Approximation Algorithms for Clustering Uncertain Data

Graham Cormode
AT&T Labs - Research
graham@research.att.com

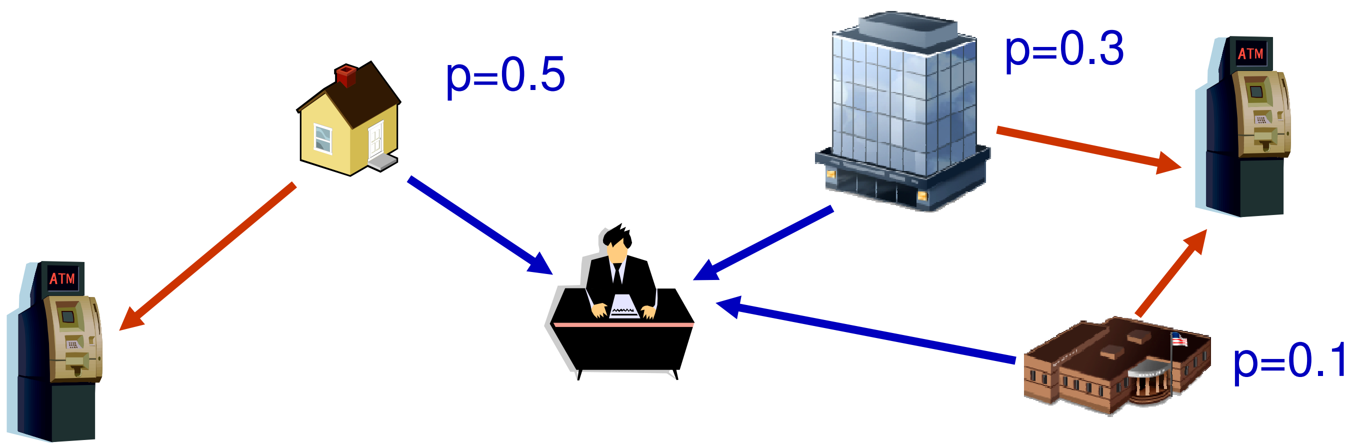
Andrew McGregor
UCSD / MSR / UMass Amherst
andrewm@ucsd.edu

Introduction

- Many applications generate data which is uncertain:
  - Quality of Record Linkages
  - Confidences of extracted rules
  - Noisy Sensor/RFID readings
- Leads us to study *probabilistic data management*
- Much recent study on uncertain data in the DBMS
  - Answering SQL style queries with probabilities
- Less work on *mining* uncertain data — equally important

Clustering Uncertain Data

- We study the core mining problem of clustering
  - Given knowledge about the *distribution* of each data point, how to locate cluster centers that optimize expected cost?
- Example: bank wants to place new locations
  - Each customer has a distribution (e.g. home, work, school)
  - Place "home branch" for each customer to minimize dist
  - Place ATMs so expected distance to any is minimized



Related Work

- Distinct from "soft clustering"
  - Soft clustering: hard location of points need soft assignment
  - Here: soft location of points, desired hard assignment
- Initial heuristics proposed for clustering uncertain data
  - Typically, treat probabilities as weights, or use traditional clustering on expected distances
  - No approximation guarantees known – no attempt to define optimization criteria

Preliminaries

- Models of data:
  1. Point probability: each point either appears with probability p_i at x_i, or else does not appear
  2. Discrete PDF: specifies $Pr[X_i = x_i]$ for a set of locations $\{x_i\}$
  3. Continuous PDF: e.g. Gaussian defined by mean and variances describes possible location
- Models of clustering:
  - Unassigned: wherever a point appears, it is associated with its closest cluster center
  - Assigned: wherever point X_i appears, it is assigned to center $\sigma(i)$. Algorithm must specify $\sigma()$

Cost Metrics

- We generalize well-known metrics from the deterministic case:
  - k-median: expected sum of distances from points to centers
  - k-means: expected sum of squared distances
  - k-center: expected max distance of a point to a center
- Expectations are taken over all possible worlds

- Given a particular set of centers and points, the cost is well-defined, hence we can try to optimize.

Our Results

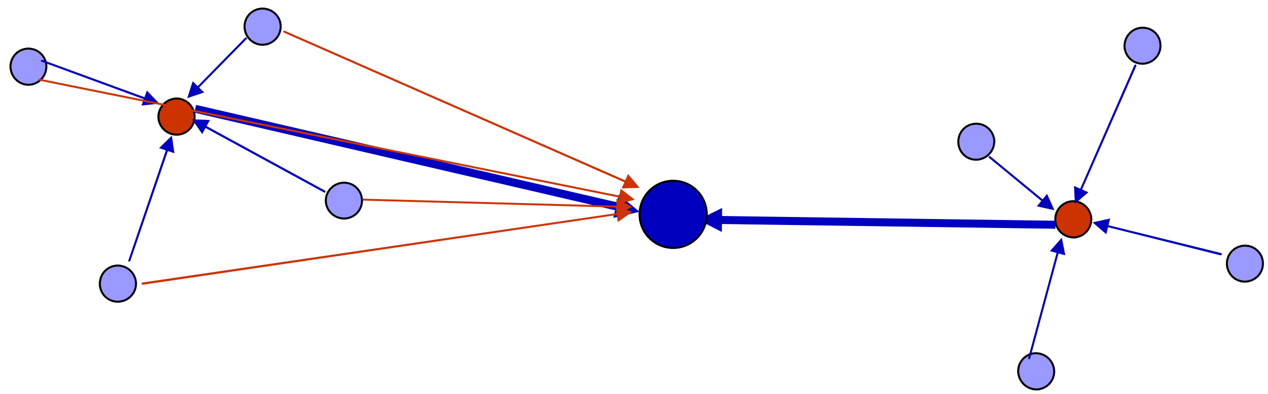
Objective	Metric	Assignment	α	β
k-center (point probability)	Any metric	Unassigned	$1+\epsilon$	$O(\epsilon^{-1}\log^2 n)$
	Any metric	Unassigned	$12+\epsilon$	2
k-center (discrete pdf)	Any metric	Unassigned	$1.582+\epsilon$	$O(\epsilon^{-1}\log^2 n)$
	Any metric	Unassigned	$18.99+\epsilon$	2
k-means	Euclidean	Unassigned	$1+\epsilon$	1
	Euclidean	Assigned	$1+\epsilon$	1
k-median	Any metric	Unassigned	$3+\epsilon$	1
	Euclidean	Unassigned	$1+\epsilon$	1
	Any metric	Assigned	$7+\epsilon$	1
	Euclidean	Assigned	$3+\epsilon$	1

- (α,β) approximations output (βk) centers to give α-approximation of best k-center clustering

k-means and k-median

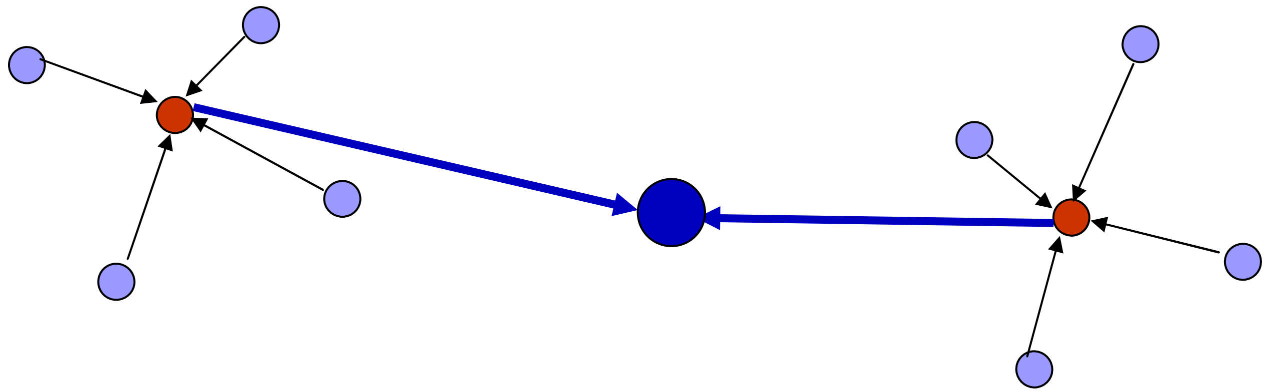
- Due to linearity, unassigned versions of k-means and k-median are quite simple:
  - By linearity of expectation, the cost is equivalent to deterministic clustering with probabilities as weights
- Assigned version is more complex, since expected distance depends which center we assign it to
- Basic idea: cluster each PDF to find best 1-cluster, then cluster these clusters

Assigned k-means



- Can show that cost of assigning a point to some center is equal to cost assigning weighted centroid of PDF to that center, plus “variance” of the PDF
  - Good homework problem (Pythagoras on each dimension)
- Since variance is positive, α-approximation of clustering centroids yields α-approximation for original problem
  - Plug in $(1+\varepsilon)$ approximation for k-means in Euclidean space

Assigned k-median



- Clustering the 1-medians is no longer approximation preserving
- Some algebra shows that given an α-approximation for weighted k-median, we obtain a $(2\alpha + 1)$ approximation
  - Plug in $(1+\varepsilon)$ approx in Euclidean space or $(3+\varepsilon)$ in arbitrary metric space
  - Similar techniques used in clustering streams of points

k-center

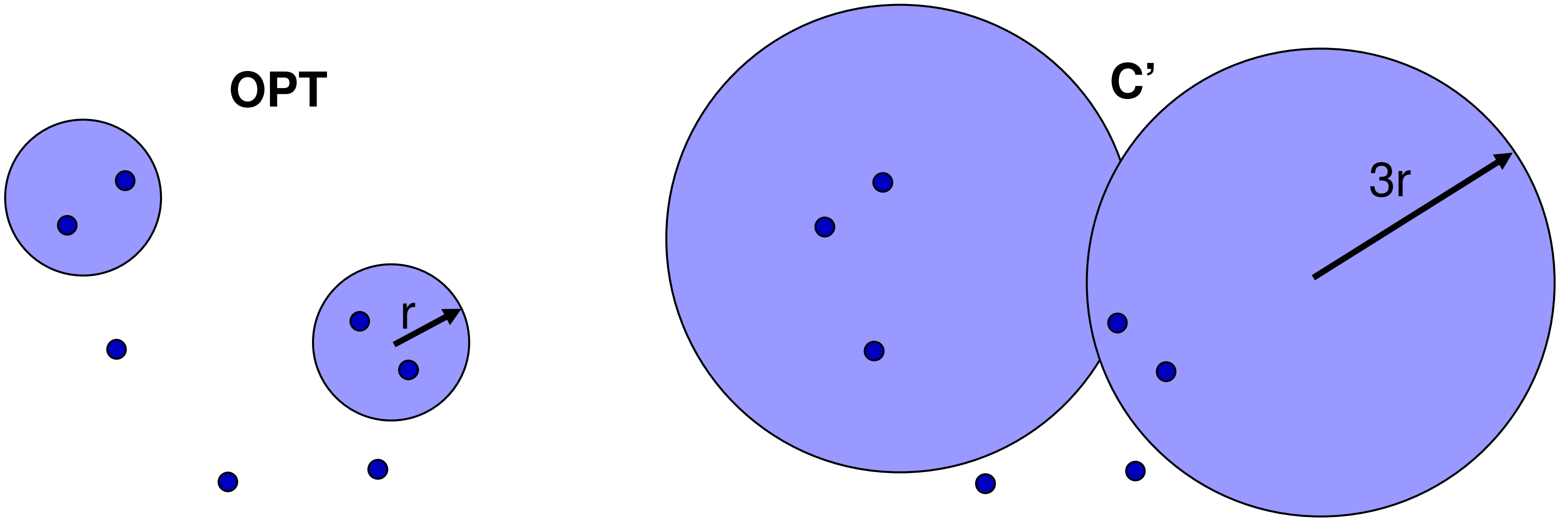
- k-center is more challenging, since cost function has 'min' inside the expectation
- Can be counterintuitive:
  - If all probabilities are close to 1, it behaves like traditional k-center
  - If all probabilities are very small, it behaves like k-median
  - An α-approximate clustering for half the points and an α-approx for the other half does not yield an α-approx for all
- Discuss only the point probability case here
  - Unassigned PDF case can be reduced to point probability up to an $(e/(e-1)) = 1.582$ factor in cost

Constant Factor Approximation

- Use a result of Charikar et al. [SODA 2001] in the deterministic case to show for the probabilistic data:
  - Given radius r, can find a clustering C’ so that

$$\Pr[\text{ cost}(C') \geq 3r\ ] < \Pr[\text{ cost}(OPT) \geq r]$$

  - Bounds the tail of the distribution of the cost function



Choosing a Radius

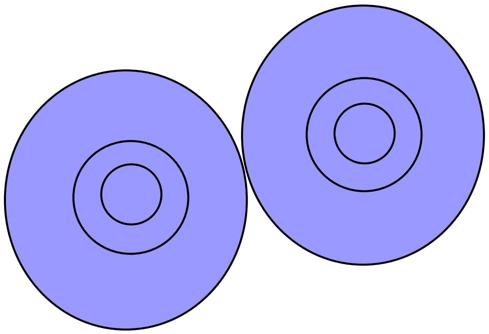
- Let $r_0 \leq r_1 \leq \ldots \leq r_t$ be the $O(n^2)$ distances in the input
- For each r_j find clustering C_j satisfying previous claim
- Pick the largest radius r_l satisfying

$$\Pr[\, \mathrm{cost}(C_l) \geq r_l] < ½$$

- Split the input into "near points" with $d(x,C_l) \leq r_l$ and "far points" with $d(x,C_l) > r_l$
  – In point probability case, each input point has only one possible location

Clustering the Near Points

- Use property of the clustering C_l to show optimal cost of clustering on the near points is at least $1/6 \text{ cost}(C_l)$
  - Write cost in terms of each "shell" of $(r_j - r_{j-1})$
  - Cost of optimal on each shell is at least $1/3$ that of C_j for that shell by construction of C_j
  - By choice of C_l and defn. of 'near', replacing C_j with C_l for each shell only affects shell cost by factor 2
  - This shows $\text{cost}(C_l)$ on the near points is a 6-approximation



$$\text{Cost} = \sum_j \Pr[\text{cost}(C) > r_j](r_j - r_{j-1})$$

"Discrete integration"

Clustering the Far Points

- The probability of seeing a point that fall more than C_l is chosen to be "small" ($\leq 1/2$), so the probability of these points must each be small
  - In particular for the far points, $\prod (1-p_i) \geq ½$
  - k-center cost can be written in terms of probability that no further points are present, as $\sum_i p_i\, d(x_i, C) \prod_{j<i} (1-p_j)$
  - So cost is at least $½ \sum_i p_i\, d(x_i, C)$ — the k-median cost
- Let C^* be a $(3+\varepsilon)$ approximation to the optimal k-*median* of the far points
- C^* is a $(6+\varepsilon)$ approximation to the optimal k-center for the far points.

Combining Clusterings

- Combine C^* and C_l to get $2k$ centers
- Cost of all points and $(C^* \cup C_l)$ is at most $cost(C^*)$ on far points and $cost(C_l)$ on the near points
- Optimal cost of a subset of points is at least as big as optimal on whole set
- Thus $C^* \cup C_l$ is at worst $(6 + 6 + \varepsilon) = 12+\varepsilon$ approximation to the best k-center clustering

$1+\varepsilon$ Factor Clustering

- We can get a much better clustering, at the expense of many more cluster centers
- Define a weight for each probability as $w_i = -\ln(1-p_i)$
- Reduce to a covering problem
  - Given radius r, define F as points further than r from C
  - $\Pr[\text{ cost} > r\ ] = 1 - \prod_{i \in F} (1-p_i) = 1 - \exp(-\sum_{i \in F} w_i)$
- Can cover at least as much "weight" as optimal algorithm by greedily picking points as centers to cover most weight
  - Picking $k \ln(w/w_{min}) = O(k \ln n)$ points cover as much as opt
  - Proof by weighted version of greedy set cover

$1+\varepsilon$ Factor Clustering

- Round all distances between points to powers of $(1+\varepsilon)$
- Find a covering for each $r \in \{1, 1+\varepsilon, (1+\varepsilon)^2 \ldots\}$
- Take the union of all centers found
- We have only given up a factor of $(1+\varepsilon)$ in the objective

- Result: We find $O(k/\varepsilon \log n \log \Delta)$ centers which $(1+\varepsilon)$ approximates the optimal k-center cost
  - Δ is ratio between closest and furthest point

Conclusions

- Can give guaranteed approximation algorithms for clustering uncertain data
  - Natural questions: can we improve approximations?
  - Assigned k-center still to be understood
- Other mining / optimization problems on uncertain data have not been much studied
  - Facility location and other generalizations of clustering
  - Other mining tasks: association rules, classification
  - Summarization – e.g. wavelets and histograms