sim(x,y)* admits an SPH scheme then *1-sim(x,y)* is isometrically embeddable in the Hamming cube.

Random Hyperplane Rounding based SPH

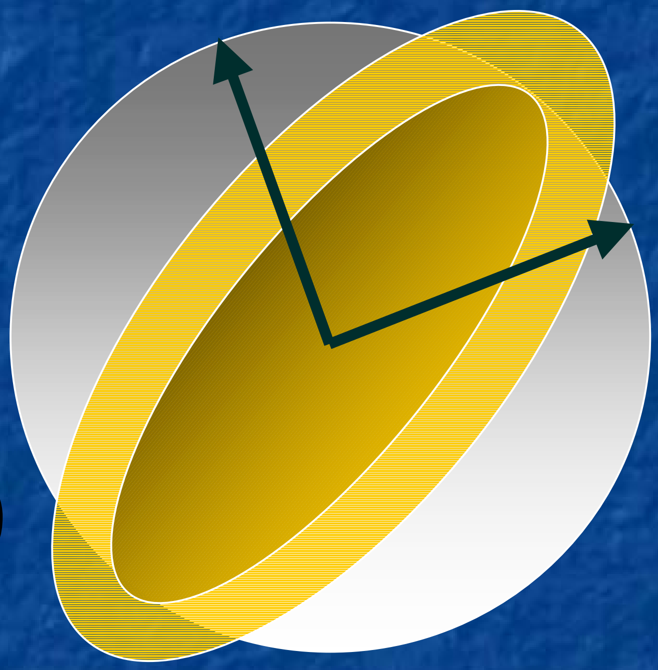
- Collection of vectors

$$sim(\vec{u},\vec{v}) = 1 - \frac{\measuredangle(\vec{u},\vec{v})}{\pi}$$

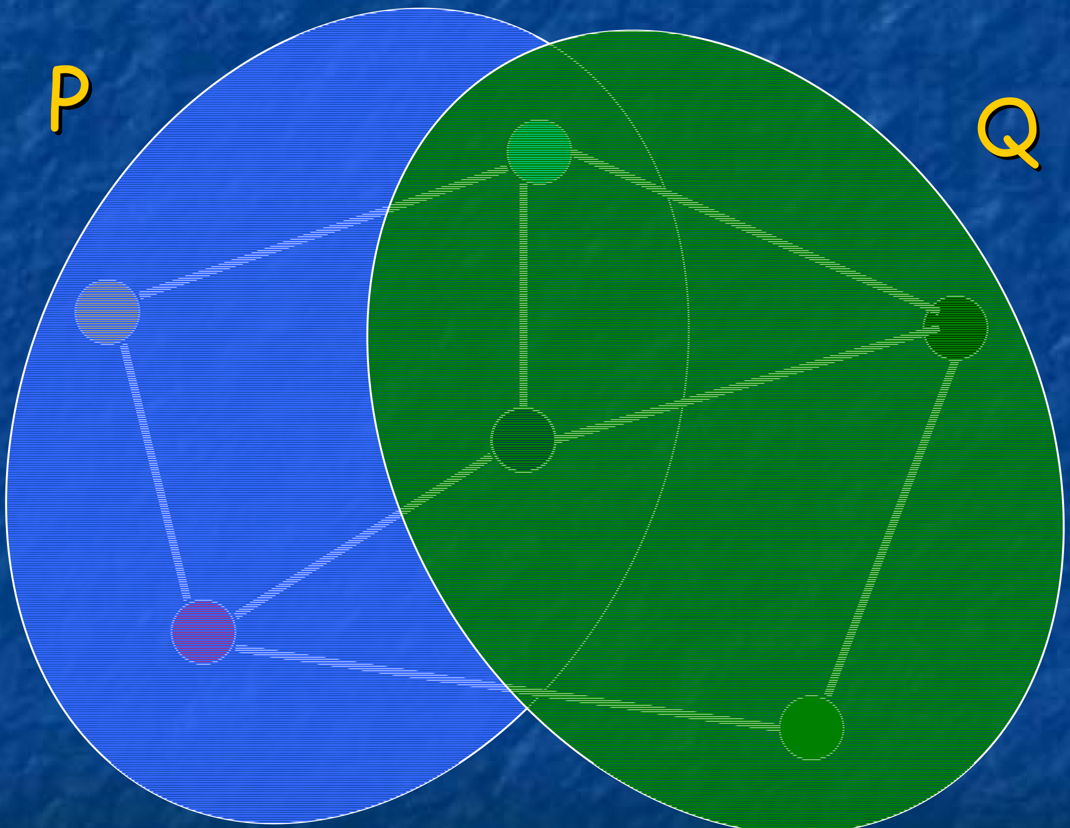
- Pick random hyperplane through origin (normal $\vec{r}$)

$$h_{\vec{r}}(\vec{u}) = \begin{cases} 1 & \text{if } \vec{r}\cdot\vec{u} \geq 0 \\ 0 & \text{if } \vec{r}\cdot\vec{u} < 0 \end{cases}$$

- [Goemans,Williamson]



Earth Mover Distance (EMD)



EMD(P,Q)

Earth Mover Distance

- Set of points $L=\{l_1,l_2,\ldots l_n\}$
- Distance function $d(i,j)$ (assume metric)
- Distribution $P(L)$: non-negative weights $(p_1,p_2,\ldots p_n)$.
- Earth Mover Distance (EMD): distance between distributions P and Q.
- Proposed as metric in graphics and vision for distance between images. [Rubner,Tomasi,Guibas]

$$\min \sum_{i,j} f_{i,j} \cdot d(i,j)$$

$$\forall i \quad \sum_{\mathrm{j}} f_{i,j} = p_i$$

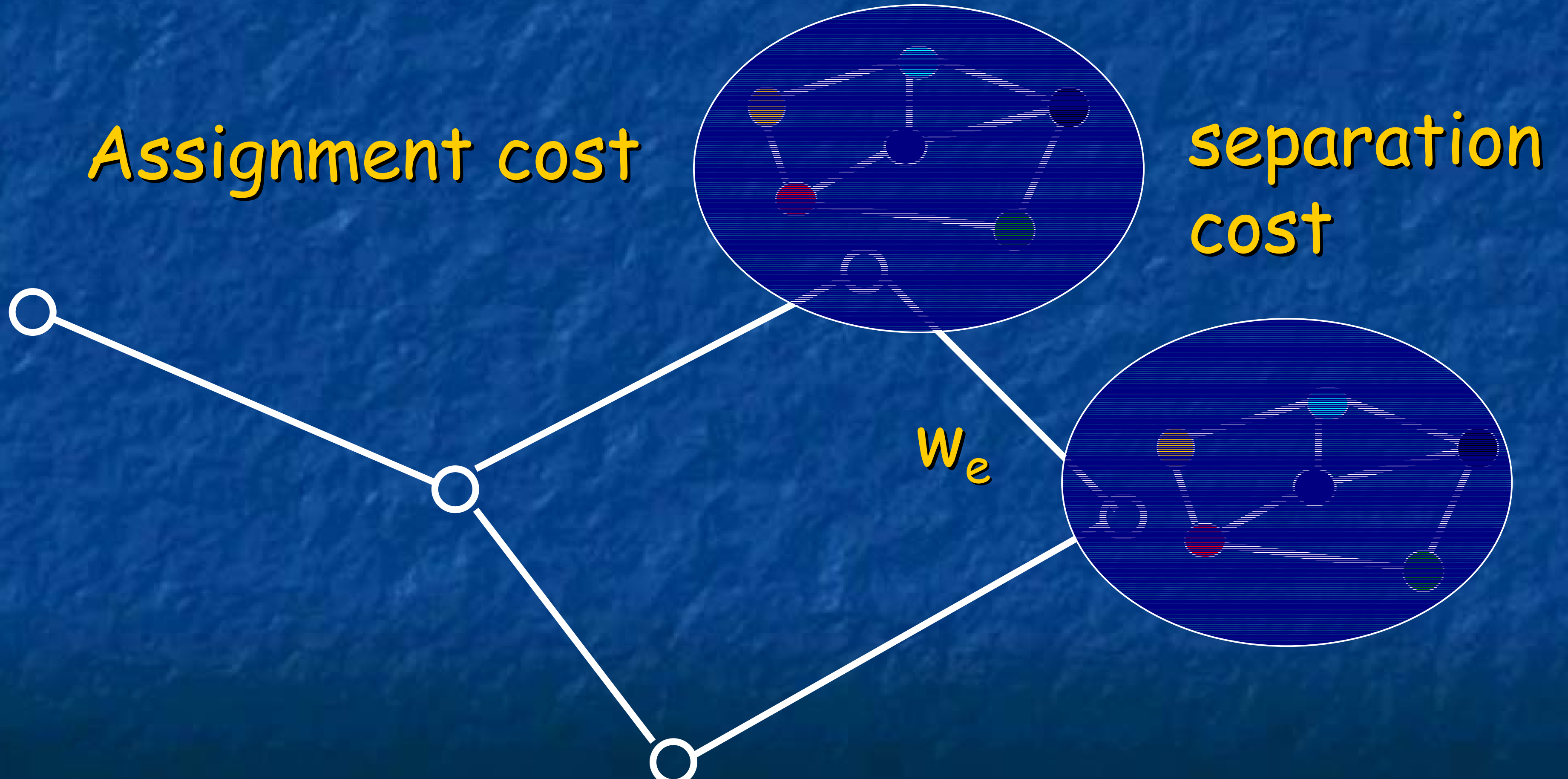
$$\forall j \quad \sum_{\mathrm{i}} f_{i,j} = q_j$$

$$\forall i,j \quad f_{i,j} \geq 0$$

Relaxation of SPH

- Estimate distance measure, not similarity measure in [0,1].
- Allow hash functions to map objects to points in metric space and measure $E[d(h(P),h(Q))]$.
  (SPH: $d(x,y) = 1$ if $x \neq y$)
- Estimator will approximate EMD.

Classification with pairwise relationships [Kleinberg,Tardos]

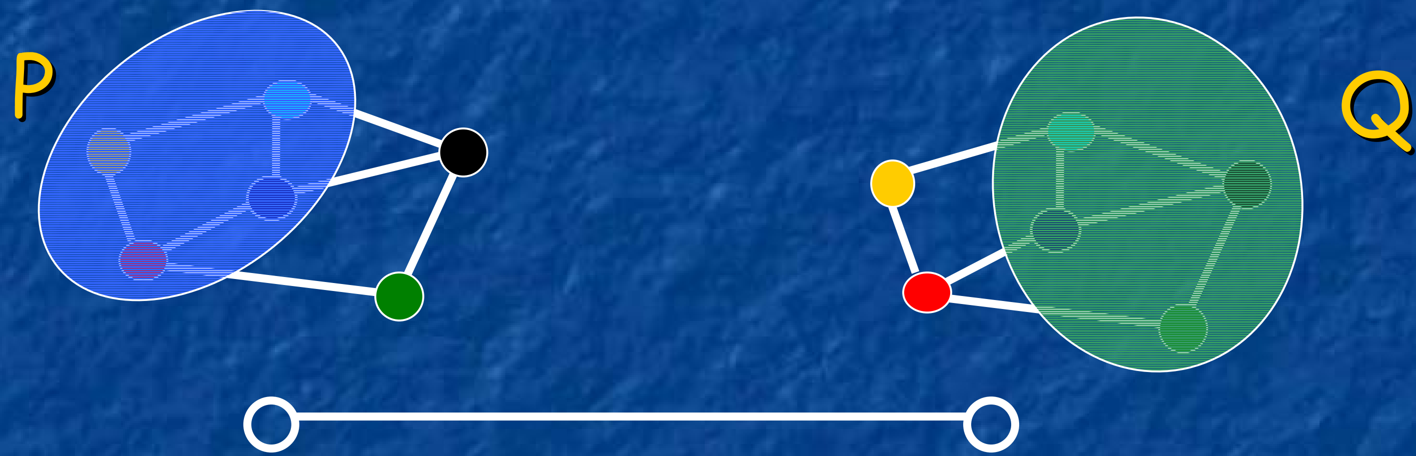


Classification with pairwise relationships

- Collection of objects V
- Labels $L=\{l_1,l_2,\ldots l_n\}$
- Assignment of labels $h : V \rightarrow L$
- Cost of assigning label to u : $c(u,h(u))$
- Graph of related objects; for edge $e=(u,v)$, cost paid: $w_e.d(h(u),h(v))$
- Find assignment of labels to minimize cost.

LP Relaxation and Rounding

[Kleinberg,Tardos]
[Chekuri,Khanna,Naor,Zosin]



Separation cost measured by EMD(P,Q)

Rounding algorithm guarantees

$\Pr[h(P)=l_i] = p_i$

$E[d(h(P),h(Q)] \leq O(\log n \log\log n)\ EMD(P,Q)$

Rounding details

- Probabilistically approximate metric on L by tree metric (HST)
- Expected distortion $O(\log n \log \log n)$
- EMD on tree metric has nice form:
- T: subtree
- $P(T)$: sum of probabilities for leaves in T
- l_T : length of edge leading up from T
- $EMD(P,Q) = \sum l_T|P(T)-Q(T)|$

Theorem: The rounding scheme gives a hashing scheme such that

$$EMD(P,Q) \leq E[d(h(P),h(Q))]$$
$$\leq O(\log n \log\log n)\ EMD(P,Q)$$

Proof: $y_{i,j}$: Probability that $h(P) = l_i, h(Q) = l_j$

$y_{i,j}$ give feasible solution to LP for EMD

Cost of this solution $= \mathrm{E}[d(h(P),h(Q))]$

Hence $EMD(P,Q) \leq \mathrm{E}[d(h(P),h(Q))]$

SPH for weighted sets

- Weighted Set: $(p_1, p_2, \ldots p_n)$, weights in [0,1]
- Kleinberg-Tardos rounding scheme for uniform metric can be thought of as a hashing scheme for weighted sets with

$$sim(P,Q) = \frac{\sum \min(p_i, q_i)}{\sum \max(p_i, q_i)}$$

- Generalization of minwise independent permutations

Conclusions and Future Work

- Interesting connection between rounding procedures for approximation algorithms and hash functions for estimating similarity.
- Better estimators for Earth Mover Distance
- Ignored variance of estimators: related to dimensionality reduction in L_1
- Study compact representation schemes in general