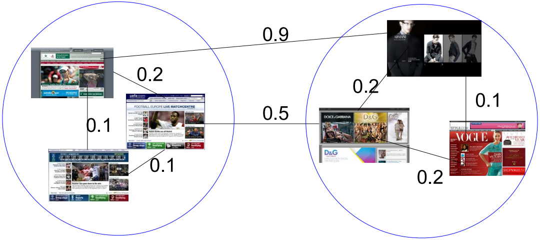
Modern Aspects of Unsupervised Learning

Yingyu Liang
Advisor: Maria Florina Balcan
Georgia Institute of Technology

August 9, 2013

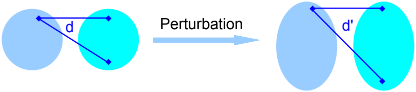
Clustering

- A set of n objects, pairwise dissimilarities/similarities
- A target clustering/cluster that has specific properties
- Goal: efficient algorithm that finds the target

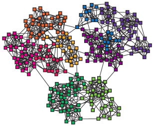


Outline

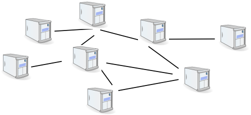
1. Perturbation Resilience: Beyond Worst Case



2. Community Hierarchies: Beyond Partitions



3. Distributed Clustering: Beyond Centralized

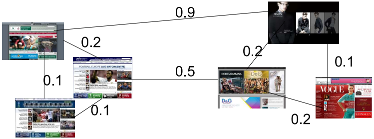


Outline

1. Clustering under Perturbation Resilience
   - α-Perturbation Resilience for k-median
   - (α, ϵ)-Perturbation Resilience for k-median
   - α-Perturbation Resilience for Min-Sum
2. Modeling and Detecting Community Hierarchies
3. Distributed Clustering

Objective-Based Clustering

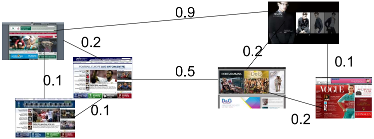
- A set S of n points, a distance function d



- Pick some objective to optimize
  - k-median: find centers $\{c_1, \ldots, c_k\} \subset S$ to minimize $\sum_i \sum_{p \in C_i} d(p, c_i)$
  - Min-sum: find partition $\{C_1, \ldots, C_k\}$ to minimize $\sum_i \sum_{p,q \in C_i} d(p, q)$
- NP-hard to optimize

Objective-Based Clustering

- A set S of n points, a distance function d



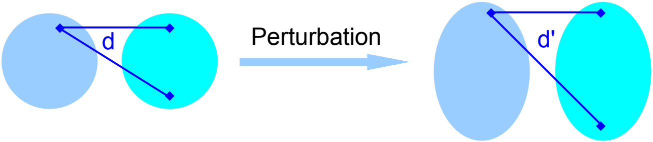
- Pick some objective to optimize
  - k-median: find centers $\{c_1, \ldots, c_k\} \subset S$ to minimize $\sum_i \sum_{p \in C_i} d(p, c_i)$
  - Min-sum: find partition $\{C_1, \ldots, C_k\}$ to minimize $\sum_i \sum_{p,q \in C_i} d(p, q)$
- NP-hard to optimize

α-Perturbation Resilience

Cool new direction: exploit additional stable property of the data

α-PR [Bilu and Linial, 2010; Awasthi, Blum and Sheffet, 2012]

A clustering instance (S, d) is α-perturbation resilient to a given objective function Φ if for any function $d' : S \times S \rightarrow R_{\geq 0}$ s.t. $\forall p, q \in S, d(p,q) \leq d'(p,q) \leq \alpha d(p,q)$, there is a unique optimal clustering $\mathcal{OPT}'$ for Φ under d' and this clustering is equal to the optimal clustering $\mathcal{OPT}$ for Φ under d.



Our Contribution [Balcan and Liang, ICALP 2012]

- Polynomial time algorithm for finding $\mathcal{OPT}$ for α-PR k-median instances when $\alpha \geq 1 + \sqrt{2}$
  - It works for any center-based objective function, e.g. k-means

- Polynomial time algorithm for a generalization (α, ϵ)-PR

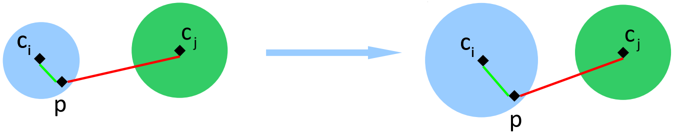
- Polynomial time algorithm for finding $\mathcal{OPT}$ for α-PR min-sum instances when $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$

Structure Properties of α-PR k-Median Instance

Claim

α-PR for k-median implies that $\forall p \in C_i, \alpha d(p, c_i) < d(p, c_j)$.

- Blow up all pairwise distances within the optimal cluster by α
- The $\mathcal{OPT}$ does not change, so $\forall p \in C_i, d'(p, c_i) < d'(p, c_j)$
- $d'(p, c_i) = \alpha d(p, c_i) < d'(p, c_j) = d(p, c_j)$



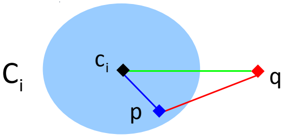
Structure Properties of α-PR k-Median Instance

Claim

α-PR for k-median implies that $\forall p \in C_i, \alpha d(p, c_i) < d(p, c_j)$.

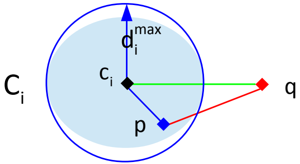
Implication:

- if $\alpha \geq 1 + \sqrt{2}, \forall p \in C_i, q \notin C_i$,
  (1) $d(c_i, p) < d(c_i, q)$ and (2) $d(c_i, p) < d(p, q)$



Structure Properties of α-PR k-Median Instance

- Let $d_i^{max} = \max_{p \in C_i} d(p, c_i)$. Construct a ball $B(c_i, d_i^{max})$
  - The ball covers exactly C_i
  - Points inside are closer to the center than to points outside, i.e. $\forall p \in B(c_i, d_i^{max}), q \notin B(c_i, d_i^{max}), d(p, c_i) < d(p, q)$

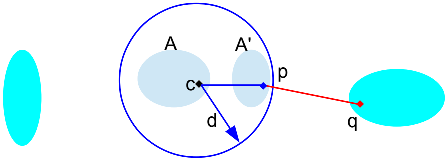


Closure Distance

Closure Distance

The closure distance $d_S(A, A')$ between two subsets A and A' is the minimum d, such that there is a point $c \in A \cup A'$ satisfying:

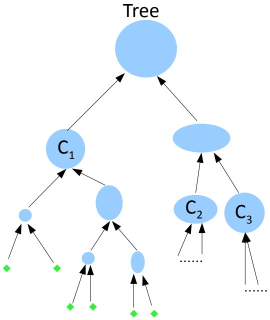
- **coverage condition**: the ball $B(c, d)$ covers $A \cup A'$;
- **margin condition**: points inside are closer to the center than to points outside, i.e. $\forall p \in B(c, d), q \notin B(c, d), d(c, p) < d(p, q)$.



Algorithm for α-PR k-median

Closure Linkage

- Begin with each point being a subset
- Repeat until one cluster remains:
  merge the two subsets with
  minimum closure distance
- Output the tree with points as leaves and merges as internal nodes



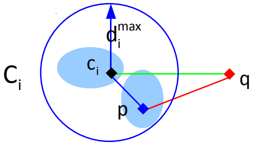
Theorem

If $\alpha \geq 1 + \sqrt{2}$, the tree output contains $\mathcal{OPT}$ as a pruning.

Proof

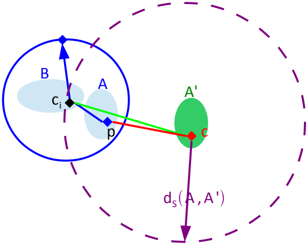
By induction, we show that the algorithm will not merge a strict subset $A \subset C_i$ with a subset A' outside C_i.

- Pick $B \subset C_i \setminus A$ such that $c_i \in A \cup B$
- $d_S(A, B) \leq d_i^{max} = \max_{p \in C_i} d(p, c_i)$
    - d_i^{max} and $c_i \in A \cup B$ satisfy the two conditions of closure distance



Proof

- $d_S(A, A') > d_i^{max}$
  - Suppose the center c for the ball defining $d_S(A, A')$ is from A'
  - Since $c \notin C_i$, $d(c_i, p) < d(p, c)$ for arbitrary $p \in A$.
    By margin condition,
    $c_i \in B(c, d_S(A, A')), i.e.\ d_S(A, A') \geq d(c_i, c)$
  - Since $c \notin C_i$, $d(c_i, c) > d_i^{max}$



  - A similar argument holds for the case $c \in A$

(α, ϵ)-Perturbation Resilience

- α-PR imposes a strong restriction that the $\mathcal{OPT}$ does not change after perturbation
- We propose a more realistic relaxation

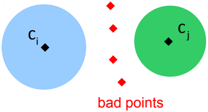
(α, ϵ)-Perturbation Resilience

A clustering instance (S, d) is (α, ϵ)-perturbation resilient to a given objective function Φ if for any function $d' : S \times S \to R_{\geq 0}$ s.t. $\forall p, q \in S, d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$, the optimal clustering $\mathcal{OPT}'$ for Φ under d' is ϵ-close to the optimal clustering $\mathcal{OPT}$ for Φ under d.

Structure Property of (α, ϵ)-PR k-median

Theorem

Assume $\min_i |C_i| = \Omega(\epsilon n)$. Except for $\leq \epsilon n$ bad points, any other point is α times closer to its own center than to other centers.



Keypoint of the Proof

- Carefully construct a perturbation that forces all the bad points move
- By (α, ϵ)-PR, there could be at most ϵn bad points

Algorithm for (α, ϵ)-PR k-median

A robust version of Closure Linkage algorithm can be used to show:

Theorem

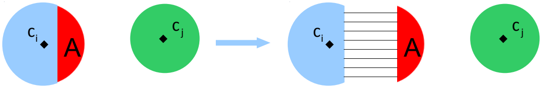
Assume $\min_i |C_i| = \Omega(\epsilon n)$. If $\alpha \geq 2 + \sqrt{7}$, then the tree output contains a pruning that is ϵ-close to the optimal clustering. Moreover, the cost of this pruning is $(1 + O(\epsilon/\rho))$-approximation where $\rho = \min_i |C_i|/n$.

Structure Property of α-PR Min-Sum

Claim

α-PR implies $\forall A \subseteq C_i, \alpha d(A, C_i \setminus A) < d(A, C_j)$.

Proof: blow up the distances between A and $C_i \setminus A$ by α



Structure Property of α-PR Min-Sum

Claim

α-PR implies $\forall A \subseteq C_i, \alpha d(A, C_i \setminus A) < d(A, C_j)$.

Implications when $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$:

1. For any point, its $\min_i |C_i|/2$ nearest neighbors are from the same optimal cluster
2. For sufficiently large subsets $A_i \subseteq C_i, A_j \subseteq C_j$, $d_{avg}(A_i, A_j) > \min\{d_{avg}(C_i \setminus A_i, A_i), d_{avg}(A_j, C_j \setminus A_j)\}$

Algorithm for α-PR Min-Sum

Algorithm for α-PR Min-Sum

- Connect each point with its $\min_i |C_i|/2$ nearest neighbors
- Perform average linkage on the components to get a tree

Theorem

If $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$, then the tree contains $\mathcal{OPT}$ as a pruning.

Future Work

1. Design algorithm for (α, ϵ)-PR min-sum

	α-PR	(α, ϵ)-PR
k-median	✓	✓
min-sum	✓	?

Current result:

- Structural property: $\tilde{O}(\epsilon n)$ bad points
- Constructed a tree with pruning close to the optimal
- Next Step: find this pruning

Future Directions

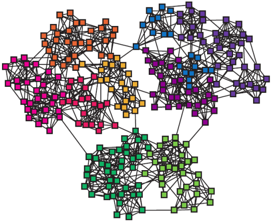
2. Combining α-PR with other stability properties
   - (α, ϵ)-approximation-stability [Balcan, Blum and Gupta, 2009]
   - center separation [Awasthi and Sheffet, 2012]

Outline

1. Clustering under Perturbation Resilience
2. Modeling and Detecting Community Hierarchies
   - Model Definition
   - Detection Algorithm
3. Distributed Clustering

Community Detection

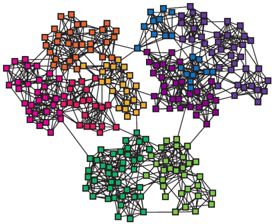
- n points, a similarity function
- Communities: meaningful groups such that connections are tighter within than with the outside



A hierarchical network [Clauset, Moore and Newman, 2008]

Community Detection

- No established consensus on definition
- Theoretical models aiming to capture common intuitions
  - Tighter connections within than with the outside world
  - Hierarchical organization



A hierarchical network [Clauset, Moore and Newman, 2008]

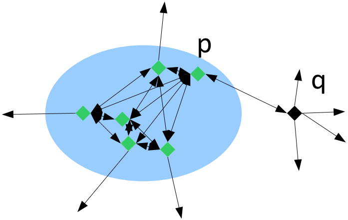
Our Contribution [Balcan and Liang, SIMBAD 2013]

- Theoretical model for community hierarchy
- Efficient algorithm with provable guarantee

Model Definition
Compact Blob

C is a compact blob if out of $|C|$ nearest neighbors,

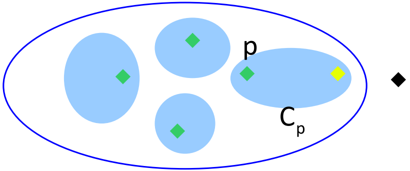
- $[\text{internal}]$ any $p \in C$ has $\leq \alpha n$ neighbors outside C
- $[\text{external}]$ any $q \notin C$ has $\leq \alpha n$ neighbors inside C



Model Definition
Stable Community

C is a stable community if

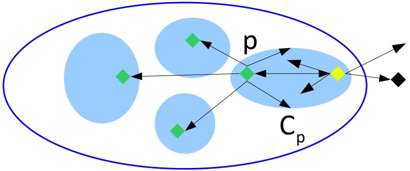
- [local] any point $p \in C$ falls into a compact blob $C_p \subseteq C$
- [between blobs] a majority of points in the blob C_p have $\leq \alpha n$ neighbors outside C out of the $|C|$ nearest neighbors
- [external] any point $q \notin C$ has $\leq \alpha n$ neighbors inside C out of the $|C|$ nearest neighbors



Model Definition
Stable Community

C is a stable community if

- [local] any point $p \in C$ falls into a compact blob $C_p \subseteq C$
- [between blobs] a majority of points in the blob C_p have $\leq \alpha n$ neighbors outside C out of the $|C|$ nearest neighbors
- [external] any point $q \notin C$ has $\leq \alpha n$ neighbors inside C out of the $|C|$ nearest neighbors

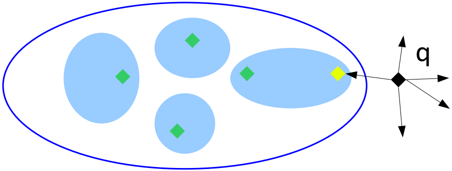


Model Definition

Stable Community

C is a stable community if

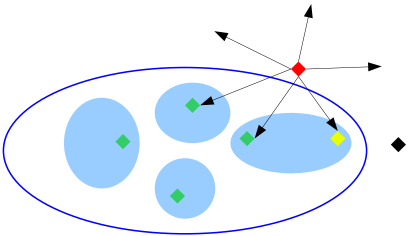
- [local] any point $p \in C$ falls into a compact blob $C_p \subseteq C$
- [between blobs] a majority of points in the blob C_p have $\leq \alpha n$ neighbors outside C out of the $|C|$ nearest neighbors
- [external] any point $q \notin C$ has $\leq \alpha n$ neighbors inside C out of the $|C|$ nearest neighbors



Model Definition
Stable Community

C is a stable community if after removing $\leq \nu n$ bad points,

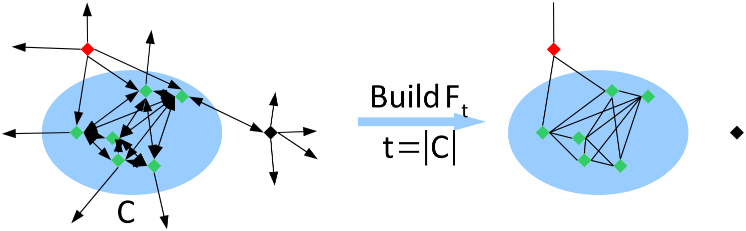
- [local] any point $p \in C$ falls into a compact blob $C_p \subseteq C$
- [between blobs] a majority of points in the blob C_p have $\leq \alpha n$ neighbors outside C out of the $|C|$ nearest neighbors
- [external] any point $q \notin C$ has $\leq \alpha n$ neighbors inside C out of the $|C|$ nearest neighbors



Detection Algorithm
Easy case: Blob With Known Size

Consider a compact blob C with $|C|$ known.

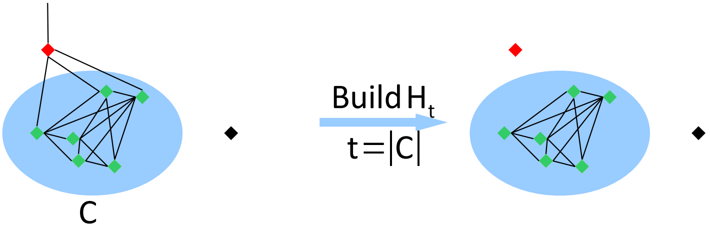
1. Build F_t by connecting points that share many neighbors out of the $t = |C|$ nearest neighbors
   - Good points in C and those outside C are disconnected
   - Good points in C are all connected



Detection Algorithm
Easy Case: Blob with Known Size

Consider a compact blob C with $|C|$ known.

2. Build H_t: connect points with many common neighbors in F_t
   - Bad point "bridges" are disconnected
3. Merge components in H_t;
   - One of the components represents C



Detection Algorithm
Easy Case: Blob with Unknown Size

Consider a compact blob C with $|C|$ unknown.

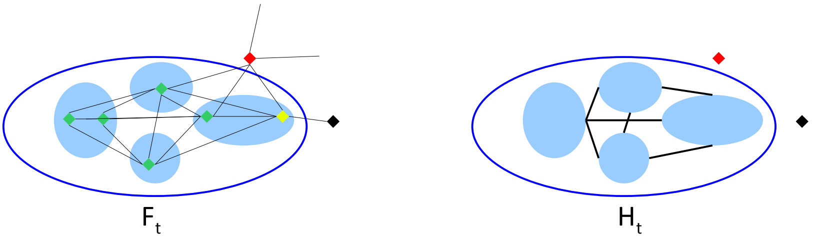
Vary the threshold t:

- Begin with a small t
- Increase t and build F_t, H_t
- When $t = |C|$, a component in H_t represents C

Consider a stable community C

Build H_t on sets of points instead of on points

- Maintain a list $\mathcal{L}$ of communities
- Build H_t on $\mathcal{L}$ to disconnect bad point "bridges" between sub-communities in C and those outside C
- Merge connected components in H_t to form C



Hierarchical Community Detection Algorithm

1. Initialize $\mathcal{L}$ to be a list of singleton points
   Initialize the threshold t to be the size of the minimum blob
2. Repeat until all points merged:
   build F_t, H_t;
   update $\mathcal{L}$ by merging large components in H_t;
   increase t
3. Output a tree with internal nodes corresponding to the merges

Theorem

Any stable community is ν-close to a node in the tree.

Future Directions

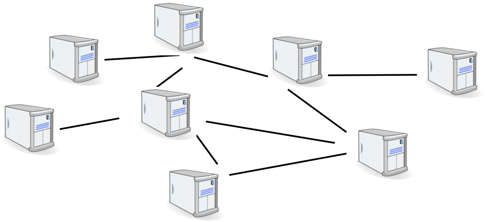
1. Local algorithm for our model
   - Speed up for Internet scale

2. Community hierarchies more general than trees
   - Weak hierarchies

Outline

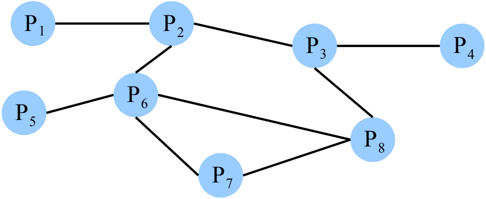
Distributed Data

- Distributed databases
- Images and videos on the Internet
- Sensor networks
- ...



Distributed Clustering

- Communication graph G on n nodes:
  an edge indicates that the two nodes can communicate
- Global data $P \subseteq \mathbf{R}^d$ is divided into local data sets $P_1, \ldots, P_n$



Goal: efficient distributed algorithm for k-median/k-means
with low communication cost

Coreset

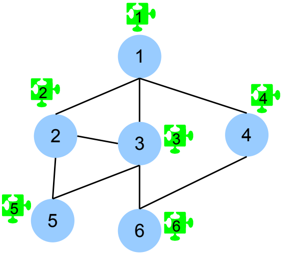
Coreset [Feldman and Langberg, 2011]

An ϵ-coreset for a set of points P with respect to a cost objective function is a set of points S and a set of weights $w\colon S \to \mathbf{R}$ such that for any set of centers $\mathbf{x}$,

$$(1-\epsilon)\mathrm{cost}(P,\mathbf{x}) \le \sum_{p\in S} w(p)\mathrm{cost}(p,\mathbf{x}) \le (1+\epsilon)\mathrm{cost}(P,\mathbf{x}).$$

Our Contribution [Balcan, Ehrlich and Liang, CoRR 2013]

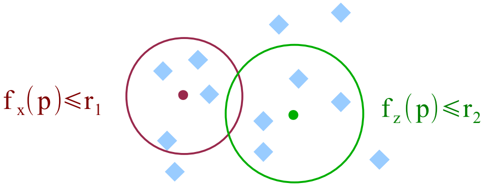
- Distributed coreset construction algo with low communication



Coreset Construction

Function Space Dimension [Feldman and Langberg, 2011]

Let $F = \{f\}$ be a set of functions from P to $\mathbf{R}_{\geq 0}$.
For any $G \subseteq P$, each pair $f \in F, r \in \mathbf{R}_{\geq 0}$ introduces a subset $\{p \in G : f(p) \leq r\}$.
$\dim(F)$ is the smallest integer t such that for any $G \subseteq P$, there are at most $|G|^t$ subsets introduced by $f \in F, r \in \mathbf{R}_{\geq 0}$.

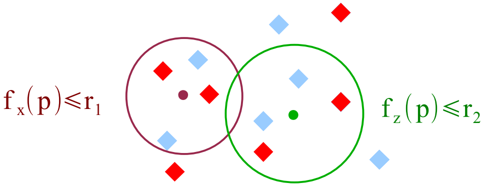


$f_x(p) = d(p, x)$

Coreset Construction

Function Space Dimension [Feldman and Langberg, 2011]

Let $F = \{f\}$ be a set of functions from P to $\mathbf{R}_{\geq 0}$.
For any $G \subseteq P$, each pair $f \in F, r \in \mathbf{R}_{\geq 0}$ introduces a subset $\{p \in G : f(p) \leq r\}$.
$\dim(F)$ is the smallest integer t such that for any $G \subseteq P$, there are at most $|G|^t$ subsets introduced by $f \in F, r \in \mathbf{R}_{\geq 0}$.



$f_x(p) = d(p, x)$

Coreset Construction

Sampling Lemma

Let $m_p = \max_{f \in F} f(p)$. Sample S from P with probability proportional to m_p, and let $w_p = \frac{\sum_q m_q}{m_p |S|}$.
If $|S| = \tilde{O}(\dim(F)/\epsilon^2)$, then w.h.p.

$$\forall f \in F, \left| \sum_{p \in P} f(p) - \sum_{p \in S} w_p f(p) \right| \leq \epsilon \sum_{p \in P} m_p.$$

Coreset Construction

Sampling Lemma

Let $m_p = \max_{f \in F} f(p)$. Sample S from P with probability proportional to m_p, and let $w_p = \frac{\sum_q m_q}{m_p |S|}$.
If $|S| = \tilde{O}(\dim(F)/\epsilon^2)$, then w.h.p.

$$\forall f \in F, \left| \sum_{p \in P} f(p) - \sum_{p \in S} w_p f(p) \right| \leq \epsilon \sum_{p \in P} m_p.$$

First attempt: $f_{\mathbf{x}}(p) = \mathrm{cost}(p, \mathbf{x})$, $\mathbf{x}$ is a set of centers
Problem: $m_p = \max f_{\mathbf{x}}(p)$ unbounded

Coreset Construction

Sampling Lemma

Let $m_p = \max_{f \in F} f(p)$. Sample S from P with probability proportional to m_p, and let $w_p = \frac{\sum_q m_q}{m_p |S|}$.
If $|S| = \tilde{O}(\dim(F)/\epsilon^2)$, then w.h.p.

$$\forall f \in F, \left| \sum_{p \in P} f(p) - \sum_{p \in S} w_p f(p) \right| \leq \epsilon \sum_{p \in P} m_p.$$

Idea: Choose a set of centers B_i for P_i.
For $p \in P_i$, let b_p denote its nearest center in B_i.
Set $f_{\mathbf{x}}(p) = \mathrm{cost}(p, \mathbf{x}) - \mathrm{cost}(b_p, \mathbf{x})$, then $|f_{\mathbf{x}}(p)| \leq \mathrm{cost}(b_p, p)$.

Coreset Construction

Communication aware distributed coreset construction

1. Compute a constant approximation solution B_i for P_i; Broadcast the costs of the local solutions.
2. Sample points S_i from P_i according to $\mathrm{cost}(p, b_p)$. Weight sampled points: $w_p = \frac{\sum_{p \in P} \mathrm{cost}(p, b_p)}{|S| \mathrm{cost}(p, b_p)}$
3. Weight each center in the local solutions: for $b \in B_i$, let P_b be points of P_i in its Voronoi region set $w_b = |P_b| - \sum_{p \in P_b \cap S} w_p$

Coreset Construction

Communication aware distributed coreset construction

1. Compute a constant approximation solution B_i for P_i; Broadcast the costs of the local solutions.
2. Sample points S_i from P_i according to $\mathrm{cost}(p, b_p)$. Weight sampled points: $w_p = \frac{\sum_{p \in P} \mathrm{cost}(p, b_p)}{|S| \mathrm{cost}(p, b_p)}$
3. Weight each center in the local solutions: for $b \in B_i$, let P_b be points of P_i in its Voronoi region set $w_b = |P_b| - \sum_{p \in P_b \cap S} w_p$

One additional detail: $\sum_{p \in P} f_p(\mathbf{x}) \neq \sum_{p \in P} \mathrm{cost}(p, \mathbf{x})$
The difference can be compensated by cost on the centers $\{b_p\}$

Results

Theorem (Distributed Coreset Construction)

With probability $\geq 1 - \delta$, our algorithm outputs an ϵ-coreset of size $O(\frac{1}{\epsilon^4}(kd + \log \frac{1}{\delta}) + nk \log \frac{nk}{\delta})$ for k-means, and of size $O(\frac{1}{\epsilon^2}(kd + \log \frac{1}{\delta}) + nk)$ for k-median.

Theorem (Distributed Clustering)

Given any α-approximation algorithm as a subroutine, we can compute a $(1 + \epsilon)\alpha$-approximation solution for distributed k-means/k-median. The total communication cost is $O(m)$ times the coreset size.

Future Directions

1. Better bounds for the size of coreset
   - Better dependence on the accuracy ϵ

2. Distributed minimum enclosing ball (MEB)
   - equivalent to L2-SVM [Tsang, Kwok and Cheung, 2005]
   - MEB has ϵ-coreset of size $O(1/\epsilon)$ [Bădoiu and Clarkson, 2008]

Some other work

Efficient Semi-supervised and Active Learning of Disjunctions, with Maria Florina Balcan, Steven Ehrlich, Christopher Berlind, In *ICML*, 2013.

- Efficient semi-supervised/active learning algorithms
- Extension to random classification noise

Thanks!

Q&A

Awasthi, P. and Sheffet, O. (2012).
Improved spectral-norm bounds for clustering.
In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*.

Bădoiu, M. and Clarkson, K. L. (2008).
Optimal core-sets for balls.
Computational Geometry.

Bandelt, H.-J. and Dress, A. (1989).
Weak hierarchies associated with similarity measuresan additive clustering technique.
Bulletin of mathematical biology.

Bilu, Y. and Linial, N. (2010).
Are stable instances easy?
In *Proceedings of the Innovations in Computer Science*.

Bryant, D. and Moulton, V. (2004).

Neighbor-net: an agglomerative method for the construction of phylogenetic networks.
Molecular biology and evolution.

Feldman, D. and Langberg, M. (2011).
A unified framework for approximating and clustering data.
In *Proceedings of the Annual ACM Symposium on Theory of Computing*.

Girvan, M. and Newman, M. E. J. (2002).
Community structure in social and biological networks.
Proceedings of the National Academy of Sciences.

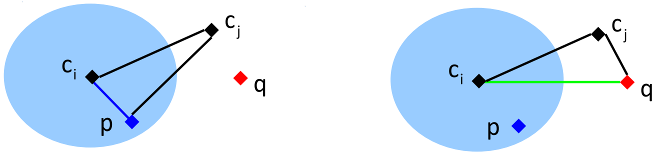
Lancichinetti, A. and Fortunato, S. (2009).
Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities.
Physical Review E.

Spielman, D. A. and Teng, S.-H. (2004).

Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems.
In *Proceedings of the Annual ACM Symposium on Theory of Computing.*

Structure Properties of α-PR k-Median Instance

- (1) If $\alpha \geq 1+\sqrt{2}, \forall p \in C_i, q \notin C_i$, $d(c_i, p) < d(c_i, q)$
  - $d(c_i, c_j) \geq d(p, c_j) - d(p, c_i) > (\alpha - 1) d(p, c_i)$
  - $d(c_i, c_j) \leq d(q, c_i) + d(q, c_j) < (1 + \frac{1}{\alpha}) d(q, c_i)$



- (2) A similar argument shows $d(c_i, p) < d(p, q)$

Proof of Property of (α, ϵ)-PR

Perturbation

- For technical reasons, for each i select $\min(|B_i|, \epsilon n + 1)$ bad points from B_i
- Blow up all pairwise distances by α, except
  - between the bad points and their second nearest centers
  - between the other points and their own centers
- Intuition: ideally, after the perturbation,
  all bad points are assigned to their second nearest center,
  all the other points stay

Proof of Property of (α, ϵ)-PR

Centers after Perturbation Let c_i' be the new center for the new i-th cluster C_i'.
Sufficient to show: $c_i' \neq c_i$ leads to a contradiction.

- C_i' differs from C_i on at most ϵn points
- c_i' is close to c_i
- $d(c_i', C_i' \cap C_i) \approx d(c_i, C_i' \cap C_i)$
- $$\begin{aligned} d'(c_i', C_i' \cap C_i) &= \alpha d(c_i', C_i' \cap C_i) \\ &\gg d'(c_i, C_i' \cap C_i) = d(c_i, C_i' \cap C_i) \end{aligned}$$
- $d'(c_i', C_i') > d'(c_i, C_i')$, a contradiction

Proof for Community Detection

Base Case: Compact Blob

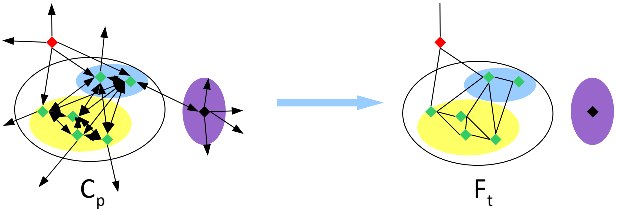
Lemma 1

For any good point p,

- when $t \le |C_p|$, good points from C_p will not be merged with good points outside C_p.

Properties of F_t:

- No good point inside C_p is connected to good points outside
- No bad point is connected to both a good point inside and a good point outside



Proof for Community Detection
Base Case: Compact Blob

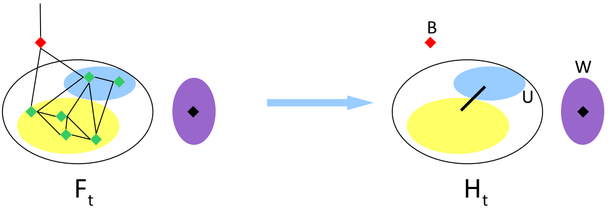
Lemma 1

For any good point p,

- when $t \leq |C_p|$, good points from C_p will not be merged with good points outside C_p.

Properties of H_t:

- U: community in $\mathcal{L}$ containing good points inside
- W: community in $\mathcal{L}$ containing good points outside
- B: community in $\mathcal{L}$ containing only bad points



Proof for Community Detection

Base Case: Compact Blob

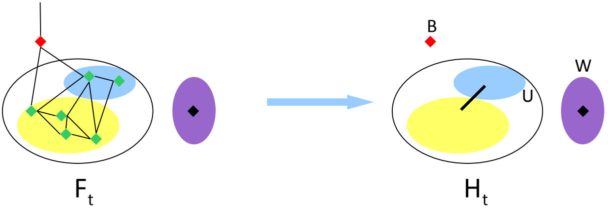
Lemma 1

For any good point p,

- when $t \leq |C_p|$, good points from C_p will not be merged with good points outside C_p.

Properties of H_t:

- U is not connected to W
- B cannot be connected to both U and W



Proof for Community Detection
Base Case: Compact Blob

Lemma 2

For any good point p,

- when $t = |C_p|$, all good points in C_p are merged into one community.

Properties of F_t, H_t:

- All good points in C_p are connected in F_t
- All communities containing good points in C_p are connected in H_t

Proof for Community Detection

General Case

Lemma 3

For any stable community C,

- when $t \leq |C|$, good points from C will not be merged with good points outside C.
- when $t = |C|$, all good points in C are merged into one community.

Proof Sketch:

- Lemma 1 and 2 show:
  compact blobs in C are formed
- Similar arguments as in Lemma 1 and 2 then show:
  these compact blobs are merged into one community

Experiment

Lift network adjacent matrix A to similarity function S

- direct lifting: $S = A$
- diffusion lifting: $S = \exp\{\lambda A\}, \lambda = 0.05$

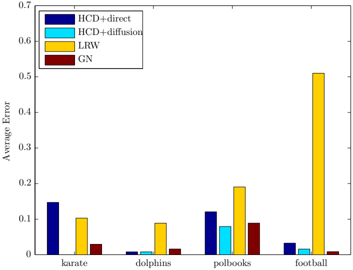
Evaluation criterion:

- Recover error of a true community C w.r.t. the tree $\mathcal{T}$

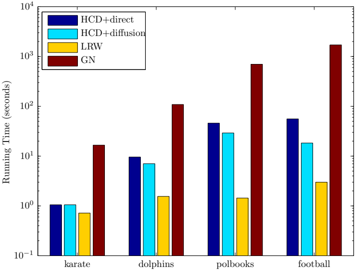
$$\mathrm{error}(C, \mathcal{T}) = \min_{C' \in \mathcal{T}} \frac{|C \oplus C'|}{n}$$

Experiment
Real World Data

Compare our algo (HCD) to:
Lazy Random Walk (LRW [Spielman and Teng, 2004]),
Girvan-Newman algo (GN [Girvan and Newman, 2002])



Average recover error



Running time (log scale)

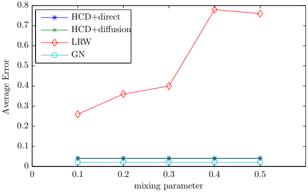
4 network type with two level community hierarchies ([Lancichinetti and Fortunato, 2009])

Data set	n	m	k	$maxk$
LF50	50	≈500	10	15
LF100	100	≈1500	15	20
LF150	150	≈3000	20	30
LF200	200	≈6000	30	40

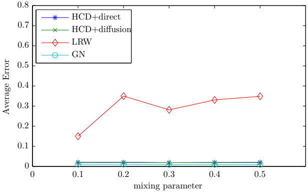
Table: The parameters of the synthetic data sets. n/m: number of nodes/edges; $k/maxk$: average/maximum degree of the nodes.

For each type, vary a mixing parameter to get 5 networks

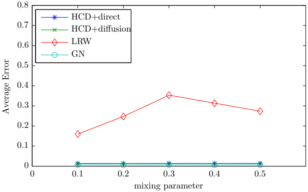
- mixing parameter: probability of connecting points inside a community to points outside
- larger parameter: more difficult to recover communities



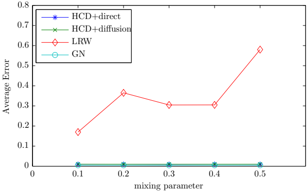
(a) LF50



(b) LF100



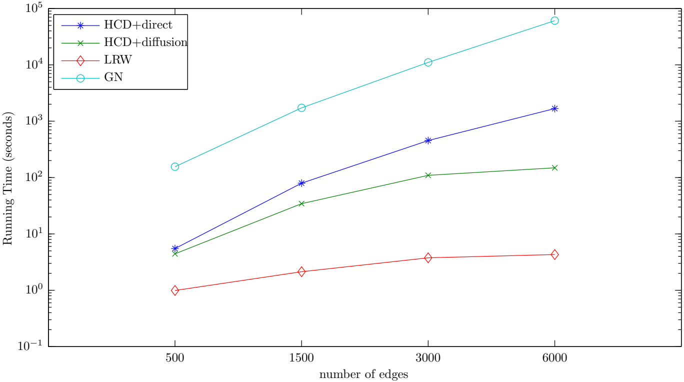
(c) LF150



(d) LF200

Average error v.s. mixing parameter

Experiment
Synthetic Data



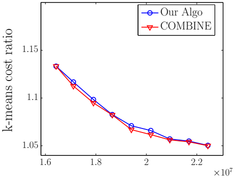
Running time (log scale) v.s. network size

Experiment for Distributed Clustering

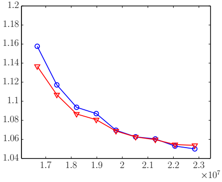
Setup

- Data set: YearPredictionMSD
- Partition into 100 local data sets:
  uniform, similarity-based, weighted, degree-based
- Communication graph: random, grid, preferential
- Evaluation criteria: k-means cost

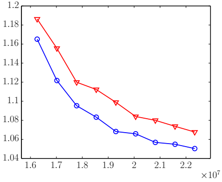
Experiment for Distributed Clustering
On Graphs



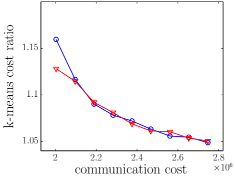
random graph, uniform



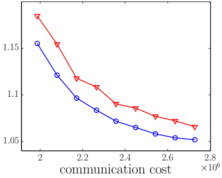
random graph, similarity-based



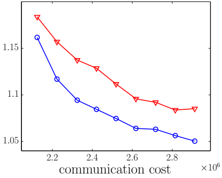
random graph, weighted



grid graph, similarity-based



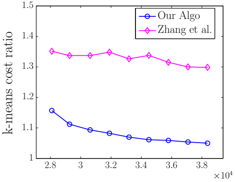
grid graph, weighted



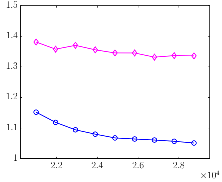
preferential graph, degree-based

Experiment for Distributed Clustering

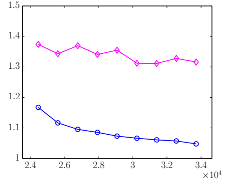
On Spanning Trees



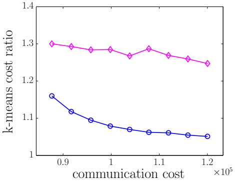
random graph, uniform



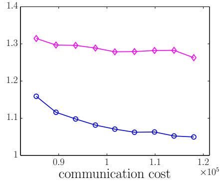
random graph, similarity-based



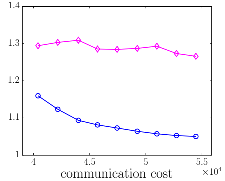
random graph, weighted



grid graph, similarity-based



grid graph, weighted



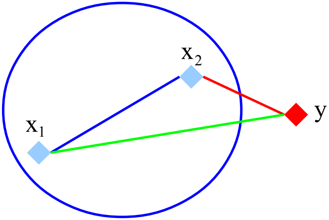
preferential graph, degree-based

Weak Clusters

Weak Clusters [Bandelt and Dress, 1989]

A set $C \subseteq S$ is called a weak cluster, if for any $x_1, x_2 \in C$, $y \notin C$,

$$d(x_1, x_2) < \max\{d(x_1, y), d(x_2, y)\}.$$



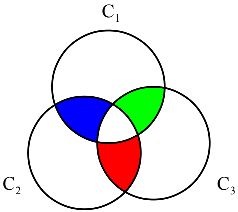
Our Contribution [Balcan and Liang, Manuscript 2013]

- More structural properties of weak clusters
- A new, faster algorithm for finding all weak clusters

Weak Hierarchies

Weak Hierarchies

A non-empty collection $\mathcal{H}$ of clusters is called a weak hierarchy, if $\forall C_1, C_2, C_3 \in \mathcal{H}, C_1 \cap C_2 \cap C_3 \in \{C_1 \cap C_2, C_2 \cap C_3, C_3 \cap C_1\}$.

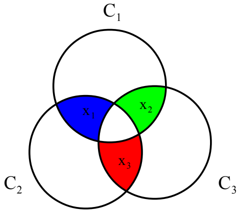


At least one of three colored areas must be empty

Weak Hierarchies

Weak Hierarchies

A non-empty collection $\mathcal{H}$ of clusters is called a weak hierarchy, if $\forall C_1, C_2, C_3 \in \mathcal{H}, C_1 \cap C_2 \cap C_3 \in \{C_1 \cap C_2, C_2 \cap C_3, C_3 \cap C_1\}$.



Lemma

A collection of weak clusters is a weak hierarchy.

Structure Properties

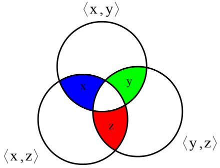
Closure

The closure $\langle A \rangle$ of a set A is the intersection of all members of $\mathcal{H}$ containing A, i.e. $\langle A \rangle = \cap_{A \subseteq C \in \mathcal{H}} C$.

Lemma (Trace)

If $\mathcal{H}$ is a weak hierarchy, then for every non-empty subset A, there exist $x, y \in A$ such that $A = \langle \{x, y\} \rangle$.

- Select $x, y = \operatorname{argmax} |\langle \{x, y\} \rangle|$
- Assume $z \in A \setminus \langle \{x, y\} \rangle$
- Then $x \notin \langle \{y, z\} \rangle, y \notin \langle \{x, z\} \rangle$



Algorithm

Maximal Expansion

Let the expansion of A be $f(A) = A \cup B$ where
$B = \{y \notin A | \exists x_1, x_2 \in A, d(x_1, x_2) > \max\{d(x_1, y), d(x_2, y)\}\}$.
Let $F(A) = f^{\infty}(A)$ be the maximal expansion of A.

Lemma (Maximal Expansion)

If (x, y) is the trace of a weak cluster C, then $F(\{x, y\}) = C$.

- $F(\{x, y\})$ is a weak cluster
- Any weak cluster containing x, y contains C: $F(\{x, y\}) \supseteq C$
- $f(\{x, y\}) \subseteq C, f^2(\{x, y\}) \subseteq C, \ldots, F(\{x, y\}) \subseteq C$