Characterization of Linkage-Based Clustering

Margareta Ackerman
Joint work with
Shai Ben-David and **David Loker**

University of Waterloo

COLT 2010

Motivation

There are a wide variety of clustering algorithms, which often produce very different clusterings.

How should a user decide which algorithm to use for a given application?

Our approach for clustering algorithm selection

- Identify properties that separate input-output behaviour of different clustering paradigms
- The properties should
  1) Be intuitive and meaningful to clustering users
  2) Distinguish between different clustering algorithms

Previous work

- Kleinberg proposes abstract properties ("Axioms") of clustering functions (NIPS, 2002)
- Bosagh Zadeh and Ben-David provide a set of properties that characterize *single linkage* clustering (UAI, 2009)

Our contributions

Characterize *linkage-based* clustering algorithms, using a set of intuitive properties

Outline

- Define linkage-based clustering
- Introduce new clustering properties
- Main result
- Sketch of proof
- Conclusions

Formal setup

For a finite domain set X, a *dissimilarity function* d over the members of X.

A Clustering Function F maps
Input: (X,d) and $k>0$
to
Output: a k-partition (clustering) of X

We require clustering functions to be representation independent and scale invariant.

Linkage-based algorithm: An informal definition

Proceed in steps:

- Start with the clustering of singletons
- At each step, merge the closest pair of clusters ?
- Repeat until only k clusters remain.

Ex. Single linkage, average linkage, complete linkage

Informally, a linkage function is
an extension of the between-point distance that applies to subsets of the domain.

- The choice of the linkage function distinguishes between different linkage-based algorithms.

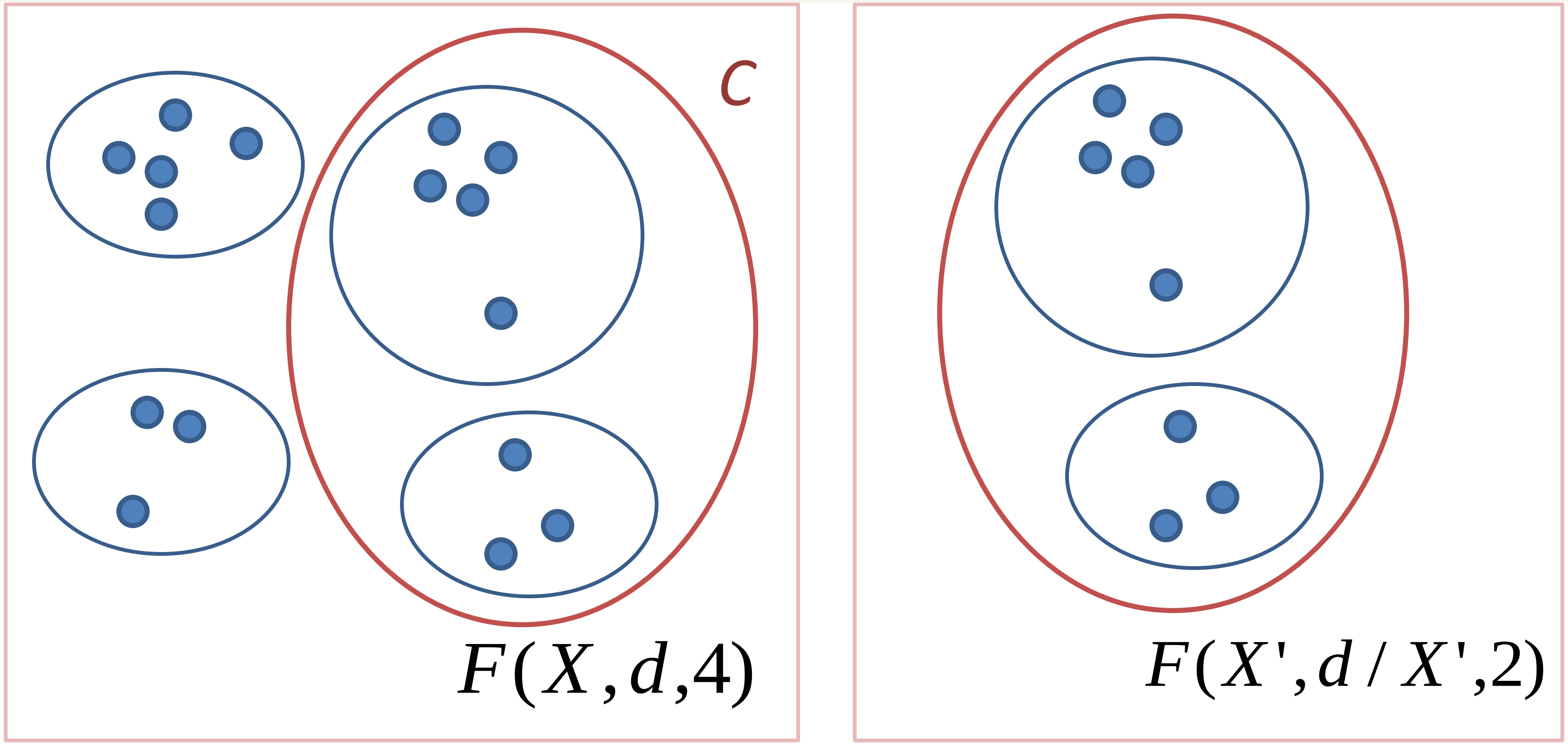
Outline

- Define linkage-based clustering
- Introduce new clustering properties
- Main result
- Sketch of proof
- Conclusions

Hierarchical clustering

- A clustering C is a *refinement* of clustering C' if every cluster in C' is a union of some clusters in C.
- A clustering function is *hierarchical* if for $\forall X \forall d$ and every $1 \le k \le k' \le |X|$ $F(X,d,k')$ is a refinement of $F(X,d,k)$.

Locality

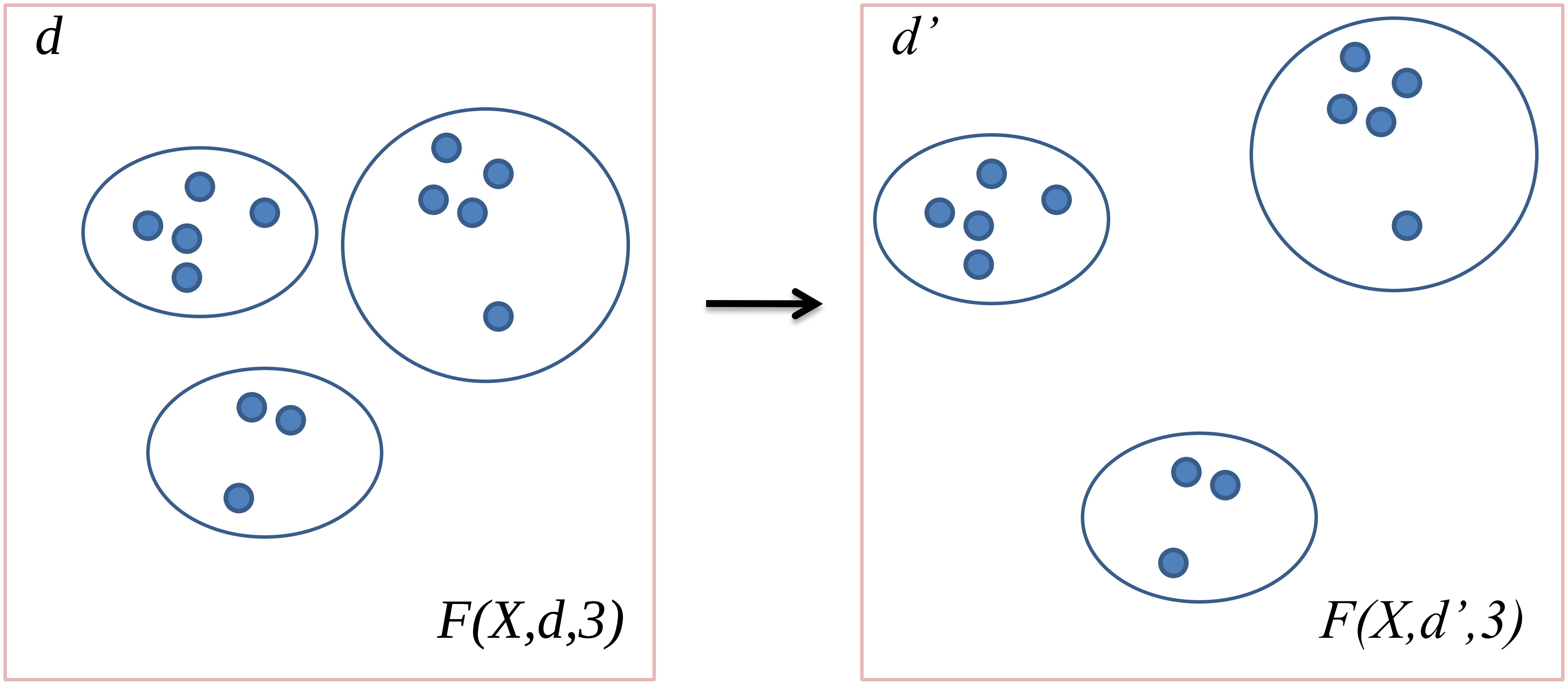


F is *local* if for any X, d, k and any $C \subseteq F(X,d,k)$,

$$C = F\left(\bigcup_{c \in C} c, d, |C|\right)$$

M. Ackerman, S. Ben-David, and D. Loker

Outer Consistency
Based on Kleinberg, 2002.

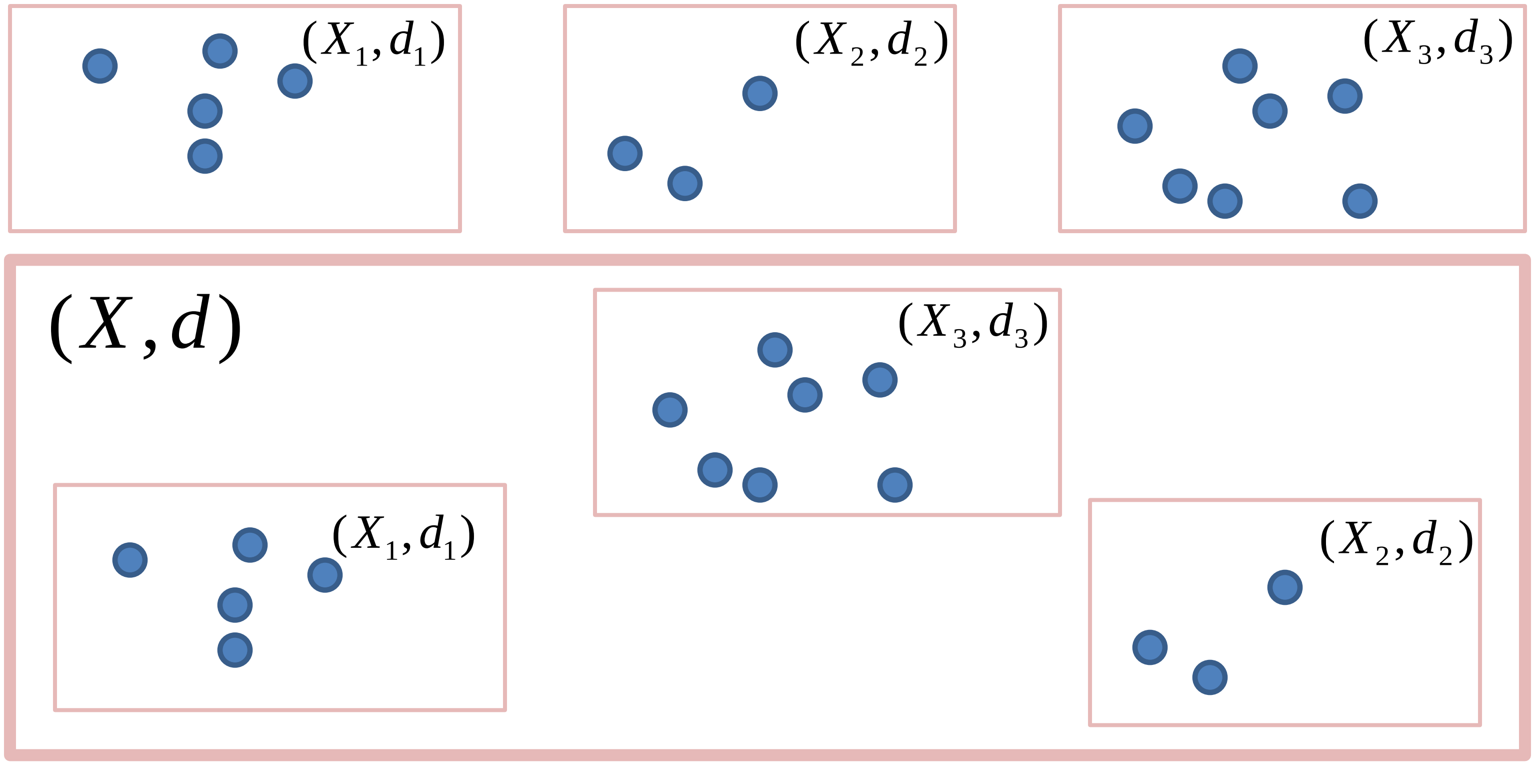


If d' equals d, except for increasing between-cluster distances, then $F(X,d,k)=F(X,d',k)$ for all d, X, and k.

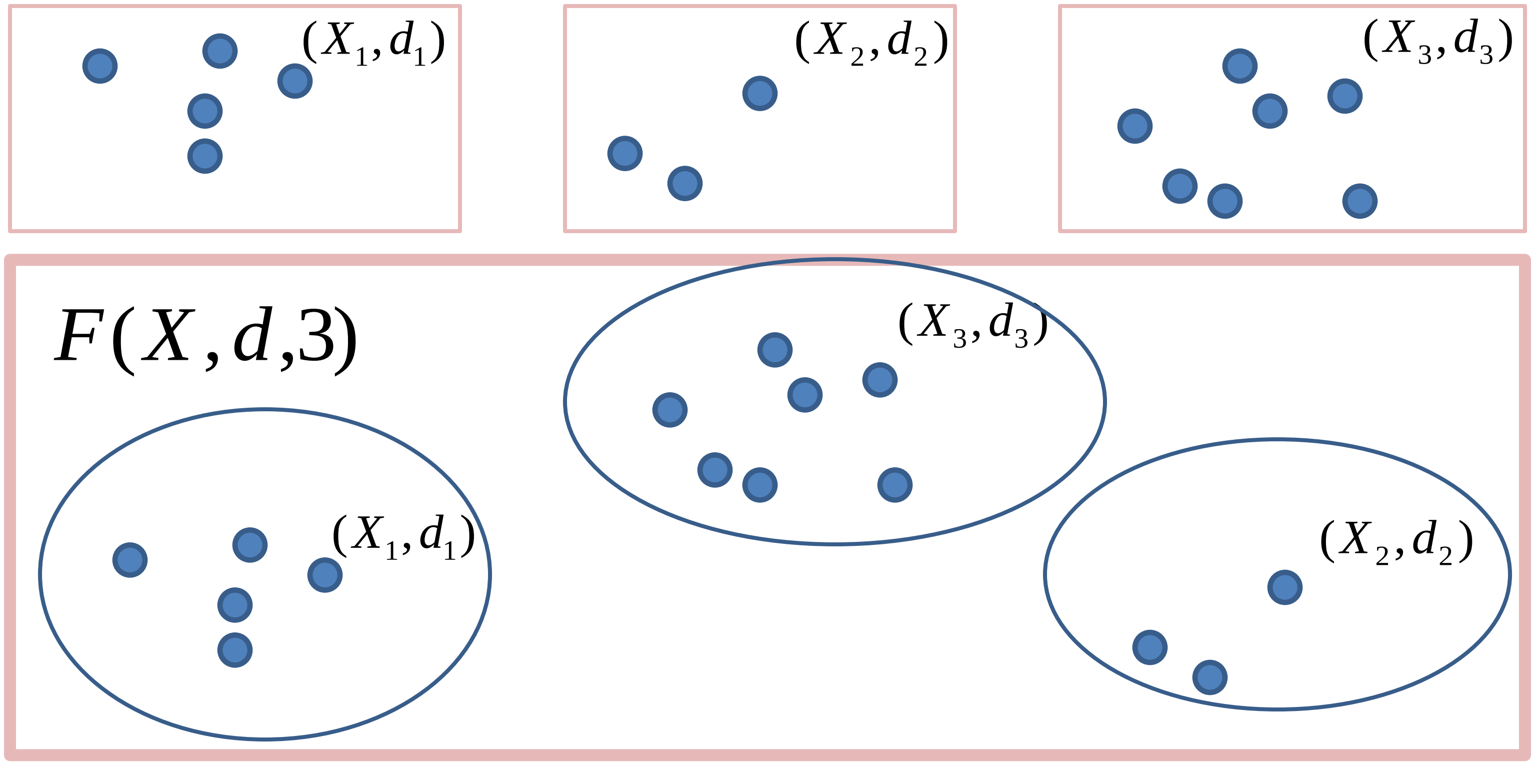
Not all algorithms are local and outer-consistent!

- Some common clustering algorithms fail locality and outer-consistency
  - Ex. Spectral clustering objectives Ratio Cut and Normalized Cut
- Locality and outer-consistency can be used to distinguish between clustering algorithms (they are not axioms).

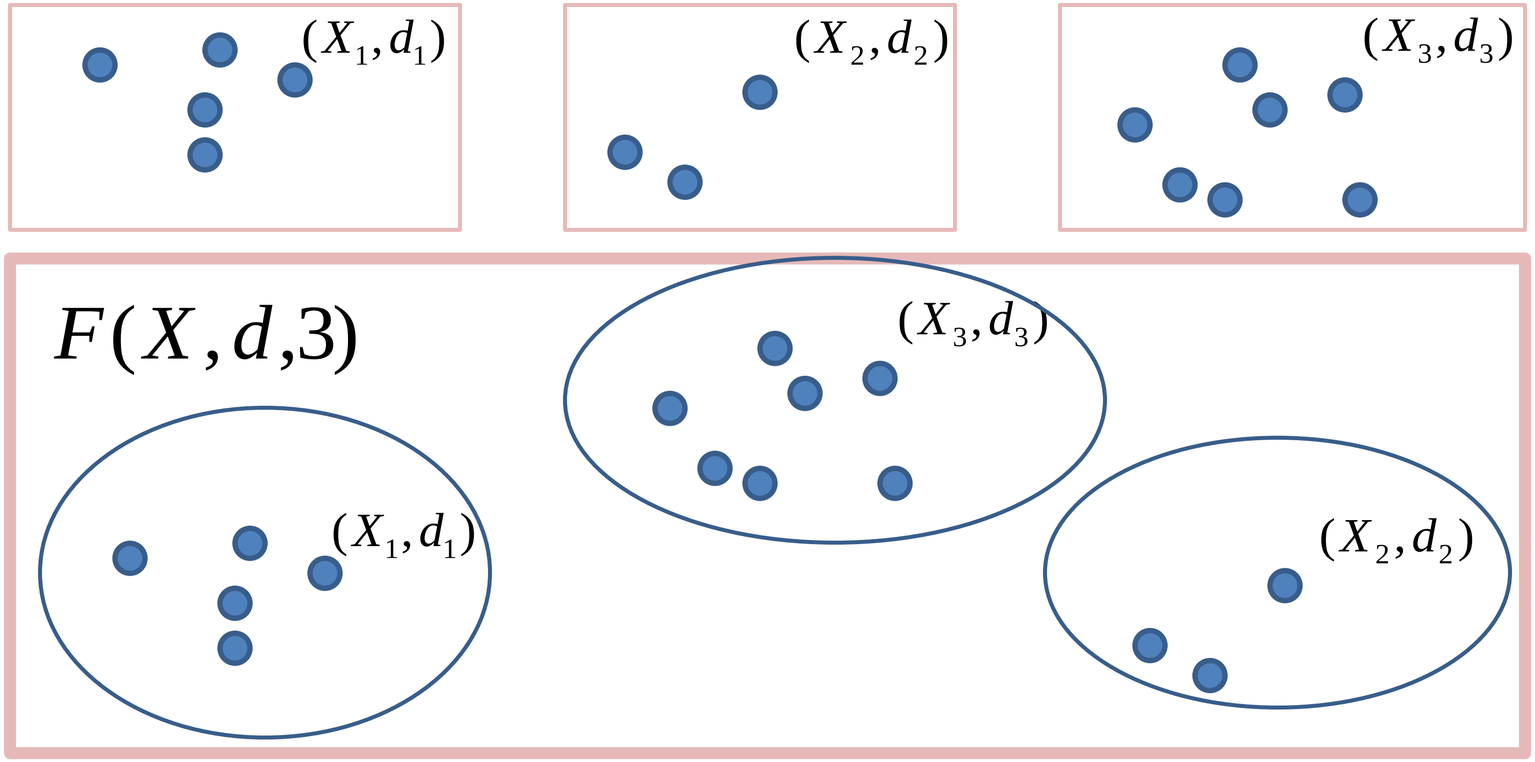
Extended Richness



Extended Richness



Extended Richness



F satisfies *extended richness* if for *any* set of domains

$$\{(X_1, d_1), (X_2, d_2), \cdots, (X_k, d_k)\}$$

there is a d over $X = \bigcup X_i$ that extends each of the d_is so that $F(X, d, k) = \{X_1, X_2, \cdots, X_k\}$.

Outline

- Define linkage-based clustering
- Our new clustering properties
- Main result
- Sketch of proof
- A taxonomy of common clustering algorithms using our properties
- Conclusions

Our main result

Theorem:

A clustering function is Linkage-Based

if and only if

it is Hierarchical, Outer-Consistent, Local and satisfies Extended Richness.

Easy direction of proof

Every Linkage-Based clustering function is Hierarchical, Local, Outer-Consistent, and satisfies Extended Richness.

The proof is quite straight-forward.

Interesting direction of proof

If *F* is Hierarchical and it satisfies Outer Consistency, Locality and Extended-Richness then *F* is Linkage-Based.

To prove this direction we first need to formalize linkage-based clustering, by formally defining what is a linkage function.

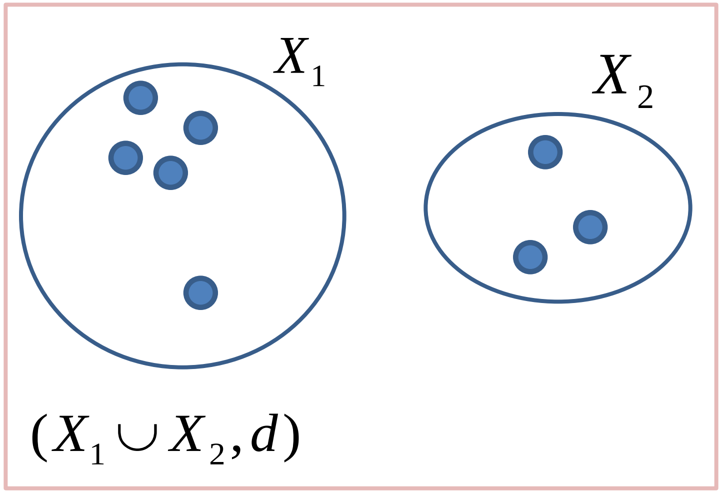
What do we expect from linkage function?

A *linkage function* is a function

$$\ell : \{(X_1, X_2, d) : d \text{ is a dissimilarity function over } X_1 \cup X_2\} \rightarrow R^+$$

that satisfies the following:

1) *Representation independent*: Doesn't change if we re-label the data
2) *Monotonic:* if we increase edges that go between X_1 and X_2, then $\ell(X_1, X_2, d)$ doesn't decrease.
3) Any pair of clusters can be made arbitrarily distant:
   By increasing edges that go between X_1 and X_2, we can make $\ell(X_1, X_2, d)$ exceed any value in the range of ℓ.



Sketch of proof

Need to prove:

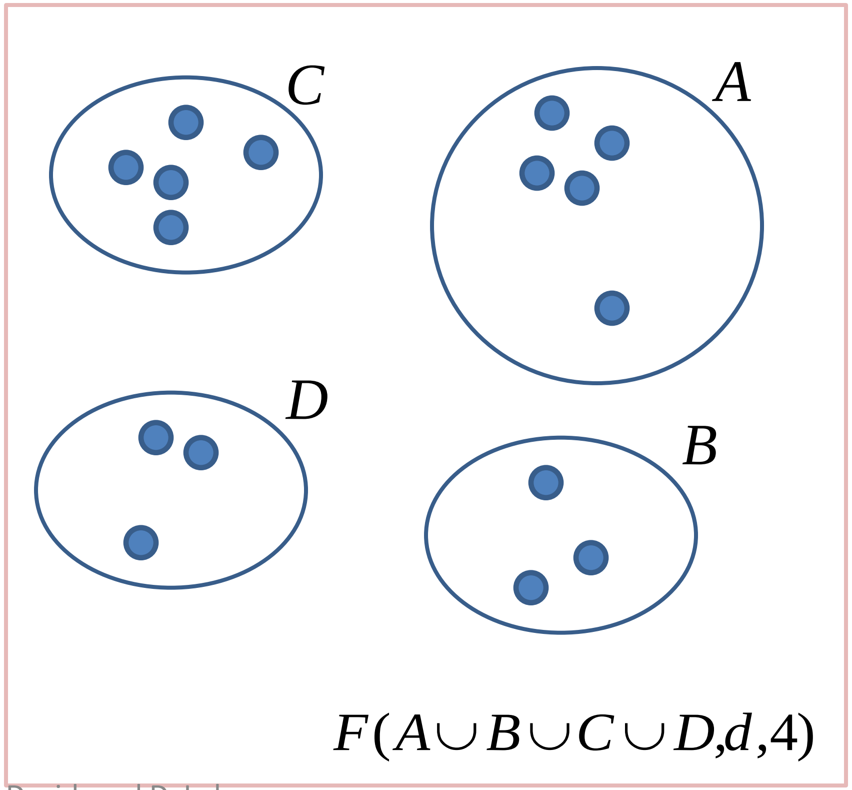
If F is a hierarchical function that satisfies the above clustering properties then F is linkage-based.

<u>Goal:</u>

Given a clustering function F that satisfies the properties, define a linkage function $\boldsymbol{\ell}$ so that the linkage-based clustering based on $\boldsymbol{\ell}$ coincides with F (for every X, d and k).

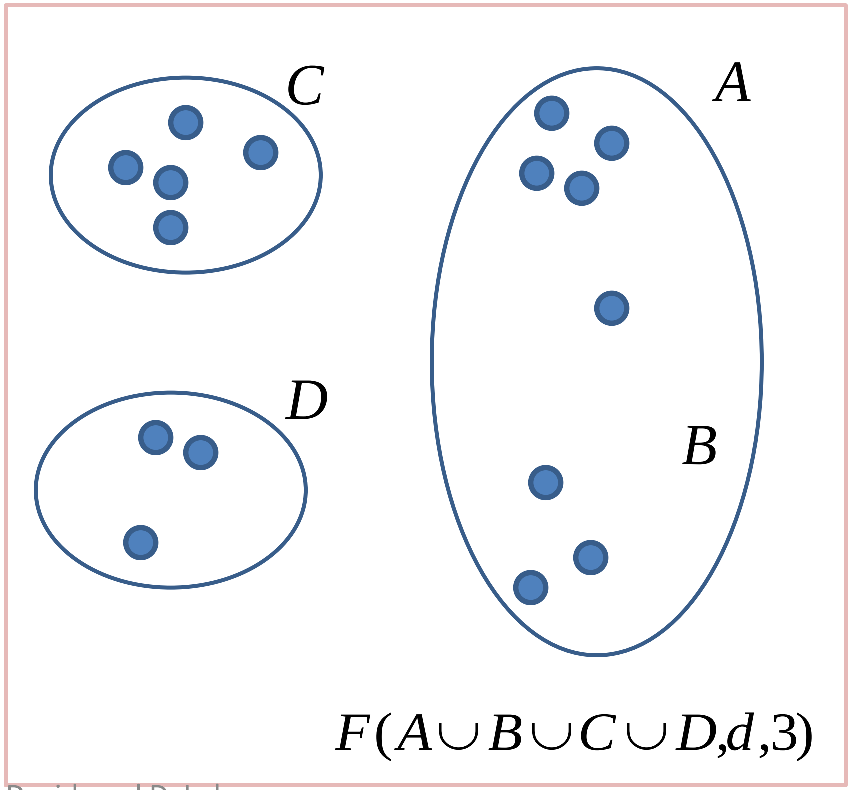
Sketch of proof (continued...)

- Define an operator $<_F$: $(A,B,d_1) <_F (C,D,d_2)$ if there exists d that extends d_1 and d_2 such that when we run F on $(A \cup B \cup C \cup D, d)$, A and B are merged before C and D.



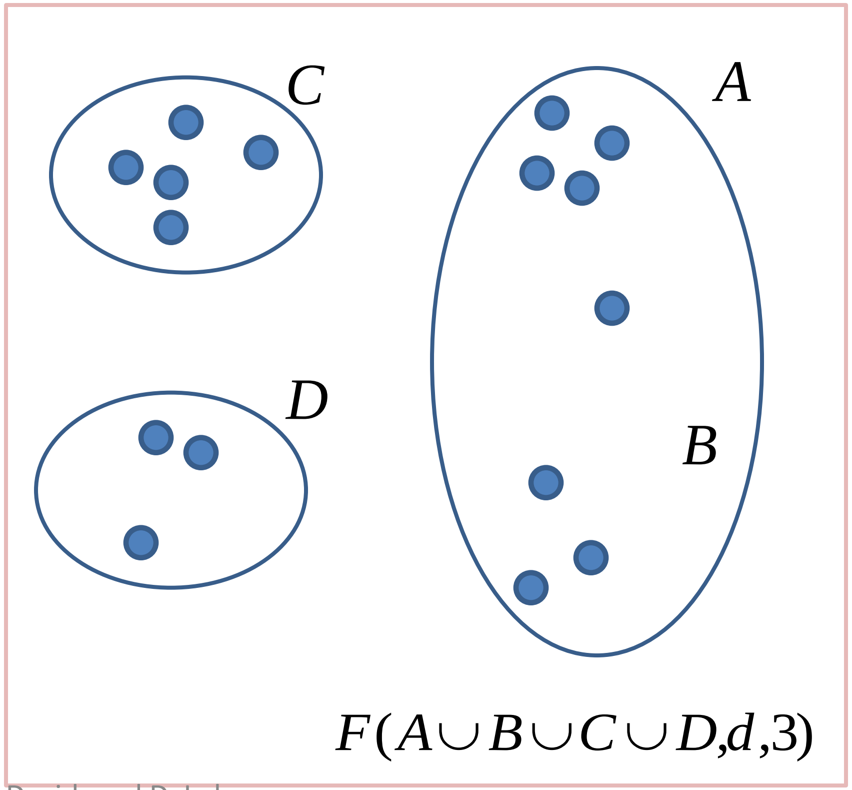
Sketch of proof (continued…)

- Define an operator $<_F$: $(A,B,d_1) <_F (C,D,d_2)$ if there exists d that extends d_1 and d_2 such that when we run F on $(A \cup B \cup C \cup D, d)$, A and B are merged before C and D.



Sketch of proof (continued…)

- Define an operator $<_F$: $(A,B,d_1) <_F (C,D,d_2)$ if there exists d that extends d_1 and d_2 such that when we run F on $(A \cup B \cup C \cup D, d)$, A and B are merged before C and D.
- Prove that $<_F$ can be extended to a partial ordering
- Use the ordering to define ℓ



Sketch of proof continue:
Show that $<_F$ is a partial ordering

We show that $<_F$ is cycle-free.

Lemma: Given a function F that is hierarchical, local, outer-consistent and satisfies extended richness, there are no $(A_1, B_1, d_1), (A_2, B_2, d_1), \cdots, (A_n, B_n, d_1)$ so that $(A_1, B_1, d_1) <_F (A_2, B_2, d_2) <_F \cdots <_F (A_n, B_n, d_n)$ and $(A_1, B_1, d_1) = (A_n, B_n, d_n)$

Sketch of proof (continued...)

- By the above Lemma, the transitive closure of $<_F$ is a partial ordering.
- This implies that there exists an order preserving function ℓ that maps pairs of data sets to R (since $<_F$ is defined over a countable set).
- It can be shown that ℓ satisfies the properties of a linkage function.

Conclusions

- We introduced new meaningful properties of clustering algorithms.
- Prove they characterize linkage-based algorithms.
- Whenever all these properties are desirable, a linkage-based algorithm should be used.