Chapter 7

Empirically Defined Semantic Relatedness and Category Judgment Time*

Nancy Jaworski Cooke

Cognitive psychologists generally believe that information is represented in memory in an organized fashion (but see Landauer, 1975) and that the basis for this organization is association. If this is the case, tasks that require memory retrieval should be affected by associations among the items. In fact, much empirical evidence has suggested that the organization of stimulus materials affects performance on memory-related tasks. It has been shown that list recall is facilitated when the list is organized according to common associations (Jenkins, Mink, & Russell, 1958; Jenkins & Russell, 1952). In addition, for cases in which common associations are not obvious, subjects will impose their own organization on the material in order to remember it (Tulving, 1962). Others have found significant correlations between recall order and judgments of pairwise relatedness, again strengthening this link between memory retrieval and association (Caramazza, Hersh, & Torgerson, 1976; Cooke, Durso, & Schvaneveldt, 1986; Schwartz & Humphreys, 1973). Finally, Meyer and Schvaneveldt (1971) found that subjects judged the lexicality of pairs of related words faster than unrelated pairs, suggesting that the related context provided by the first word primed or activated associated concepts in memory, thus facilitating retrieval of the second word.

In recent years various models of memory organization have been proposed to describe how concepts and relations could be represented in order to account for relatedness effects in memory tasks. In network models of memory, concepts are represented as nodes in a graph structure, and relations between concepts as links between nodes (Anderson, 1983; Anderson & Bower, 1973; Collins & Loftus, 1975; Collins & Quillian, 1969; Glass & Holyoak, 1975; Quillian, 1969; Sowa, 1984). The degree of relatedness between two concepts can be represented by either the strength of a link, or the number of shared links. Links are also labeled according to the type of semantic relation (e.g., property relations are represented by *has-a* links and superordinate relations by *is-a* links). Although specific instantiations of these network models differ, the general network structure has received empirical support (e.g., Collins & Loftus, 1975; Collins & Quillian, 1969).

Alternative models of memory have been proposed by other investigators (McCloskey & Glucksberg, 1979; Smith, Shoben, & Rips, 1974) that involve a comparison process operating over feature sets. In their model, Smith et al. (1974) represent a concept by a set of features, some of which are defining features possessed by all category members (e.g., the presence of *wings* is a defining feature of *birds*). Characteristic features, on the other hand, are shared by many, but not all, category members (e.g., *flying* is a characteristic feature of *birds*). In this model specific relations, such as *is-a* and *has-a*, are not

*I would like to thank Russell Branaghan and Roger Schvaneveldt for their assistance in designing and carrying out this study.

represented explicitly as they are in network models, but relations are derived by assessing the overlap among feature sets. Like network models, feature comparison models can also account for a variety of empirical findings, such as typicality judgments (Rosch, 1973).

In summary, empirical evidence exists that supports both network and feature comparison models. Furthermore, in cases in which the models do not adequately explain the data, they can be altered slightly by changing the structure (i.e., adding features or links) or the processing assumptions. Which model presents the most accurate description of how memory is actually organized? Unfortunately, if these models are intuitively or logically based, it is difficult to evaluate them in terms of psychological meaningfulness. That is, if a particular model fails to account for certain experimental results it could be due to misguided intuitions on the part of the creator of the specific model, a discrepancy between relations used by humans and logical relations, or inadequacies of the general model.

One solution to the problem of evaluating various arbitrary models of memory organization is to generate the models empirically. Psychological scaling techniques, such as multidimensional scaling (MDS) and Pathfinder network scaling, generate various structures from subjects' judgments of relatedness and it is assumed that these judgments, and consequently the resulting structure, reflect information about memory organization. To the extent that the empirically derived structure adheres to the definitions and assumptions of the general memory model, it can be used to evaluate the psychological meaningfulness of that model.

Empirically Derived Structures

MDS algorithms take pairwise proximity estimates for a set of concepts and generate *d*-dimensional spatial layouts of those concepts, where the value of *d* is decided by the experimenter (Kruskal, 1977; Kruskal & Wish, 1978; Shepard, 1962). Dimensions are assumed to reflect features along which the concepts vary, and psychological proximity is represented by distance between concepts in the spatial representation. MDS solutions share many of the characteristics of feature comparison models and in fact have been used by theorists who adhere to this type of model. For instance, several investigators have used MDS procedures as means of empirically identifying semantic features (Rips, Shoben, & Smith, 1973; Rumelhart & Abrahamson, 1973; Shoben, 1976). Additionally, results from various studies have indicated that the distances derived from MDS solutions are psychologically meaningful in that they correspond to free recall order (Caramazza et al., 1976), analogy completion (Rips et al., 1973; Rumelhart & Abrahamson, 1973), similarity judgment time (Hutchinson & Lockhead, 1977), categorical judgment time (Caramazza et al., 1976; Rips et al., 1973; Shoben, 1976), and judgments in an inductive reasoning task (Rips, 1975).

Recent development of techniques like Pathfinder for empirically deriving network structures (Chi & Koeske, 1983; Fillenbaum & Rapoport, 1971, Friendly, 1977; Hutchinson, 1989; Schvaneveldt & Durso, 1981; Schvaneveldt, Durso, & Dearholt, 1985) has enabled researchers to investigate the psychological meaningfulness of network representations of memory. It is important to emphasize the fact that Pathfinder can use the same type of input as MDS. As a result, the same set of proximity estimates can be used to derive both spatial and network representations which can then be compared in terms of psychological meaningfulness. Because the models are algorithmically based, the definitions and assumptions of each model are well-specified and thus, testable.

There have been several studies conducted (many of which are discussed in this book) that have compared MDS structures to Pathfinder network structures by evaluating the effectiveness of the respective structures in terms of a particular application, such as training, selection, or interface design (e.g., Roske-Hofstrand & Paap, 1986b; Schvaneveldt, Durso, Goldsmith, Breen, Cooke, Tucker, & DeMaio, 1985). For the most part, results have indicated that each representation has particular strengths and weaknesses depending on the application. In this chapter, a theoretically driven comparison of MDS and Pathfinder representations is presented. By evaluating the two types of structures in terms of their correspondence to basic memory-related tasks, one may gain an understanding of the differences between the two scaling techniques that impacts on their usefulness in certain applications. In addition, such an understanding should shed light on the basic question of memory organization as defined by network and feature comparison theories.

Empirically Derived Structures and Recall

In a recent study, Cooke et al. (1986) compared MDS and Pathfinder structures in terms of their correspondence to performance on serial and free recall tasks. For the serial recall task, lists of 13 words were constructed so that they corresponded to either a Pathfinder network or MDS representation of those words. That is, successive list items were either linked in the network and distant in MDS, or close in MDS and not linked in the network. Unorganized control lists were also constructed in which successive items were both unlinked and distant in MDS. Results indicated that subjects learned the network list in fewer trials than its control, but there was no difference between the MDS list and its control. Furthermore, subjects learned the network list in fewer trials than the MDS list. These results suggested that the network organization captured information about memory organization that was particularly useful for serial recall.

In a second study, Cooke et al. (1986) extended these findings to a free-recall paradigm in which subjects recalled a list of 13 items in any order. Proximities derived from recall order and averaged across all subjects correlated significantly with similarity ratings, proximities derived from the MDS solution, and proximities derived from the Pathfinder network. However, partial correlations revealed that unlike the MDS proximities, the Pathfinder proximities correlated significantly with recall order even with the effect of the original similarity ratings partialled out. These results suggest that although both Pathfinder and MDS structures correspond to performance in recall tasks, Pathfinder extracted psychologically valid information about the structure of memory that was not explicit in the original ratings.

In speculating about the extra information that Pathfinder captures, Cooke et al. (1986) pointed out that the Pathfinder algorithm weights judgments about related pairs more heavily than unrelated pairs. MDS on the other hand, weights all judgments equally. Secondly, it is possible that subjects provide more accurate estimates of relatedness for related pairs than unrelated pairs (cf. Roske-Hofstrand & Paap, Chapter 4, this volume). Thus, the Pathfinder advantage could be due to an emphasis on relatedness judgments that are the most accurate. In general, Pathfinder tends to emphasize local relations between pairs of concepts, and MDS stresses global relations among a set of concepts in the form of dimensions. Whereas the local information represented by Pathfinder corresponds well to performance in recall tasks, it is possible that performance in other tasks (e.g., analogy completions, categorical judgments) is better captured by the global information emphasized by MDS.

Empirically Derived Structures and Category Judgment

Indeed, numerous studies involving category judgment have indicated that semantic relatedness affects the time it takes subjects to decide whether two items belong to the same category or different categories (Caramazza et al., 1976; Herrmann, Shoben, Klun, & Smith, 1975; Rips et al., 1973; Schaeffer & Wallace, 1969; Schvaneveldt, Durso, & Mukherji, 1982). Relatedness has been defined in a variety of ways. For example, Schaeffer and Wallace (1969) had subjects judge items to be the same if they were both living things or both nonliving things. Stimulus items consisted of types of mammals, flowers, metals, or fabrics. Subjects were faster to make a "same" judgment if both items were members of the same category (e.g., both flowers or both metals). Thus, in this case, relatedness was defined in terms of logical relations.

Rips et al. (1973) defined semantic relatedness according to either subjects' ratings of similarity or derived distance in an MDS solution. Items were either birds or mammals and subjects were asked to make a "same" judgment whenever the two items presented referred to one of the categories (half of the subjects responded "same" to two birds and the other half responded "same" to two mammals). They found that both the original ratings and the MDS distances predicted reaction time for this task, although the MDS distances predicted slightly better than ratings for mammals and slightly worse than ratings for birds. Interestingly, neither interitem ratings nor interitem MDS distances accounted for reaction time, however, ratings and distances between the instance and the category did. Similar results were obtained by Herrmann et al. (1975) who found that similarity as represented in a hierarchical cluster analysis was predictive of same/different reaction time.

In summary, several studies have supported the relationship between time to judge that two items are from the same category and degree of similarity or semantic relatedness. This relationship has been demonstrated using various measures of semantic relatedness, including logical relations, similarity ratings, and derived distance from a multidimensional scaling solution or cluster analysis. As mentioned previously, whereas Pathfinder networks are better than MDS at representing relations useful for recall, MDS solutions may be better than Pathfinder networks at representing relations that are useful for tasks involving categorical judgments. Therefore, in the following study MDS and Pathfinder representations were compared in terms of their relation to judgment time in a categorization task. The first step involved identifying a concept set and generating the Pathfinder and MDS structures.

Construction of the Structures

The 30 stimulus items that were selected for this experiment consisted of the superordinate category *animal,* six categories subordinate to animal (e.g., *pet, bird, fish*), and 23 instances subordinate to those categories (e.g., *dog, sparrow, bass*). These items are presented in Table 1. Care was taken to ensure that the total set of 30 items comprised a fairly broad category (i.e., *animals*) in order to increase variance among the pairwise proximities. Rips et al. (1973) noted that their failure to obtain a significant effect of interitem distance may have been due to the fact that the categories were fairly narrow (i.e., *mammals, birds*) and thus, within-category variance was small.

Table 1. Items used to construct the Pathfinder and MDS structures.

animal	dog	horse	trout
farm animal	cow	lion	wolf
reptile	turtle	dolphin	robin
pet	donkey	bat	mouse
bird	lamb	rat	bull
fish	bass	cat	sheep
wild animal	sparrow	tiger	whale
	eagle	lizard	

Fifteen introductory psychology students at New Mexico State University voluntarily participated in the rating part of this study in order to fulfill partial course credit. They were seated in front of an IBM PC and presented with instructions about the rating task, followed by a randomized list of the 30 items so that they would get an idea of the scope of the items that they would be rating. They were then presented with all 435 pairs of concepts, one pair at a time. The pairs were randomized for each subject and the order of the items in each pair was counterbalanced across subjects. A rating scale appeared on the screen along with the pair. The values on the scale ranged from 1 (slightly related) to 5 (highly related) and were selected by moving a bar marker over the value and entering the response by pressing the SPACE BAR. Subjects also had the option of entering a "U" (unrelated) for a pair that they felt was completely unrelated. This "U" option was included to reduce variance in the ratings at the lower end of the scale.

The mean intersubject rating correlation was .571. Relatedness ratings were inverted (ratings of U translated to a 6) and averaged across the 15 subjects. These average dissimilarities were then submitted to Kruskal MDS (Kruskal, 1977; Kruskal & Wish, 1978) and Pathfinder network algorithms. The resulting MDS and network representations are presented in Figures 1 and 2, respectively. Three dimensions were chosen for the MDS solution based on the Isaac and Poor (1974) procedure. Also, the stress tended to level out at .149 for three dimensions. Spatial proximities used to predict judgment time in the next part of the study were derived by taking the euclidean distance between each pair of items in this three-dimensional space. In generating the Pathfinder network the options $r = \infty$ and $q = 2$ were selected because these parameter values required only ordinal assumptions about the data, yet they yielded a network of greater density (38 links) than the tree (29 links) that was generated when q was set to $n-1$. Pathfinder proximities were equal to the length of the shortest path between each pair of nodes in the network where path length was computed by assigning ranks to each link weight and summing the ranks of the links in the path. Summing ranks required only ordinal assumptions about the link weights. In this way, three sets of 435 proximity estimates (one estimate per pair of 30 items) were generated in this part of the study. One set was obtained by averaging the original relatedness across subjects and the other two sets were obtained from the MDS and network structures derived from these average ratings.

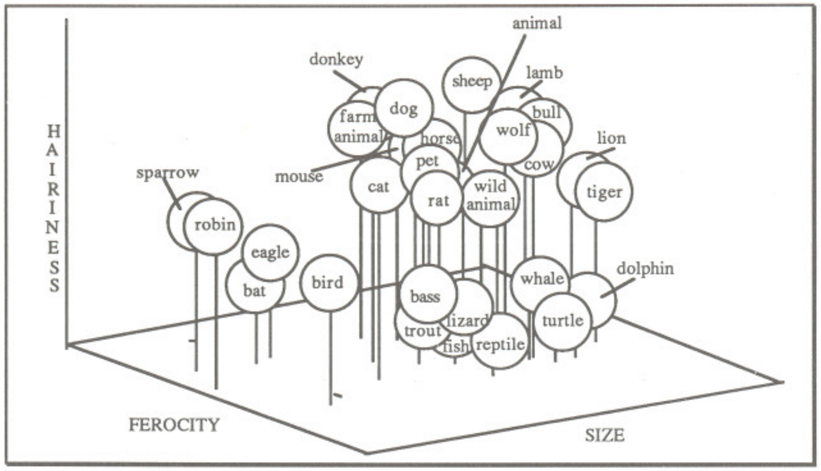

Figure 1. The Kruskal three-dimensional MDS representation. The dimension labels correspond to the author's interpretation of the space.

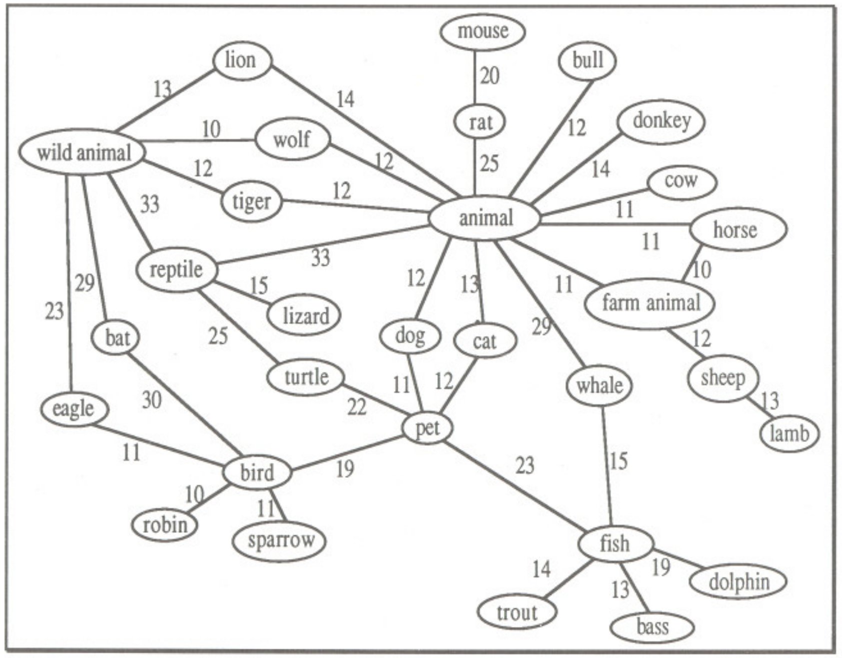

Figure 2. The Pathfinder network representation: PFNET (∞, 2).

Categorical Judgment Study

Method

Subjects. Fifty-four introductory psychology students from New Mexico State University volunteered in partial fulfillment of a research familiarization requirement.

Materials. The materials used in this task consisted of the 23 instances from the total set of 30 previously scaled. Because of the nature of this task the superordinate concepts were not used in this study. An additional set of distractor items was chosen that contained nonanimals (e.g., *carrot, carpet, road*), as well as some additional animals not included in the set of 30.

Procedure. Subjects were seated individually in front of a Terak microcomputer on which instructions about the task were presented. They were told that pairs of words would be displayed and that they were to respond by pressing the key marked "YES" if both words referred to animals. Otherwise, they were to respond by pressing the key marked "NO." Subjects were then presented with 25 practice trials. Animal concepts that were used in the practice trials were taken from the distractor item set. After the practice session subjects were able to ask the experimenter questions.

Out of the total 392 trials, 253 pairs (63%) required a "YES" response. These pairs were constructed from all possible pairs of the 23 animal instances. The remaining 149 (37%) of the trials required a "NO" response. Seventy of these pairs consisted of 2 nonanimals and 69 consisted of an animal from the set of 23 paired with a nonanimal. Therefore, subjects could not simply base responses on the presence of an item from the set of 23. Trials were divided into 8 blocks of 49 with a filler trial for warm-up inserted at the beginning of every block. After each block subjects were presented with feedback concerning their error rate and average judgment time for that block. In addition, they could take a break in between blocks if they wished.

Results

Response times to the 253 "YES" pairs were included in the analysis along with the 253 corresponding proximities from each of the three sets. A $\log_{10}$ transform was conducted on the correct response times. Average error rate for these trials was quite low (1.4%), so only correct trials were included in response time averages.

In order to determine whether the interitem proximity estimates could account for differences between long and short judgment times, a median split was performed on each of the three sets of proximities (ratings, MDS, and network). Then, for each proximity set, the average judgment times for each subject associated with items having either high or low proximities were calculated. It was expected that times corresponding to high interitem proximities would be longer than times corresponding to low proximities in accord with previous results investigating similarity and same category judgments. Differences between "high" times and "low" times are significant for all three comparisons at the .01 level: ratings, $t\,(53) = 4.8$, $SE = .002$; MDS, $t\,(53) = 4.17$, $SE = .002$; and network, $t\,(53) = 6.16$, $SE = .002$. Mean response times (converted from log to original scale) are presented in Table 2. These results do not discriminate between the spatial and network structures, but it is encouraging that each representation captures information that corresponds to either long or short judgment times.

Table 2. Mean response times (msec) associated with high and low proximities.

Proximity Measure	*Less Related*	*More Related*
Ratings	652	638
MDS	650	639
Network	653	637

A fine grain analysis was also conducted by investigating correlations and partial correlations between proximities and response time. Correlations for the three sets of proximities and average log response times are all significant at the .01 level with 251 degrees of freedom. The correlation matrix is presented in Table 3. When split-half reliability of response times ($r = .502$) is taken into account, the resulting corrected correlations between proximities and response times are .320, .336, and .413 for ratings, MDS, and network, respectively. The magnitude of these correlations is slightly lower than those found in similar studies. For instance, Rips et al. (1973) obtained multiple correlations between MDS distances and same response time of .55 and .72; however, these distances were based on an MDS solution that was rescaled using only the subset of rating estimates relevant to the categorization task. Because rescaling makes use of all information in the solution, it may result in distortions due to contextual changes.

Results of partial correlations replicated those results found by Cooke et al. (1986). The correlation between network interitem proximities and response time after removing the effect of the ratings is significant, $r\ (250) = .208$, $p < .01$, but the correlation between MDS proximities and response times without the rating effect is not, $r\ (250) = .087$. In other words, the network-derived proximities tend to capture information relevant to response times that is linearly independent of the original relatedness ratings used to generate the network structure. The MDS proximities are not independently predictive of response times as would be expected given the high (.853) correlation between MDS proximities and ratings. Again, it seems that the network representation extracts some information that is not directly represented in the original ratings.

Table 3. Correlation matrix for average ratings, MDS interitem proximities, network interitem proximities, and average response time.

	Ratings	*MDS*	*Network*	*Response Time*
Ratings	1.000	.853	.544	.227
MDS	.853	1.000	.500	.238
Network	.544	.500	1.000	.293
Response Time	.227	.238	.293	1.000

Unlike the results found previously using free recall (Cooke et al., 1986), the partial correlation between ratings and response time after removing the network effect ($r\ (250) = .084$) or the MDS effect ($r\ (250) = .047$) is not significant. Also, beyond the shared ratings there is virtually no overlap between the MDS and network solution

(r network, MDS/ratings $(250) = .082$). Thus, although the MDS solution represented the critical rating information, the network captured all of this information plus additional information critical to the categorical judgment task that is not directly represented in the ratings. The possible content of this additional information is discussed below.

Whereas the above analysis was performed on interitem distances, an analysis based on the proximity between the first item in a pair and the superordinate *animal*, or between the second item presented and *animal*, did not produce radically different results. Correlations between response times and either first item-*animal* proximity or second item-*animal* proximity were of a lesser magnitude than interitem distance correlations and ranged from $r = -.091$ to .189. Consequently, multiple correlations based on all three types of proximity (interitem, first item-*animal*, and second item-*animal*) were of magnitude comparable to interitem correlations alone.

Conclusions

The results of this study support the psychological meaningfulness of Pathfinder structures. Although long or short response times corresponded to gross distinctions (high vs. low proximity) in the ratings, MDS solution, or Pathfinder network, a more detailed analysis revealed that the Pathfinder network accounted for approximately 17% of the variance in the response times, whereas the ratings and MDS accounted for only 10% and 11% of the variance, respectively. Furthermore, partial correlations reveal a pattern of results similar to those found using a free recall task (Cooke et al., 1986). That is, unlike MDS, the Pathfinder network is predictive of category judgment time even when the rating effect is partialled out. In particular these results are interesting, given that a categorization task seems compatible with a spatial representation. In general, these results lend some support to the generalizability of the Cooke et al. (1986) results. That is, the advantage of the network representation over the spatial one appears to be more than an artifact of the recall task.

It seems that the Pathfinder algorithm captures and represents information in the original ratings that is not represented explicitly in the ratings or in the multidimensional scaling solution. It was previously mentioned that this "extra" information may arise from Pathfinder's emphasis on local, instead of global, relationships. This local/global distinction between the two scaling procedures parallels a distinction made by Lorch (1981) between retrieval (network) models and feature comparison models of memory. He points out that most comparison models describe relatedness effects in terms of semantic overlap among features and in doing so emphasize the comparison process that is required to compute relations. On the other hand, retrieval models emphasize the direct storage of relations and explain relatedness in terms of either semantic overlap (number of shared connections) or relation strength, "the strength of the most accessible subject-predicate connection which is sufficient to determine a response" (Lorch, 1981, p. 595). However, in retrieval models, as in the Pathfinder procedure, strength is emphasized over semantic overlap.

Lorch (1981) suggested that these two accounts of relatedness, semantic overlap and relation strength, could explain different patterns of results for false items in a category verification task. In one of his experiments subjects were required to respond "true" if the first item of a pair was a member of the *category* indicated by the second item (e.g., *bee-insect*). False pairs consisted of items that had strong (e.g., *bee-wings*) or weak (e.g., *bee-stinger*) *property* relations. Subjects were faster to make false responses to strongly related pairs than weakly related pairs. However, this finding was reversed in an experiment in which unrelated false items (e.g., *bee-hair*) were included. Lorch (1981) concluded that subjects based their judgments on different information in each case. In the first case, judgments were based on relation strength and thus, strong property relations resulted in rapid judgments. In the second case, subjects based their decisions on overall semantic overlap, and accordingly, low overlap resulted in a quick one-stage response, whereas high overlap often resulted in a second stage of additional processing.

Assuming that relatedness does consist of semantic overlap and relation strength, then it is possible that network and spatial representations differ to the extent that they capture information about each of them. MDS representations tend to define relatedness in terms of semantic overlap and not specific relation strength. The Pathfinder proximities, however, were based on relation strength (i.e., link weights) as opposed to the number of shared connections. To the extent that relation strength is relevant to the task, Pathfinder's emphasis on this component of relatedness could provide it with a predictive advantage over MDS.

It would be interesting to investigate the correspondence between response time and a measure of semantic overlap based on the Pathfinder network. Such a measure of semantic overlap might be calculated by counting the number of paths consisting of q or fewer links that connect two nodes, where q is the maximum number of links allowable in any path. In other words, instead of calculating the shortest path between two items (i.e., relation strength), the number of short paths would be counted. Based on the network in Figure 2, the semantic overlap ($q = 2$) between *wild animal* and *animal* equals 4 and is greater than the overlap of 2 between *farm animal* and *animal.* However, the relation strength based on the sum of the ranked link weights equals 15 and 6.5 for *wild animal-animal* and *farm animal-animal* respectively. Keeping in mind that smaller link weights correspond to a higher degree of relatedness, then these two measures of relatedness would make opposite predictions about response times to these pairs.

It would also be useful to compare MDS and Pathfinder representations in terms of some other tasks, particularly some that are likely to require interpreting relatedness in terms of semantic overlap (e.g., analogy completion). It is also possible that the method of obtaining proximities from multidimensional space does not adequately represent the concept of semantic overlap captured by MDS. An alternative metric may be better suited in describing the strengths of MDS. These are all issues requiring future research.