SYLLABUS V1

HUDM5124: MULTIDIMENSIONAL SCALING, CLUSTERING, AND NETWORK MODELS

Spring 2020 Sec. 001 CRN = 51563

TIME: Thursdays 1:00-2:40 PLACE: GDH 535
INSTRUCTOR: James Corter, office 551-C Dodge Hall
OFFICE HOURS: Tuesday 2:45-4 Thursday 2:45-4:30
Office Tel. 212 678-3843 (email: corter@tc.edu)

This course covers methods used to analyze *proximity data*; i.e. data that can be thought of as measuring the similarity, distance, or association between pairs of conceptual entities. Some of the methods can also be applied to ordinary multivariate data (e.g., variables x individuals) or to *preference/dominance data*, resulting from the ordering of entities or choices among them. The types of models that we will employ include spatial models, clustering and discrete-feature models, and graphs / networks. Note that in machine learning / data mining contexts, clustering algorithms are sometimes referred to as "unsupervised learning" techniques, because they are data-driven exploratory techniques that form groups of entities.

REQUIRED TEXTS: NOTE: ** = required * = strongly recommended

Most readings will be made available as course E-reserves. However, the following texts will be used:

**Borg, I. & Groenen, P. (2005; 2ND Ed.) *Modern Multidimensional Scaling: Theory and Applications*. New York/Berlin: Springer. (NOTE: I have designated this text as merely "recommended" at the bookstore, because it is available as an e-book from CU libraries).

**Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis.* (Sage University Papers series: Quantitative Applications in the Social Sciences, series no. 07-44). Thousand Oaks CA: Sage.

**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.

R Resources: RECOMMENDED (if needed)

* 1) Dalgaard, P. (2002). *Introductory Statistics with R.* New York: Springer.

* 2) Braun, W. J., & Murdoch, D. J. (2007). *A First Course in Statistical Programming with R.* Cambridge: Cambridge University Press.

OTHER USEFUL TEXTS AND GENERAL REFERENCES:

Borg, I., Groenen, P. J., & Mair, P. (2013). *Applied Multidimensional Scaling* (SpringerBriefs in Statistics). New York/Berlin: Springer. A relatively elementary introduction to MDS, by the authors of MMDS above. Also available as an e-book from CU libraries.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. (Sage University Papers series: Quantitative Applications in the Social Sciences, series no. 07-11). Thousand Oaks CA: Sage. A classical intro to MDS.

(e-book) Hartigan, John A. (1975). *Clustering Algorithms.* New York: Wiley.

(e-book) Jain, A.K., & Dubes, R.C. (1988). *Algorithms for Clustering Data*. Englewood, NJ: Prentice Hall.

Chartrand, G. (1977). *Introductory Graph Theory.* New York: Dover.

Kassambara, A. (2017). Practical guide to cluster analysis in R. STHDA.

COURSE REQUIREMENTS:

There will be assignments every week. Sometimes this will be a summary of an assigned reading, but often it will be a homework assignment. Over Spring break, I will give a take-home review assignment, with a few discussion questions. Finally, there is a required course project. Ideally, this would be an empirical project, in which you collect and analyze a set of proximity data using the methods discussed in the course. Or, you can take a set of existing proximity data from some source and analyze it using several of the methods, with a careful evaluation and discussion of which method is best and why. In some cases projects may consist of a literature review only, focused on some topic that we touch on only briefly in the course. All project plans should be approved in advance by me – I will require a 1-page project proposal, due by April 1st.

We will use several statistics packages (chiefly R, with some SPSS) and stand-alone software programs.

Grades will be determined by the following: weekly assignments (approx.. 60%); midterm review questions (10%); final project (30%).

OUTLINE OF TOPICS AND READINGS:

Topic 1: Introduction: Overview of MDSCN Techniques and Applications (JAN 23)

Course overview; A taxonomy of data for scaling; Overview of models for similarity data; examples of multivariate, proximity, preference and dominance data; the use of similarity as an explanatory concept in psychology and related fields; the definition of distance.

A. Overview: MDSCN models for EDA and for modeling similarity relations

**Ch. 1 of MMDS NOTE: ** = required * = strongly recommended

**Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*, 390-398. [an overview of MDSC models, including some lesser-known techniques]

*Carroll, J. D. & Arabie, P. (1980). Multidimensional scaling and clustering. *Annual Review of Psychology, 31*, 607-649. [EXTRACT: a taxonomy of data types for scaling]

B. Using similarity as an explanatory concept in psychology and related fields

On the relationship between similarity and confusability:

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Critiques of similarity as an explanatory concept in philosophy and psychology:

Goodman, N. (1977). Seven strictures on similarity. In *Problems and Projects* (1972), pp. 437-447.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

C. Distance models; the concept of a metric space

**MMDS sec. 2.5

*Shreider, Y. A. (1974). *What is Distance?* Chicago: University of Chicago Press. Ch. 1-4, 7, 9

D. USING R

-handouts

Topic 2: BACKGROUND: Review of Matrix Algebra (Cont.) & Principal Components Analysis (PCA) (Jan 30)

READINGS:

**Ch. 7 from MMDS – matrix algebra

**Section 24.1 from MMDS – PCA

**Ch. 6 from Braun & Murdoch (2007): Computational Linear Algebra (on E-Reserves)

-examples using R

Topic 3: Metric Multidimensional Scaling (MDS) (Feb 6)

Metric MDS

We will wrap-up our discussion of PCA, and begin our discussion of MDS by looking at Torgerson's "classical MDS" algorithm (Ch. 12). If time permits, we will begin discussing the current state-of-the-art method for fitting metric (and non-metric) MDS models -- the majorization method embodied in programs like SMACOF (R) and PROXSCAL (SSPS).

1) Torgerson's algorithm for metric MDS

Torgerson, W. S. (1957). *Theory and methods of scaling* (Ch. 11: Multidimensional Scaling). NY: Wiley.

2) Majorization method to fit metric MDS models.

De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification,* 5(2), 163-180.

Groenen, P. J. F, Mathar, R. & Heiser, W. J. (1995). The majorization approach to multidimensional scaling for Minkowski distances. *Journal of Classification* 12, 3-19.

Topic 4 Nonmetric Multidimensional Scaling (FEB 13)

Nonmetric MDS; geometric distance vs. psychological distance

**Kruskal, J. B. (1964a). Nonmetric multidimensional scaling: a numerical method. *Psychometrika, 29*, 115-129.

Kruskal, J. B. (1964b). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1-27.

(later algorithms: Max Likelihood approaches; SMACOF)

Topic 4 (cont). Practical issues and methods for nonmetric MDS (FEB 2)

Topic 5: Three-way (individual differences) MDS (FEB 27)

Insight: individuals may weight psychological dimensions differently

(separable vs. integral; Torgerson article..)

**Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 283-319.

**Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika, 42, 7-67. [ref: ALSCAL method]

Arabie, P., Carroll, J. D., & DeSarbo, W. S. (1987). Three-way scaling and clustering (Sage University Papers series: Quantitative Applications in the Social Sciences, no. 07-65). Thousand Oaks:Sage

Topic 6: Spatial Unfolding (Spatial Models for 3-way proximity data) (MAR 5)

Models for the analysis of choice and preference data

**Carroll, J.D. (1980). Models and methods for multidimensional analysis of preferential choice (or other dominance) data. In E.D. Lanterman & H. Feger (Eds.), Similarity and Choice. Bern: Hans Huber.

*Busing, F. M. T. A., Groenen, P. J. K., & Heiser, W. J. (2005). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika, 70(1), 71-98.

De Soete, G., & Carroll, J. D. (1983). A maximum likelihood method for fitting the wandering vector model. Psychometrika, 48-4, 553-566.

De Soete, G., Carroll, J. D., & DeSarbo, G. W. (1986). The wandering ideal point model: a probabilistic multidimensional unfolding model for paired comparisons data. *Journal of Mathematical Psychology, 30*, 28-41.

Topic 7: Interpretations and Mathematical Models of Psychological Similarity (MAR 12)

Is similarity = spatial distance, feature matching, or a higher-order / derived cognitive task?

The geometric view of similarity (similarity as maps):

*Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review, 75-2*, 127-142.

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology, 7-3*, 572-596.

*Garner, W. R. (1978). Selective attention to attributes and to stimuli. *Journal of Experimental Psychology, 107*, 287-308.

*Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Similarity as feature-matching:

**Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Similarity as the comparison of structured objects:

Larkey, L., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science, 29(6)*, 1061-1076.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 325-340.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

**Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32*, 517-535.

Similarity as transformation:

graph edit distance (ref: TBD)

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition, 87*, 1-32.

Topic 8: Introduction to Cluster Analysis (MAR 12)

Overview; using multivariate data to compute proximity: similarity and association coefficients; variable selection, weighting, and standardization

**Aldenderfer, M.S., & Blashfield, R.K. (1984). *Cluster Analysis.* (Sage University Papers series: Quantitative Applications in the Social Sciences, series no. 07-44). Thousand Oaks CA: Sage.

Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman.

Hartigan, J. A. (1975). Clustering Algorithms. New York: John Wiley & Sons.

Arabie, P., & Hubert, L.J. (1996). An overview of combinatorial data analysis. In P. Arabie, L. Hubert, & G. De Soete (Eds.), *Clustering and Classification.* River Edge NJ: World Scientific.

Gower, J.C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3*, 5-48.

WEEK OF MARCH 16-20: TC SPRING BREAK

Topic 9: Partitioning methods (MAR 26)

Partitioning methods; Assessing cluster solution fit and validity**;** Latent class/Latent cluster models

**Steinley (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology. 29:1*.

Jain, A.K., & Dubes, R.C. (1988). *Algorithms for Clustering Data.* Englewood Cliffs NJ: Prentice-Hall. [extract from Ch. 3]

Milligan, G.W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. Hubert, & G. De Soete (Eds.), *Clustering and Classification.* River Edge NJ: World Scientific.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193-218.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5*, 181-204.

Fowlkes, E.B., Gnadadesikan, R., & Kettenring, J.R. (1988). Variable selection in clustering. *Journal of Classification, 5*, 205-228.

Brusco, M. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods, 9(4)*, 510–523.

Topic 10: Hierarchical Clustering (APR 2)

Techniques for hierarchical clustering (fitting ultrametric trees); Variable selection, weighting, and standardization

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32-3*, 241-254.

**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.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-170.

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.

Topic 11: Ultrametric and Additive trees (APR 9)

**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.

Topic 13: Nonhierarchical Clustering Methods (APR 16)

<u>Overlapping clustering; Extended trees; Multiple trees; Block Clustering</u>

**Shepard, R. N., & Arabie, P. (1979). Additive clustering: representation of similarities as combinations of discrete overlapping properties. Psychological Review, 86, 87-123.

**Corter, J.E., & Tversky, A. (1986). Extended similarity trees. Psychometrika, 51, 429-451.

Carroll, J.D., & Corter, J.E. (1995). A graph-theoretic method for organizing overlapping clusters into trees, multiple trees, or extended trees. Journal of Classification, 12, 283-314.

Lee, M. D. (2002). A simple method for generating additive clustering models with limited complexity. *Machine Learning, 49*, 39–58.

<u>“Block Clustering”: Simultaneous Clustering of the Rows and Columns of a Rectangular Data Set</u>

Mirkin, B.G., 1987. The method of principal clusters. *Automatic Remote Control*, 10, 131–143.

Van Mechelen, I., Bock, H-H., De Boeck, P. (2004). Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research, 13*, 363-394.

Depril, D., Van Mechelen, I., Mirkin, B. (2008). Algorithms for additive clustering of rectangular data tables. *Computational Statistics and Data Analysis, 52*, 4923–4938.

Topic 14: Graph and Network Models of Proximity (modeling approaches) (APR 23)

Schvaneveld, R.W. (Ed.). (1990). PATHFINDER Associative Networks: Studies in Knowledge Organization. Norwood NJ: Ablex.

**Hutchinson, J.W. (1989). NETSCAL: A network scaling algorithm for nonsymmetric proximity data. Psychometrika, 54, 25-51.

Klauer, K.C. (1989). Ordinal network representation: Representing proximities by graphs. Psychometrika, 54, 737-750.

**Klauer, K.C., & Carroll, J.D. (1989). A mathematical programming approach to fitting general graphs. Journal of Classification, 6, 247-70.

Klauer, K.C., & Carroll, J.D. (1991). A comparison of two approaches to fitting directed graphs to nonsymmetric proximity measures. Journal of Classification, 8, 258-268.

Topic 15: Applications of Graph and Network Models to Social Networks (APR 30)

Burt, R. S. (1980). Models of network structure. *Annual Review of Sociology, 6,* 79-141.

**Scott, J. (1991). *Social Network Analysis*. Newbury Park CA: Sage.

***Social Networks* (journal)

**Ronald Breiger, Kathleen Carley, and Philippa Pattison (2003). *Dynamic Social Network Modeling and Analysis: Workshop summary and papers, Workshop on Dynamic Social Network Modeling and Analysis (2002 : Washington, D.C.)* [electronic resource] Washington, D.C. : National Academies Press, c2003.

Watts, D. J. (1998). Collective dynamics of 'small-world' networks. *Nature,* 6(393), 440-442.

Barabási, B. A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*. May 2003, 50-59.

**Watts, D. J. (2004). The "new" science of networks. *Annual Review of Sociology, 30*, 243-270.

**Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports, 2* (doi:10.1038/srep00335).

Nobi Hanaki, Alex Peterhansl, Peter S. Dodds, and Duncan J. Watts (2007). Cooperation in evolving social networks. *Management Science, 7*.

Gueorgi Kossinets and Duncan J. Watts (2006). Empirical analysis of evolving social networks. *Science,*

Memon, Nasrullah (2010). *Data Mining for Social Network Data* [electronic resource]. New York : Springer.

Ajith Abraham, Aboul-Ella Hassanien, Vaclav Snasel (2010). *Computational social network analysis: Trends, tools and research advances [electronic resource].* Dordrecht: Springer.

Martin G. Everett, & David Krackhardt (2012). A second look at Krackhardt's graph theoretical dimensions of informal organizations. *Social Networks, 34,* 159–163.

Final Session: Class Presentations of Projects (written versions of projects due) **(MAY 7)**

Each student will give a 7-minute presentation of their final project (or a status report if it is not completed). Please plan on about a 5-minute presentation, with 2 minutes for questions and comments. You should prepare either presentation slides, or a 2-page handout with tables and figures. Plan to submit your slides in advance (by noon), so that I can have them on a stick or downloadable.

NOTICES:

1. **Accommodations** – The College will make reasonable accommodations for persons with documented disabilities. Students are encouraged to contact the Office of Access and Services for Individuals with Disabilities (OASID) for information about registration. You can reach OASID by email at oasid@tc.columbia.edu, stop by 163 Thorndike Hall or call 212-678-3689. Services are available only to students who have registered and submit appropriate documentation. As your instructor, I am happy to discuss specific needs with you as well. Please report any access related concerns about instructional material to OASID and to me as your instructor.
2. **Incomplete Grades** – For the full text of the Incomplete Grade policy please refer to <http://www.tc.columbia.edu/policylibrary/Incomplete Grades *Revised 7/22/2019*>
3. **Student Responsibility for Monitoring TC email account** – Students are expected to monitor their TC email accounts. For the full text of the Student Responsibility for Monitoring TC email account please refer to http://www.tc.columbia.edu/policylibrary/Student Responsibility for Monitoring TC Email Account
4. **Religious Observance** – For the full text of the Religious Observance policy, please refer to http://www.tc.columbia.edu/policylibrary/provost/religious-observance/
5. **Sexual Harassment and Violence Reporting** – Teachers College is committed to maintaining a safe environment for students. Because of this commitment and because of federal and state regulations, we must advise you that if you tell any of your instructors about sexual harassment or gender-based misconduct involving a member of the campus community, your instructor is required to report this information to the Title IX Coordinator, Janice Robinson. She will treat this information as private, but will need to follow up with you and possibly look into the matter. The Ombuds officer for Gender-Based Misconduct is a confidential resource available for students, staff and faculty. "Gender-based misconduct" includes sexual assault, stalking, sexual harassment, dating violence, domestic violence, sexual exploitation, and gender-based harassment. For more information, see http://sexualrespect.columbia.edu/gender-based-misconduct-policy-students.
6. **Emergency Plan** – TC is prepared for a wide range of emergencies. After declaring an emergency situation, the President/Provost will provide the community with critical information on procedures and available assistance. If travel to campus is not feasible, instructors will facilitate academic continuity through Canvas and other technologies, if possible.
   1. It is the student's responsibility to ensure that they are set to receive email notifications from TC and communications from their instructor at their TC email address.
   2. Within the first two sessions for the course, students are expected to review and be prepared to follow the instructions stated in the emergency plan.
   3. The plan may consist of downloading or obtaining all available readings for the course or the instructor may provide other instructions.
7. **Academic Integrity --** Students who intentionally submit work either not their own or without clear attribution to the original source, fabricate data or other information, engage in cheating, or misrepresentation of academic records may be subject to charges. Sanctions may include dismissal from the college for violation of the TC principles of academic and professional integrity fundamental to the purpose of the College.