COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–36

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12299

Graph-Theoretic Properties of Networks Based on Word Association Norms: Implications for Models of Lexical Semantic Memory

Thomas M. Gruenenfelder, Gabriel Recchia, Tim Rubin, Michael N. Jones

Department of Psychological and Brain Sciences, Indiana University

Received 15 August 2013; received in revised form 17 May 2015; accepted 23 May 2015

Abstract

We compared the ability of three different contextual models of lexical semantic memory (BEAGLE, Latent Semantic Analysis, and the Topic model) and of a simple associative model (POC) to predict the properties of semantic networks derived from word association norms. None of the semantic models were able to accurately predict all of the network properties. All three contextual models over-predicted clustering in the norms, whereas the associative model under-predicted clustering. Only a hybrid model that assumed that some of the responses were based on a contextual model and others on an associative network (POC) successfully predicted all of the network properties and predicted a word's top five associates as well as or better than the better of the two constituent models. The results suggest that participants switch between a contextual representation and an associative network when generating free associations. We discuss the role that each of these representations may play in lexical semantic memory. Concordant with recent multicomponent theories of semantic memory, the associative network may encode coordinate relations between concepts (e.g., the relation between pea and bean, or between sparrow and robin), and contextual representations may be used to process information about more abstract concepts.

Keywords: Lexical semantic memory; Word association; Graph theory; Semantic memory; Semantic networks; Coordinate relations; Concrete concepts; Abstract concepts

1. Introduction

Understanding the meaning of a spoken utterance or written phrase entails knowing the meanings of the individual words used. Theories of lexical semantic memory are

Correspondence should be sent to Thomas M. Gruenenfelder, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405. E-mail: tgruenen@indiana.edu

concerned with how people represent the meanings of words as well as the processes that operate on those representations. Investigators have used a number of different laboratory tasks to discriminate among and refine these theories. A non-exhaustive list includes category verification (e.g., Collins & Quillan, 1969), feature or property verification (e.g., Conrad, 1972), priming in word recognition (e.g., Meyer & Schvaneveldt, 1971), and performance of theoretical models on synonymy tests (e.g., Landauer & Dumais, 1997). The present paper focuses on word association data (e.g., Deese, 1965; Nelson, McEvoy, & Schreiber, 1999) as a means of exploring lexical semantic memory. Word association data are frequently used to evaluate corpus-based models of lexical semantics (e.g., De Deyne & Storms, 2008; Griffiths, Steyvers, & Tenenbaum, 2007; Maki & Buchanan, 2008).

In a word association task, the participant is presented a word, referred to as the *stimulus* or *cue word*, and asked to respond with the first word that comes to mind. The word that the participant produces is referred to as the *response*.[1] Building on Griffiths et al. (2007), the specific approach we take here is to build networks based on word association data and to then compare the structural properties of those networks with the properties of networks constructed from different models of lexical semantic memory. If a particular model accurately characterizes how humans represent lexical semantics in memory, then a network generated from that model would have properties similar to those seen in word association norms.

The relatedness between two concepts in memory (which is assumed to drive word association data, e.g., De Deyne & Storms, 2008; Steyvers & Tenenbaum, 2005) has been approximated in the literature with at least three statistical measures based on theories of semantic representation. Perhaps the simplest is the assumption that conceptual relatedness is strengthened via the principles of associative learning (Deese, 1965). Because temporal contiguity is considered fundamental to associative learning, various measures of the frequency with which two words co-occur in the text have been used as proxies of how related mental concepts are (see, for example, Spence & Owens, 1990; Wettler, Rapp, & Sedlmeier, 2005), giving them strong connections to compound cuing models of memory (McKoon & Ratcliff, 1992). We refer to any simple measure of similarity based on direct co-occurrence as *associative similarity*.

A competing theoretical construct is that of *featural similarity*, referring to the extent to which two concepts share semantic features.[2] These features are usually the discrete properties used in some early models of semantic memory (Gellatly & Gregg, 1975, 1977; McCloskey & Glucksberg, 1979; Rips, Shoben, & Smith, 1973; Smith, Rips, & Shoben, 1974; Smith, Shoben, & Rips, 1974) as well as in the direct descendants of those models (Masson, 1995; McRae, Cree, Seidenberg, & McNorgan, 2005; McRae, de Sa, & Seidenberg, 1997; Moss, Hare, Day, & Tyler, 1994; Rogers & McClelland, 2004). Theoretically, two concepts can be high in associative similarity but low in featural similarity, or vice versa. However, featural and associative similarities are highly correlated and can be difficult to disentangle operationally (Hutchison, 2003; Lucas, 2000; McNamara, 2005).

A third proxy of conceptual relatedness is *contextual similarity*, referring to the similarity between two words in a high-dimensional spatial representation of semantic

memory, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) or Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007). Similarity in these models reflects the extent to which two concepts appear in similar linguistic contexts, but it does not require direct co-occurrence. Statistical abstraction mechanisms allow these models to capitalize on higher-order statistical relationships above and beyond direct co-occurrence embodied by associative similarity measures. For example, *bee* and *honey* may have a primarily syntagmatic relationship, which would be reflected in associative similarity. In contrast, *bee* and *wasp* have a primarily paradigmatic/taxonomic relationship, which would be reflected in featural similarity. *Contextual similarity* could reflect either or both of these two types of relationships, depending on the statistical regularity of the words' linguistic behaviors (Jones, Kintsch, & Mewhort, 2006).

Although it does not rely on a spatial representation of word meaning, the Topic model (Griffiths et al., 2007) shares LSA and BEAGLE's emphasis on contextual similarity. In the Topic model, a word's meaning is represented by the probability distribution of its use across a set of topics. Hence, words that are used in conjunction with the same topics —that is, appear in similar contexts—would be related in meaning. As is the case with BEAGLE and LSA, two words can be semantically related in the Topic model without ever co-occurring with one another. Collectively, we refer to these three models as contextual models, since they model word similarities based on the similarity of contexts in which they occur rather than on their direct co-occurrence.

There is a tacit assumption, particularly in the lexical decision priming literature (for reviews, see Hutchison, 2003; Lucas, 2000; McNamara, 2005), that word association data reflect associative similarity. However, it is just as likely that performance in a word association task could be based on featural or contextual similarity as on associative similarity, or on some combination of the three. There has been little research directly testing the assumption that word associations reflect associative relatedness rather than featural relatedness or contextual similarity, largely because controlling confounds in stimulus selection is practically impossible. That research which does exist (Enguix, Rapp, & Zock, 2014; Spence & Owens, 1990; Wettler et al., 2005) has found that co-occurrence frequency was correlated with the probability that one word was produced as a response to the other in a word association task, a finding interpreted as supporting the hypothesis that word associations reflect associative similarity. None of these studies, however, attempted to control for the confound between associative similarity and featural or contextual similarity, making their results difficult to interpret.

It is well known that simple associative learning is not sufficient to explain linguistic behavior (e.g., Chomsky, 1959). That is not to say, however, that simple associative learning has no role at all in language. Recent work with semantic space models has questioned whether their "deep" learning of latent contextual similarity is necessary to simulate human data. For example, Louwerse (2011) demonstrated that simple associative similarity is sufficient to explain a wide range of human semantic data previously used to argue that models need deep contextual similarity, provided that the corpus on which the model is trained is sufficiently large. Recchia and Jones (2009) found a very similar

pattern—LSA was outperformed by a very simple associative metric based on mutual information on several tasks when the associative model was provided with a sufficiently large training corpus. These results and others (e.g., Baayen, Hendrix, & Ramscar, 2013; Ramscar, Dye, & Klein, 2013; Stone, Dennis, & Kwantes, 2011) have renewed interest in simple models that learn from direct associative similarity. One goal of the current work is to examine the extent to which such simple associative learning may contribute to the development of the mental lexicon.

The present work used graph theoretic analyses of networks based on association norms to evaluate the relative ability of three contextual models (BEAGLE, LSA, and the Topic model) and an associative model (Proportion of Co-occurrence or POC) to predict the pattern of human word associations. We first built a set of networks based on the Nelson et al. (1999) word association norms. We then used Luce's choice axiom (R. D. Luce, 1959) along with model-specific measures of similarity to predict the word association responses that would be made by each of the models. Each set of a model's predicted norms was then used to build a set of networks analogous to those constructed from the Nelson et al. (1999) norms. A model's ability to capture the network structure of human semantic memory was evaluated with respect to how closely the properties of the networks constructed from the model's predictions matched the properties of the networks constructed from the human norms. The specific network properties measured and the rationale for focusing on them are discussed later. In addition, we evaluated the models with respect to their ability to directly predict human word-association responses.

The four models examined were BEAGLE (Jones & Mewhort, 2007), LSA (Landauer & Dumais, 1997), the Topic model (Griffiths et al., 2007), and *proportion of co-occurrence* (POC), a member of the family of mutual information metrics. BEAGLE was selected because in an earlier, much less extensive study, it performed fairly well in predicting the graph theoretic properties of association norms (Jones, Gruenenfelder, & Recchia, 2011). LSA was chosen because it is perhaps the most widely used and influential model of lexical semantic memory based on contextual similarity (see Landauer, McNamara, Dennis, & Kintsch, 2007 for a review.). The Topic model has also been extensively used in the literature on semantic memory, and it has had success in predicting characteristics of a network based on word associations, particularly the degree distribution (Griffiths et al., 2007). For the psychological assumptions underlying these models, the reader is referred to the original sources or to the recent review by Jones, Willits, and Dennis (2014). POC was intended to reflect the ability of a simple associative network, such as those proposed by Collins and Quillan (1969) or Glass and Holyoak (1975), to predict the pattern of word associations without hand-coding. BEAGLE, LSA, and the Topic model all reflect contextual similarity; POC reflects associative similarity. Ideally, we would also have included a model based on featural similarity. Unfortunately, despite the concerted efforts of various investigators (McRae et al., 2005; Vinson & Vigliocco, 2008), we did not feel that feature norms existed for a large enough set of words to adequately evaluate the role of feature overlap in producing word associations.

2. Generating the semantic networks

Following earlier investigators (De Deyne & Storms, 2008; Griffiths et al., 2007; Steyvers & Tenenbaum, 2005; Utsumi, 2014), each node in the norm-based networks or graphs (i.e., the graphs based on word association norms) represented a word. All graphs were undirected. A *link* or *edge* was placed between the nodes representing two words if the proportion of participants who produced one of those words as an associate of the other exceeded a particular threshold. No weights were placed on the edges. Non-weighted graphs were used because we wished to follow as closely as possible the procedures of earlier investigators studying the ability of contextual models to predict the network structure of word associations, particularly those of Griffiths et al. (2007), in order to make our results as comparable to theirs as possible. We were also concerned about the commensurability of similarity measures across different models. For the most part, though not exclusively, those investigators also used non-directed graphs (cf. Borge-Holthoefer & Arenas, 2010). Without the use of Luce's Choice Axiom or some similar tool, BEAGLE, LSA, and POC all produce symmetric, non-directional similarity measures, forcing the use of undirected graphs. Unlike those earlier studies, we did use Luce's Choice Axiom and hence could have produced directed graphs. We chose to use undirected graphs in order to keep our methods as comparable as possible to those earlier studies.

Our method of constructing the norm-based graphs did differ from earlier studies in one important respect. All earlier studies used just a single threshold for determining whether to place an edge between the nodes representing two words. In particular, they placed an edge between two words if one word was produced as the first associate by at least two participants in the Nelson et al. (1999) studies. Thresholds, of course, are arbitrary, and there is no guarantee that the qualitative characteristics of a graph will match the characteristics of another graph defined using the same nodes but with a different threshold for creating edges between the nodes (Butts, 2009). To help mitigate this problem, we used the norms to build multiple graphs, each with a different threshold. Thresholds were increased monotonically, producing a function relating the change in graph properties to the change in threshold for placing edges between nodes (i.e., the proportion of participants who produced the response word as an associate of the stimulus word [cf. Hills, Maouene, Maouene, Sheya, & Smith, 2009a]).

Griffiths et al. (2007) used an approach similar to ours to compare the relative abilities of LSA and the Topic model to predict the graph-theoretic properties of association norms. Our approach differed from theirs in two important ways. First, as we did with the norms themselves, we used multiple thresholds to build multiple networks for each model. Second, when creating the graphs based on LSA, Griffiths et al. (2007) used an absolute measure of similarity. In particular, they chose a value of cosine similarity such that the mean degree[3] of the graph would be approximately equal to the mean degree of the graph based on the Nelson et al. (1999) association norms. All pairs of words with a cosine similarity above that threshold were connected by an edge in the LSA-based

graph; all pairs of words with a cosine similarity less than that value were left unconnected. Such an approach asks the question, "According to the model, what pairs of words are most strongly related?" The question asked in a word association task, however, is "What words are most strongly related to a given stimulus word?" (Jones et al., 2011). The semantic models are theories of how concepts are related to each other in memory, but they require a process mechanism to produce a response to a cue from this memory structure. Accordingly, we used Luce's choice axiom as a simple process account to construct the graphs based on theoretical measures of similarity. (Use of the Luce Choice Axiom is implicit in the computation of word similarities used by the Topic model; see, for example, Eq. 7 of Griffiths et al.) Luce's (1959) rule was selected due to its ubiquity in models of cognitive phenomena: it has been successfully applied to phenomena ranging from low-level neural networks to high-level economic models of group choice behavior. In particular, we assumed that the probability that word R_j is produced as the first associate to word S_i is the ratio of the relatedness of S_i and R_j relative to the sum of the relatedness of S_i to every other word in the set of possibilities:

$$p(R_j|S_i) = \frac{\beta_j \eta_{i,j}}{\sum_k \beta_k \eta_{i,k}} \tag{1}$$

In Eq. 1, $\eta_{i,j}$ is the (model-specific) relatedness of stimulus i and response j. $ß_j$ corresponds to the bias for producing response j, independent of the presented stimulus.

In order to avoid an unmanageable number of free parameters, the bias parameter was a function of the word's log-frequency of occurrence in the language (cf Nelson et al., 1999). In particular, we assumed that

$$\beta_j = (\log(freq_j))^W \tag{2}$$

where $freq_j$ is the word's frequency in the corpus being used. The parameter w indicates the strength of the response bias parameter relative to the strength of relatedness parameter $\eta_{i,j}$. For a given model, w is fixed; it does not vary from one response word to another. When $w = 0$, there is no response bias.

3. Study 1: Graphs generated by simple models

In our first study, we evaluated the ability of four simple models to predict the structure of graphs built from the Nelson et al. (1999) word association norms. Those models were versions of BEAGLE, LSA, the Topic model, and POC. We began by building several different graphs based on the Nelson et al. (1999) norms by varying the threshold (in terms of the proportion of participants who produced word j in response to cue i) required for placing an edge between words i and j. As the threshold is increased, the number of edges and possibly the number of nodes in the graph decreases. We then generated

predictions from each model, using Eqs. (1) and (2), as the threshold was increased for each model.[4] As the threshold is raised, the mean number of edges per node (the mean degree or k_{mean}) tends to decrease. The thresholds used to generate model predictions were chosen such that the mean degree of the graphs based on any particular model spanned approximately the same range as the mean degree of the graphs based on Nelson et al. (1999)'s norms.

Details of the semantic representations used in LSA, BEAGLE, and the Topic model can be found in Landauer and Dumais (1997), Jones and Mewhort (2007), and Griffiths et al. (2007), respectively. The POC measure was intended to be a measure of simple associative strength as determined by the classical laws of associative learning. As such, it is a measure of how frequently two words co-occur in text relative to their total number of occurrences. Specifically, POC was computed as follows:

$$\text{POC}_{i,j} = f_{i,j}/(f_i + f_i - f_{i,j}) \tag{3}$$

where f_i is the frequency with which word i occurs in the corpus and $f_{i,j}$ is the frequency with which words i and j co-occur in the corpus (i.e., occur in the same document). When building graphs based on the POC measure, that value was then substituted into Eq. 1 for $\eta_{i,j}$.

Many different measures of co-occurrence frequency exist (see Jones et al., 2014, for a list). Our intent with POC was not to create yet another such measure to compete with existing measures. Rather, the intent was to use a simple measure that we felt was consistent with the principles that have grown out of an extensive literature on associative learning (e.g., Rescorla, 1988). That literature indicates that the important factor in determining whether an association develops between the stimulus and response is how well the stimulus predicts the response and how well the response is predicted by the stimulus, not the absolute number of times that a stimulus and response co-occur. Two events can frequently co-occur but no association between the two forms if they each also frequently occur in the absence of the other. POC is an easily calculated statistic that increases as two words co-occur and decreases as the two words occur separately from one another.

3.1. Dependent measures

In comparing the graphs based on model predictions to those based on the actual norms, we examined three dependent variables: (a) the shape and slope of the degree distribution; (b) the mean shortest path length; and (c) the clustering coefficient. These dependent variables, particularly those involving the degree distribution, are those most frequently examined by earlier investigators (e.g., Griffiths et al., 2007; Gruenenfelder & Pisoni, 2009; Hills et al., 2009a; Morais, Olsson, & Schooler, 2013; Steyvers & Tenenbaum, 2005). In addition, each has both psychological and theoretical significance, especially the clustering coefficient. We first describe each of the variables and then elaborate on their significance.

3.1.1. Shape and slope of the degree distribution

A node's degree is the number of connections it has to other nodes. A graph's degree distribution is the frequency distribution of degrees across all nodes in that graph. The degree distribution can be of interest because it provides hints concerning how the graph evolved over time, for example, in our case, how new words were added to the lexicon as a child learned its native language (Hills, Maouene, Maouene, Sheya, & Smith, 2009b; Steyvers & Tenenbaum, 2005). We asked if each degree distribution had a heavy right-hand tail, in which most nodes have a relatively small number of edges but a few have a great deal of edges. Such a degree distribution would result if new nodes (newly learned words) connect to existing nodes with a probability dependent upon the existing node's degree. The tails of such degree distributions are well fit by power law functions, $N(k) \sim k^{-\gamma}$, where N(k) is the degree distribution, k the degree, and the exponent, γ, is usually between 2 and 4 (Albert & Barabási, 2002; Barabási & Albert, 1999). In contrast, if a new node forms an edge with an existing node with some fixed probability independent of the existing node's degree, then the tail of the degree distribution is steeper and well fit by an exponential distribution (Barabási & Albert, 1999). Therefore, we compared the relative ability of power functions and exponential functions to fit the right-hand tail of the degree distribution of each of the graphs that we examined. We defined the right-hand tail as all degrees greater than or equal to the median degree. For the sake of conciseness, when we refer to the function best fitting a degree distribution, we are referring to the function that best fits its right hand tail.

In addition to the shape of the degree distribution, we also examined the parameter γ of the best fitting power law for each degree distribution. We refer to this parameter as the slope of the degree distribution since, when plotted on log-log coordinates, a power function is linear with a slope of $-\gamma$. The word association norms themselves were consistently better fit by power functions than by exponentials. Consequently, when examining the model fits, we always looked at the slope of the model's best fitting power function, even in those cases where the degree distribution based on model predictions was better fit by an exponential than by a power law.

3.1.2. Mean shortest path length

A path is a sequence of edges that can be traversed to move from one node to another. The path length is the number of edges in that path. The shortest path length is the path with the shortest length connecting two given nodes, out of all the possible paths connecting those two nodes. The mean shortest path length (MSPL) is the mean length of that shortest path across all possible pairs of nodes in the graph.

The actual measure that we used was the ratio of MSPL to MSPLer, where MSPLer is the mean shortest path length in an Erdos-Rényi random graph (Erdos & Rényi, 1960). An Erdos-Rényi graph begins with a fixed number of nodes; links are then placed between nodes with some fixed probability. MSPLer is approximately equal to

$$\text{MSPLer} \approx \ln(n)/\ln(k_{\text{mean}}) \tag{4}$$

where k_{mean} is the mean degree across all nodes in the graph and n is the number of nodes (Albert & Barabási, 2002).

3.1.3. Clustering coefficient

Nodes that have an edge between them are frequently referred to as *neighbors*. A node's clustering coefficient (CC) measures the tendency of a node's neighbors to themselves be neighbors of one another. In particular, a node's CC is the number of its neighbors that themselves are neighbors divided by the total possible number of such neighbor pairs. For example, if a node has three neighbors, two of which themselves are neighbors, then there is one neighbor pair out of a possible three, yielding a CC of 1/3.

Similar to the case for MSPL, the actual measure that we used was the ratio of the CC to the CC expected in an Erdos-Rényi graph (Erdos & Rényi, 1960), CCer. CCer is approximately equal to

$$\text{CCer} \approx k_{mean}/n \tag{5}$$

where k_{mean} and n are defined as above for MSPLer (Albert & Barabási, 2002; Watts & Strogatz, 1998).

3.2. Psychological relevance of the network measures

Our dependent measures may be considered somewhat crude, operating at what Borge-Holthoefer and Arenas (2010) have termed the macro-level of network analysis. There are numerous other summary statistics that could be derived from association norms. The measures we examined, however, do have both theoretical and psychological significance. Fundamental characteristics of the models under consideration here lead to different expectations concerning the amount of clustering in semantic graphs. In contextual models, such as BEAGLE, LSA, and the Topic model, two words can become related even when they do not co-occur with one another. For example, if the words "wine" and "apple" rarely co-occur with one another but each does co-occur with "red," "taste," and "healthy," then, within a contextual model, "wine" and "apple" become neighbors of one another, increasing the clustering coefficient for "red," "taste," and "healthy." In a simple associative model like POC, in contrast, two words that do not directly co-occur would not become neighbors on one another. Hence, the clustering coefficients for "red," "taste," and "healthy" would all be depressed. In general, contextual models should lead to more clustering than do simple associative models. The question then becomes which class more closely corresponds to the amount of clustering actually observed in lexical semantic memory.

Clustering also affects the dynamics of how activation could spread through a network; those dynamics in turn have empirical consequences (Borge-Holthoefer & Arenas, 2010; Nematzadeh, Ferrara, Flammini, & Ahn, 2014; Vitevitch, Ercal, & Adagarla, 2011). Vitevitch, Ercal, & Adagarla (2011), for example, showed via simulation that activation spreading from a source node with a high CC tended to stay concentrated in the source

node's local neighborhood. Little activation spread back to the source node or presumably to more distant nodes in the network. For nodes with a lower value of clustering, activation was able to spread throughout the network, back to the source node as well as to more distant nodes in the network. A similar pattern of results was reported by Nematzadeh et al. (2014).[5] Too much clustering resulted in activation being confined to a node's local neighborhood. Too little clustering resulted in activation diffusing throughout the network, but at the expense of little activation spreading throughout a node's local neighborhood—nodes closely related to the node from which the spread of activation started were themselves not activated. At intermediate levels of clustering, a balance was found where both a node's neighbors were highly activated and activation was able to spread to more distant areas of the network. The extent to which the results of these simulations generalize beyond the particular networks they studied is at present unknown. These results do suggest, though, that the amount of clustering observed in a network can significantly affect both the speed of retrieval from that network as well as what information can be retrieved, given a particular starting point to the search. These considerations would seem to apply regardless of whether retrieval involves activation spreading along links in an associative network or activation diffusing through a high-dimensional space, such as in BEAGLE, LSA, or the Topic model.

Clustering has also been shown to affect semantic priming (Nelson & Goodmon, 2002)—priming effects are larger for words with higher clustering. Such a result would be expected if priming is at least in part due to spreading activation and clustering affects the spread of activation, as argued above. Several studies have also found that both recognition and recall are higher for words with a larger CC than words with a smaller CC (e.g., Nelson, Bennett, Gee, Schreiber, & McKinney, 1993; Nelson & Goodmon, 2002; Nelson, McKinney, Gee, & Janczura, 1998; Nelson, Zhang, & McKinley, 2001).

To the extent that determining whether and how two words are related is important to language comprehension, path length is of obvious relevance to associative network representations—shorter paths mean those relations can be more quickly retrieved. Perhaps just as important, though, is that short paths mean that searches unlikely to be successful can be quickly terminated. When comprehending a sentence, a listener's failure to quickly retrieve a relation between two key words in that sentence may serve as a cue that a more metaphorical interpretation is necessary (Kintsch, 2000).[6]

As noted earlier, the degree distribution can provide indications of how a child's mental lexicon develops over time. A word's degree has also been shown to affect its recall and recognition in memory studies (Nelson, Schreiber, & McEvoy, 1992; Nelson et al., 1993), and other studies have used an analogous construct referred to as "number of semantic neighbors" to investigate lexical decision, semantic categorization, and other phenomena (Buchanan, Westbury, & Burgess, 2001; Hargreaves & Pexman, 2014; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Recchia & Jones, 2012; Shaoul & Westbury, 2010; Yap, Tan, Pexman, & Hargreaves, 2011).

Finally, we note that whereas models like POC have a natural representation as a network, the contextual models lends themselves more to spatial representations, in which a word is represented as a vector of values that define a point in a very high-dimensional

space. The concepts of degree and clustering, though, have meaning within that high-dimensional space as well as within a network. Degree corresponds to the number of neighbors of a word, that is, the number of other words within some threshold distance of the target word. Clustering, in turn, is simply the number of those neighbors within the same threshold distance of each other. Before the incorporation of the graph theory into psychological theories, these concepts, under different names, have long been important in psychological studies of both the phonological mental lexicon (P. A. Luce & Pisoni, 1998; Mathey, Robert, & Zagar, 2004; Mathey & Zagar, 2000) and the semantic mental lexicon (e.g., the studies of Nelson and colleagues, Pexman and colleagues, and Westbury and colleagues, cited just above). Graph theory simply provides a convenient tool for measuring and analyzing these concepts.

3.3. Method

3.3.1. Constructing the word association graphs

Multiple graphs were constructed from the Nelson et al. (1999) word association norms by varying the threshold forward relatedness strength (the proportion of participants who produced word R as the first response to cue word S) required for placing a link between two nodes representing words. In constructing the graphs, only response words that also served as stimulus words were included (cf. De Deyne & Storms, 2008). That is, only *normed* words were included in the graph. Any word that was on a standard stop list of function words was also excluded from the graph. The six threshold values used were *All*, .02, .04, .05, .08, and .13, where *All* corresponds to the case of all responses made by at least two participants.

3.3.2. Constructing graphs based on models

All four models (BEAGLE, LSA, the Topic model, and POC) were trained using the Touchstone Applied Science Associates (TASA) corpus (Landauer & Dumais, 1997). When constructing the graphs, only words that had been normed (i.e., used as cue words) by Nelson et al. (1999) were considered. Any words on the same stop list used to filter the association norms were also excluded, as were words that did not occur with at least a frequency of 10 in the TASA corpus. This procedure closely follows that of Griffiths et al. (2007).

To construct the graphs, for each normed word, its relatedness strength to every other normed word was first calculated according to Eq. 1, with Eq. 2 substituted for β and the appropriate measure of relatedness strength (cosine similarity for BEAGLE and LSA, the Griffiths et al. (2007) similarity measure—their Eq. 9—for the Topic model, Eq. 3 for POC) substituted for η. When calculating the denominator of Eq. 1 for a given word X, only the 200 words most strongly associated to X (according to the relevant measure of relatedness strength) were used. Limiting the number of terms in the denominator is one of several appropriate ways of dealing with negative cosines in Eq. 1. Pilot work showed that varying the number of terms used in the denominator from 100 to 200 to 500 had little effect on the resulting graphs. For each model, the parameter w in Eq. 2 was initially

Table 1
Values of *w* tested for the various models

Values of *w* Tested Simple Models	
Model	Values of *w* Tested
BEAGLE	0.0, 0.5, 1.0, 2.0
LSA	0.0, 1.0, 1.5, 2.0, 2.5
Topic	0.0, 1.0, 2.0, 3.0, 4.0
POC	0.0, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0

Hybrid Models		
	Values of *w* Tested	
	Model 1	Model 2
BEAGLE/LSA	0.0, 0.5, 1.0, 2.0	1.5, 2.0, 2.5, 3.0
BEAGLE/Topic	0.0, 0.5, 1.0	0.0, 1.0, 2.0
LSA/Topic	1.0, 1.5, 2.0	0.0, 1.0, 2.0
BEAGLE/POC	0.0, 0.5, 1.0	3.0, 3.5, 4.0
LSA/POC	1.0, 1.5, 2.0	3.0, 3.5, 4.0
Topic/POC	0.0, 1.0, 2.0	3.0, 3.5, 4.0

Note. For hybrid models, Model 1 refers to the first mentioned model in the hybrid (e.g., BEAGLE for the hybrid BEAGLE/LSA) and Model 2 refers to the second mentioned model (e.g., LSA for the hybrid BEAGLE/LSA).

set to 0 (no response bias) and then gradually increased until fits reached a plateau or began to consistently decline. Hence, the exact values of *w* used differed from model to model. Those values are shown in Table 1.

For each combination of model and value of *w*, multiple graphs were constructed by varying the threshold for including an edge between two words. Five or six different threshold levels were used for each value of *w* for each of the models.

For BEAGLE, only the context vector, and not the order vector, was used. The parameters were the same as those in Jones and Mewhort (2007). For LSA, we used 300 dimensions, a dimensionality which has been shown to optimize performance on various tasks (Landauer & Dumais, 1997; Landauer, Laham, & Derr, 2004). The parameters used for the Topic model were the same as the model with 1,700 topics used by Griffiths et al. (2007), with the exception that we used one additional chain and over six times as many samples. For POC, co-occurrences were counted across the entire document.

3.3.3. Measuring the dependent variables

For words with only a single neighbor, a value of 0 was used for the CC. The shapes (i.e., the relative fits of exponential and power functions) of the degree distributions were fit using linear regression. Slopes of the best fitting power functions were determined using the algorithm described in Clauset, Shaliza, and Newman (2009), with the

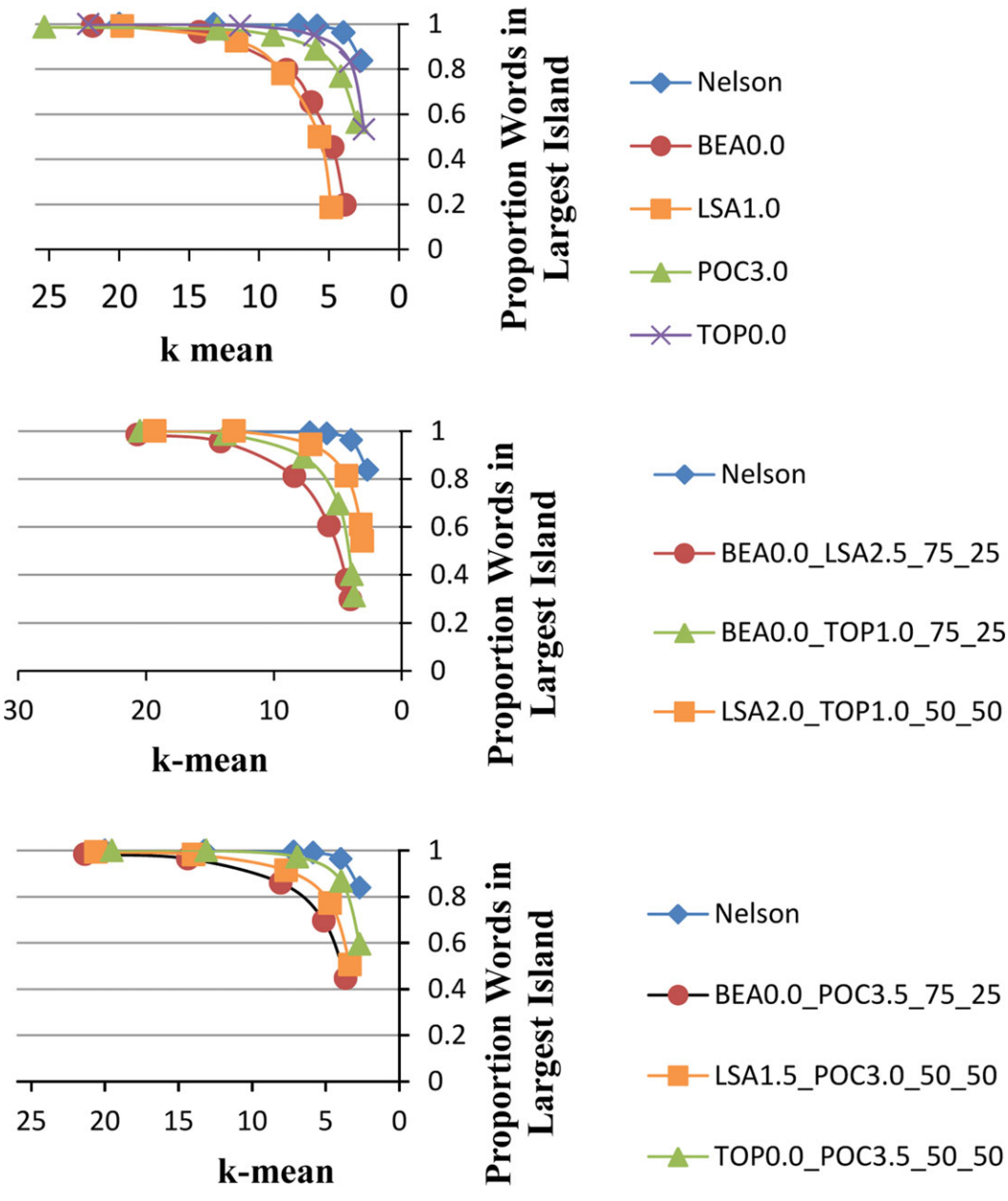
minimum degree included in the fits fixed at the median of the distribution. MSPLer and CCer were approximated using Eqs. 4 and 5, respectively.

3.4. Results

We present here the results for models using the value of w that provided the best fit to the CC data for each individual model. The supplementary material contains results for all values of w for all the simple models (see Figs. S-1 through S-4). We emphasize the CC data since in our view the CC is the dependent variable with the greatest psychological and theoretical significance. In the event that for a given model multiple values of w produced approximately equally good fits to the CC data, we next examined the fits to the MSPL data, as that dependent variable is of perhaps the second greatest psychological and theoretical significance. If multiple model-exemplars were still in contention, we used the fit to the slope of the degree distribution as the final criterion for determining the value of w to display for a particular model.

For both the norms and the models, the graphs that were constructed based on the lowest threshold (i.e., the highest number of edges) consisted of a single large island, where an *island* is a set of nodes, each of which can be reached from every other node in the island by traversing one or more edges. As the threshold was increased, the graphs became more fragmented, consisting of one large island but also a number of smaller islands, each with usually fewer than five nodes. Since not all the network statistics in which we are interested (in particular, the mean shortest path length) can be meaningfully computed across different islands, the statistics reported here are for the largest island.[7] The top panel of Fig. 1 shows the proportion of words that are included in the largest island as a function of the threshold for the norms and for each of the simple models for that value of w that provided the best fit to the CC data. This and subsequent figures share a number of common characteristics. First, each figure shows the observed results for the Nelson et al. (1999) norms as well as the predicted results. The observed results are, of course, the same regardless of the model being tested, resulting in some redundancy across the figures. However, repeating the observed results in each figure makes it simpler to visually evaluate the fit of any given model. The value plotted on the abscissa is k_{mean}, the mean degree of the corresponding graph. Different values of k_{mean} were generated by varying the threshold, as previously described. The data are plotted using k_{mean} rather than the absolute threshold because k_{mean} is simply a more meaningful variable than the absolute value of the thresholds. In general, as the threshold is increased, k_{mean} decreases. Hence, as we move from left to right along the abscissa, the threshold increases but k_{mean} decreases. For expository clarity, we frequently use the term *density* in lieu of k_{mean}.[8]

As can be seen in Fig. 1, for both the models and the norms, as the threshold increased, the proportion of nodes included in the largest island decreased. Such a result is mathematically required—as the threshold is increased, nodes can drop out of the largest island (or out of the network entirely) but none can be added. The models and norms, though, did differ in terms of how rapidly that decrease occurred. The



norms are somewhat remarkable in that the largest island included nearly all words except at the highest thresholds (lowest values of k-mean). For the models, the rapid fall off in the proportion of nodes included in the largest island began at higher densities. BEAGLE and LSA included a reasonably large number of nodes in the largest islands down to a value of k-mean of approximately 7. The Topic model and POC

Fig. 1. Proportion of words in the largest island as a function of k-mean. The proportion is relative to the total number of words in the network generated for the Nelson norms with a threshold of All. The top panel shows the results for an exemplar of each simple model. The middle panel shows the results for an exemplar of each possible contextual × contextual hybrid model. The bottom panel shows the results for an exemplar for each possible contextual × POC model. See the text for a description of how the exemplars were chosen. The following notation is used in the legends in this and subsequent figures. A simple model is designated by XXXn.m, where XXX is a three-character abbreviation of base model and n.m is the value of the parameter *w* used in this particular exemplar. For example, BEA0.0 is a BEAGLE model with $w = 0$. For hybrid models, XXXn.m_YYYl.k_p_q, where XXX is a three-character abbreviation of Model 1 of the hybrid, YYY is a three-character abbreviation of Model 2 of the hybrid, n.m is the value of *w* used in conjunction with Model 1, l.k is the value of *w* used in conjunction with Model 2, p is the percent of responses generated by Model 1 and q (= 100 – p) is the percent of responses generated by Model 2. For example, BEA0.0_LSA2.5_75_25 is a hybrid in which the first component model is BEAGLE with $w = 0$, the second component model is LSA with $w = 2.5$, and 75% of the responses come from BEAGLE and 25% come from LSA.

included a large number of nodes in the largest island to somewhat lower values of k-mean, 5 or below. This pattern was not dependent upon the particular values of *w* chosen. In order to avoid issues in comparing a relatively small island produced by a model to a much larger island produced by the norms, we included only the three lowest thresholds when evaluating BEAGLE and LSA, and the four lowest when evaluating the Topic model and POC. Only these values are shown in subsequent figures. The five lowest thresholds were included for the norms, as the largest island continues to include a quite large proportion of the words across these five thresholds. The Supplementary Material does show results across all threshold levels for both the norms and for all models (Figs. S1–S4).

Fig. 2 compares the best exemplar of each model to the norms on each of the dependent variables. Figs. 2, 3, and 4 all follow a common structure. Panel (a) shows the goodness of fit, as measured by R^2, of a power function to the degree distribution. Panel (b) shows the goodness of fit of the power function to the degree distribution minus the goodness of fit of an exponential function. Values above 0 indicate better relative fits of the power law, values below 0 indicate better relative fits of an exponential, and values near 0 indicate that the degree distribution is about equally well fit by power and exponential functions. Panel (c) shows the slope of the best fitting power law to the degree distribution. Panels (d) and (e) show, respectively, the CC/CCer and MSPL/MSPLer ratios. The legend for all panels is shown in the upper left of the figure.

3.4.1. Association norms

The results for the association norms are shown as the blue line with diamonds in Fig. 2 (and subsequent figures). At all thresholds examined, the right-hand tail of the degree distribution was well fit by a power function (Fig. 2a), and that fit was clearly superior to the fit of an exponential function (Fig. 2b). The slope of that best fitting power function was approximately −2.80 (Fig. 2c) decreasing from −2.98 to −2.74 as the threshold increased. All the graphs showed strong clustering, with the strength of that clustering generally increasing as the threshold was increased (and the density decreased),

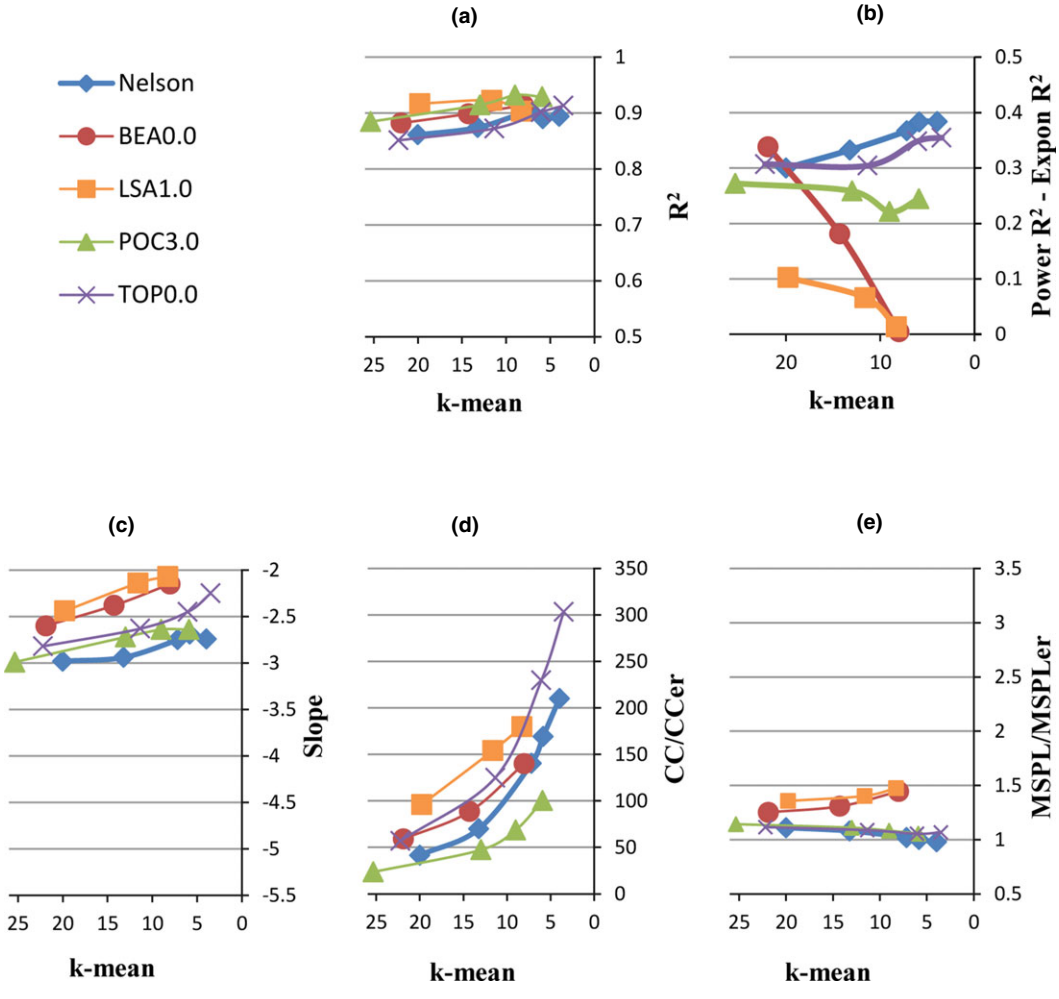


Fig. 2. Fits of simple models to the association norms. Observed values from the norms and model predictions are shown in each panel. Panel (a) shows the R^2 of the best-fitting power function to the right-hand tail of the degree distribution as a function of mean degree (k-mean). Panel (b) shows the difference between the R^2's of the best fitting power and exponential functions to the right-hand tail of the degree distribution as a function of mean degree. Panel (c) shows the slope of the best fitting power function to the right-hand tail of the degree distribution as a function of mean degree. Panel (d) shows the ratio of the clustering coefficient to the clustering coefficient expected in an equivalent Erdos-Rényi graph (CC/CCer) as a function of mean degree. Panel (e) shows the ratio of the Mean Shortest Path Length to the Mean Shortest Path Length expected in an equivalent Erdos-Rényi graph as a function of mean degree. The notation used in the legend is described in the caption of Fig. 1.

as indicated by the CC/CCer ratio, shown in Fig. 2d. Finally, the shortest path between any two nodes tended to be small. As shown in Fig. 2e, the MSPL/MSPLer ratio was about 1 and remained flat or decreased slightly as the threshold increased.

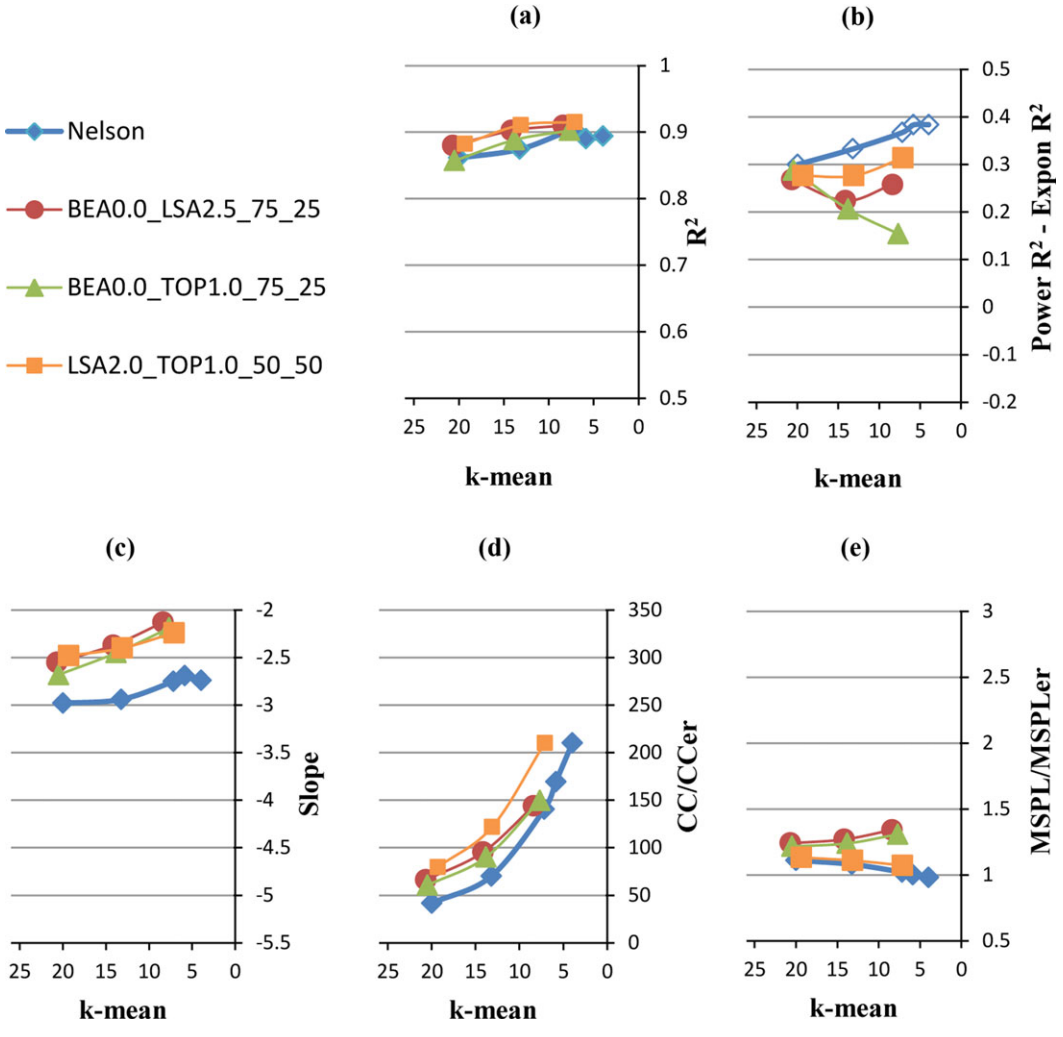


Fig. 3. Fits of contextual × contextual hybrid models to the association norms. Observed values from the norms and model predictions are shown in each panel. The panels follow those shown in Fig. 2. The notation used in the legend is described in the caption of Fig. 1.

Steyvers and Tenenbaum (2005) reported statistics for a graph based on the same set of association norms used here (see also Griffiths et al., 2007), using a single threshold, corresponding to the case of *All* in the present study. Our results for the associations norms closely parallel theirs as well as those of De Deyne and Storms (2008) for the Nelson et al. (1999) norms.

3.4.2. Fits of the models to the norms

Fig. 2 shows the fit to the norms of one exemplar of each of the models we tested (where exemplars differ based on the value of *w*) for each of the dependent variables.

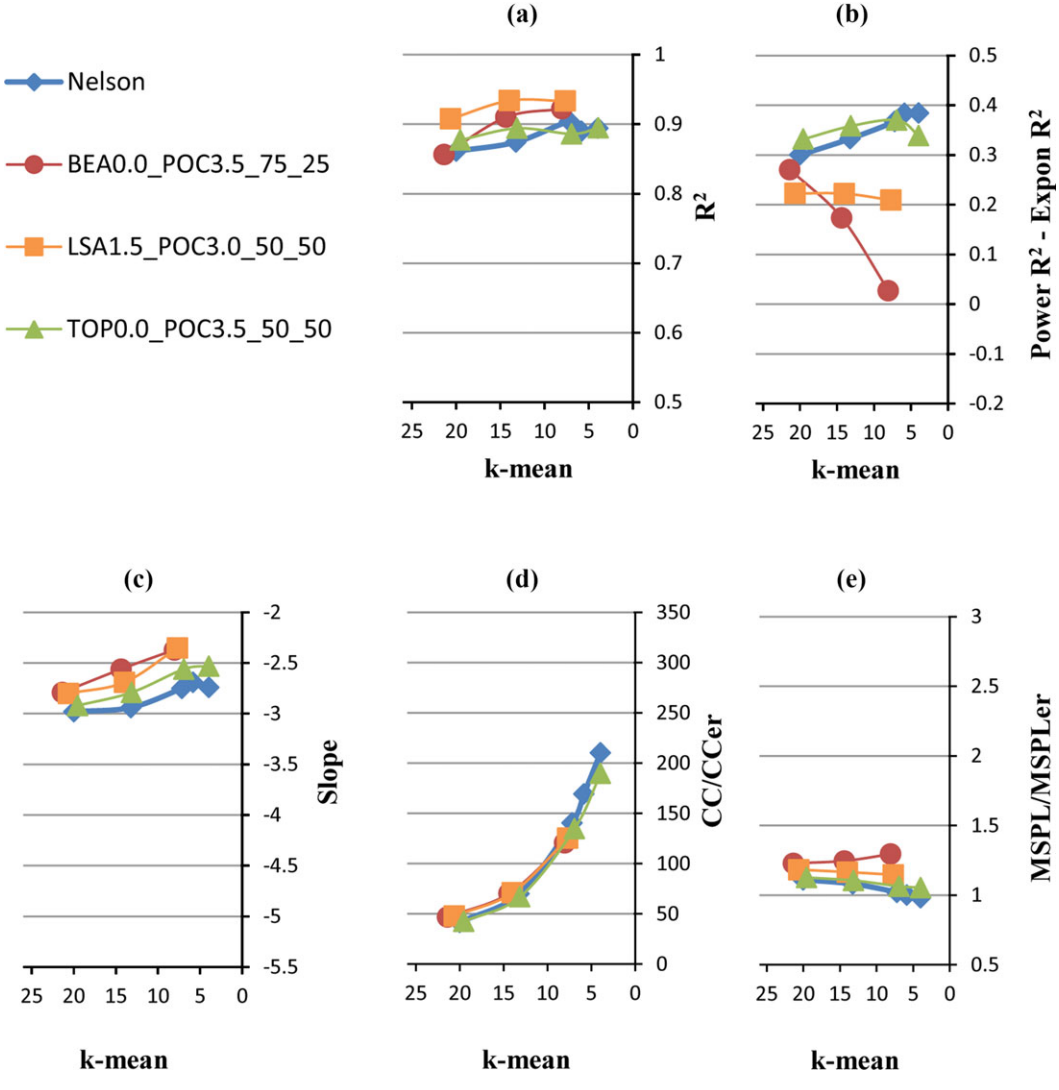


Fig. 4. Fits of contextual × POC hybrid models to the association norms. Observed values from the norms and model predictions are shown in each panel. The panels follow those shown in Fig. 2. The notation used in the legend is described in the caption of Fig. 1.

Recall that we chose the exemplar for which the value of *w* best fit the CC/CCer ratio. The values of *w* used in the exemplars shown in Fig. 2 were 0, 1, 0, and 3, respectively, for BEAGLE, LSA, Topic, and POC. As is evident in Fig. 2d, none of the models were able to accurately predict the amount of clustering in the norms. Their behavior, though, was quite consistent. The three contextual models consistently over-predicted clustering, regardless of the value of *w*. Utsumi (2014) also recently examined the ability of LSA to predict the network properties of the Nelson norms. That study used only a single

threshold but varied the parameters used to construct the LSA space. The same tendency for LSA to over-predict clustering that we observed is also evident in Utsumi's data. POC, in contrast, consistently under-predicted clustering, again regardless of the value of *w*. For both the Topic model and POC, the value of *w* had little effect on the model's ability to predict clustering (with the exception that at $w = 0$, POC under-predicted clustering more than it did at other values of *w*). For BEAGLE and LSA, the amount of over-prediction increased as *w* increased beyond the value used in Fig. 2.

For reasons discussed earlier, fundamental assumptions of the models lead to predictions of higher clustering in contextual models than in associative models. The results confirm that reasoning. The fact that clustering is too high in the contextual models and too low in the associative model suggests that neither class of model alone is able to accurately predict clustering.

Both the Topic model and POC were able to predict the mean shortest path of the norms, as shown in Fig. 2e.[9] For these two models, those fits remained good across the values of *w* tested. BEAGLE and LSA, in contrast, tended to predict path lengths that were somewhat too long. This tendency is especially noticeable at the higher thresholds (see the Supplementary Material), where only stronger associations are included and weaker associations are excluded. Hence, weaker links are necessary for BEAGLE and LSA to produce short paths, a finding that fits nicely with that of De Deyne, Navarro, and Storms (2013), who found that including weaker links improved the ability to predict human relatedness judgments from word association data. The fits of both BEAGLE and LSA to the path length data did improve at higher values of *w*, but at the expense of worse fits to the CC/CCer ratio.

Although all four models produced reasonably good fits to power functions, only the Topic model and POC consistently produced fits that were superior to fits of exponentials (see Figs. 2a and b). As the threshold increased, the degree distributions produced by BEAGLE and LSA were equally well fit by exponentials. That situation can be improved for both these latter models by choosing different values of *w*, but only at the expense of worsening the fits to the clustering data.

Finally, for the model exemplars shown in Fig. 2c, the Topic model and POC were able to reasonably predict the slopes of the degree distribution. For the Topic model, this fact was true across values of *w*, though the predicted slopes were consistently a bit shallower than the observed slopes. For POC, these fits were not strongly affected by the value of *w*, with the exception that at $w = 0$, the slopes tended to be too steep across thresholds. For both BEAGLE and LSA, across the values of *w* tested, the pattern evident in Fig. 2c was observed—predicted slopes were consistently shallower than the observed slopes.

3.5. Discussion

None of the models examined here was able to adequately predict the observed network properties of the association norms. We did some limited testing (the parameter *w* was fixed at 0) of a Pointwise Mutual Information (PMI) metric (Church & Hanks, 1990), an alternative to POC as an associative model. In PMI, the similarity parameter η

in Eq. 1 is set to the logarithm to the base 2 of the ratio of the frequency with which the two words co-occur to the product of the frequency of occurrence of each individual word. The results for the PMI model were similar to those for the POC model with $w = 0$. Results were also similar for POC models using window sizes of 5 and 15 instead of the whole document and with w varying—path lengths and degree distributions were well predicted, clustering was under-predicted. We also tested four models (BEAGLE, LSA, PMI, and POC) trained on a Wikipedia corpus (Recchia & Jones, 2009), again with w fixed at 0. The results were quite similar to those obtained for the analogous models using the TASA corpus. Finally, we tested those same four models with $w = 0$ trained on both the TASA and Wikipedia corpora but using the Edinburgh Word Association norms (Kiss, Armstrong, Milroy, & Piper, 1973) instead of the Nelson norms. The results were similar to the analogous models using the Nelson norms. Hence, our results do not seem to be an artifact of either the corpus or the set of norms we used.

Our overall conclusion from Study 1 is that, on their own, the contextual models BEAGLE, LSA, and the Topic model, and the associative model POC are unable to account for the graph-theoretic properties of association norms. Adjusting parameters to fit one pattern in the data inevitably puts each model off target for another pattern in the data, illustrating how difficult it is for these models to simultaneously explain all the network properties observed in the association norms. In fact, none was able to accurately predict the amount of clustering observed in the norms at any of the parameter values that we tested. Clustering, as noted earlier, appears to be of particular theoretical and psychological significance.

4. Study 2: Graphs generated by hybrid models

Deese (1965) speculated that the responses participants make in word association tasks are determined by multiple processes. The response made by one participant to one word could be determined by one process, whereas the response made by that same participant to a second word (or by a second participant to the same word) could be determined by a different process (cf. Morais et al., 2013). If in fact responses made in a word association task reflect multiple representation/process pairs, then it is not surprising that none of the models examined above was able to predict the network structure observed in the norms, since each model reflects a single representation/process pair.

In Study 2, we explored the hypothesis that first responses in word association are generated by multiple representation/process pairs. Suppose that on some trials, an associative network determines a participant's response. Those responses would form a network like that generated by the POC model. On other trials, the participant generates a response using a contextual representation, like those used in BEAGLE, LSA, or the Topic model. Those responses would form a network like that generated by the corresponding model. The observed graph will be a composite of these two graphs.

In the context of our models, such composite graphs are easy to construct, given one simplifying assumption, described below. Suppose that on a proportion p of trials,

participants use an associative network, like that which POC is intended to model, to generate responses, and on a proportion 1-p of trials they use an LSA representation. Then a proportion p of the word pairs would be those generating the highest values of Eq. 1 as determined by a POC similarity measure, and a proportion 1-p pairs would be those generating the highest values of Eq. 1 as determined by an LSA measure. For example, suppose that $p = .7$ and we set our threshold such that 50,000 word pairs would be used to build the graph. Then, we use Eq. 1 to generate the relative strengths of word pairs according to POC and select the 35,000 ($.7 \times 50{,}000$) most related pairs in the POC model, and we use Eq. 1 to generate the relative strengths of word pairs according to LSA and select the 15,000 ($.3 \times 50{,}000$) most related pairs in the LSA model.[10] The graphs are then constructed using those 50,000 word pairs (which may include some duplicates, since POC and LSA could both produce the same associate to a word). That is the strategy that we followed in testing hybrid models; it directly parallels the method used to construct networks for the simple models.

The necessary simplifying assumption is that the strengths of the associates generated by Model 1 and the strengths of the associates generated by Model 2 (where Models 1 and 2 are the two simple or base models comprising the hybrid) are evenly distributed across the overall distribution of strengths. Essentially, the simplifying assumption allows us to hold p constant as we vary the threshold level.

Study 2 investigated hybrid models involving all six possible combinations of the four base models. Hence, the hybrid models we tested were BEAGLE/LSA hybrids, BEAGLE/Topic hybrids, BEAGLE/POC hybrids, LSA/Topic hybrids, LSA/POC hybrids, and POC/Topic hybrids. Multiple versions of each hybrid were tested, where the versions differed in the values of the parameters p (the proportion of responses generated by a given model) and w (the weight given to the word frequency bias parameter in a model). A subscript is used to indicate which of the three models the parameter p refers to: p_b, p_l, p_t and p_p refer, respectively, to the values of p for the BEAGLE, LSA, Topic and POC models. Note that for a given hybrid pair, setting the value of p_x determines the value of p_y: $p_y = 1 - p_x$. In a hybrid pair, w was allowed to take on different values for the two base models in that pair.[11] Subscripts are thus also used when necessary to indicate the model to which the parameter w refers.

Our decision to test hybrids involving two contextual models (BEAGLE/LSA, BEAGLE/Topic, LSA/Topic) may seem somewhat strange. POC measures associative similarity, whereas BEAGLE, LSA, and the Topic model are measures of contextual similarity. Hence, POC could be considered complementary to the three contextual models, but BEAGLE, LSA, and the Topic model are more competing models than complementary models. If BEAGLE and LSA are competing models, why test a hybrid of the two? We would suggest that the fact that they are competing models is precisely the reason why a thorough testing of hybrids involving two contextual models is necessary. Contextual models have been reasonably successful at predicting human performance in a variety of semantic tasks, suggesting that they are approximating some true aspect of semantic representation. If they are competing models, that is, if they are all approximating the same true aspect of semantic representation, then hybridizing two contextual models should fail

to produce a model that does a good job of predicting the structure in the norms. If such a hybrid is found that can predict the structure of the norms, then the success of that hybrid (and by implication other hybrids) is likely simply due to the fact that they have one more parameter (p) than the simple models that we tested. In other words, success of such a hybrid would suggest that the success of hybrids in general might merely reflect the fact that they include an additional free parameter. Accordingly, we examined hybrids of two contextual models as a control condition.

4.1. Method

4.1.1. Constructing the word association graphs

The same graphs built from the word association norms that were used in Study 1 were also used in Study 2.

4.1.2. Constructing graphs based on model predictions

Six sets of graphs based on hybrid models were constructed, corresponding to the six possible combinations of the base models (BEAGLE/LSA, BEAGLE/Topic, BEAGLE/POC, LSA/Topic, LSA/POC, and Topic/POC). For each base model, the same calculations of the word-pair similarity parameter, $\eta_{i,j}$, were used in Study 2 as in Study 1. Individual graphs in the set corresponding to each hybrid differed in terms of the values of p and w used to create them. For all six hybrid combinations, p was varied from 0.25 to 0.5 to 0.75. We explored values of w around those values that provided the better fits in the case of simple models. The values of w used for each hybrid are shown in Table 1. For all hybrid combinations, graphs were constructed for all factorial combinations of the relevant values of p and w. Similar to the case for Study 1, the graphs for a particular combination of parameter values differed according to the threshold on the value of Eq. 1 used to determine whether to place an edge between two words. Several different threshold values were used for each combination of parameter values tested.

4.2. Results

All hybrids across all values of p and w tested and the results across all thresholds are shown in the supplementary material, Figs. S-5 through S-61. Here, we show a single exemplar of each hybrid pair. The criteria used for selecting the shown hybrid were the same as those used in Study 1—we first found those parameters that resulted in the best fit to the CC/CCer ratio. In the event of a "tie," we next examined the MSPL/MSPLer ratio. If multiple exemplars still fit approximately equally well, we then looked at the slope of the degree distribution.

4.2.1. Contextual × contextual hybrids

The results for hybrids consisting of two contextual base models are shown in Fig. 3. The middle panel of Fig. 1 shows the proportion of words included in the largest island of each contextual x contextual hybrid (BEAGLE/LSA, BEAGLE/Topic, and LSA/Topic)

shown in Fig. 3 as a function of k-mean. At higher thresholds (lower densities) that proportion dropped off much more quickly for the hybrids than for the norms themselves. For that reason, similar to the case for Study 1, the results shown in Fig. 3 are restricted to the three lowest thresholds.

As can be seen in Fig. 3d, all the contextual x contextual hybrids over-predicted the clustering coefficient, with the two hybrids involving BEAGLE coming the closest to the norms. These two hybrids in turn, though, over-predicted path lengths, whereas the LSA/Topic hybrid accurately predicted path lengths, as seen in Fig. 3e. There are BEAGLE/LSA and BEAGLE/Topic hybrids that do fit the path length data somewhat better, but at the expense of worse fits to the clustering data.

All of these hybrid pairs generally produce degree distributions that are better fit by power functions than by exponentials across all thresholds (see Figs. 3a and 3b). The slopes for the exemplars shown, however, are all much shallower than the slope of the degree distribution for the norms (see Fig. 3c). Parameter values can be found where all these hybrids provide much better fits to the slopes. However, those better fits come at the expense of worse fits to the clustering data and the path length data.

4.2.2. Contextual × POC hybrids

Fig. 4 shows the results for one example of each of those hybrids in which one base model is POC and the other is one of the three contextual models. The bottom panel of Fig. 1 shows the proportion of words included in the largest island of each of those exemplars. At lower densities, the size of the largest island began dropping earlier for BEAGLE/POC and LSA/POC hybrids than for the norms. This behavior was typical across all BEAGLE/POC and LSA/POC hybrids. The size of the largest island remained relatively high to lower densities for the Topic/POC hybrid. Again, this pattern was typical across all Topic/POC hybrids. Based on those results, the results in Fig. 4 show the data for the highest three densities for BEAGLE/POC and LSA/POC hybrids and the highest four for Topic/POC hybrids.

As seen in Fig. 4d, all these hybrids are capable of quite accurately predicting the amount of clustering in the norms. These hybrids are also quite capable of predicting path lengths as well, though the BEAGLE/POC hybrid does slightly over-predict them (see Fig. 4e). These hybrids also generally correctly predict degree distributions better fitted by power functions than by exponentials, though at the lowest density depicted, the degree distribution produced by BEAGLE is about equally well fit by an exponential (see Figs. 4a and 4b). Finally, as seen in Fig. 4c, all three of these hybrids reasonably accurately reproduce the slope of the degree distributions.[12]

As noted earlier, the hybrid networks may contain duplicate word pairs—word pairs that are included in the top associates of both component models. Intuitively, more duplicates might be expected in contextual × contextual hybrids than in contextual × POC hybrids as the former include only measures of contextual similarity whereas the latter reflect measures of both contextual similarity and associative similarity. Consequently, the contextual × POC hybrids might do better at predicting the graph-theoretic properties simply because they are more different from the individual component models than are

the contextual × contextual hybrids. Since the value of the parameter p can also be expected to affect the number of duplicates, we examined this question separately for the cases shown in Fig. 3 and Fig. 4 where $p = .75$ (the three hybrids involving the BEAGLE model) for one model and where $p = .5$ for both models. The proportion of duplicates was approximately the same in the BEAGLE × POC hybrids (0.10 for all three threshold levels shown in Fig. 4) as in the BEAGLE x LSA hybrids (ranging from 0.10 to 0.12 across the three threshold levels shown in Fig. 3) and in the BEAGLE x Topic hybrids (ranging from 0.11 to 0.12 across the three threshold levels shown in Fig. 3). In the case of hybrids with $p = .5$, the LSA × POC hybrids did have a smaller proportion of duplicates (ranging from 0.12 to 0.14 across the three threshold levels shown in Fig. 4) than did the LSA x Topic hybrids (ranging from 0.15 to 0.17 across the three threshold levels shown in Fig. 3). However, the proportion of duplicates was about the same in the Topic × POC hybrids (ranging from 0.16 to 0.17 across the three thresholds shown in Fig. 4) as in the LSA × Topic hybrids. Nevertheless, the Topic × POC hybrids performed better. Further, they performed as well as the LSA × POC hybrids, even though the latter had a smaller number of duplicates. It is unlikely, then, that the superior performance of the POC hybrids is simply the result of them containing a larger number of different word pairs than the other hybrids.

4.3. Discussion

Hybrid models constructed from two base contextual models were unable to predict the graph theoretic properties of the word association norms. A failure common to all these hybrids was their consistent over-prediction of clustering, a weakness shared by the individual base models comprising these hybrids.

In contrast, hybrids consisting of one contextual model and the simple associative model POC were able to predict the properties of the word association networks, including the amount of clustering, but also the shape and slope of the degree distributions and the mean shortest path lengths. Apparently, at least for the set of models examined here, to successfully emulate the network structure of human lexical semantic memory, a blend of a simple associative model with a more sophisticated contextual model is necessary.

5. Study 3: Predicting raw responses in the word association task

We also examined the relative ability of the simple and hybrid models to predict the associates actually produced in the norms by determining the probability that the top five associates of each stimulus word in the norms were included in the top t associates produced by each model to that stimulus word. For simple models, for each stimulus word, we simply selected the strongest t associates as determined by Eqs. 1 and 2. For hybrid models, we selected the strongest p_1t associates from Model 1 and the strongest $(1 - p_1)t$ associates from Model 2, again as determined by Eqs. 1 and 2. In the event that the same stimulus—response pair was selected from each component model, only one occurrence

was kept; the other was replaced with a new pair from Model 1 (with probability p_1) or Model 2 (with probability $1 - p_1$). We began by examining those hybrids displayed in Figs. 3 and 4 and the simple models comprising them, that is, the hybrids that provided the best fit to the network properties for each pair-wise combination of simple models.

5.1. Results

The results are shown in Table 2. The following notation is used in that table: XXXw.w_YYYw.w_p_q, where XXX is a three-letter abbreviation indicating the first

Table 2
Proportion of the top five associates in the norms included in the top t associates predicted by each model

	Number of Model Top Associates Considered						
Model	8	16	24	32	40	48	56
Group 1							
BEA0.0	0.185	0.246	0.284	0.313	0.337	0.358	0.376
LSA2.5	0.154	0.219	0.263	0.296	0.324	0.347	0.37
BEA0.0_LSA2.5_75_25	0.195	0.261	0.311	0.333	0.357	0.379	0.398
Group 2							
BEA0.0	0.185	0.246	0.284	0.313	0.337	0.358	0.376
TOP1.0	0.246	0.321	0.367	0.398	0.423	0.443	0.461
BEA0.0_TOP1.0_75_25	0.223	0.298	0.343	0.376	0.402	0.424	0.443
Group 3							
LSA2.0	0.161	0.227	0.271	0.304	0.331	0.357	0.379
TOP1.0	0.246	0.321	0.367	0.398	0.423	0.443	0.461
LSA2.0_TOP1.0_50_50	0.201	0.272	0.315	0.346	0.369	0.389	0.405
Group 4							
BEA0.0	0.185	0.246	0.284	0.313	0.337	0.358	0.376
POC3.5	0.162	0.225	0.268	0.3	0.326	0.349	0.367
BEA0.0_POC3.5_75_25	0.193	0.263	0.307	0.34	0.366	0.39	0.412
Group 5							
LSA1.5	0.166	0.234	0.279	0.311	0.339	0.363	0.383
POC3.0	0.161	0.226	0.272	0.304	0.332	0.354	0.373
LSA1.5_POC3.0_50_50	0.166	0.234	0.279	0.311	0.34	0.36	0.38
Group 6							
TOP0.0	0.253	0.331	0.377	0.409	0.436	0.455	0.472
POC3.5	0.162	0.225	0.268	0.3	0.326	0.349	0.367
TOP0.0_POC3.5_50_50	0.234	0.321	0.372	0.406	0.434	0.459	0.478
Group 7							
TOP0.0	0.253	0.331	0.377	0.409	0.436	0.455	0.472
POC3.0	0.161	0.226	0.272	0.304	0.332	0.354	0.373
TOP0.0_POC3.0_50_50	0.24	0.321	0.373	0.409	0.438	0.461	0.483
Group 8							
TOP1.0	0.246	0.321	0.367	0.398	0.423	0.443	0.461
POC3.5	0.162	0.225	0.268	0.3	0.326	0.349	0.367
TOP1.0_POC3.5_50_50	0.23	0.312	0.364	0.4	0.429	0.453	0.473

base model, YYY is a three-letter abbreviation indicating the second base model, w.w is the value of *w* used with the immediately preceding base model, *p* is the percent of responses generated by the first base model, and q (= 100 – *p*) is the percent of responses generated by the second base model. Thus, BEA0.0_LSA2.5_75_25 means that for the BEAGLE component *w* was set to 0, for the LSA component *w* was set to 2.5, and that 75% of the pairs were drawn from BEAGLE and 25% from LSA. The table is divided into groups of three, where the first two elements of a group are simple models and the third is the hybrid composed of those two simple models. The first six groups of models in the table show the results for the six hybrids displayed in Figs. 3 and 4.

Of the simple models, the Topic model clearly does the best job predicting the top five associates in the Nelson norms. That superiority is evident for the simple models in Table 2 and also for other values of *w* that we examined but that are not shown in Table 2. Griffiths et al. (2007) found a similar superiority of the Topic model over LSA; we can add that the Topic model is also superior to BEAGLE and POC.

For contextual × POC hybrids (groupings 4 through 6 in Table 2), the BEA0.0_POC3.5 hybrid performed better than both its simple models. For the LSA1.5_POC3.0_50_50 and TOP0.0_POC3.5_50_50 models, the hybrid and the better performing simple model (LSA for the LSA/POC hybrid and the Topic model for the Topic/POC hybrid) performed approximately equivalently in predicting the top 5 observed associations. We also examined the TOP0.0_POC3.0_50_50 hybrid, as this model's predictions of the network properties differed only negligibly from the TOP0.0_POC3.5_50_50 model. Those results are shown in the seventh group in Table 2. This hybrid also performed approximately equivalently to its better component. In order to equate the value of *w* used with the Topic model across its hybrids with BEAGLE, LSA, and POC, we examined the TOP1.0_POC3.5_50_50 model (see the last group in Table 2). This hybrid model also performed approximately the same as its Topic component.

The better performance with POC hybrids does not simply reflect the fact that the hybrids are selecting the top associates from each component model and hence will inevitably do better. Although the BEAGLE_LSA hybrid does perform somewhat better than its better component (BEAGLE, see Group 1 in Table 2), the BEAGLE_Topic and LSA_Topic hybrids perform worse than their better component (Topic in both cases, see Groups 2 and 3 in Table 2). Creating a hybrid clearly does not necessarily result in performance as good as or better than the better component.

If attention is restricted to only the hybrid models, as can be seen in Table 2, five of the six hybrids group fairly closely together with respect to performance, Groups 1 through 5 in Table 2. The one hybrid that stands out as superior is the Topic_POC hybrid, a result suggesting that of the models we have examined, the Topic_POC hybrid is the one most capable of explaining the word association data.

5.2. Discussion

The importance of the findings across Studies 2 and 3 is as follows. When the graph-theoretic properties are included in the analyses, we find that hybrids that include POC as

one component are able to adequately predict the network properties while doing as well as or better at predicting the top associates than the better of the two component models. None of the simple models, however, is able to predict those graph-theoretic properties. This pattern of findings supports the hypothesis that word associations are the result of a hybrid model, one component of which is a contextual representation and the other component of which is an associative network. Finally, although an analysis of only the graph-theoretic properties was ambiguous concerning which contextual model was the best candidate for that hybrid model, the analysis of predicted associates suggests that the Topic model best fills that role.

6. General discussion

None of the simple BEAGLE, LSA, TOPIC, or POC models that we tested were able to predict the graph-theoretic properties of the Nelson et al. (1999) word association norms. The contextual-by-contextual hybrids (BEAGLE/LSA, BEAGLE/Topic, LSA/Topic) that we examined were also unable to predict these properties. The fact that hybrids consisting of two contextual models did no better than the simple models indicates that hybrids do not automatically produce better fits simply because they include an additional parameter (p), or are in some way more robust. BEAGLE/POC hybrids were perhaps able to predict these properties well enough that such models of how word associations are produced should not be excluded from future consideration. LSA/POC hybrids and Topic/POC hybrids, in contrast to the simple models and to the other hybrids, were able to simultaneously predict all graph-theoretic properties of the word association norms that we examined. All those hybrids involving a contextual model plus POC were also able to predict the top five associates of each stimulus in the norms as well as or better than the better of their two component models, with the Topic/POC hybrid performing best on this measure. Overall, then, the results suggest that two components of lexical semantic memory contribute to first responses in word association, with one of those components being an associative network and the other a contextual representation, with the Topic model being the leading contender for the contextual representation.

There are of course a wide range of models of lexical semantic memory and measures of relatedness (see Bullinaria & Levy, 2007; Holyoak, 2008; Jones et al., 2014; Riordan & Jones, 2011 for lists and taxonomies). There are also many possible variants of the four base models used here. It is not feasible to test all possible variants of all possible models. Conceivably, some simple, non-hybrid model (including possibly a variant of one of those studied here) may someday be shown to be able to predict the graph-theoretic properties of word association data. All four models examined here learn their semantic representations based on the co-occurrences of word pairs in text. Potentially, a model based on sensorimotor features (e.g., McRae et al., 1997, 2005; Vigliocco, Vinson, Lewis, & Garrett, 2004; Vinson & Vigliocco, 2008) or a model that statistically combines co-occurrence information with sensorimotor features (Andrews, Vigliocco, & Vinson,

2005, 2009; Johns & Jones, 2012; Steyvers, 2010) could do as well as or better than the contextual-by-POC hybrids at predicting the network structure of association norms.

Despite these limitations, our results show that predicting graph-theoretic properties of association norms, in particular, the amount of clustering in those norms, can be accomplished by hybrids of a contextual model and an associative model. To our knowledge, no simple model or other combination of hybrid models has been shown capable of making those same predictions. Our hybrid models assume that first responses in word association tasks are sometimes generated by processes operating on an associative network and are at other times generated by processes operating on a representation of word meaning reflecting contextual similarity. Recent evidence (Morais et al., 2013) suggests that the network properties of word association data differ from one individual to another. Conceivably, such individual differences could arise from the relative reliance that different people place on one of those two representations versus the other.

7. Multi-component views of lexical semantic memory

Our results join an increasing number of studies suggesting that there are multiple components to lexical semantic memory, with the component evident in any given task dependent upon the demands of that task, the information available in each component, and the relative speed of access to the information in that component (Barasalou, Santos, Simmons, & Wilson, 2008; De Deyne & Storms, 2008; Gruenenfelder, 1986; Hampton, 1997; Lorch, 1981; Louwerse, 2011; Maki & Buchanan, 2008; McRae et al., 1997).

For example, McRae et al. (1997) argued that relatively fast access to featural representations is important when initially determining a word's meaning, but that higher level conceptual representations or theory-based knowledge is required when reasoning with concepts. Barasalou et al. (2008) argued in their LASS (Language and Situated Simulation) framework for a division of semantic memory into a linguistic level and a level involving situated simulations. The linguistic level is accessed relatively quickly but is responsible for only superficial processing. True conceptual processing occurs at the deeper level of situated mental simulations, analogous to both the featural representations and the higher-level conceptual representations of McRae et al.

De Deyne and Storms (2008), using a multi-response word association task, noted a shift in the type of conceptual relations reflected in the second and third responses as compared to the first response. Based on these shifts, De Deyne and Storms argued that initial responses reflected relatively quick access to a linguistic, LSA-like representation, whereas second and third responses reflected slower access to representations at a more conceptual level. The present results do not conflict with either the LASS framework or with De Deyne and Storms's results. Both representations that we are suggesting play a role in word association—the associative network and the contextual representation—operate, in terms of LASS, at the linguistic level. As such, our proposal can be considered a refinement of LASS at the linguistic level more than as an alternative to LASS. Similarly, De Deyne and Storms and we seem to agree that initial responses in a word association

task reflect processes operating at the linguistic level. We offer a more detailed view of how the first response to a cue word is made in a word association task, but we do not argue with their view that second and third responses reflect a shift to another form of representation at a more conceptual level.

Why have two components—an associative network and a high-dimensional spatial representation reflecting contextual similarity—of lexical semantic memory at the linguistic level? Two non-mutually exclusive alternatives suggest themselves. The two different components could better lend themselves to representing different aspects of a word's meaning. In addition, the two components could predominantly be used to represent information about different types of concepts.

Consider the first possibility, that the different components are used to represent different aspects of a word's meaning. Olson (1970) has argued that knowledge of coordinate relations (e.g., the relation between "red" and "blue" or between "sparrow" and "robin") is critical to resolving reference. There is evidence from the category verification task (in which participants verify as quickly as possible statements of the form, "An S is a P," e.g., "A cucumber is a vegetable") that coordinate information is represented in an associative network, whereas featural information (or perhaps contextual similarity—the data do not discriminate between these two possibilities) is used to make fine category judgments. Compared to the case where all false sentences pair semantically unrelated concepts ("A cucumber is a car"), the size of the typicality effect[13] for true items increases when false sentences pair concepts with a category coordinate to their appropriate category ("A cucumber is a fruit") (McCloskey & Glucksberg, 1979; Experiment 2; see also Hampton, 1997, Experiment 2). Such a result follows from models that assume that category verification involves a comparison of the semantic features of the two concepts in the to-be-verified sentence (e.g., Gellatly & Gregg, 1975, 1977; Hampton, 1997; McCloskey & Glucksberg, 1979). Such a result would also seem to follow from models that use cosine similarity within a high-dimensional spatial semantic representation to explain performance in category verification (Louwerse, Cai, Hu, Ventura, & Jeuniaux, 2006).

In contrast, the size of the typicality effect for true items is invariant when coordinate false items ("cucumber—okra") are used compared to the case where only anomalous false items are used (Gruenenfelder, 1986). Presumably, retrieving an association between the two concepts from an associative network is sufficient to determine that an item is true when all the false items pair unrelated concepts. When false items include coordinate items, that retrieval must be followed by a process that evaluates the labels on the retrieved association in order to discriminate category and coordinate statements (cf. Glass & Holyoak, 1975; Holyoak & Glass, 1975). That evaluation process is presumably independent of typicality. The overall result is no change in the size of the typicality effect. In addition, both Gruenenfelder (1986) and Lorch (1981) found that, contrary to the usual pattern, false items pairing coordinates were rejected more quickly in a category verification task the more strongly related the two concepts were (but see also Lorch, 1978), consistent with the hypothesis that rejecting these statements as false involves retrieving an association between the two words (see also Glass & Holyoak, 1975; Holyoak & Glass, 1975). These results, then, suggest that people rely on an associative

network to determine coordinate relations, but on featural or contextual information to make fine-grained category judgments.

The second, non-mutually exclusive reason for both a contextual representation and an associative network at the lexical level is that the two could be used to represent information about different types of concepts, perhaps because the way those concepts are encountered in the world lends itself to one form of learning over another. Recchia (2012) has recently produced evidence consistent with the hypothesis that people rely on information in an associative network-like structure to perform a variety of semantic tasks with concrete concepts but rely on a contextual representation to perform those same tasks with abstract concepts. For example, the number of semantic neighbors, a measure reflecting contextual similarity, predicted performance in lexical decision for words denoting abstract concepts but not for words denoting highly concrete concepts. In contrast, number of features predicted lexical decision performance for highly concrete but not abstract words (Recchia & Jones, 2012). Recchia (2012) noted that these features tended to denote visual properties and physical components of concepts denoted by concrete words, and these properties in turn had high associative relatedness with concrete words. Similarly, LSA cosines, a measure of contextual similarity, predicted priming for abstract words but not for concrete words. Finally, Recchia examined features generated to cue words in a feature generation task. He found that a hybrid model of PMI (similar to the POC measure) and LSA better predicted the features generated to a particular cue word than did either PMI alone or LSA alone. Furthermore, associative strengths between cues and generated features were higher for concrete cues than for abstract cues, while contextual similarities were higher for abstract cues than for concrete cues. The overall pattern is consistent with the use of an associative network for tasks involving early processing of concrete words, but with more reliance on a contextual representation when performing those same tasks on abstract words.

To summarize, our results indicate that two different forms of representations, each operating at the linguistic level, are sufficient to explain word association data. These two different representations could be used to solve different problems faced in language comprehension and production. An associative network appears well suited for representing information about coordinate relations, and for encoding information about the physical properties of concrete words. Contextual similarity may be more appropriate for making fine-grained semantic decisions and for representing information about abstract words.

Acknowledgments

This work was supported by National Science Foundation grant BCS-1056744 and by a grant from Google Research to MNJ. GR is now at the Cambridge Centre for Digital Knowledge, University of Cambridge, Alison Richard Building, 7 West Road, Cambridge, CB3 9DT, United Kingdom.

Notes

1. In Nelson et al.'s (1999) terminology, the response word is referred to as the *target.* We follow the clearer terminology of most subsequent articles that refer to this elicited word as the *response*.
2. The term *semantic relatedness* is sometimes used in a similar way in the literature. However, we do not wish to preclude the possibility that two concepts can be semantically related while sharing few if any features.
3. In undirected graphs, which are the focus of the present work, a node's *degree* is simply the number of edges connected to that node.
4. Strictly speaking, because the Topic model directly predicts the probability of response R, given stimulus S, there was no need to apply Eq. 1 to the values generated by Eq. 9 of Griffiths et al. (2007) to this model. Doing so, and including all words in the denominator of our Eq. 1, results in the same values as applying only the Griffiths et al. (2007) Eq. 9. However, for reasons discussed below, we included only the 200 most strongly related word pairs in the denominator of Eq. 1 for the other three models. Hence, to make the comparisons amongst all four models as alike as possible, we did the same for the Topic model. Doing so does not change the relative order of the probabilities of different words being generated as responses to a particular stimulus word.
5. Their study involved information diffusion within a social network; the situation, though, is analogous to activation spreading in a memory network.
6. The concept of a path, and hence of path length, is most obviously applicable to associative network models of semantic memory. One of the most frequently cited reasons for examining path length is to determine the degree to which an associative network model has a small-world structure similar to that observed in human associative networks, as the combination of short path lengths and highly clustered neighborhoods in a small-world network is presumed to facilitate efficient retrieval (Morais et al., 2013; Steyvers & Tenenbaum, 2005). For similar reasons, the concept also has relevance to high-dimensional spatial models (Steyvers & Tenenbaum, 2005). Such models, after all, need a search mechanism. That mechanism might be a random walk from one point in the space representing a word to another point in the space representing an adjacent word. Alternatively, it might involve activation diffusing through space. As words are encountered, they are evaluated in order to determine if they meet the criteria of the search. In either case, the concepts of path and of path length would seem to be psychologically relevant.
7. We did examine the clustering coefficient and the degree distribution for all nodes in a given network, including those at all thresholds tested. For these measures, including nodes from all islands had at most a negligible effect on the results.
8. Technically, density and k_{mean} are linear transformations of one another.
9. The line for POC lies so close to that for the norms that it may be hard to discern in the figure.

10. This process is directly analogous to the method used for creating simple models, where we set a threshold on the strength of word pairs used to construct the network such that a particular value of k-mean was approximated.
11. Because different models have different mechanisms for incorporating word frequency into their base representations, we did not expect *w* to be constant across all models.
12. For the slope of the degree distribution, the Topic/POC hybrid does appear to produce better fits than either the BEAGLE/POC hybrid or the LSA/POC hybrid. Particularly in the case of the LSA/POC hybrid, we caution the reader against interpreting that observation as favoring Topic/POC hybrids over LSA/POC hybrids. As mentioned earlier, the Clauset et al. (2009) algorithm allows the user to fix the value of the lowest degree included in the fit, or allows that variable to be a free parameter. In the fits reported here and in the supplementary material, we fixed that parameter to be the median of the degree distribution. We also fit the degree distributions allowing that parameter to be a free parameter. In that case, the LSA/POC and Topic/POC hybrids continued to fit the slope reasonably well across thresholds. (The fit of the BEAGLE/POC hybrid became worse.) However, in that case, the LSA/POC fit was somewhat better than the Topic/POC fit.
13. The typicality effect refers to the increase in time needed to verify less typical members of a category ("okra—vegetable") as compared to more typical members of the category ("peas—vegetable").

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Andrews, M., Vigliocco, G., & Vinson, D. (2005). The role of attributional and distributional information in semantic representation. In B. Barra, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 127–132). Austin, TX: Cognitive Science Society.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 464–498.

Baayen, H. R., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language & Speech*, *56*, 329–347.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Barasalou, L., Santos, A., Simmons, K., & Wilson, C. (2008). Language and simulation in conceptual processing. In M. D. Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–283). New York: Oxford University Press.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy*, *12*, 1264–1302. doi:10.3390/e12051265

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*, 531–544.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science*, *325*, 414–416.

Chomsky, N. (1959). Review of *Verbal Behavior* by B. F. Skinner. *Language*, *35*, 26–58.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Clauset, A., Shaliza, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. arXiv:0706.1062v0702 [physics.data-an] 0702 Feb 2009.

Collins, A. M., & Quillan, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental Psychology*, *92*, 149–154.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*, 480–498.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*, 213–231.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, MD: Johns Hopkins University Press.

Enguix, G. B., Rapp, R., & Zock, M. (2014). How well can a corpus-derived co-occurrence network simulate human associative behavior? In A. Lenci, M. Padra, T. Poibeau, & A. Villavicencio (Eds.), *Proceedings of the 5th workshop on cognitive aspects of computational language learning* (pp. 43–48). Gothenburg, Sweden: Association for Computational Linguistics.

Erdos, P., & Rényi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, *5*, 17–61.

Gellatly, A. R. H., & Gregg, V. H. (1975). The effects of negative relatedness upon word-picture and word-word comparisons and subsequent recall. *British Journal of Psychology*, *66*, 311–323.

Gellatly, A. R. H., & Gregg, V. H. (1977). Intercategory distance and categorization times: Effects of negative-probe relatedness. *Journal of Verbal Learning and Verbal Behavior*, *16*, 505–518.

Glass, A. L., & Holyoak, K. J. (1975). Alternative conceptions of semantic theory. *Cognition*, *3*, 313–339.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Gruenenfelder, T. M. (1986). Relational similarity and context effects in category verification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 587–599.

Gruenenfelder, T. M., & Pisoni, D. B. (2009). The lexical restructuring hypothesis and graph theoretic analyses of networks based on random lexicons. *Journal of Speech, Language, & Hearing Research*, *52*, 596–609.

Hampton, J. A. (1997). Associative and similarity-based process in categorization decisions. *Memory & Cognition*, *25*, 625–640.

Hargreaves, I. S., & Pexman, P. M. (2014). Get rich quick: The signal to respond procedure reveals the time course of semantic richness effects during visual word recognition. *Cognition*, *131*, 216–242.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009a). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, *112*, 381–396.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009b). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*, 729–739.

Holyoak, K. J. (2008). Relations in semantic memory: Still puzzling after all these years. In A. Lesgold, B. Ross, E. A. Loftus, & W. K. Estes (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (pp. 141–158). New York: Lawrence Erlbaum Associates.

Holyoak, K. J., & Glass, A. L. (1975). The role of contradictions and counterexamples in the rejection of false sentences. *Journal of Verbal Learning and Verbal Behavior*, *14*, 215–239.

Hutchison, K. A. (2003). Is semantic priming due to associaiton strength or feature overlap? A *micro*analytic review. *Psychnomic Bulletin & Review*, *10*, 785–813.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*, 103–120.

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2011). In defense of spatial models of spatial models of lexical semantics. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3444–3449). Austin, TX: Cognitive Science Society.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Jones, M. N., Willits, J., & Dennis, S. (2014). *Models of semantic memory*. In D. Reisberg (Ed.), Oxford handbook of mathematical and computational psychology. New York: Oxford University Press.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychnomic Bulletin & Review*, *7*, 257–266.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: Edinburgh University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5214–5219.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analsysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lorch, R. F. J. (1978). The role of two types of semantic information in the processing of false sentences. *Journal of Verbal Learning and Verbal Behavior*, *17*, 523–537.

Lorch, R. F. J. (1981). Effects of relation strength and semantic overlap in retrieval and comparison processes during sentence verification. *Journal of Verbal Learning and Verbal Behavior*, *20*, 593–610.

Louwerse, M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Louwerse, M., Cai, Z., Hu, X., Ventura, M., & Jeuniaux, P. (2006). Cognitively-inspired NLP-based knowledge representations: Further explorations of Latent Semantic Analysis. *International Journal on Artificial Intelligence Tools*, *15*, 1021–1039.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychnomic Bulletin & Review*, *7*, 618–630.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Luce, P. A., & Pisoni, D. B. (1998). Recognzing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*, 1–36.

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychnomic Bulletin & Review*, *15*, 598–603.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 3–23.

Mathey, S., Robert, C., & Zagar, D. (2004). Neighbourhood distribution interacts with orthographic priming in the lexical decision task. *Language & Cognitive Processes*, *19*, 533–559.

Mathey, S., & Zagar, D. (2000). The neighborhood distribution effect in visual word recognition: Words with single and twin neighbors. *Journal of experimental Psychology: Human Perception & Performance*, *26*, 184–205.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, *11*, 1–37.

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisted. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1155–1172.

McNamara, T. (2005). *Semantic priming*. New York, NY: Psychology Press.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behavior Research Methods, Instruments, & Computers*, *37*, 547–559.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representation of word meaning. *Journal of Experimental Psychology: General*, *126*, 99–130.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, *37*, 125–145. doi:10.1111/cogs.12013

Moss, H. E., Hare, M. L., Day, P., & Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, *6*, 413–427.

Nelson, D. L., Bennett, D. J., Gee, N. R., Schreiber, T. A., & McKinney, V. M. (1993). Implicit memory: Effects of network size and interconnectivity on cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 747–764.

Nelson, D. L., & Goodmon, L. B. (2002). Experiencing a word can prime its accessibility and its associations to related words. *Memory & Cognition*, *30*, 380–398.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). The University of South Florida word association norms. Available at: http://w3.usf.edu/FreeAssociation. Accessed October 14, 2008.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*, 299–324.

Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explict associations. *Psychological Review*, *99*, 322–348.

Nelson, D. L., Zhang, N., & McKinley, L. N. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*, 1147–1159. doi:10.1037MK78-7393.27.5.1147

Nematzadeh, A., Ferrara, E., Flammini, A., & Ahn, Y.-Y. (2014). Optimal network clustering for information diffusion. *Physical Review Letters*, *113*, 1–5.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, *77*, 257–273.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*, 161–167.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*, 1017–1023. doi:10.1177/0956797612460691

Recchia, G. (2012). *Investigating the semantics of abstract concepts: Evidence from a property generation game. Unpublished doctoral disseration*. Bloomington, IN: Indiana University.

Recchia, G., & Jones, M. N. (2009). More data trumpts smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647–656.

Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, *6*(315), 1–16. doi:10.3389/fnhum.2012.00315

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*, 151–160.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*, 303–345.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1–20.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic knowledge: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*, 393–413.

Smith, E. E., Rips, L. J., & Shoben, E. J. (1974a). Semantic memory and psychological semantics. In G. H. Bower (Ed.), *The psychology of learning and motivation*. Vol. 8. New York: Academic Press.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974b). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*, 317–330.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*, 234–243.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analysis and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Stone, B., Dennis, S., & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science*, *3*, 92–122. doi:10.1111/j.1756-8765.2010.01108.x

Utsumi, A. (2014). Complex network analysis of distributional semantic models. In P. Bello, M. Guarine, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 3008–30013). Austin, TX: Cognitive Science Society.

Vigliocco, G., Vinson, D., Lewis, W., & Garrett, M. F. (2004). Representing the meaning of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*, 422–488.

Vinson, D., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*, 183–190.

Vitevitch, M. S., Ercal, G., & Adagarla, B. (2011). Simulating retrieval from a highly clustered network: Implications for spoken word recognition. *Frontiers in Psychology*, *2*, Article 369. doi:10.3389/fpsyg.2011.00369

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*, 440–442.

Wettler, M., Rapp, R., & Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in text. *Journal of Quantitative Linguistics*, *12*, 111–122.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*, 742–750.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Supplementary material Figures and Table.