HUDM5124 Session 12: Tree Models of Proximity

Overview:

- Trees as graphs
- Trees as metric spaces
- Two types of tree models:
    - ultrametric trees
    - additive trees

Trees as graphs

Def. A graph G consists of a set of nodes (vertices) and arcs (edges): G = <V,E>, where the edges correspond to a binary relationship defined on the elements of V (i.e., defined as a subset of the set V x V).

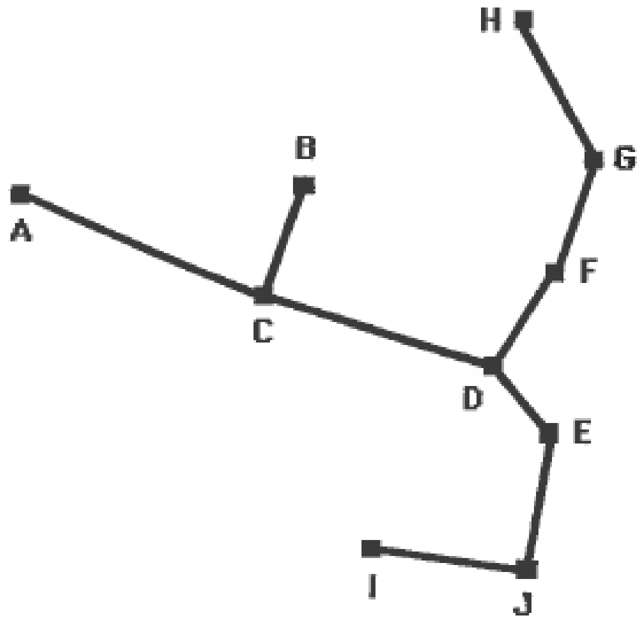
Def. A tree is a connected graph without cycles.

Def. A weighted graph (tree) associates a nonnegative weight w with each arc in E:

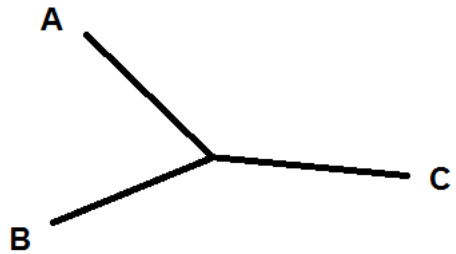
Gw = <V,E,W>

Minimal Spanning Tree

→not particularly useful to <u>model proximities</u>, rather it solves a *distribution or path problem*: it is the tree graph of minimal length that connects all points in a fixed set of nodes or points. Note that in the MST each internal (non-leaf) node in the tree corresponds to an object.



Note: the *Steiner tree* generalizes the MST, by allowing new internal nodes ("Steiner points") to be added that do not correspond to one of the "leaf" nodes or conceptual objects being modeled.

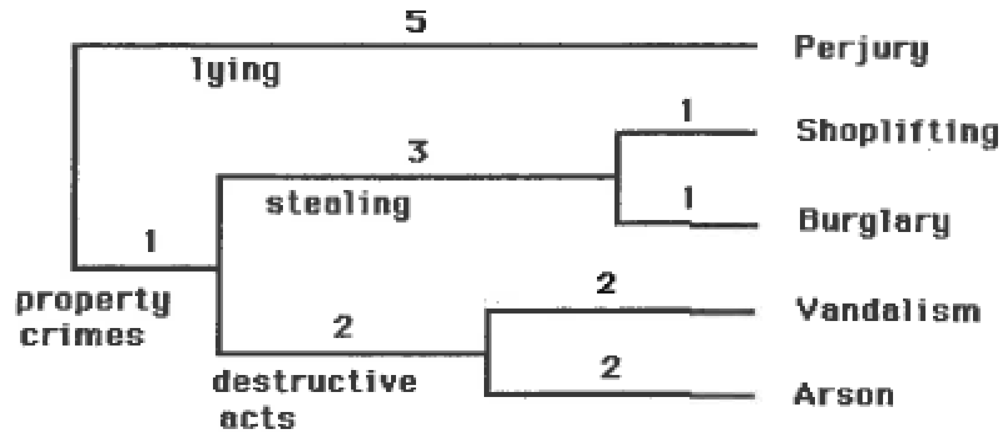


Hypothetical data on dissimilarity of crimes: ultrametric tree

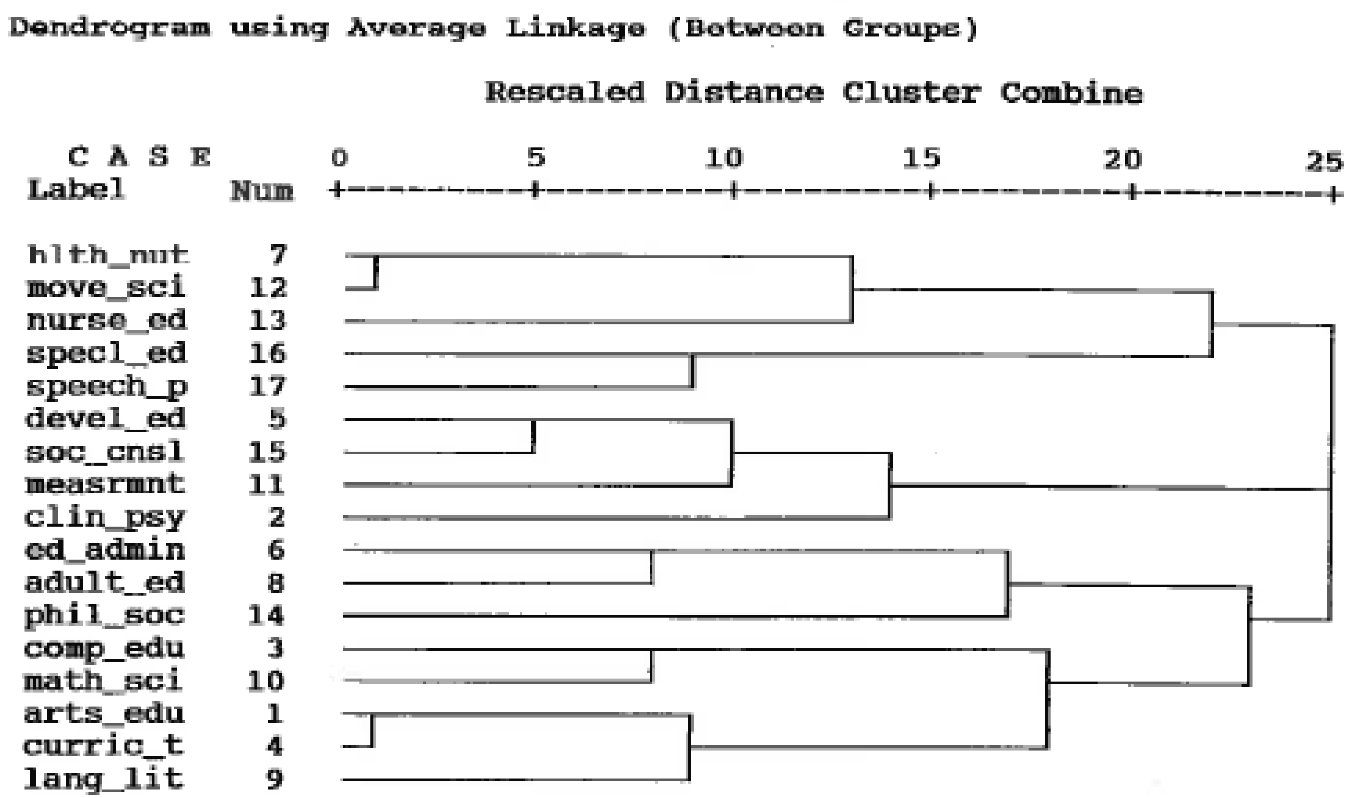
a. Dissimilarities

	Ar	Bu	Pe	Sh	Va
Arson	--				
Burglary	8	--			
Perjury	10	10	--		
Shoplifting	8	2	10	--	
Vandalism	4	8	10	8	--

b. Ultrametric Tree



The dendrogram representing a hierarchical clustering can be viewed as an ultrametric tree (but fit is not optimized)

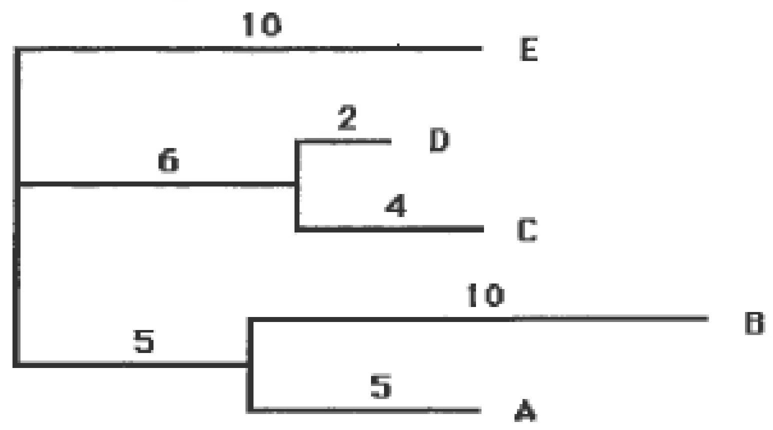


additive tree of hypothetical data on worker communication frequencies (inverted to dissimilarities)

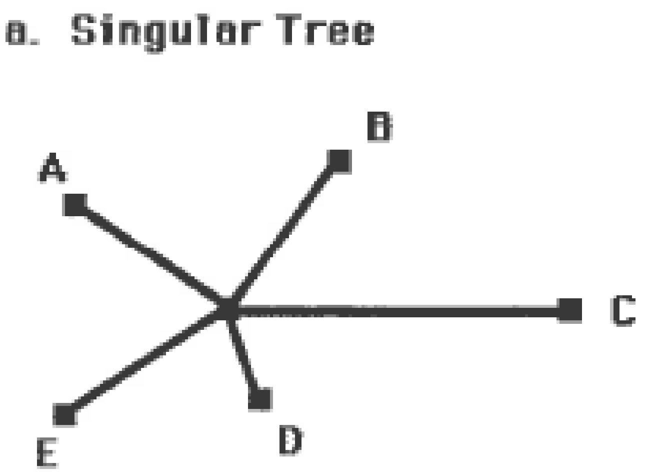
a. Dissimiliarities

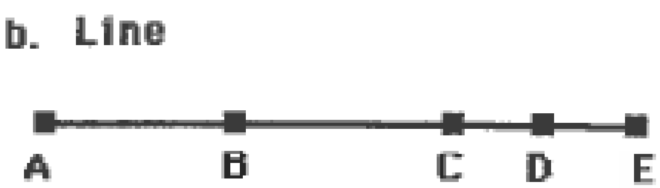
	A	B	C	D
Worker A	--			
Worker B	15	--		
Worker C	20	25	--	
Worker D	18	23	6	--
Worker E	20	25	20	18

b. Additive Tree

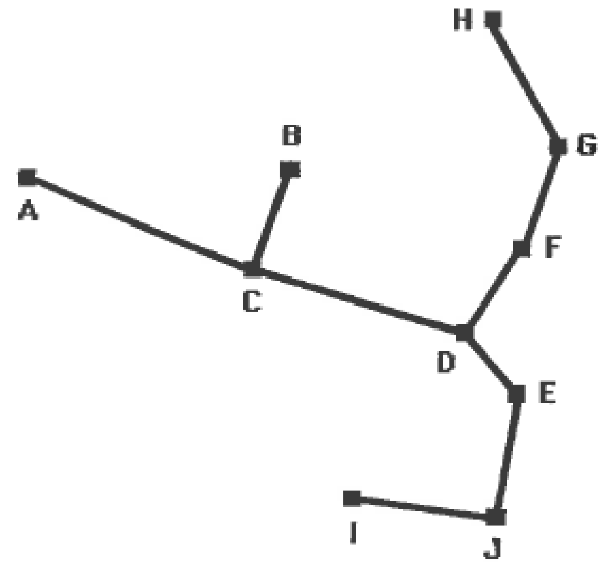


Special cases of additive trees

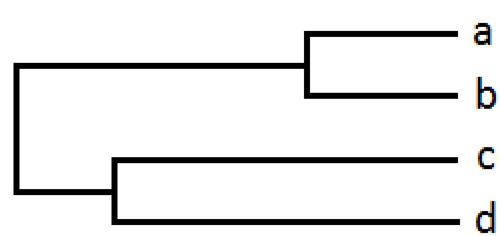




c. Spanning tree

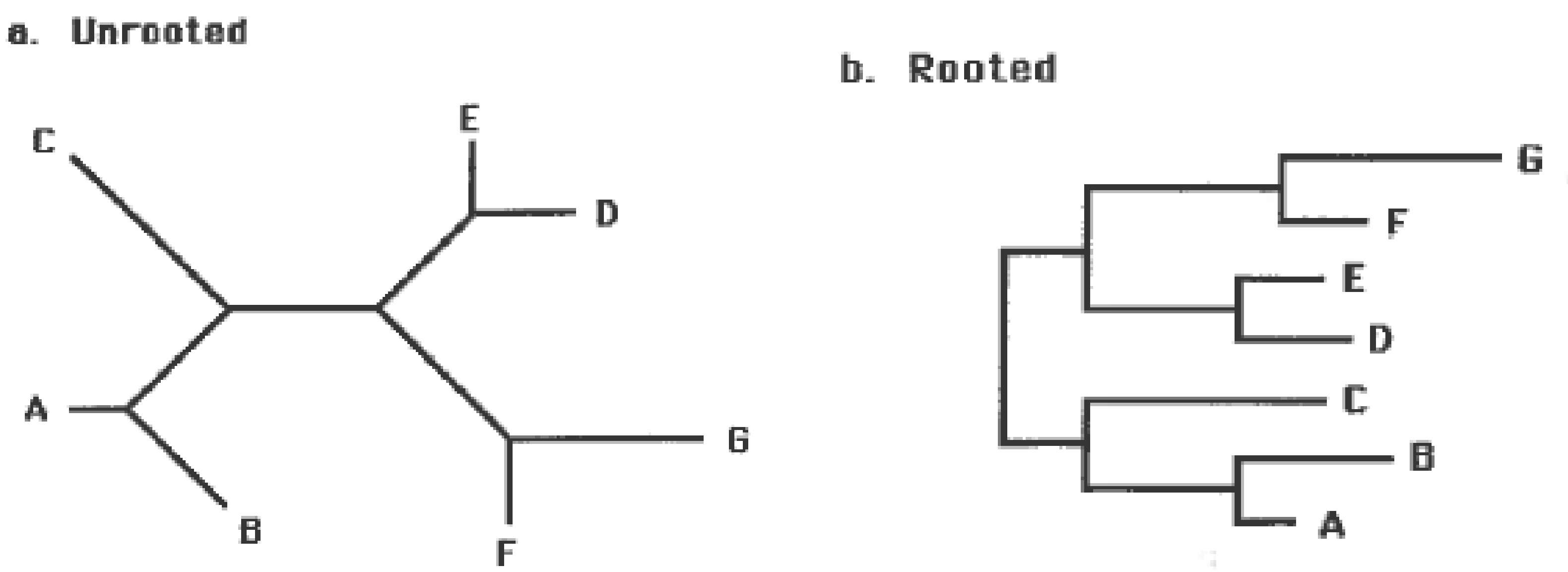


d. Ultrametric tree



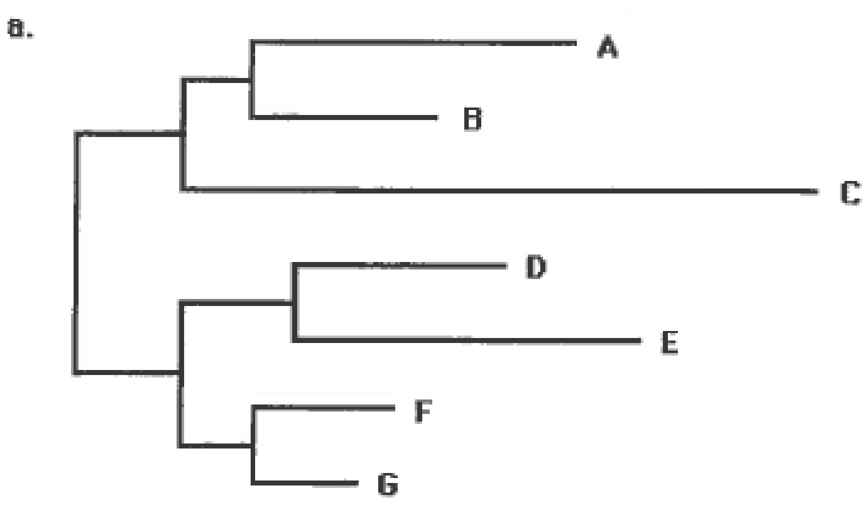
An additive tree (path-length tree) may be displayed in rooted or unrooted form. Placement of the root is arbitrary

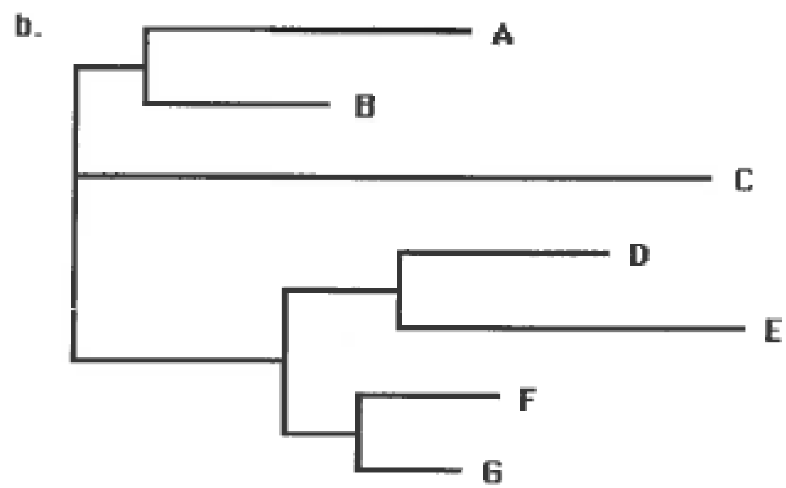
(but Sattath &Tversky suggested placing the root to minimize the variance of the distances from the root to the leaves)



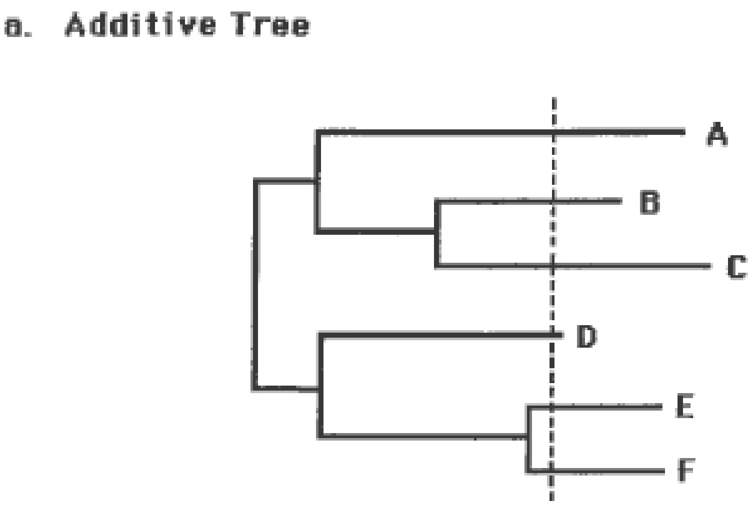
Alternate rootings of an additive tree can change our interpretation of the tree in terms of clusters or common features

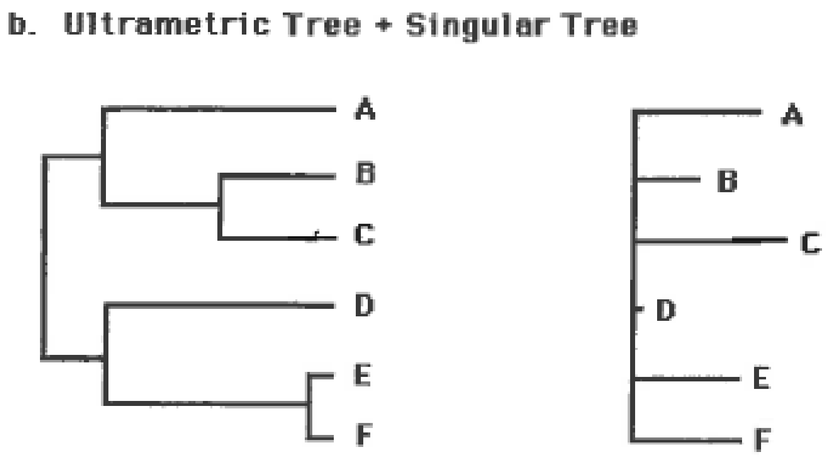
(but tree distances do not change!)





A rooted additive tree can be decomposed into the sum of an ultrametric tree + singular tree (→ common features + unique features)





Mathematical characterization of ultrametric and additive trees, based on path-length distances in tree:

Distances in an ultrametric tree satisfy the <u>ultrametric inequality</u>: $d(x,y) \leq d(x,z) = d(y,z)$ if x & y are "neighbors" in the tree.

(or) $d(i,j) \leq \text{MAX}[\, d(i,k), d(j,k) \,]$ for all i,j,k

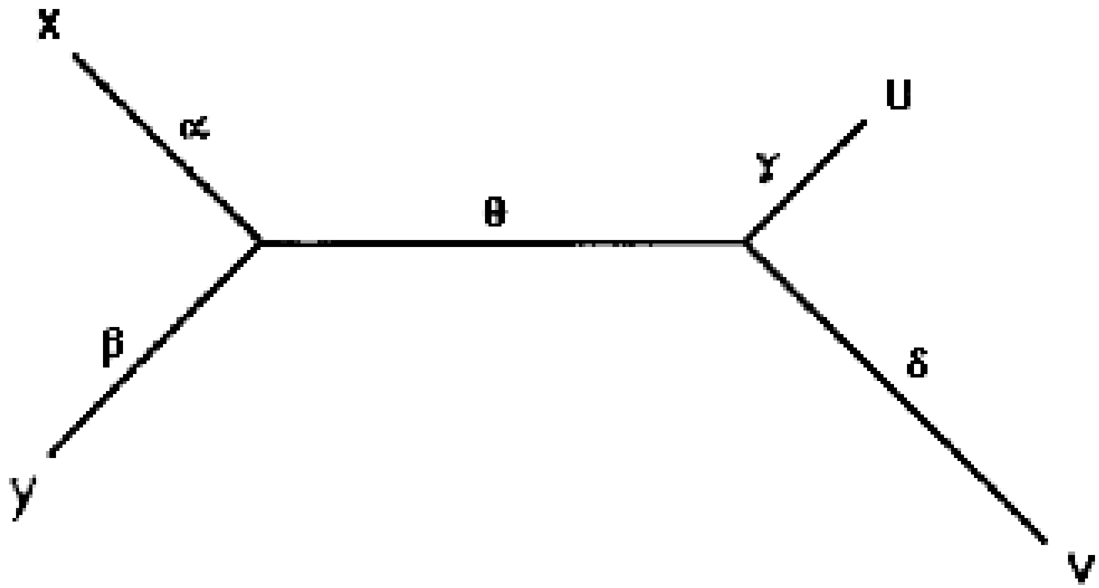
Distances in an additive tree satisfy the <u>additive (or tree) inequality</u>:

$d(x,y) + d(u,v) \leq d(x,u) + d(y,v) = d(x,v) + d(y,u)$

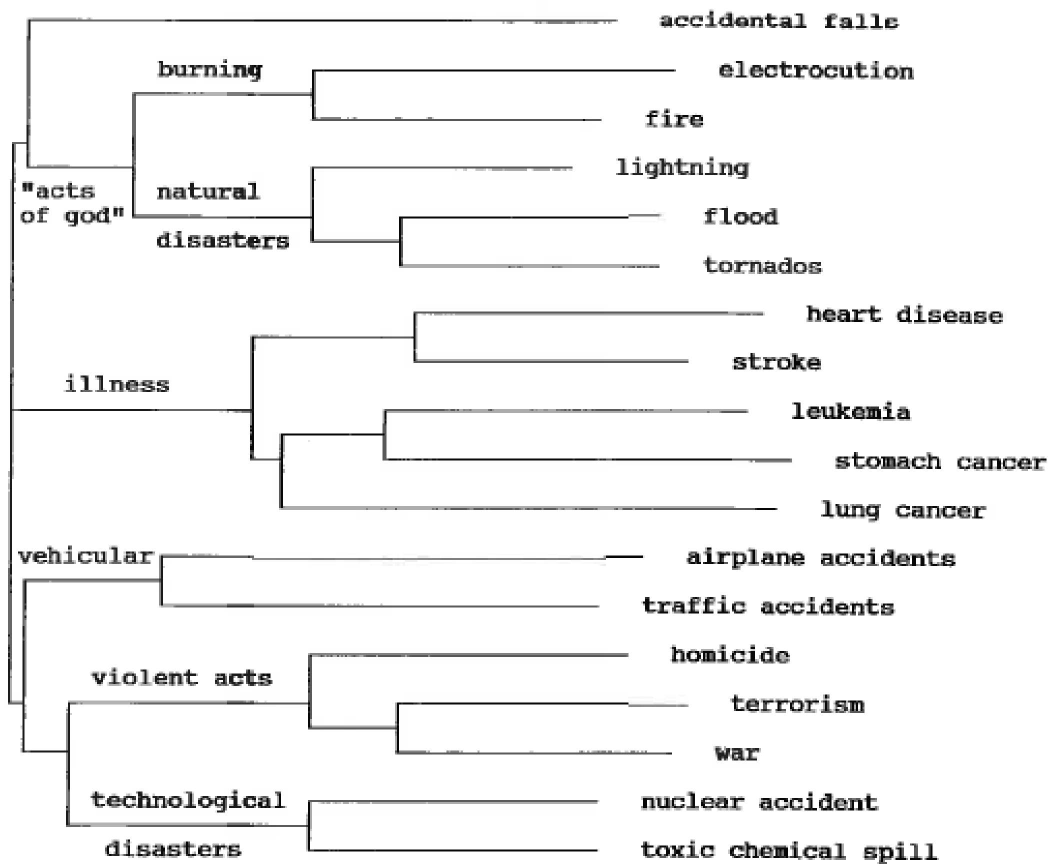
if x & y (and u & v) are "neighbors" in the tree structure

(or) $d(i,j)+d(k,l) \leq \text{MAX}[\, d(x,u)+d(y,v), d(x,v)+d(y,u) \,]$

An additive tree on 4 objects, with labeled parameters



Additive tree of generalization data on societal risks (Johnson & Tversky, 1987)



Social nominations data: finding an optimal organizational hierarchy (from Corter, 1996)

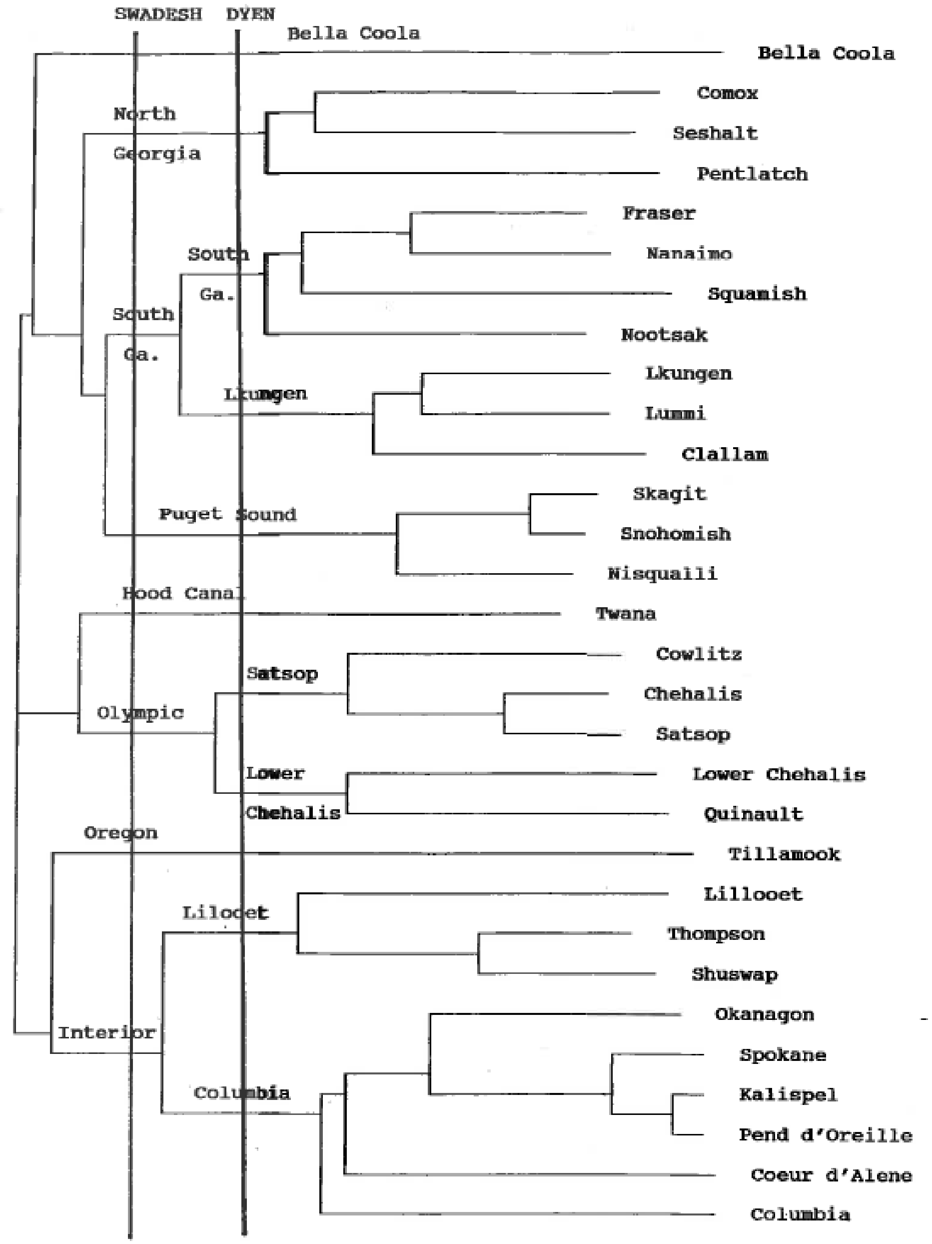
TABLE 3: Summed number of nominations of column departments by respondents in row department.

dept:	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	n
01	0	0	1	2	1	1	0	0	2	1	1	1	0	2	0	1	0	2
02	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	2
03	0	0	0	2	1	1	0	0	0	1	1	0	0	1	0	0	0	2
04	3	0	0	0	1	2	0	1	3	3	0	2	0	1	0	3	0	3
05	0	3	3	2	1	0	0	0	1	1	3	1	0	0	4	2	3	5
06	0	0	1	1	0	0	0	4	0	0	1	0	0	3	0	1	0	4
07	0	0	0	0	0	0	0	0	0	1	0	2	1	0	0	0	1	2
08	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	2
09	2	0	1	3	1	0	0	2	1	0	0	0	0	2	0	1	1	6
10	2	1	3	1	1	0	2	1	2	0	2	2	1	1	1	1	1	3
11	0	3	0	1	3	1	0	0	1	2	0	0	1	0	3	0	0	4
12	1	0	0	1	0	0	3	0	1	3	0	0	1	0	0	1	1	3
13	0	2	0	1	0	2	2	1	0	0	2	1	0	0	2	1	1	2
14	0	0	0	1	0	1	1	1	1	0	1	0	1	0	1	0	0	4
15	0	6	0	0	6	0	0	1	0	0	4	0	0	0	0	0	0	7
16	0	0	0	3	3	2	0	0	1	1	0	0	0	0	0	0	2	5
17	0	1	0	0	1	0	1	0	0	0	1	1	0	0	1	2	0	2

num	label	div	full name of department
01	arts_edu	IV	The Arts in Education
02	clin_psy	II	Clinical Psychology
03	comput_e	IV	Communication, Computing, and Technology in Education
04	curric_t	III	Curriculum and Teaching
05	devel_ed	II	Developmental and Educational Psychology
06	ed_admin	III	Educational Administration
07	hlth_nut	V	Health and Nutrition Education
08	adult_ed	III	Higher and Adult Education
09	lang_lit	IV	Languages, Literature, and Social Studies in Education
10	math_sci	IV	Mathematics and Science Education
11	measrmnt	II	Measurement, Evaluation, and Applied Statistics
12	move_sci	IV	Movement Sciences and Education
13	nurse_ed	V	Nursing Education
14	phil_soc	I	Philosophy and the Social Sciences
15	soc_cnsl	II	Social, Organizational, and Counseling Psychology
16	specl_ed	III	Special Education
17	speech_p	II	Speech and Language Pathology and Audiology

Fitting an additive tree to derive useful partitions and compare them

Data are linguistic similarities (% cognates) of Native American languages, Pacific Northwest area (taxonomies of Swadesh and Dyen shown).



Fitting Ultrametric Trees: references

NOTE: any hierarchical clustering procedure can be seen as fitting an ultrametric tree to proximity data. Node heights could be viewed as determined by the "fusion coefficient" (→non-optimal), or estimated post-hoc via least-squares (Corter, 1996) (→optimal given the tree structure). BUT ideally we would <u>simultaneously</u> estimate the tree structure and the real-valued parameters, which are the node heights, or (constrained) arc lengths. There are (n-1) real-valued parameters in an ultrametric tree.

- Corter, J. E. (1996). *Tree Models of Similarity and Association.* (Sage University Papers series: Quantitative Applications in the Social Sciences, series no. 07-112). Thousand Oaks CA: Sage.
- De Soete, G. (1984). Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification, 1*, 235- 242.
- De Soete, G., Desarbo, W.S., & Carroll, J.D. (1985). Optimal variable weighting for hierarchical clustering: An alternating least-squares algorithm. *Journal of Classification, 2,* 173-192.
- De Soete, G. (1986). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters, 2*, 133-137.

Fitting Additive Trees: references

- Corter, J. E. (1996). *Tree Models of Similarity and Association.* (Sage University Papers series: Quantitative Applications in the Social Sciences, series no. 07-112). Thousand Oaks CA: Sage.
- Sattath, S., & Tversky, A. (1977). Additive similarity trees. Psychometrika, 42, 319-345.
- Corter, J. E. (1982). ADDTREE/P: a PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. *Behavior Research Methods & Instrumentation, 14*, 353-354.
- De Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika, 48*, 621 626.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology & Evolution, 4(4),* 406-425.
- Studier, J. A., & Keppler, K. J. (1988). A note on the neighbor-joining algorithm of Saitou & Nei. *Molecular Biology & Evolution, 5(6),* 729-731.
- Hubert, L. & Arabie, P. (1995). Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology, 48(2),* 281–317.
- Corter, J.E. (1998). An efficient metric combinatorial algorithm for fitting additive trees. *Multivariate Behavioral Research, 33,* 249-272.

Software for least-squares fit of ultrametric trees

Algorithm	Authors / reference	Implementation	Available from
SUMT (sequential unconstrained minimization technique)	De Soete (1986); Fiacco & McCormick (1968)	ls_fit_ultrametric	R {clue}
Iterative projection	Hubert & Arabie (1995)	ls_fit_ultrametric	R {clue}
Iterative reduction	Roux (1988)	ls_fit_ultrametric	R {clue}
(various)	(various)	MatLab code	MATLAB

Software to fit additive trees

Program	Authors / reference	Type	Available from
ADDTREE	Sattath & Tversky (1977)	FORTRAN	---
ADDTREE/P	Corter (1982)	1) PASCAL 2) SYSTAT	columbia.edu/~jec34; netlib.sandia.gov/mds/ SPSS
Neighbor-joining	Saitou & Nei (1987); Studier & Kepler (1988)	MatLab code	MATLAB
GTREE	Corter (1996)	PASCAL	columbia.edu/~jec34; netlib.sandia.gov/mds
SUMT (sequential unconstrained minimization technique)	De Soete (1986); Fiacco & McCormick (1968)	ls_fit_addtree	R {clue}
Iterative projection	Hubert & Arabie (1995)	1) ls_fit_addtree; 2) MatLab code	R {clue} MATLAB
Iterative reduction	Roux (1988)	ls_fit_addtree	R {clue}

What is the proper model?

On the distribution of distances in trees versus spaces

- Holman, E.W. (1972). The relation between hierarchical and Euclidean models for psychological distances. *Psychometrika, 37,* 417-423.
- Critchley, F., & Heiser, W. (1988). Hierarchical trees can be perfectly scaled in one dimension. *Journal of Classification, 5,* 5-20.
- Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial vs. tree representations of proximity data. *Psychometrika, 47*, 3-24.
- Schwartz, G. & Tversky, A. (1980). On the reciprocity of proximity relations. *Journal of Mathematical Psychology,* 22-3, 157-175.

→ (possible) practical significance: Analyzing the distributional properties of your data might be diagnostic between models

→ Research area: hybrid models might be useful, have been proposed