HUDM5124 Session 13: Graph / network models of proximity

Overview:

- graph theory
- modeling proximities with graphs
- directed graphs (for asymmetric relations)
- algorithms for fitting graphs to proximity data
- applications

Some graph theory:

- *A graph* consists of a set of *nodes* (vertices) and *arcs* (edges) connecting those nodes (e.g., defined on the power set of V, VxV):

  G = <V,E>

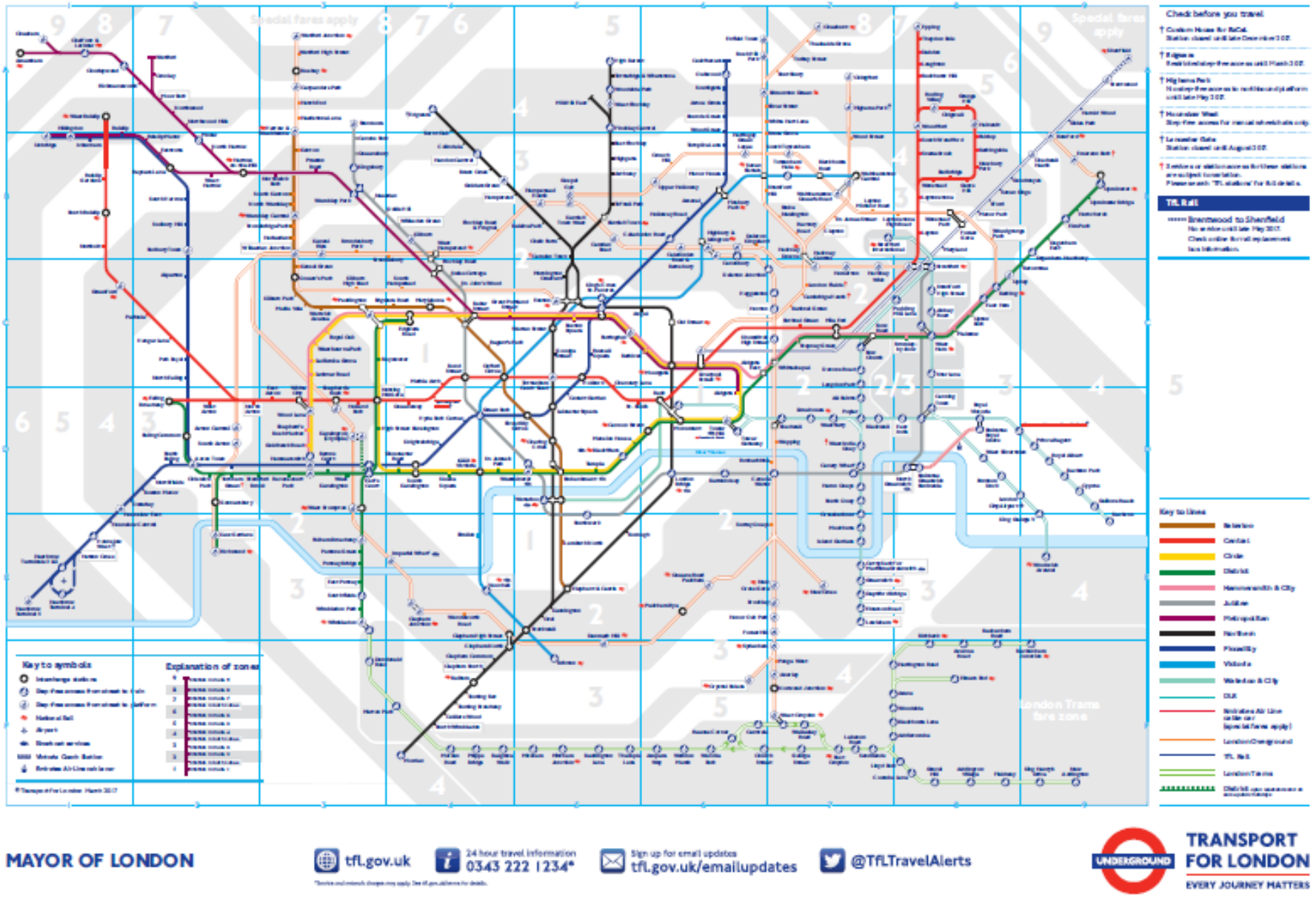
- A *weighted graph (network)* also has a set of weights B , each weight associated with an arc:

  N = <V,E,B>

  The b are weights, usually interpreted as arc lengths (in which case they should be nonnegative)

- If the arcs are directional (i.e., x→y does not necessarily imply y→x), the structure is referred to as a *directed graph (*or, if weighted, a *weighted directional graph)*

London underground – official day map



An application of graphs in education: “Conceptual maps”

(Chi & Koeske, 1983)

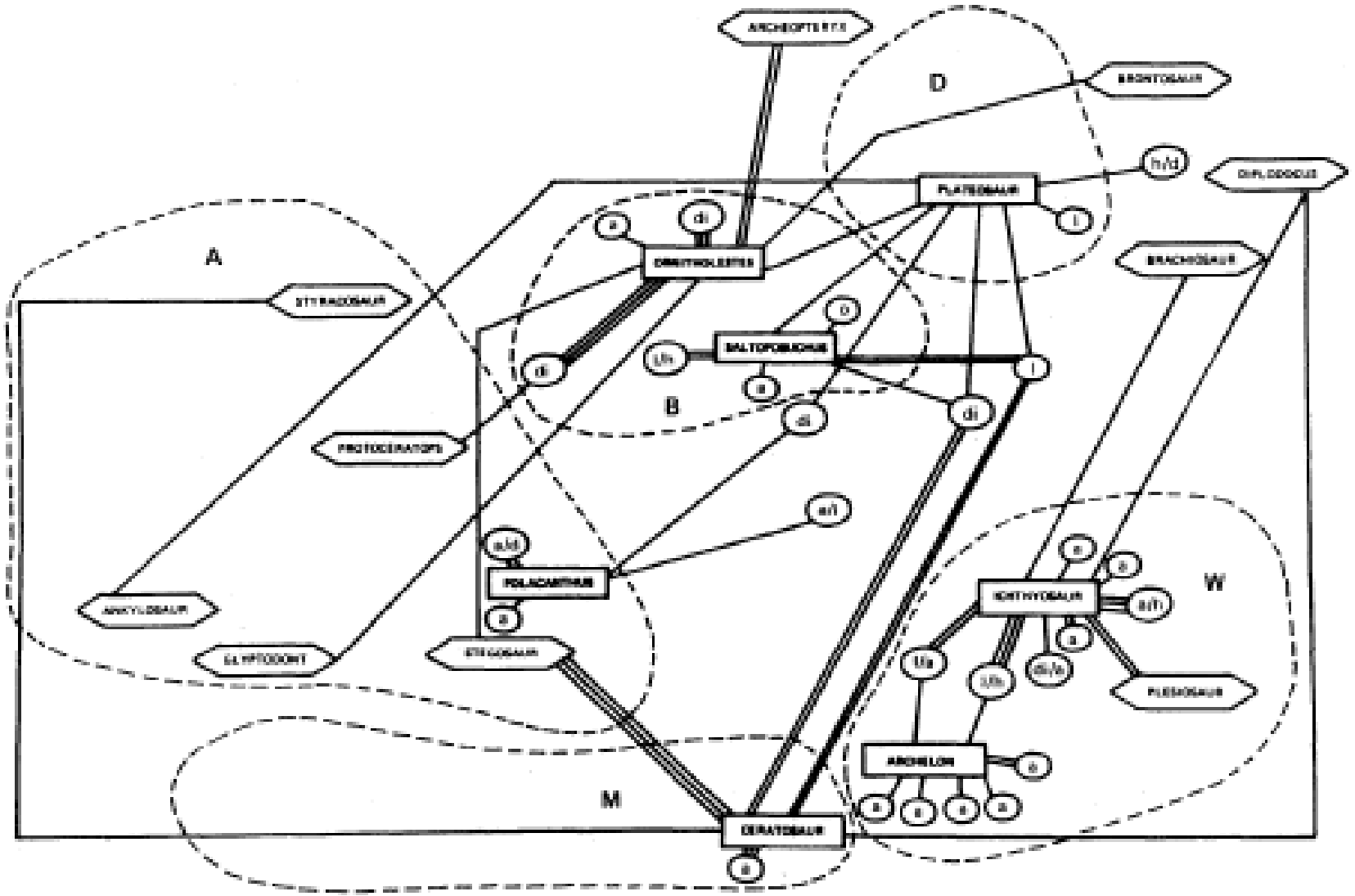
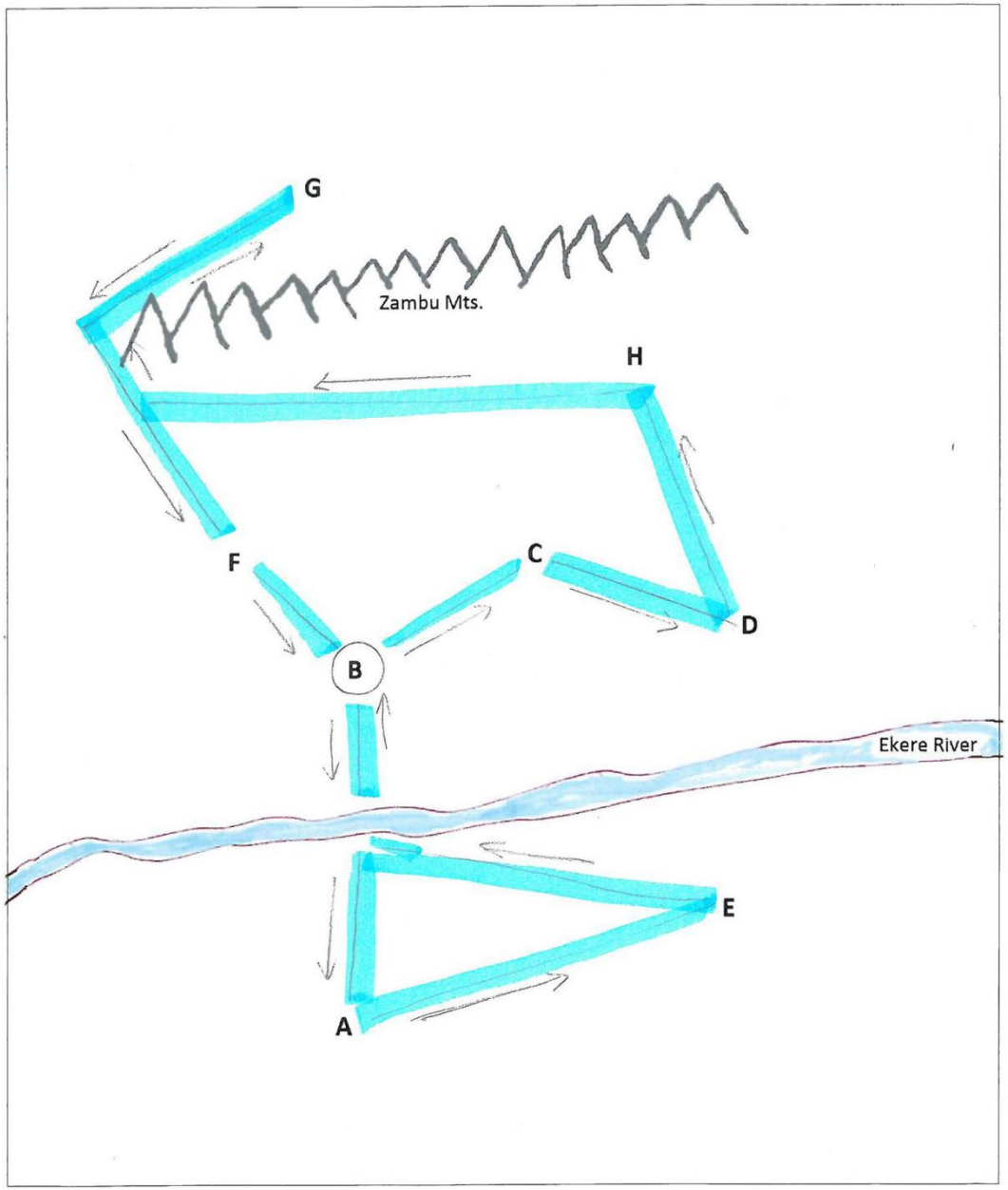


Figure 2. Network representation for the target dinosaurs from the lesser known list. (A = armored; B = bird or egg eaters; D = duckbills; M = giant meat eaters; W = water dwellers; a = appearance; d = defense mechanism; di = diet; h = habitat; l = locomotion; n = nickname; o = other.)

Route planning (Corter et al., 2015)



Transitions between stages of solving mathematical word problems (Zahner & Corter, 1990)

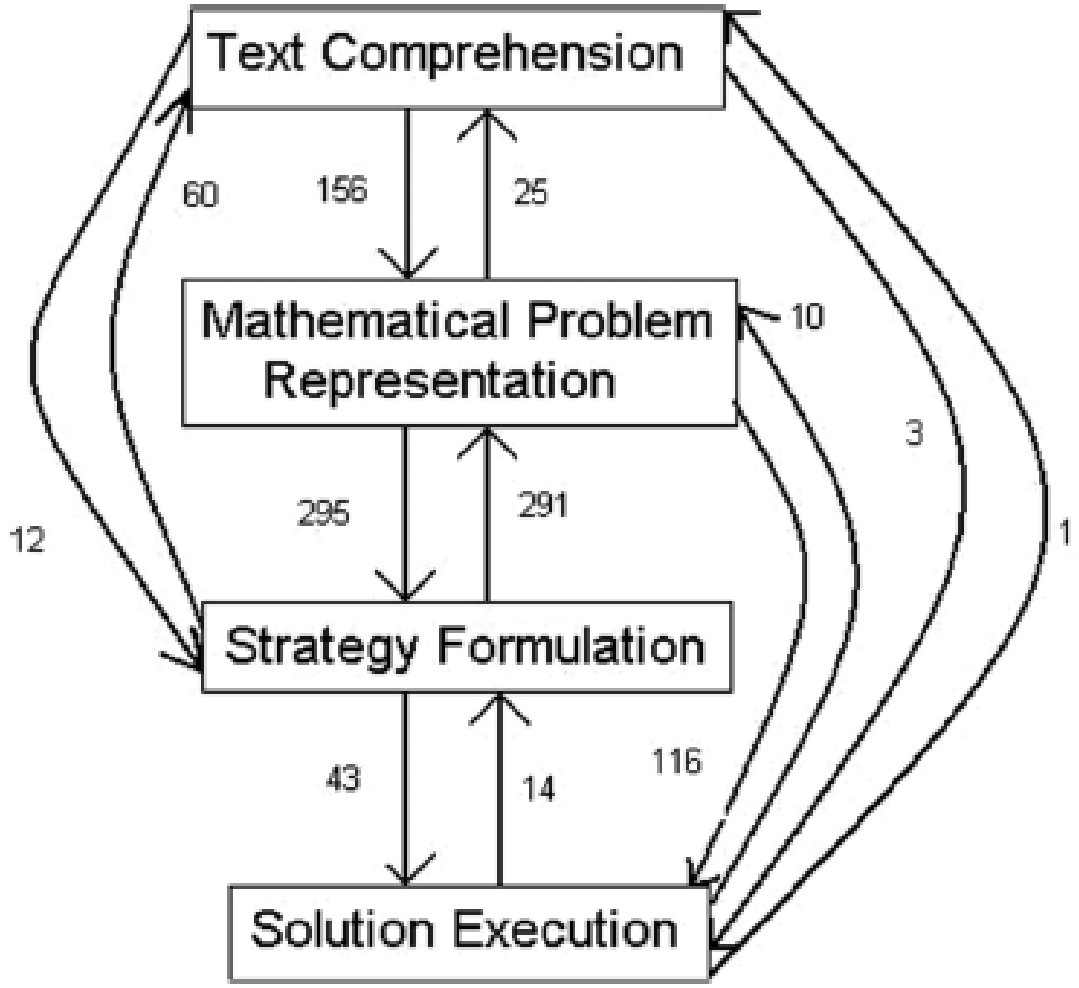
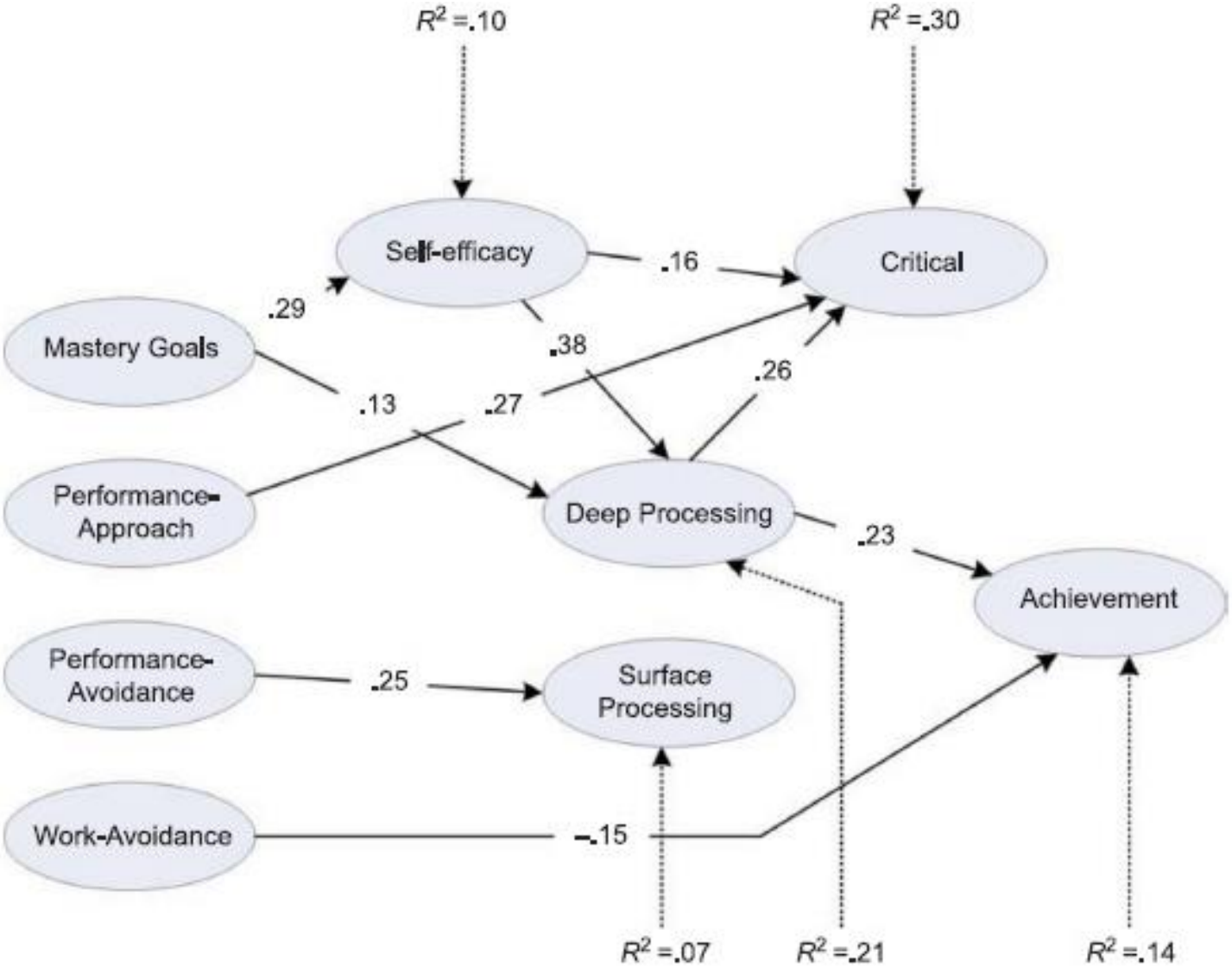


Fig 2 from: Phan, H. P. (2009). Relations between goals, self-efficacy, critical thinking and deep processing strategies: A path analysis. *Educational Psychology, 29(7)*, 777-799.

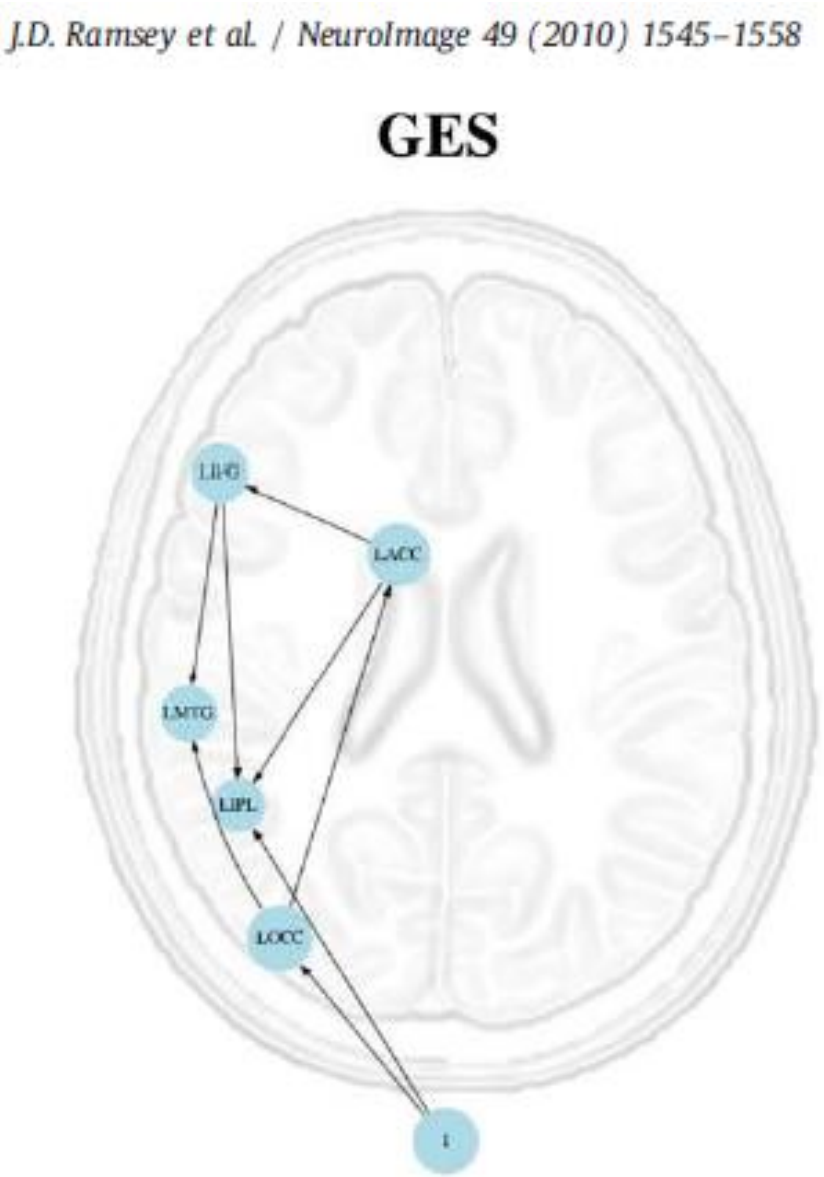


Modeling functional connections of brain areas

..Meek's Greedy Equivalence Search (GES) (Meek, 1997) begins with an empty graph whose vertices are the recorded variables and proceeds to search forward, one new connection at a time, over Markov Equivalence classes of DAGs. Each class of models with an additional edge is scored using BIC (Schwarz, 1978): -2ln(ML) + k ln (n), where ML is the maximum likelihood estimate, k is the dimension of the model (in our cases, the number of directed edges plus the number of variables), and n is the sample size. The algorithm searches forwards from the empty graph until no improvement in BIC score is possible, and then backwards, and outputs a description of a Markov Equivalence class.

..The software we used is available as freeware in the TETRAD IV suite of algorithms with a graphical user interface at *www.phil.cmu.edu/projects/tetrad*.

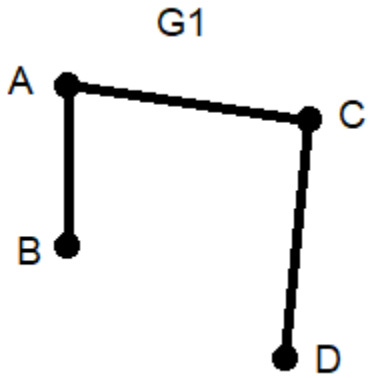
From: Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage, 49*, 1545–1558.

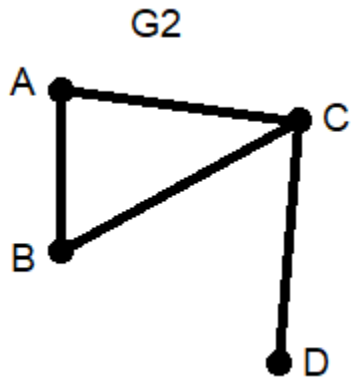


undirected graphs

- Consider a *graph* G = (V,E). Each arc is a pair of nodes (*x,y*) representing a bidirectional (symmetric) connection from *x* to y.
- The *degree* of node *x* is the number of distinct arcs incident to x.
- A *path* between node *x* and node y is a sequence of arcs (x,z)(z,w)...(v,y) that one can follow through the graph from *x* to y (or vice-versa). If a path exists between every pair of nodes (x,y) in the graph, the graph is said to be *connected*.
- A (maximal) completely connected subgraph of G is called a *clique.*
- A *cycle* is a path (length >2) that begins and ends on the same node.

Examples:
G1 = <V,E1>
V = {A,B,C,D}
E1 = {ab,ac,cd}
G2 = <V,E2}
E2 = {ab,bc,ac,cd}





undirected graphs (cont.)

- In an unweighted graph, the *path length* between *x* and y is simply the number of arcs in the path. In a weighted graph or network, the length of a path between x and y may be defined as the sum of the weights of the arcs in that path.
- In a graphs that contain cycles, there may be more than one path connecting two nodes. Usually, we define the distance between two nodes in a graph as the *minimal (shortest) path* between them. This type of graph is referred to as a *geodetic* graph (and each such path is a *geodesic)*

directed graphs

- Consider a *directed graph* G = (V,E), E = VxV. Each arc is an ordered pair of nodes (*x,y*) representing a directional connection from *x* to y.
- The *out-degree* of a node *x* is the number of distinct arcs *(x,*) (i.e.,* the number of links from *x*). The *in-degree* of node *x* is the number of distinct arcs (**,x*) (i.e., the number of links to *x*).
- A path from node *x* to node y is a (directed) sequence of arcs (x,z)(z,w)...(v,y) that one can follow through the graph from *x* to y. In a directed graph, the existence of a path from *x* to y does not imply a path from y to *x*.
- In an unweighted graph, the path length from *x* to y is simply the number of arcs in the path. In a weighted graph or network, the distance of a path between x and y may be defined as the sum of the weights of the arcs in the path.
- A *cycle* is a directed path (length >2) that begins and ends on the same node
- Usually, we define the distance between two nodes in a graph as the *minimal (shortest) path* between them. If a path does not exist from x to y, we may define the distance as =infinity for mathematical convenience.
- *A strongly connected component* of directed graph G is a set of nodes such that for any pair of nodes *x* and y in the set there is a path from *x* to y.

Graph Processing / Algorithms

We generally model the proximity between x and y with the *shortest path (or geodesic)* connecting x and y in the weighted graph.

- Finding the shortest path between two nodes in a (weighted or unweighted) graph G is a standard problem in graph theory. It is non-trivial when G is large (e.g., Goldberg & Harrelson, 2003). One of the most efficient algorithms is known as “A* search” (Hart, Nilsson & Raphael, 1968).
- Another standard problem is finding all the *cliques* of a graph (e.g., Bron & Kerbosch, 1973). A clique is a set of nodes that are all connected to each other with direct links. Formally, it is a maximal complete subgraph of G.
- There has been much recent work on computing statistical properties of graphs and networks, in the (somewhat) new field of network science

Modeling proximities with a (weighted) graph

- Finding a weighted graph G that best approximates a proximity matrix is computationally intensive. Klauer (1989) investigated this problem for undirected graphs (for symmetric proximity data) and directed graphs.

  →What is the network structure? What are the optimal arc lengths?

- Hutchinson (1989) presented an effective algorithm, NETSCAL, to fit directed graphs to asymmetric proximities.

  →NETSCAL merely identifies "needed" arcs, then must attempt to estimate optimal(?) arc lengths

- Schvaneveld (1990) generalized Hutchinson's algorithm for both undirected and directed graphs, and termed the generalized algorithm "Pathfinder" (see also Schvaneveldt, Durso & Dearholt, 1989)

Software for fitting graph / network models to proximity data

Algorithm	Authors / reference	Implementation	Available from
MAPNET	Klauer & Carroll (1991)		??
NETSCAL	Hutchinson (1989)	SAS macro	Hutchinson
Pathfinder	Schvaneveld (1990)	various	R ??

REFERENCES:

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

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

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*, 247-270.

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.

Some general references on graph theory

Harary, F. (1969). *Graph theory*. Reading, Mass: Addison-Wesley.

Chartrand, G. (1985). *Elementary Graph Theory.* New York: Dover.

Gibbons, A. (1985). *Algorithmic Graph Theory*. Cambridge: Cambridge University Press.

Jørgen Bang-Jensen, J., & Gutin, G. (2008). *Digraphs: Theory, Algorithms and Applications (2nd Ed.), Springer Monographs in Mathematics*. London: Springer-Verlag.

Newman (2010). *Networks: An Introduction*. Oxford University Press.

Algorithms:

Goldberg, A. V. & Harrelson, C. (2003). Computing the shortest path: A* search meets graph theory. *Microsoft Research Technical Report MSR-TR-2004-24.*

Bron, C. & Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communcations of the ACM*, 16(9), 575–577.

The NETSCALE algorithm (Hutchinson, 1989)

Goals: to offer a data-driven method to derive network structures from proximities; to offer a model for asymmetric proximity data

- In modeling a distance matrix, we are generally interested in parsimonious representations.
- Different schemes can be proposed for defining the path length between two nodes in a network. The most common is probably the *minimum pathlength metric*; structures with this distance function are said to be *geodetic*.
- This distance function, and the metric axioms, have implications for (parsimonious) network structure. For example, an arc is called *redundant* if its deletion changes no distances in the network. A network is *irreducible* if it contains no redundant arcs.

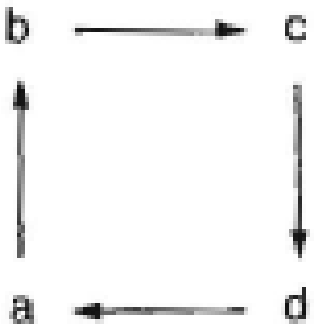
Examples: modeling distances

RANK ORDERED DISTANCES :

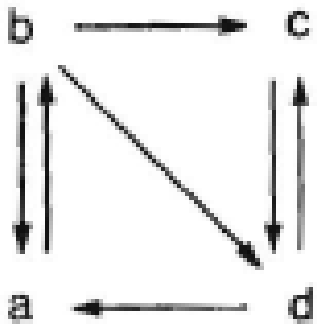
	1	2	3
ab	1.0•	1.0•	1.10•
bc	1.5•	1.5•	1.15•
ac	2.5	2.5	1.25•
cd	3.0•	3.0•	1.30•
da	4.0•	4.0•	1.40•
bd	4.5	4.3•	1.45•
db	5.0	5.0	1.50•
ad	5.5	5.3	1.55•
dc	6.5	6.4•	1.65•
ca	7.0	7.0	1.70•
cb	8.0	8.0	1.80•
ba	8.5	8.4•	1.85•

Network realizations:

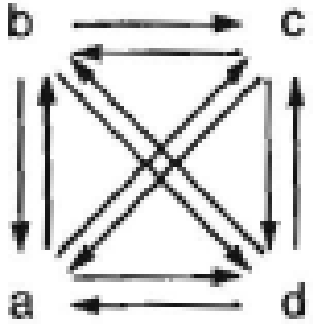
N1



N2



N3



Definitions & Terminology (Hutchinson , 1989)

- The *adjacency matrix* of a network contains ones in each cell associated with an ordered pair for which there is an arc connecting the first vertex of the pair to the second. Zeroes are in all other cells.
- The *reachability* matrix of a network contains ones in each cell associated with an ordered pair for which there is a path from the first vertex of the pair to the second. Again, zeroes are in all other cells.
- The *distance matrix* of a network contains the distances associated with each ordered pair of vertices. If there is no path connecting one vertex to another, the corresponding distance is considered to be infinite. Thus, the distance matrix can be mapped into the reachability matrix by setting all finite values equal to one and all infinite values equal to zero.
- A real two-way one-mode matrix D is called *realizable* if there is a network G such that the values of D are equal to the corresponding distances in G.

Theoretical Results: NETSCAL (Hutchinson, 1989)

Theorem 1. Let D be a real one-mode matrix with entries d_{xy}. The necessary and sufficient conditions for D to be realizable are

$$d_{xx} = 0, \tag{1}$$

$$d_{xy} > 0, \qquad \text{for all } x \neq y, \tag{2}$$

$$d_{xy} \leq d_{xz} + d_{zy}. \tag{3}$$

Theorem 2. Let G be an irreducible representation of some matrix D satisfying (1) through (3).

The arc (x, y) is present in G if, and only if,

$$x \neq y \quad \text{and} \quad d_{xy} < \min \{d_{xz} + d_{zy} : z \neq x, y\}. \tag{4}$$

Corollary 1. Let G be an irreducible representation of some matrix D satisfying (1) through (3).

If

$$d_{xy} \leq \min \{\max \{d_{xz}, d_{zy}\} : z \neq x, y\}, \qquad \text{then } (x, y) \text{ is an arc in } G. \tag{5}$$

Steps in NETSCAL algorithm (overview)

1) Using corollary 1, the structure of the network is determined (i.e., the necessary arcs)
2) The arc lengths are estimated by fitting the model distances to the dissimilarities via a generalized power function for distances: $d_{xy} = (\gamma + D_{xy})^{\lambda}$.

(where D is the linearly transformed raw data)

The optimization of parameters lambda and gamma is accomplished by STEPIT (Chandler, 1973)

Pathfinder algorithm
(Schvaneveld et al., 1988)

Extension of NETSCAL concepts to data of other measurement types, with a generalized definition of path-length; directed or undirected graphs

1) Path length: uses Minkowski r-metric

   r=1: additive path-length distances

   r=∞: path distance defined as MAX link wt.

   (this corresponds to NETSCAL definition)

2) *q* parameter: max # links used in path-length computation

A trivial representation of a set of distances (dissims) by a graph:

Schvaneveld et al.: “datanet” = define node for each “object”, define arc between each pair of nodes; set arc length = dissim between the two objects

How to “prune” (i.e., eliminate arcs)?

General Goal: can we find a parsimonious representation that perfectly (or imperfectly) represents the dissimilarities?

PATHFINDER, NETSCAL give reasonable (heuristic) solutions. Klauer & Carroll (1991) offer a method to *optimize* structure and weights.

References

Graph / Network models of proximity:

- 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.

Application: word associations

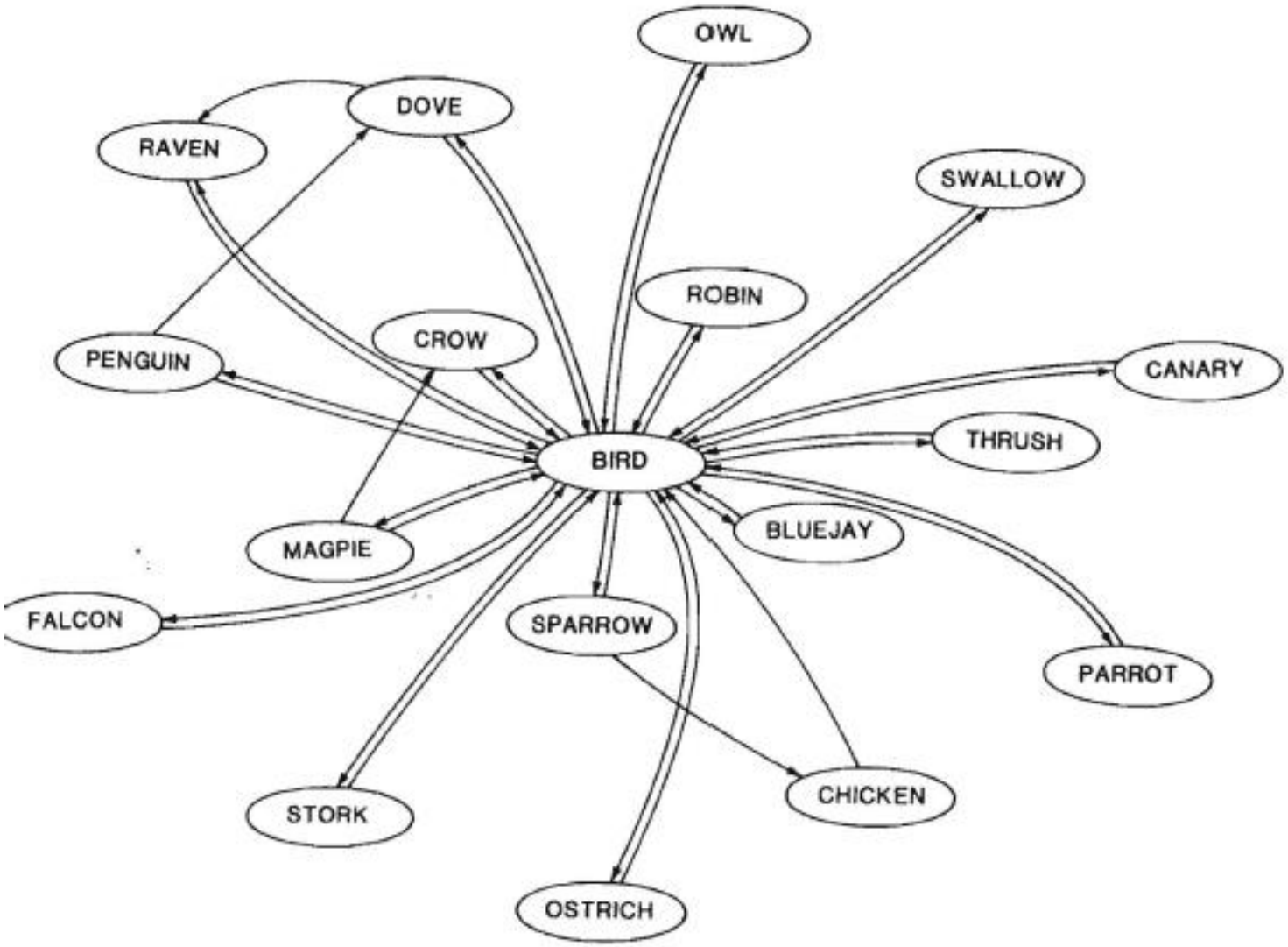


FIGURE 8
NETSCAL solution for word association data for birds.

Application: word associations

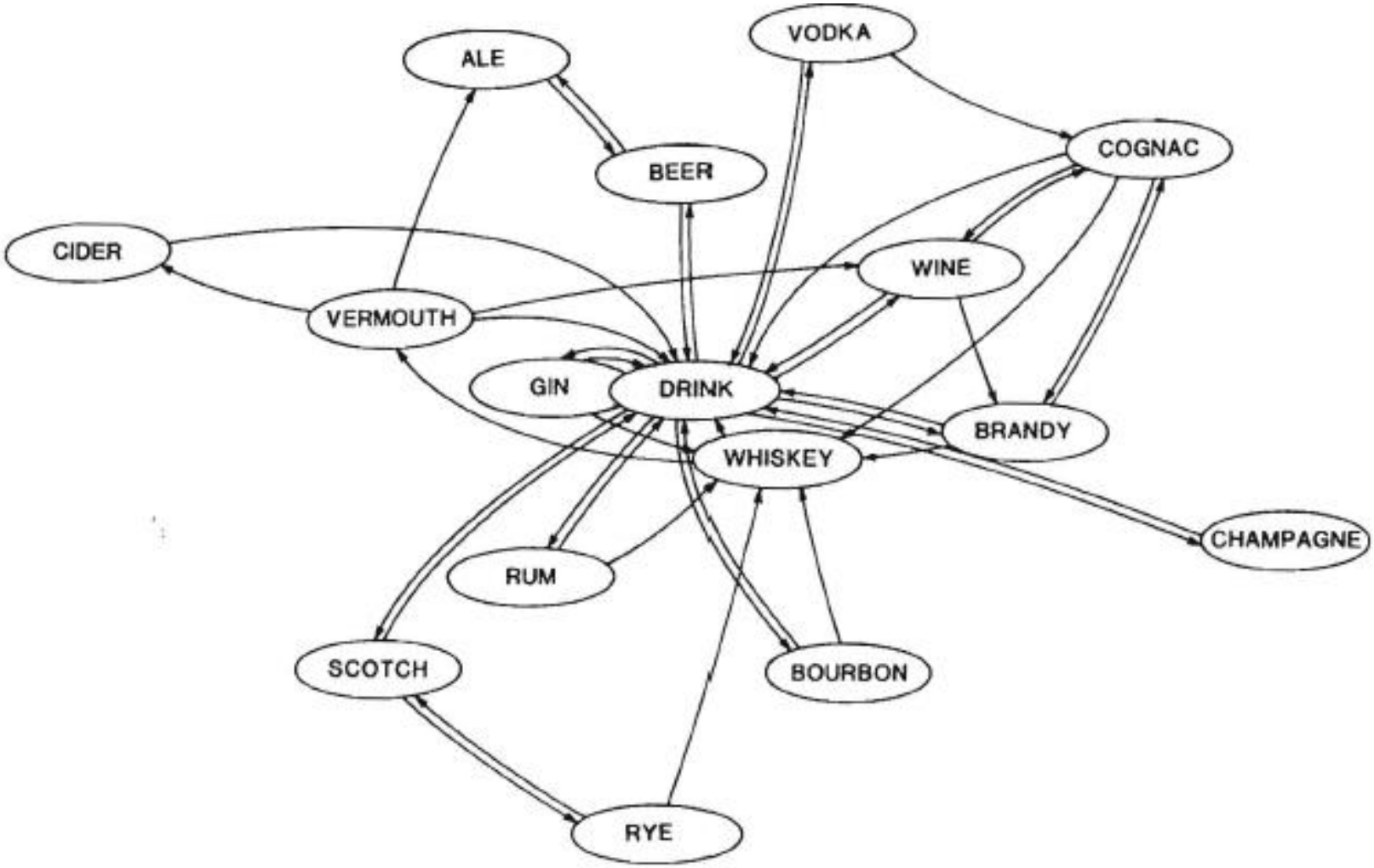


FIGURE 9
NETSCAL solution for word association data for drinks.

Social Networks:

- For social networks, structure is about *connections*, not shared features or (purely) spatial position
- Social networks are naturally represented as a graph, with nodes representing actors and lines between nodes (arcs) representing social connections
- Social network data is usually binary: either a pair of nodes is directly connected, or is not. Thus, “fitting” of the network is usually trivial (→“construction”)
- Thus, we usually do not assess fit to a proximity matrix, rather the focus is on computing and interpreting general graph properties of the constructed network

An application of graph models: Social Networks (more next week)

References:

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.

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).

→ For current work: *Social Networks* (journal)

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

GRAPH / NETWORK MODELS EXAMPLE (fitting directed graphs):

Apply the NETSCAL algorithm to the social nominations data below.
NOTE: first convert these proportions to dissimilarities by subtracting each entry from 1.

Extract from Table 3 of Corter (1995): Social nominations data for a subset of academic departments. Entries are the summed number of nominations of column departments by respondents in row department, converted to proportions by dividing by total number of nominations from a department.

	02: clin_p	05: dev_p	10: math	11: meas	15: soc_p	N
02: clin_psy	--	1/2	0	0	1/2	2
05: devel_ed	3/5	--	1/5	3/5	4/5	5
10: math_sci	1/3	1/3	--	2/3	1/3	3
11: measrmnt	3/4	3/4	2/4	--	3/4	4
15: soc_cnsl	6/7	6/7	0	4/7	--	7