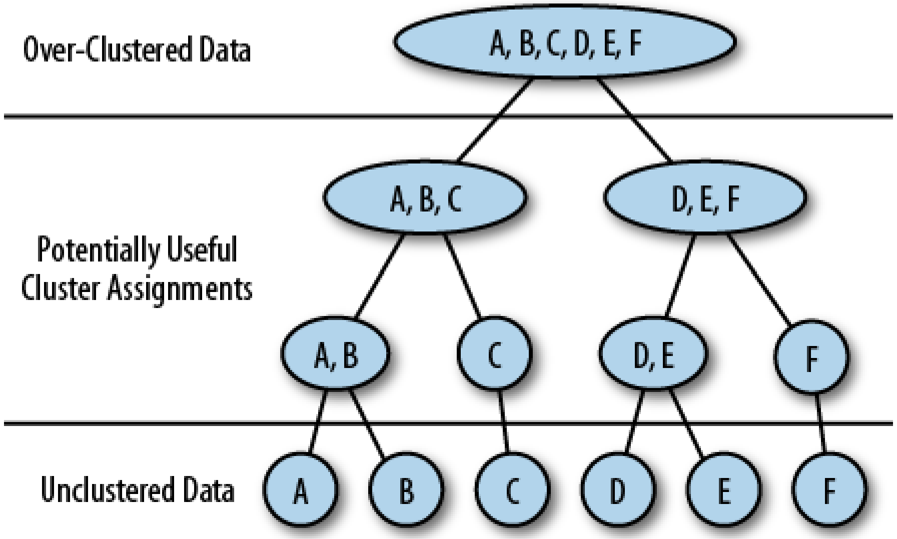
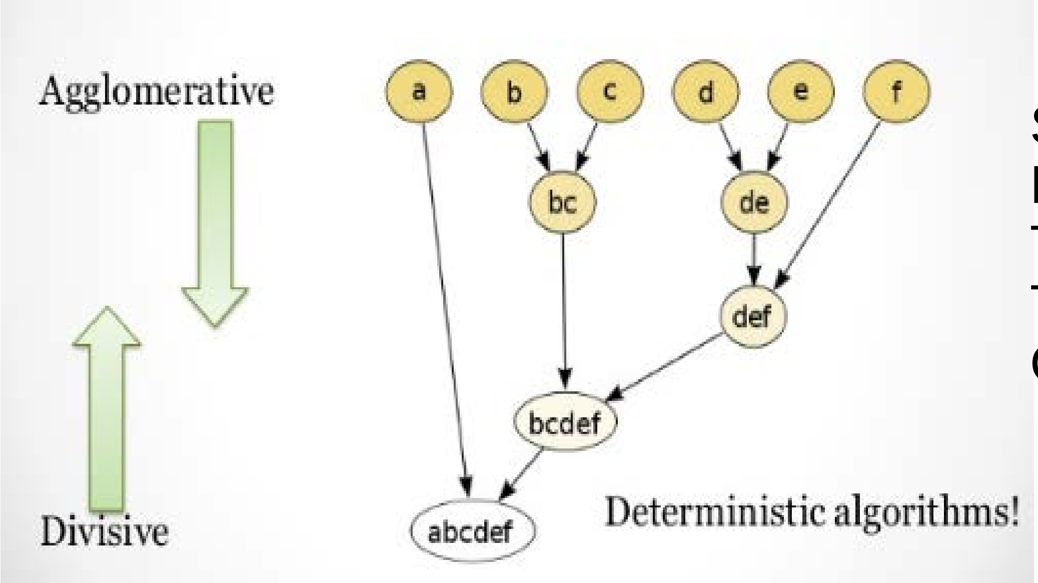
Hierarchical Clustering



Hierarchical Clustering

- **Previously we stated that there are two types of clustering n data points**
  - **Partitioning** such as k-means which we have already presented
  - **Hierarchical** which consists of two methods
    - **Agglomerative methods**: In this method, we begin with each data point being a cluster. We combine clusters based on some distance function until we have a single cluster
    - **Divisive methods**: We begin with all data in a *single cluste*r and divide until each data point is in a single cluster
  - On the next slide, an image from Wikipedia shows the difference.
  - ***We have not specified the decision process for forming clusters! That will come later.***

Agglomerative versus Divisive



Steps in Agglomerative

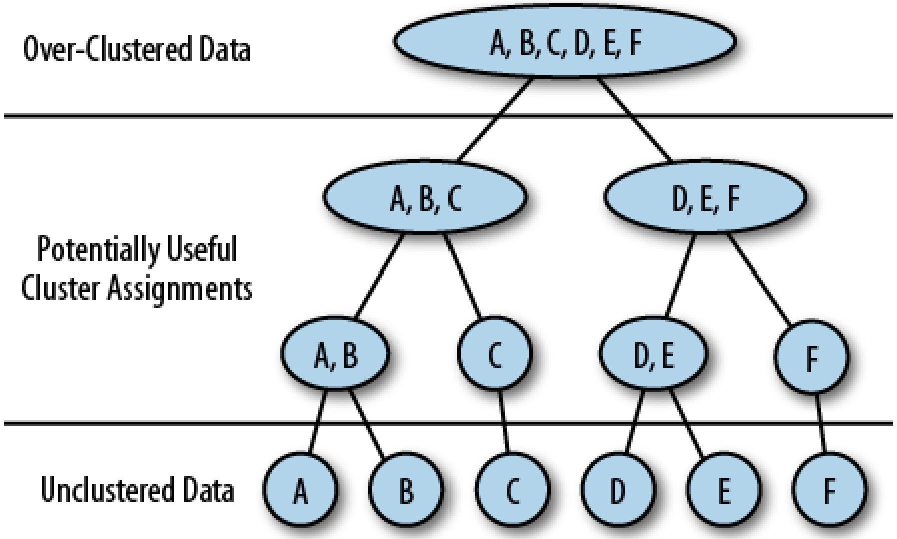
Six clusters = {a}, {b}, {c}, {d}, {e}. {f}
Four clusters = {a}, {bc}, {de}, {f}
Three clusters = {a},{b,c}, {d,e,f}
Two clusters = {a}, {b,c,d,e,f}
One cluster = {a,b,c,d,e,f}

Steps in Divisive

One cluster = {a,b,c,d,e,f}
Two clusters = {a}, {b,c,d,e,f}
Three clusters = {a},{b,c}, {d,e,f}
Four clusters = {a}, {bc}, {de}, {f}
Six clusters = {a}, {b}, {c}, {d}, {e}. {f}

Usable Clusters

- **As we have stated, having one cluster or n clusters is not informative**
- **The diagram below shows how a subject matter expert could determine meaningful clusters**



Determining “distance” between clusters

Before we can start forming clusters we need to define the distance between two clusters

There are several ways to do this

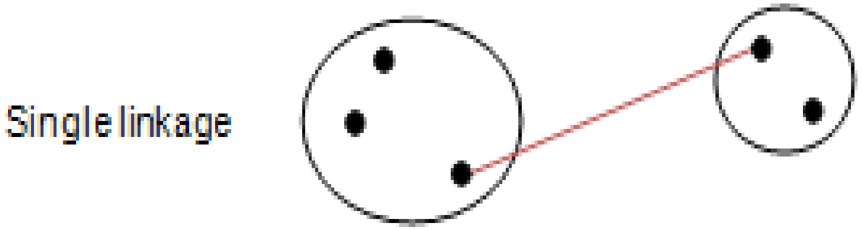
- Some actually involve a distance such as Euclidean distance
- Others involve similarity of clusters

Those that involve notions of distance include

- Single link clustering
- Complete link clustering
- Average link clustering

http://www.analytictech.com/networks/hiclus.htm

Single Linkage



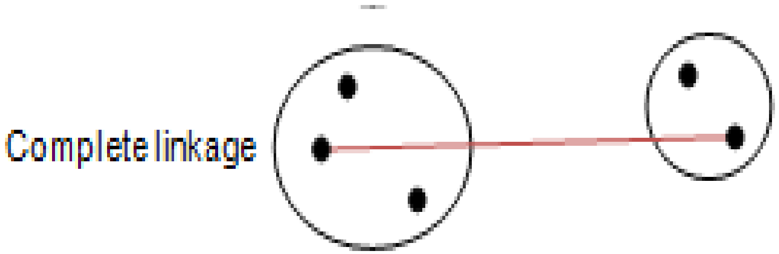
In **single linkage** (SL), the distance between two clusters A and B is defined to be the minimum of all the distances between any element in A and any element in B.

In symbols, D(A,B) = min{d(x,y) where min is over all x in A and all y in B

If A has n elements and B has m elements, how many distances would we have to compute? There is actually a way to reduce the number of such calculations

When we cluster using SL, we join clusters with the the smallest SL distance.

Complete Linkage

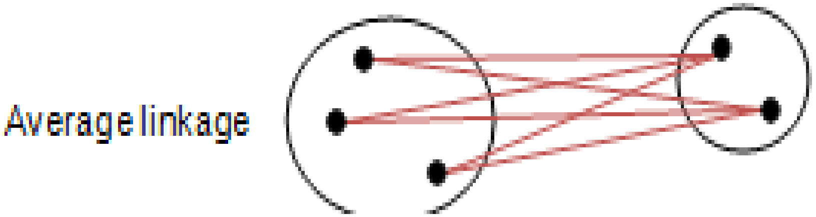


In **complete linkage** (CL) the distance between two clusters A and B is defined to be the maximum distance between any element in A and any element in B.

D(A,B) = max{d(x,y) where max is over all x in A and all y in B.

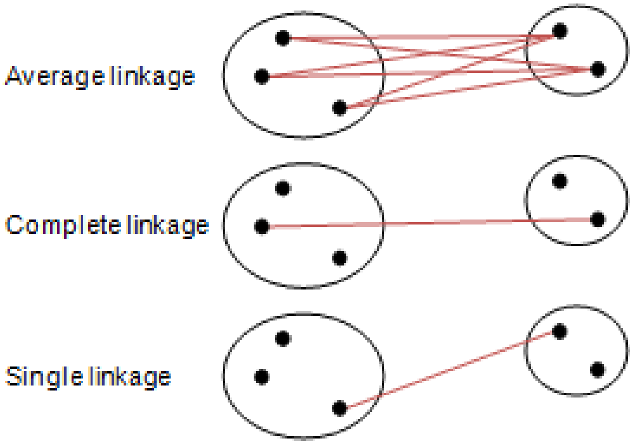
When we cluster using complete linkage, we join clusters with the smallest CL distance.

Average Linkage

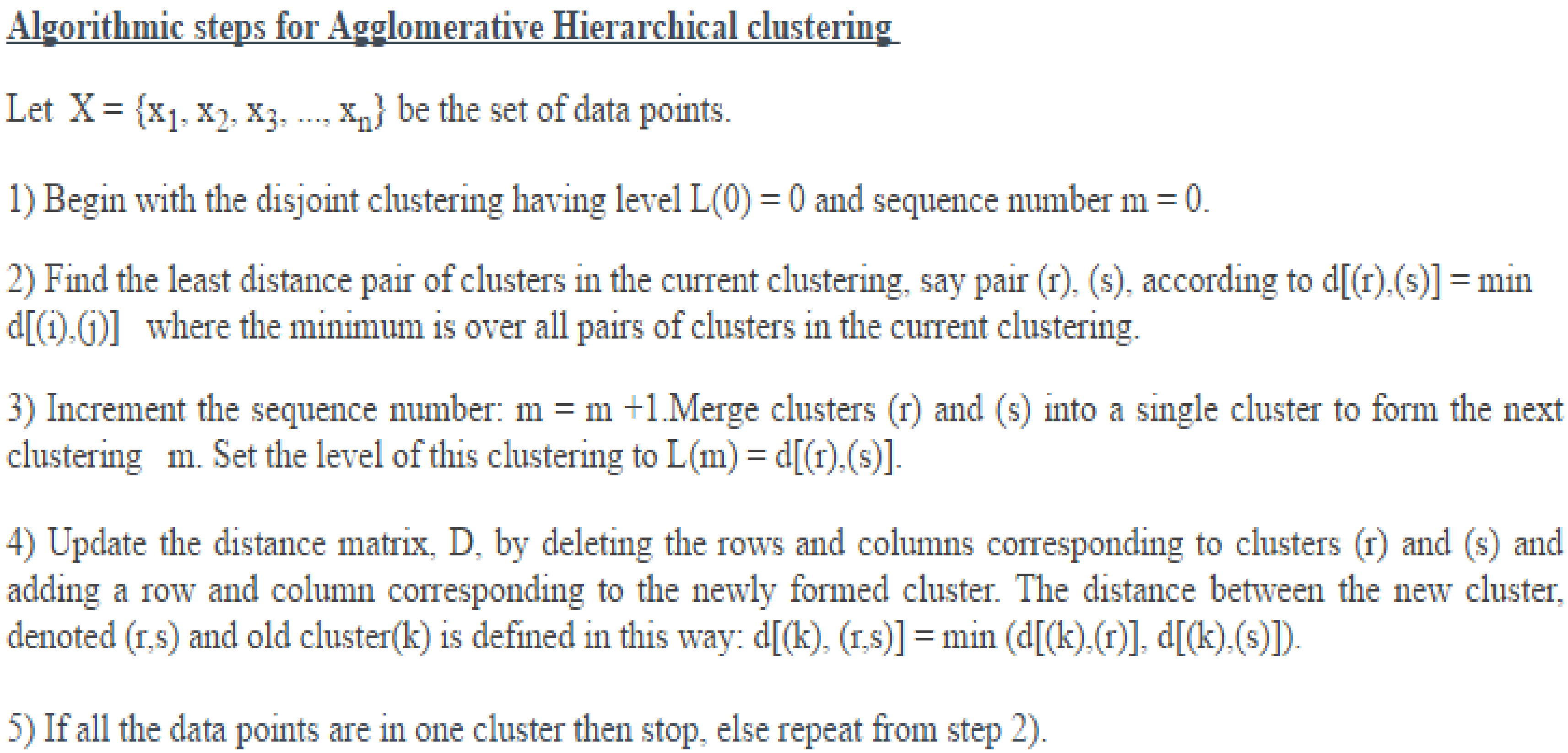


In Average Linkage (AL), the distance between clusters A and B is defined to be the average distance between all elements x of A and y of B.

Below, we see all methods side by side.



Algorithm for Agglomerative Clustering



Algorithmic steps for Agglomerative Hierarchical clustering

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points.

1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].

4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

5) If all the data points are in one cluster then stop, else repeat from step 2).

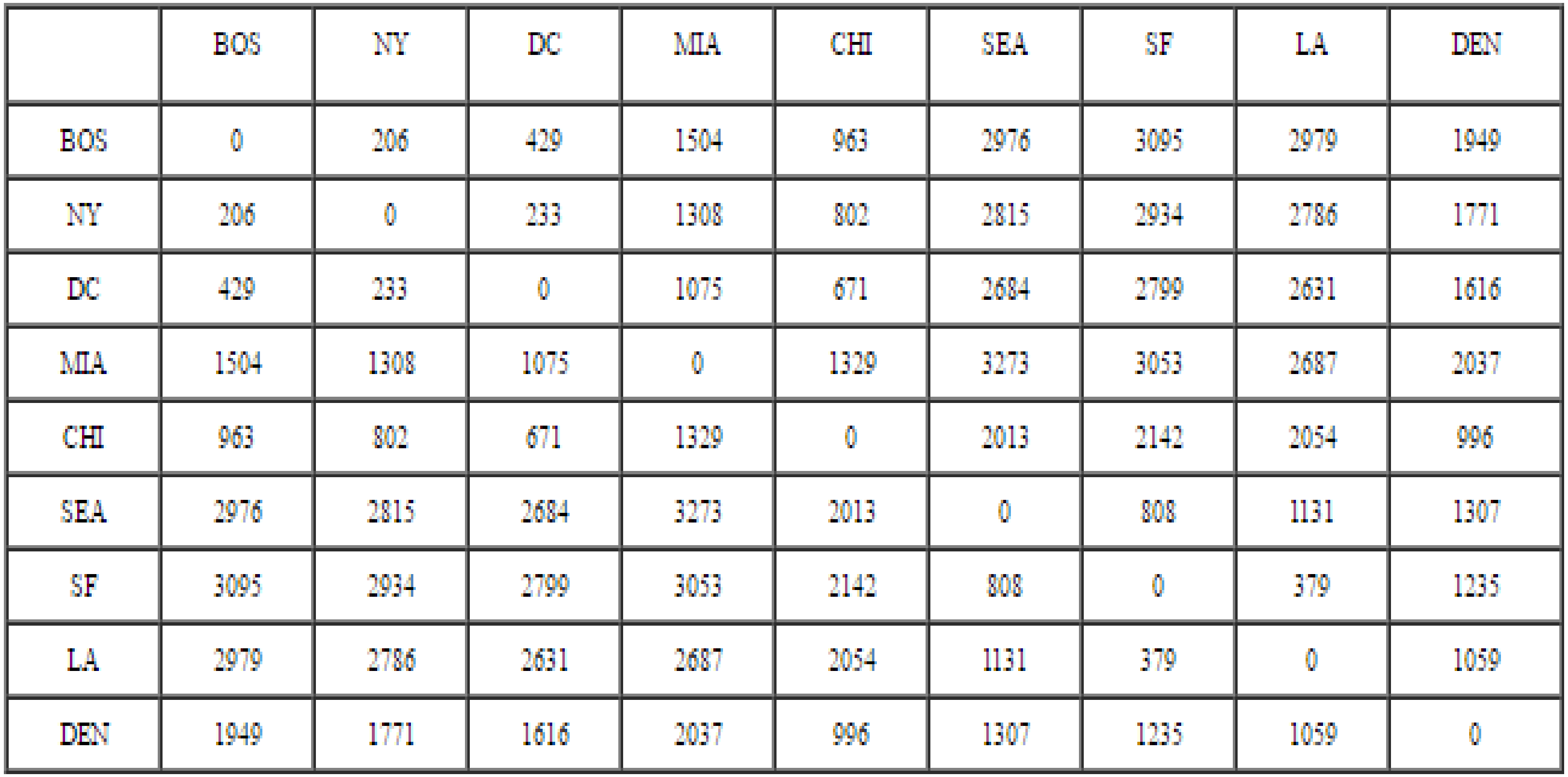
https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm

Questions for previous slide

- **On the previous slide, there are variables m and L. How could these be used to assist in Cluster Analysis?**

- **On the next slide, we begin an example. The data items are US Cities. The distance between any two cities is the distance in air miles between them.**
- **We will use these distances to cluster the cities**
- **We will attempt to determine a “usable” number of clusters and give names to each cluster.**

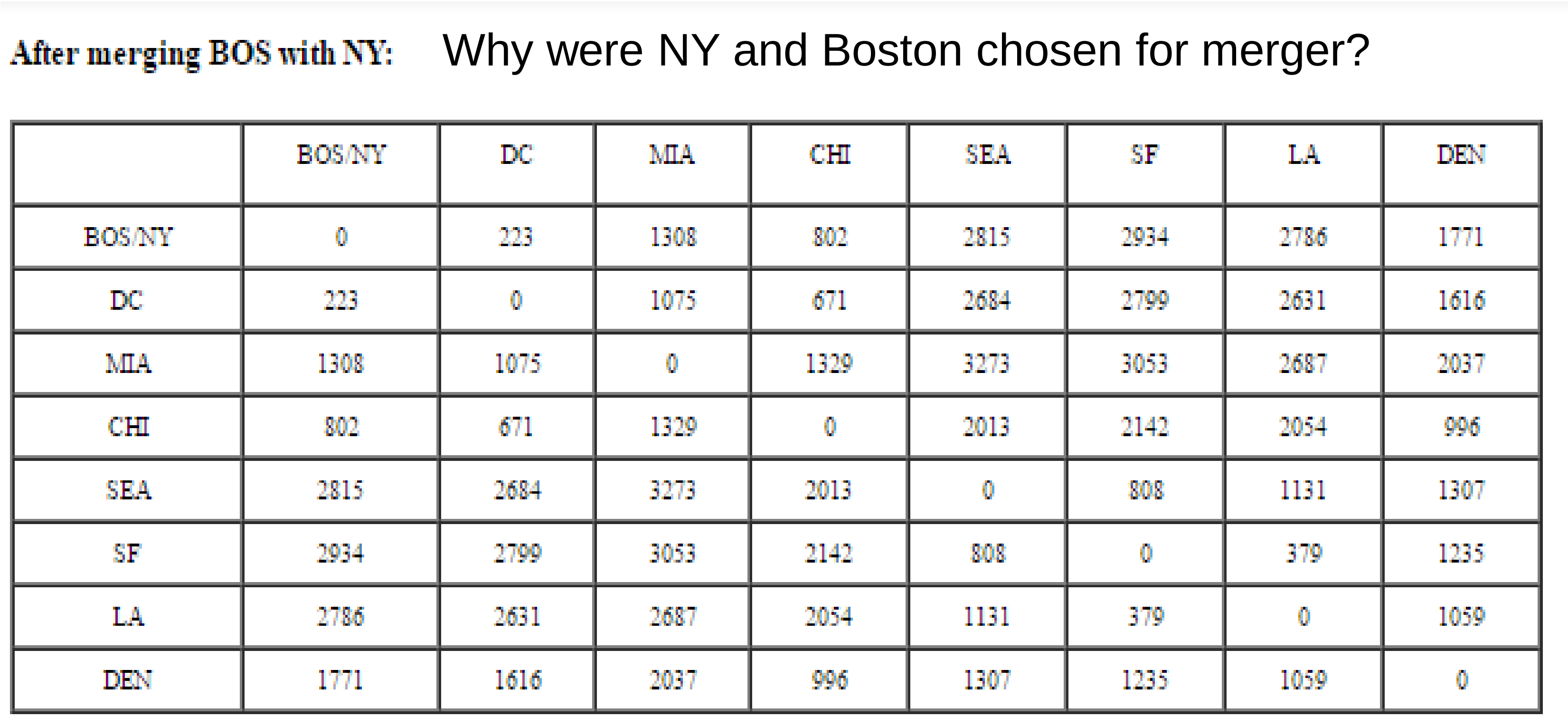
Clustering Air Flights



	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Using SL, we see that the nearest pair of cities is Boston and NY. On the next slide, we will merge them into a single cluster and update the distances.
http://www.analytictech.com/networks/hiclus.htm

One merger m=1



After merging BOS with NY:

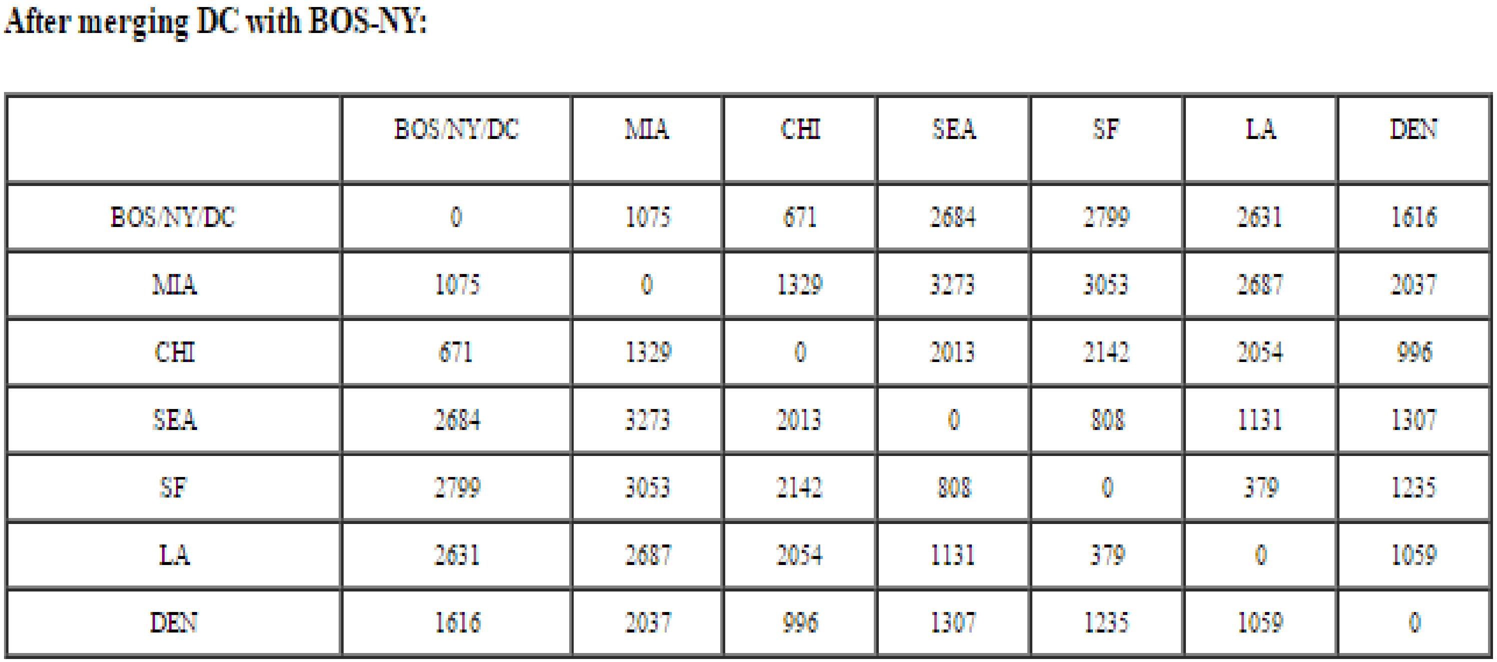
Why were NY and Boston chosen for merger?

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	0	223	1308	802	2815	2934	2786	1771
DC	223	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

Which entries in the matrix needed to be updated if we use Single Linkage? How were they calculated? For example, why is the distance from {Bos, NY} to {MIA} = 1308?

Which clusters will be merged next?

m=2

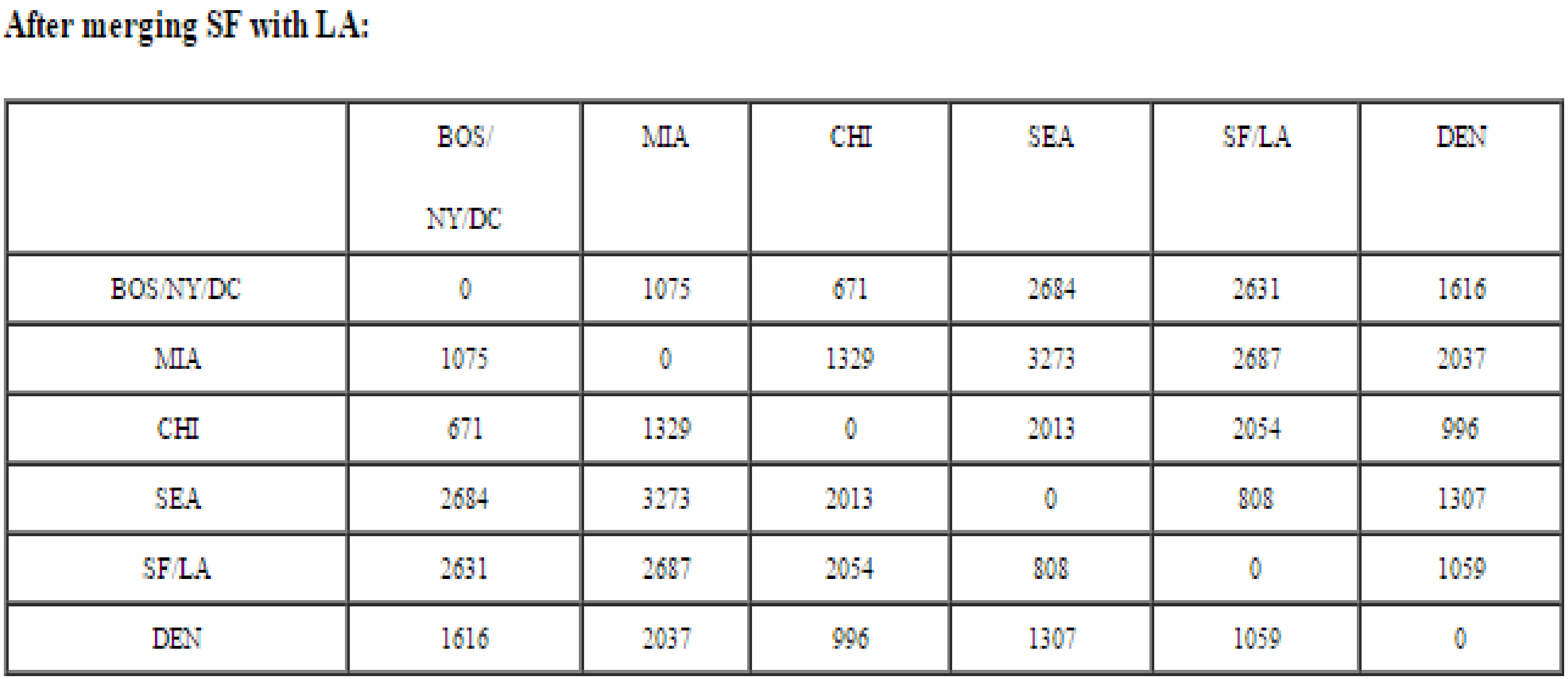


After merging DC with BOS-NY:

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

Which clusters will be merged next? Why?

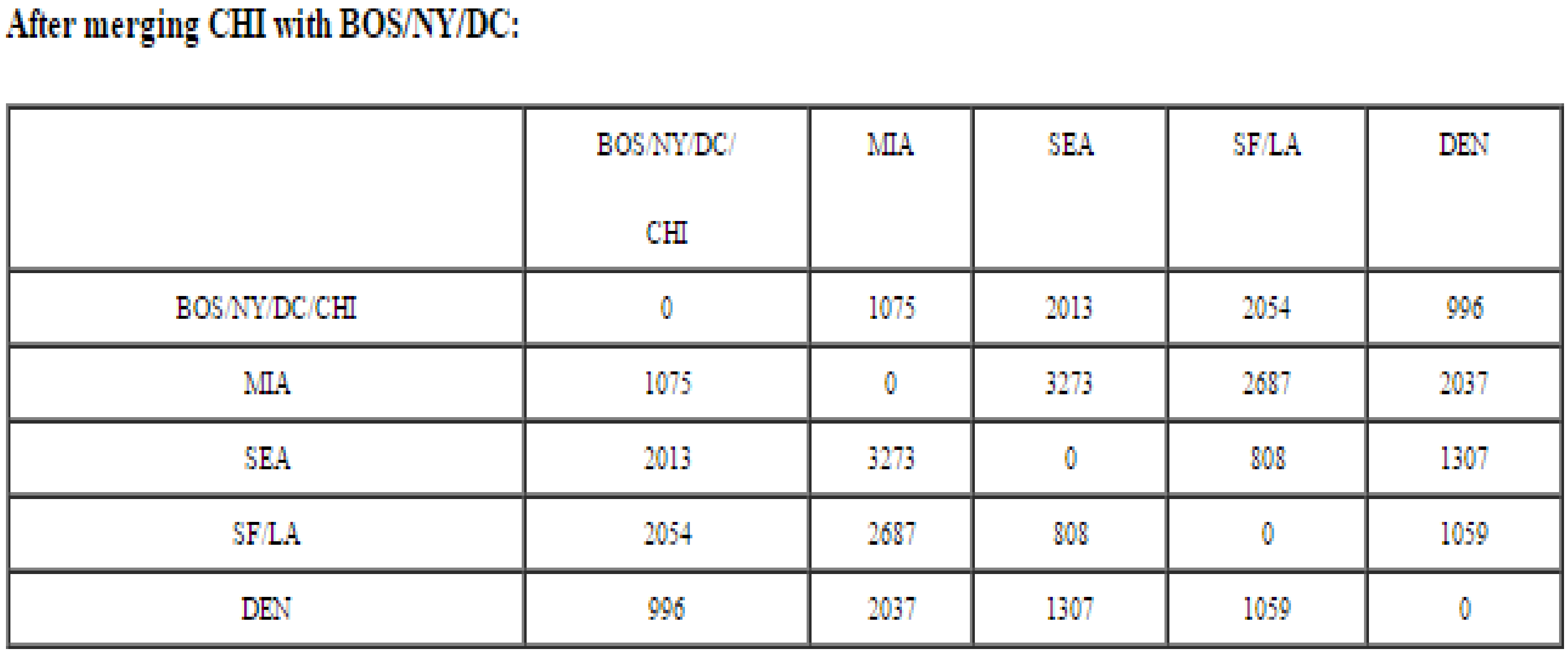
m=3



After merging SF with LA:

	BOS/ NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0

m=4



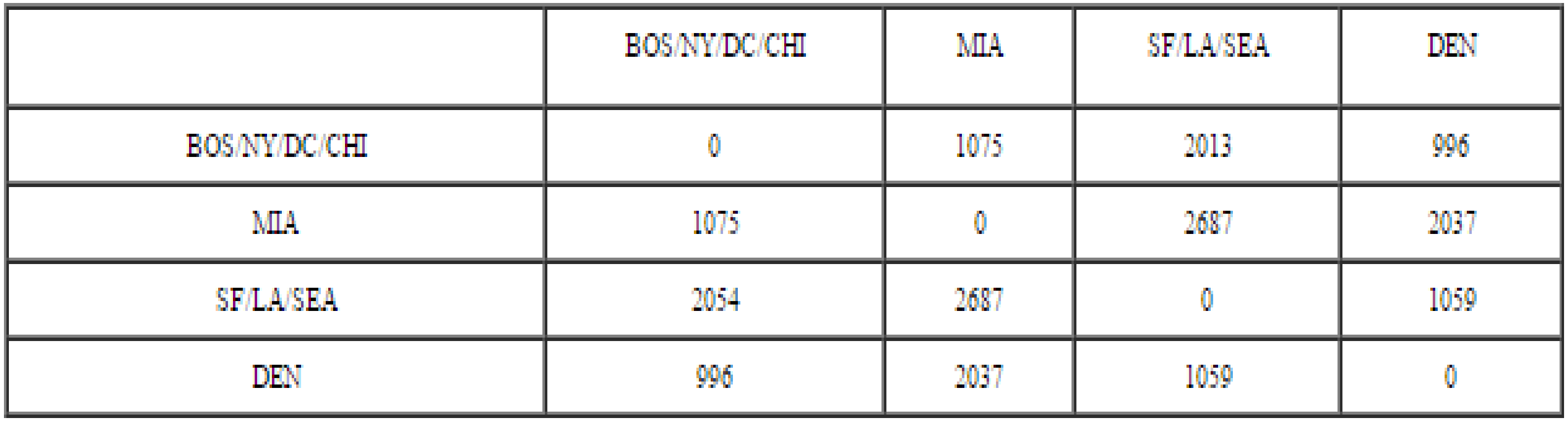
After merging CHI with BOS/NY/DC:

	BOS/NY/DC/ CHI	MIA	SEA	SF/LA	DEN
BOS/NY/DC/CHI	0	1075	2013	2054	996
MIA	1075	0	3273	2687	2037
SEA	2013	3273	0	808	1307
SF/LA	2054	2687	808	0	1059
DEN	996	2037	1307	1059	0

What is the next step? Why?

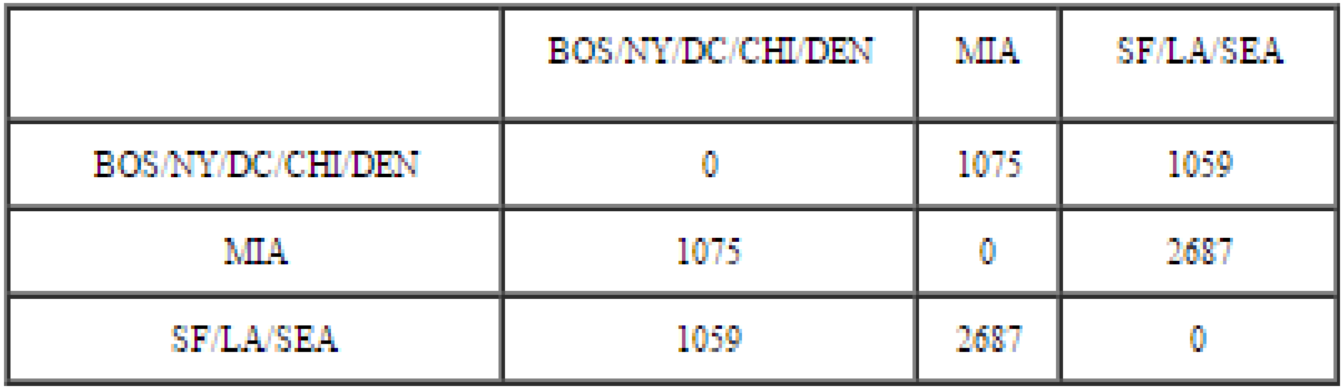
m=5 and m=6

After merging SEA with SF/LA:



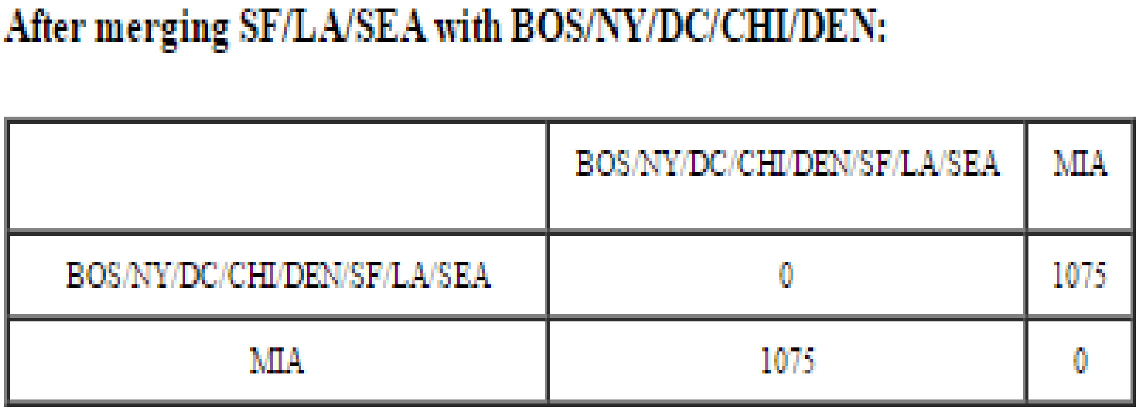
	BOS/NY/DC/CHI	MIA	SF/LA/SEA	DEN
BOS/NY/DC/CHI	0	1075	2013	996
MIA	1075	0	2687	2037
SF/LA/SEA	2054	2687	0	1059
DEN	996	2037	1059	0

After merging DEN with BOS/NY/DC/CHI:



	BOS/NY/DC/CHI/DEN	MIA	SF/LA/SEA
BOS/NY/DC/CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0

m=7

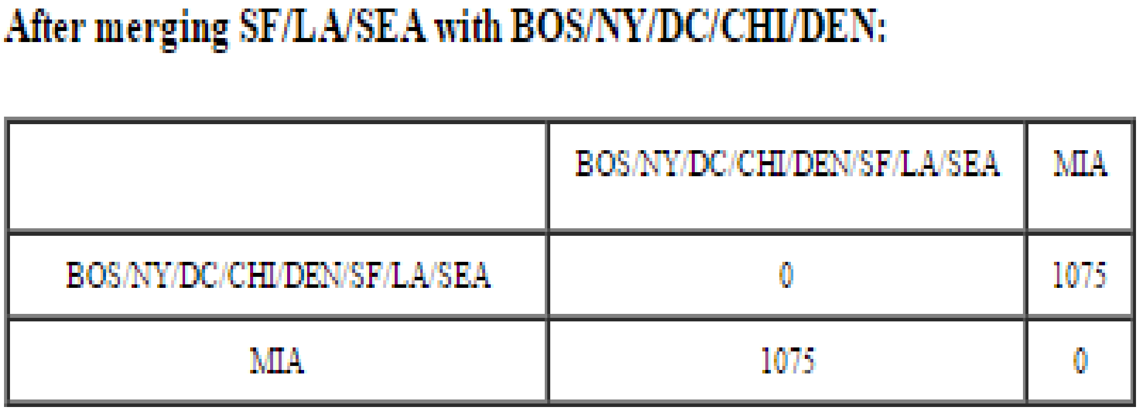


After merging SF/LA/SEA with BOS/NY/DC/CHI/DEN:

	BOS/NY/DC/CHI/DEN/SF/LA/SEA	MIA
BOS/NY/DC/CHI/DEN/SF/LA/SEA	0	1075
MIA	1075	0

When m = 8, we will have constructed a single cluster BOS/NY/DC/CHI/DEN/SF/LA/SEA/MIA

m=7



After merging SF/LA/SEA with BOS/NY/DC/CHI/DEN:

	BOS/NY/DC/CHI/DEN/SF/LA/SEA	MIA
BOS/NY/DC/CHI/DEN/SF/LA/SEA	0	1075
MIA	1075	0

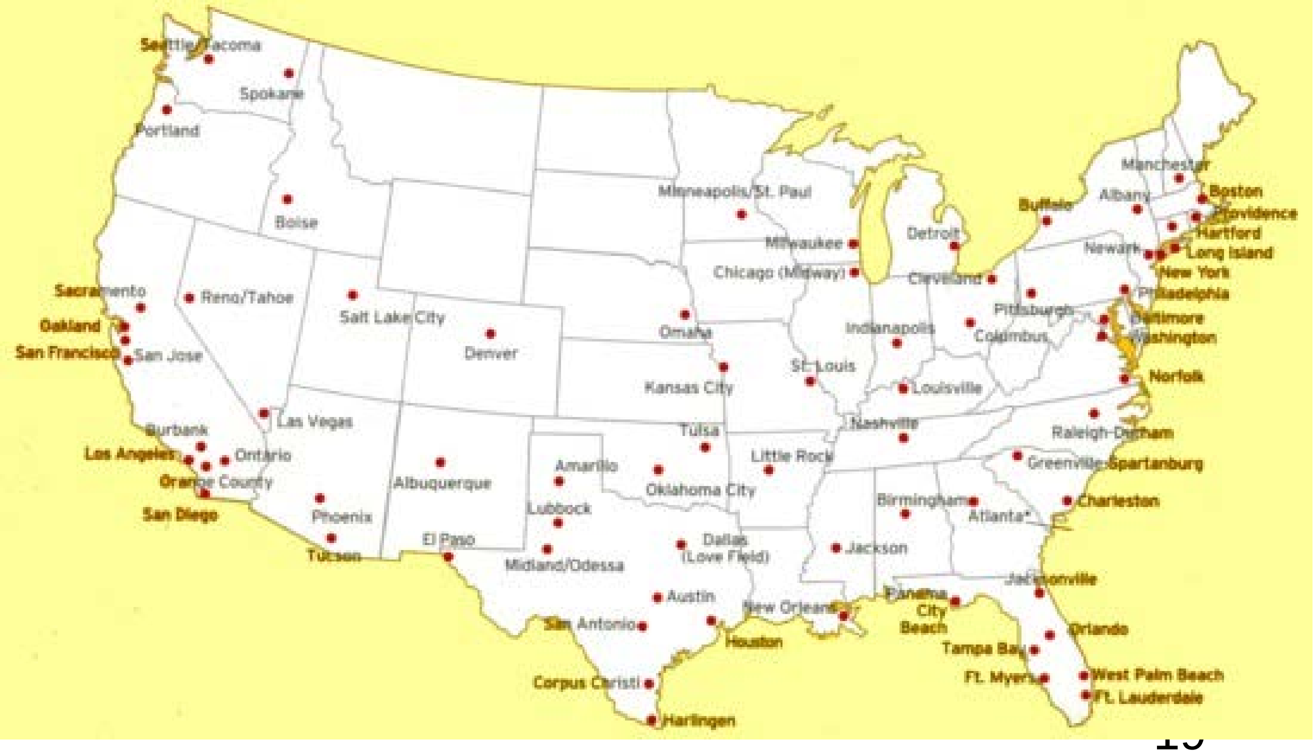
When m = 8, we will have constructed a single cluster BOS/NY/DC/CHI/DEN/SF/LA/SEA/MIA

Analyzing the Clusters

As we have seen, one cluster and n clusters are not usable

At which level do you (as a subject matter expert) think the clusters are useful

Give names to the clusters and defend your answer.



Seattle/Tacoma
Spokane
Portland
Boise
Sacramento
Reno/Tahoe
Salt Lake City
Oakland
San Francisco
San Jose
Denver
Las Vegas
Burbank
Los Angeles
Ontario
Orange County
San Diego
Phoenix
Tucson
Albuquerque
El Paso
Amarillo
Lubbock
Midland/Odessa
Minneapolis/St. Paul
Milwaukee
Chicago (Midway)
Omaha
Kansas City
St. Louis
Tulsa
Oklahoma City
Little Rock
Dallas
(Love Field)
Austin
San Antonio
Houston
Corpus Christi
Harlingen
New Orleans
Jackson
Detroit
Cleveland
Indianapolis
Columbus
Louisville
Nashville
Pittsburgh
Birmingham
Atlanta
Panama
City
Beach
Buffalo
Albany
Manchester
Boston
Providence
Hartford
Newark
Long Island
New York
Philadelphia
Baltimore
Washington
Norfolk
Raleigh-Durham
Greenville-Spartanburg
Charleston
Jacksonville
Orlando
Tampa Bay
Ft. Myers
West Palm Beach
Ft. Lauderdale