New Ideas in Psychology xxx (2017) 1–7

Contents lists available at ScienceDirect

New Ideas in Psychology

journal homepage: www.elsevier.com/locate/newideapsych

In defense of spatial models of semantic representation☆

Michael N. Jones [a, *], Thomas M. Gruenenfelder [a], Gabriel Recchia [b]

[a] Indiana University, USA
[b] University of Cambridge, United Kingdom

ARTICLE INFO

Article history:
Received 6 January 2017
Received in revised form
22 August 2017
Accepted 25 August 2017
Available online xxx

ABSTRACT

Recent semantic space models learn vector representations for word meanings by observing statistical redundancies across a text corpus. A word's meaning is represented as a point in a high-dimensional semantic space, and semantic similarity between words is quantified by a function of their spatial proximity (typically the cosine of the angle between their corresponding vector representations). Recently, Griffiths, Steyvers, and Tenenbaum (2007) demonstrated that spatial models are unable to simulate human free association data due to the constraints placed upon them by metric axioms which appear to be violated in association norms. However, it is important to note that free association data is the product of a retrieval process operating on a semantic representation, and the failures of spatial models are likely be due to mistaking the similarity metric (cosine) for an appropriate process model of the association task—cosine is not what people do with a memory representation. Here, we test the ability of spatial semantic models to simulate association data when they are fused with a simple Luce choice rule to simulate the process of selecting a response in free association. The results provide an existence proof that spatial models can produce the patterns of data in free association previously thought to be problematic for this class of models.

1. Introduction

A longstanding belief in theories of lexical semantics, dating back at least to Osgood (1952) is that words can be represented as points in a multidimensional semantic space. Similarity between word meanings is then defined as some function of their distance in space. This classic notion of mental space has had an obvious impact on modern computational semantic space models, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). Models such as LSA borrow techniques from linear algebra to infer the semantic representation for words from their contextual co-occurrences in linguistic corpora. In the resulting space, a word's meaning is represented by a vector over latent dimensions. Inter-word similarity is based on Euclidean geometry: Words that are more similar are more proximal in the learned space. Virtually all distributional models of semantic memory adhere to the spatial notion of semantics (for a review, see Jones, Willits, & Dennis, 2015), including recent popular neural embedding models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

In contrast to spatial models, the popularity of probabilistic models of cognition has led to the development of Bayesian models of semantic representation, such as the LDA-based Topic models explored by Griffiths et al. (2007). In a Topic model, a word's representation is a probability distribution over latent semantic "topics." When a word is processed, its semantic representation is the predicted probability across latent topics. Hence while LSA represents a word as a point in high-dimensional space and requires a spatial metric of similarity between two words, a Topic model represents a word as a probability distribution and computes the association between words as the probability of one word given the other. This allows Topic models to make very different predictions depending on which word is being conditioned upon, in contrast to LSA in which similarity is identical regardless of which word is "first." In addition, the issue of whether humans represent meaning as a coordinate in space or as a conditional probability is a fundamental question in cognitive science, and has implications for downstream models that make use of these representations.

Tversky (1977) has noted that spatial models must respect several metric axioms. Firstly, in a metric space the distance between a point and itself must be zero by any Euclidean metric,

☆ This work was supported by NSF BCS-1056744 and IES R305A150546.
* Corresponding author. Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA.
E-mail address: jonesmn@indiana.edu (M.N. Jones).

http://dx.doi.org/10.1016/j.newideapsych.2017.08.001
0732-118X/

$d(x,x) = 0$ (*non-negativity*). Secondly, distance must respect *symmetry*: $d(x,y) = d(y,x)$. Thirdly, distance must respect the *triangle inequality*: If x and y are proximal and y and z are proximal, then x and z are likely to be proximal points as well (specifically, $d(x,z) \leq d(x,y) + d(y,z)$). As Tversky and Gati (1982) have demonstrated, human judgments of similarity routinely violate these axioms—specifically, symmetry and the triangle inequality. Tversky used human violations of the metric axioms to argue against spatial models of similarity, and instead proposed an additive feature comparison model. The spatial debate, however, has a long history in cognitive science, with Tversky's work being followed by explanations of how metric spaces could produce violations of metric axioms (e.g., Krumhansl's (1978) notion of density or Holman's (1979) similarity and bias model).

Griffiths et al. (2007) note that human free association norms also violate metric axioms, making them problematic for semantic space models such as LSA. In a free association task, the participant is asked to respond to a cue word with the first associated word that comes to mind (Nelson, McEvoy, & Schreiber, 2004). Word association norms contain a significant number of asymmetric associations: For example, the probability of generating *baby* as a response to *stork* as a cue is much greater than the reverse. Part of this effect is due to a bias to respond with a high frequency target independent of the cue, but part appears to be due to some sort of asymmetry in the computation of similarity. In addition, word association norms contain apparent violations of the triangle inequality axiom: To use the example from Griffiths et al., *asteroid* is strongly associated with *belt*, and *belt* is strongly associated with *buckle*, but *asteroid* and *buckle* have no association. Finally, Griffiths et al (see also Steyvers & Tenenbaum, 2005). have demonstrated that association norms contain neighborhood structure that is incompatible with spatial models. If one constructs an associative network with nodes representing words and connecting edges based on nonzero association probabilities, the resulting networks are *scale-free*: they have power law degree distributions and high clustering coefficients.[1]

Griffiths et al. (2007) note, however, that probabilistic representations are not subject to the same metric restrictions as spatial representations, and they provide an elegant demonstration of how Topic models can naturally account for the qualitative nature of violations in asymmetry and the triangle inequality that LSA cannot. Griffiths et al. further demonstrate that while LSA (based on a thresholded cosine) cannot reproduce the scale-free and small-world network structure seen in word association norms, this structure naturally emerges in a Topic model.

However, it is important to note that an observable behavior such as free association is the product of a cognitive process operating on a memorial representation (Anderson, 1978; Estes, 1975). This notion is ubiquitous in cognitive science. For example, Nosofsky (1986) uses a spatial representation of stimuli, but the complex classification behavior of his model is the result of applying a simple choice rule to this spatial representation, not spatial distance itself. Similarly, semantic space models are models of memory structure; the structural model should not be expected to simulate a complex behavior like memory retrieval without the benefit of a process account to explain how the memory structure is used in a particular task. While the cosine between two word vectors is often used as a measure of their semantic similarity, it is a measure of the similarity of memory structures rather than an appropriate process model of the task—*a* cosine *is not what people do in a task*, and should not be used as an estimate of behavioral data (see Jones, Hills, & Todd, 2015). A cosine, or similar metric, should be the input to a process model if one is interested in simulating behavioral data. This also enhances the models' generalizability across different tasks that tap semantic structure, and is particularly appealing given the low correlation in responses between different tasks thought to utilize the same semantic structure (Maki & Buchanan, 2008), and the fact that different semantic space models give the best fit to different behavioral tasks even though all tasks are thought to tap the same semantic memory structure (Mandera, Keuleers, & Brysbaert, 2017).

Griffiths et al. (2007, p. 224) imply that a "more complex" spatial metric based on LSA (similar to Nosofsky's 1986, 1991 use of a similarity-choice function) could potentially account for the metric axiom violations in association norms. We return to the issue of complexity with regard to spatial and probabilistic models in the discussion. The bulk of this paper will be focused on evaluating this suggestion by fusing spatial semantic models with a parameter-free version of Luce's (1959) similarity-choice model to evaluate their ability to account for the problematic data identified by Griffiths et al. In doing so, we provide an existence proof that semantic space models can indeed produce asymmetries, violations of the triangle inequality, and scale-free network structure with an appropriate process rule. It is premature to reject spatial models of semantic representation based on violations of metric axioms in association data.

2. A generic spatial choice model

In this paper, we evaluate the application of Luce's (1959) classic choice rule to simulate the cognitive process involved in the task of free association when applied to three (metric) semantic space models, gradually increasing in complexity. Although similarity and distance in the semantic spaces respect the metric axioms, the behavior of the choice rule applied to these spaces need not (cf. Nosofsky, 1991). The Luce choice rule was selected as our generic output model here due to its ubiquity in models of cognitive phenomena—it has been successfully applied to choice behavior ranging from low-level neural networks to high-level economic models of group choice behavior.

The Luce choice rule simulates how humans select from possible choice alternatives given a stimulus similarity space, governed by probabilities conditioned on the choice set. Hence, its input is metric space, but its output is a probability of a given response. Given a set of stimulus similarities (where similarity is defined as an inverse monotonic function of psychological distance) the Luce choice rule states that the probability of responding to stimulus S_i with response R_j is defined as:

$$p(R_j|S_i) = \frac{\beta_j \eta_{i,j}}{\sum_{k \epsilon M} \beta_k \eta_{i,k}} \quad (1)$$

where β_j is the response bias for item j, and $\eta_{i,j}$ is the similarity between stimuli i and j. Given the restrictions of metric spaces, the total probability over all responses sums to one. Most applications of the choice rule include exponential scaling of similarity based on Shepard's (1987) universal law of distance and perceived similarity. Hence, this general formula is often referred to as the Shepard-Luce choice axiom:

$$p(R_j|S_i) = \frac{\beta_j e^{-\lambda d(S_i,R_j)}}{\sum_{k \epsilon M} \beta_k e^{-\lambda d(S_i,R_k)}} \quad (2)$$

[1] Utsumi (2015) has revisited the Steyvers and Tenenbaum (2005) work and demonstrated that while scale-free and small-word structure is unobtainable by LSA, several other variants of the model, all spatial models, naturally produce the correct structure from association norms.

Please cite this article in press as: Jones, M. N., et al., In defense of spatial models of semantic representation, New Ideas in Psychology (2017), http://dx.doi.org/10.1016/j.newideapsych.2017.08.001

where d is a psychological distance function, and λ is a free parameter for the slope of the exponential (indicating a subject's sensitivity to stimulus differences).

Due to computational complexity that would be required to fit free parameters in the choice rule for our simulations, we evaluate a very simple parameter-free version of the choice rule here. Firstly we fix $\lambda = 1$, and ignore exponential scaling. Secondly, although it is reasonable to fix β to normative log word frequency for each word in the lexicon, we also ignore bias in our application here to make the similarities easily comparable to previous work. Hence, given a semantic similarity matrix for all words in the lexicon (for example, using LSA cosines) we simulate the probability of producing a target word in response to a cue word in free association as:

$$p(target|cue) = \frac{cos(cue,\ target)}{\sum_{i=1}^{\tau} cos(cue,\ word[i])} \quad (3)$$

where τ is a minimum similarity threshold parameter. Hence, this is a very simple version of the Luce choice rule, and performance should only be seen as a baseline—the model could obviously produce better predictions with parameter optimization.

3. Testing the semantic choice model

In this section, we test the ability of the simple Luce choice rule (free of parameters except for a minimum similarity threshold in the denominator) to account for violations of the metric axioms. Each of the metric spaces conform to the metric axioms, but the simple behavior of making a choice in this space does not.

3.1. Training corpus

We trained each semantic space model on the standard TASA corpus (Landauer & Dumais, 1997), and duplicated the modifications to the corpus made by Griffiths et al. (2007) for easy comparison to their results. The models were restricted to words that occurred with a frequency of at least 10, and were not contained on the standard LSA stoplist of function words. This reduced the model vocabularies to 26,240 words with ~4.2 million tokens in the modified TASA, consistent with the version used by Griffiths et al.

3.2. Semantic space models

We tested three semantic space models, ranging in assumptions and complexity: the Jaccard Index, LSA, and BEAGLE. The semantic spaces produced by each of the models conform to the three metric axioms (non-negativity, symmetry, and the triangle inequality).

Jaccard Index: The simplest mechanism we tested to create a semantic space was a basic and direct co-occurrence metric based on mutual information (intersection over union), as recent work has demonstrated superior performance on some semantic tasks using simple models based only on the surface form of language (e.g., Louwerse, 2008; Recchia & Jones, 2009), suggesting that "deep" models like LSA may be over-thinking the problem of human semantic learning. Here we use the classic Jaccard Index (Jaccard, 1901, 1912) from information retrieval, a variant of mutual information metrics:

$$J_{i,j} = \frac{|word_i \cap word_j|}{|word_i \cup word_j|} = \frac{|word_i \cap word_j|}{|word_i| + |word_j| - |word_i \cap word_j|} = \frac{f_{ij}}{f_i + f_j - f_{ij}} \quad (4)$$

where f_i and f_j are the marginal frequencies of words i and j, respectively, and f_{ij} is the co-occurrence frequency of i and j together in the same document. The Jaccard Index is essentially the intersection of the Venn, and can be applied efficiently to the same W x D matrix LSA learns from, but without the complexity of inferring latent semantic dimensions. The index is also referred to simply as "Proportion of Co-occurrence" in the psychological literature (Gruenenfelder, Recchia, Rubin, & Jones, 2015).

Latent Semantic Analysis (LSA): LSA spaces were constructed based on a word-by-document (W x D) co-occurrence matrix, in which each word is initially represented as a frequency distribution over documents. Entries were then weighted inversely proportionate to the word's frequency and entropy ($-\sum p \log p$) over documents. Singular value decomposition was applied to this matrix, and only the components with the largest eigenvalues were retained. The resulting word vectors are thought to represent the latent semantic dimensions which best explain the co-occurrence pattern of words over contexts.

We constructed LSA spaces using both 300 and 700 dimensions. Similar to Griffiths et al. (2007), we found little difference in performance on the association task as a function of dimensionality. Our 300-dimensional version matched the version on the LSA website. However, we use the 700-dimensional version here so our results are directly comparable to Griffiths et al.

BEAGLE: In addition to LSA and the Jaccard Index, we use a model intermediate to the two in complexity—the context learning mechanism from the BEAGLE model of Jones and Mewhort (2007), which is similar in spirit to other random accumulation models (Kanerva, 2009). BEAGLE begins by assigning initial random vectors to all words in the corpus, with elements sampled randomly from $N\left(0, \frac{1}{\sqrt{D}}\right)$, where D is an arbitrary vector dimensionality. As BEAGLE experiences sentences, the model updates the memory vectors for each word in the sentence as the sum of the random initial vectors representing each other word in the sentence. Across learning, semantically similar words naturally develop similar distributed vector patterns because they have had common random vectors summed into their memory representations. This has the effect that words which frequently co-occur develop similar vectors (a pattern reflected by the Jaccard Index), but also that words which occur in similar sentences develop similar vectors (a pattern learned by LSA), even if they never directly co-occurred. Note that the original BEAGLE model of Jones and Mewhort (2007) also uses holographic binding to learn grammatical information about word usage—however, here we just use the very simple random vector summation to learn semantic structure in TASA (the convolution-based holographic learning mechanism would introduce unnecessary complexity, as grammatical similarity is unlikely to play a large role in free association). Here, we use BEAGLE trained with 1024 dimensions trained on TASA using context learning only.

3.3. Simulating asymmetric associations

We do not bother with simulations based on the raw semantic spaces here, as they are obviously unable to simulate asymmetries in free association (i.e., cos [A,B] = cos [B,A]). However, a choice rule applied to these spaces to simulate the process of free association need not respect symmetry. The reason for this is very similar to Krumhansl's (1978) notion of similarity density. The density asymmetry in semantic space models has been previously identified and discussed elsewhere (Burgess & Lund, 2000; Hare, Jones, Thomson, Kelley, & McRae, 2009; Jones & Kintsch, 2006).

Although the distance between *baby* and *stork* is equal in either direction, the density of the landscape is not. If one computes and ranks the similarity of every word in the lexicon to *baby* and *stork*, *baby* is the 22nd most similar word to *stork*, but *stork* is the 9,279th most similar word to *baby* (cosines from BEAGLE). Hence, while the

Please cite this article in press as: Jones, M. N., et al., In defense of spatial models of semantic representation, New Ideas in Psychology (2017), http://dx.doi.org/10.1016/j.newideapsych.2017.08.001

numerator of the choice rule is the same for both *baby-stork* and *stork-baby*, the denominator changes dramatically depending on the ratio of similarity to other competitors. When a simple choice rule is applied to a metric space, *baby* comes to mind easily when cued with *stork*, but it is extremely unlikely to respond with *stork* when cued with *baby* due to strong competition from the many other words that come to mind more easily—the same pattern seen in human free association.

We reproduced Griffiths et al.'s (2007) method of selecting asymmetric pairs from Nelson's association norms. Two words were asymmetrically associated in the norms if 1) one word was produced as a response to the other with greater than zero probability, and 2) the forward-backward ratio of cue-response probability was greater than an order of magnitude. This procedure produced 38,740 asymmetric associations.

We then tested the ability of the choice rule to correctly predict the direction of the asymmetry in these pairs. Note that the raw semantic space models would produce baseline chance of ~50%. For each model we varied the threshold parameter τ in the denominator of the choice rule. This represents the τ most similar words to the cue considered as competitors to the target—τ was fixed across pairs within a given simulation (so all had the same threshold). Rather than using a threshold, the same effect could be obtained with an exponential similarity gradient (Shephard's Law), but it would be much more computationally expensive. Performance did not vary widely depending on τ regardless, so we present only performance with the best τ per model here (with τ hand fit).

Fig. 1 shows the percentage of asymmetric pairs for which the choice model predicted the correct direction, varying semantic space. For comparison, the horizontal line is chance performance without a choice model, and we have inserted Griffiths et al.'s (2007) Topics model performance for the same pairs, and raw frequency of the target word.

The first pattern to notice in Fig. 1 is that LSA did not perform much better with a choice rule than it could without. We found this puzzling, but consistent across a wide range of τ (and the model often did *worse* than chance). While this could be taken as evidence against spatial models in isolation, notice that both the Jaccard Index and BEAGLE improve considerably with the choice rule; both perform as well as word frequency and the Topic model. This is particularly intriguing given that the Jaccard Index is not a "deep" inductive model, but is more reflective of simple Rescorla-Wagner discrimination learning. When fused with an appropriate process model to simulate the task of free association, however, it easily predicts the correct pattern of asymmetry in the association norms.

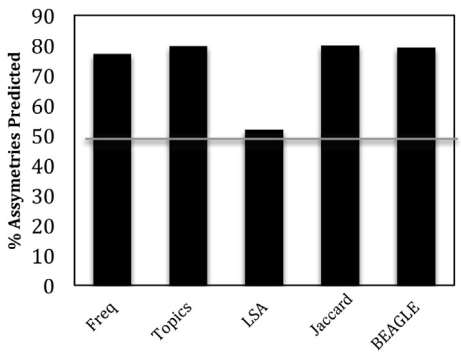

Fig. 1. Percentage of asymmetries in association norms predicted by each choice model (horizontal line is chance).

3.4. The triangle inequality

The triangle inequality is more difficult to test because there is disagreement about what constraints it places on a semantic similarity space, and how these constraints should manifest in a free association task. The triangle inequality comes from Euclidean geometry, in which the shortest path between two points is a line. Given this observation, the inequality states that the length of any side of a triangle must be less than the sum of the other two sides. Hence, when translated to proximities among three words in a metric semantic space, the distance between any pair of words in a triple may be no greater than the sum of the distances of the other two pairs, $d(x,z) \leq d(x,y) + d(y,z)$.

Tversky has demonstrated violations of the triangle inequality with similarity judgments of low-dimensional stimuli, in which humans weight feature matches more heavily than geometry suggests they should. However, it is difficult to determine what hard constraints the triangle inequality places on semantic similarity spaces. Griffiths et al. (2007) interpret the triangle inequality as implying that if x is similar to y and y is similar to z, then x must be similar to z. In word association, this leads to the expectation that if $P(w_2|w_1)$ is high and $P(w_3|w_2)$ is high, then $P(w_3|w_1)$ must be high. However, they note that this constraint is violated in free association norms—as $P(w_2|w_1)$ and $P(w_3|w_2)$ are selected to exceed an increasing threshold, $P(w_3|w_1)$ remains unaffected. To use their example, although *asteroid* is highly associated with *belt*, and *belt* is highly associated with *buckle*, *asteroid* and *buckle* have no association.

It is important to note, however, that the triangle inequality is difficult to explore (and impossible to test) with association data. The inequality does not state that if x and y are close points and y and z are close points, then x and z must also be close points; it simply states that x and z may be *no further* apart than the sum of the distances between x-y and y-z. Hence, the triple *asteroid-belt-buckle* in free association may conform to the triangle inequality (rather than being a violation). *Asteroid* and *buckle* need not be similar under the inequality, just not dissimilar.

It is difficult to determine from free association data whether the triangle inequality has been violated because association is a coarse indicator of similarity—a word is produced in response to a cue word or not. But the fact that a target is not produced in response to a cue is not evidence that they have no similarity, nor is it evidence of violating the triangle inequality. Griffiths et al. (2007) demonstrate that even as $P(w_2|w_1)$ and $P(w_3|w_2)$ increase in the norms, there are still many cases in which $P(w_3|w_1) = 0$. While they are careful to note that this only *suggests* a violation of the triangle inequality, we worry about the reliance on zero probabilities in this type of analysis. A zero probability association simply means that the event did not occur. It can be problematic to make inferences based largely on events that were unobserved. In addition, the practice assumes that all word pairs with zero probability (unobserved) have equal similarity, an assumption that is certain to be wrong.

We duplicated the thresholding analysis conducted by Griffiths et al. (2007). However, to avoid interpretation issues with unobserved data, we instead used only triples for which all three pairs exist in the association norms. Hence, all probabilities in our analysis are nonzero, and we can examine whether $P(w_3|w_1)$ is related to systematic increases in $P(w_2|w_1)$ and $P(w_3|w_2)$, relying on variance of observed events only. Our selection resulted in 80,212 triples. We systematically increased the threshold τ above which $P(w_2|w_1)$ and $P(w_3|w_2)$ were required to lie, and examined the distribution of $P(w_3|w_1)$ values. In the analysis by Griffiths et al. (2007) which included zero probabilities, they essentially found that $P(w_3|w_1)$ was uncorrelated with τ. However, in our data

Please cite this article in press as: Jones, M. N., et al., In defense of spatial models of semantic representation, New Ideas in Psychology (2017), http://dx.doi.org/10.1016/j.newideapsych.2017.08.001

(which excluded zero probabilities), we observed a significant correlation between the median of the $P(w_3|w_1)$ distribution and τ, $r = 0.42$. This indicates that the triangle inequality may indeed apply to association data when missing values (zero probabilities) are removed.

With the Luce choice rule applied to simulate the process of selecting a response in a free association task given a spatial semantic similarity space, metric models can produce violations of the triangle inequality, regardless of whether missing values are included. However, given that it is unclear whether humans violate this axiom in free association, it is important to note that metric models also can conform to the inequality. This is particularly important given that we are still uncertain as to whether or not human free associations actually contain evidence of a mental space that violates the inequality. In addition, it would seem from other types of semantic data that the triangle inequality is alive and well in the head. For example, mediated priming is a well-established semantic phenomenon that relies on triangulation: priming with *lion* facilitates recognition of *stripes* due to their mediated relationship through *tiger* (McNamara & Altarriba, 1988).

3.5. Semantic network structure

In addition to constraints from metric axioms, the neighborhood structure of semantic spaces (specifically LSA) has been shown to be inconsistent with what is suggested from word association. To create the mental connectivity structure necessary to produce association norms, LSA would need more words with extremely dense semantic neighborhoods than it appears to have. For example, Steyvers and Tenenbaum (2005) created network graphs based on free association norms and then investigated the ability of different growth models to produce this structure, as well as the network graphs of WordNet and various thesauri.

Steyvers and Tenenbaum (2005) created graphs based on association norms in which each word is a node and nodes are connected if they have nonzero probability of association. The resulting graphs are *scale-free*, a common property of connectivity in nature. If a word's degree k is defined as the number of other words connected to it, a scale-free network is one in which the distribution of degrees over all nodes follows a power law, $P(k) \sim k^{-\gamma}$ where γ is the constant rate of the power function. If both $P(k)$ and k are plotted on a log scale, the result is a straight line with a slope of $-\gamma$. More recently, Utsumi (2015) has reanalyzed the association norms and found that networks are better described by a truncated power law with initially exponential decay (cf. Heathcote, Brown, & Mewhort, 2000).

In addition, Steyvers and Tenenbaum (2005) found that association networks had much higher clustering of interconnected nodes than would be expected in a randomly constructed network. LSA was unable to reproduce this scale-free small-world structure for a variety of generation methods attempted by Steyvers and Tenenbaum: LSA produces degree distributions that fall off too slowly for small values of k and then too steeply as k increases, and LSA's clustering properties are both too high and are qualitatively distinct from association networks. This pattern for LSA was replicated by Utsumi (2015).

In contrast, Griffiths et al. (2007) found that networks created from the Topic model produced power law degree distributions and clustering properties that closely matched association networks. It is unclear, however, whether LSA's failure to reproduce the structure of the association network is common to all spatial models, or whether LSA would fail to produce the correct structure if it had the benefit of the Luce choice rule to simulate the process of free association.

We constructed semantic networks in the same way as Griffiths et al. (2007) both for LSA based on raw cosines as they did, but also for LSA, Jaccard Index, and BEAGLE with the addition of the Luce choice rule to simulate free association. Here, we discuss only undirected networks. Only normed words were used to create the networks. For each model, a threshold was set to determine whether to connect two nodes in the network (based either on cosine for raw LSA, or cue-target probability predictions from the Luce rule for the others). For each network, we fit the degree distribution to both a power and exponential function, and computed the clustering coefficient (Watts & Strogatz, 1998). The results are displayed in Table 1 (LC = Luce choice rule applied to a similarity space). For comparison, we have also added the network properties from the free association norms in the first row of Table 1.

Although the degree distribution for raw LSA was slightly better fit by a power function than an exponential, it shows little preference between the two, and the clustering properties of LSA are several orders of magnitude greater than the association network. The final column in Table 1 gives the ratio of the clustering coefficient in the model's network to the clustering coefficient expected in a random Erdos-Rényi graph constructed with the same density (Barabási & Albert, 1999). The CC/CCer ratio for raw LSA is much greater than that observed in the association network. As with the asymmetry simulation, the Luce choice rule integrated with LSA actually produces network structure more incompatible with the association network than did the raw LSA space, producing an exponentially distributed degree distribution. In contrast, JI-LC produces relatively weak clustering.

When fused with the Luce choice rule, BEAGLE produces network structure that is remarkably similar to the structure observed in the association network. The degree distributions show a strong preference for a power function over an exponential, and the slope of the power function for BEAGLE ($\gamma = 2.22$) is very close to that of the association network ($\gamma = 2.25$). For comparison, the slope of the power fit for LSA-LC was $\gamma = 3.96$. Fig. 2 plots the log-log degree distributions for the Luce choice version of LSA (left

Table 1
Network structure statistics for word association norms, raw LSA, and spatial + choice models (LC).

Network	Power R^2	Exp R^2	CC	CC/CCer
Association	0.877	0.571	0.187	42.59
LSA-Raw	0.882	0.872	0.449	85.41
LSA-LC	0.830	0.909	0.352	72.58
Jaccard-LC	0.952	0.939	0.092	18.81
BEAG-LC	0.882	0.550	0.290	59.03

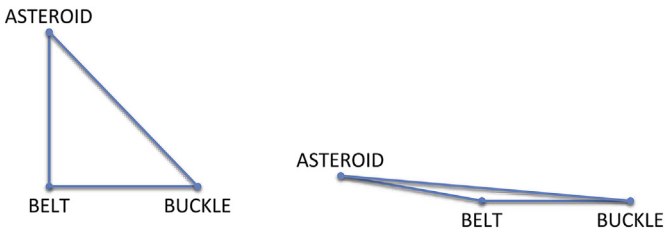

Fig. 2. The triangle inequality (Tversky, 1977) states simply that the distance between two points on a triangle in metric space can be no greater than the sum of the distances between the other two points, $d(x,z) \leq d(x,y) + d(y,z)$. Although a strong association may exist between two sets of pairs (e.g., *asteroid-belt*, and *belt-buckle*), this does not imply that the remaining pair (*asteroid-buckle*) should be expected to show an association in behavioral data. Such a pattern would be true only in the case of an equilateral triangle. The left panel shows the more likely pattern in human semantic memory, and the right panel shows an arrangement in a spatial model that would be closer to the extreme, but still consistent with the triangle inequality. In the extreme, the points would all lie along a single line, and $d(x,z) = d(x,y) + d(y,z)$.

Please cite this article in press as: Jones, M. N., et al., In defense of spatial models of semantic representation, New Ideas in Psychology (2017), http://dx.doi.org/10.1016/j.newideapsych.2017.08.001

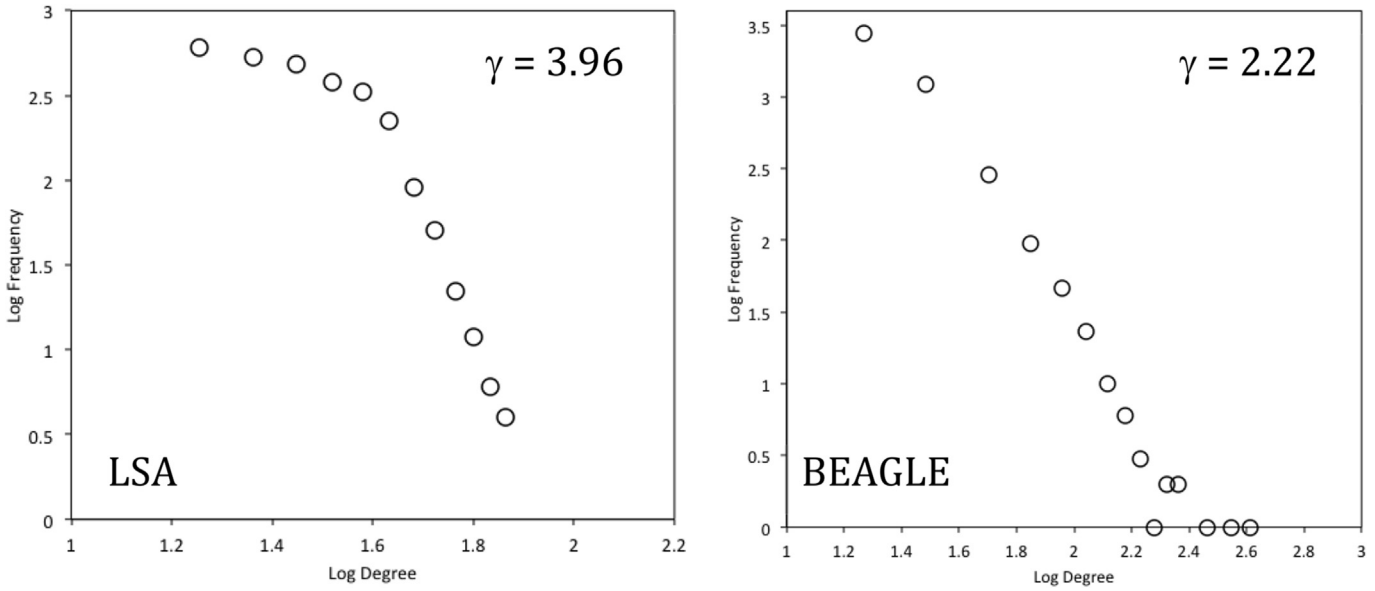

Fig. 3. Log-log degree distribution for Luce-LSA (left panel) and Luce-BEAGLE (right panel).

panel) and BEAGLE (right panel). Recall that the log-log degree distribution of the association network is linear with a slope of $\gamma = 2.25$. Hence, while network connectivity structure is a more difficult test for these models, BEAGLE demonstrates that it is certainly possible for a spatial model to produce the connectivity structure observed in association norms with the benefit of a process model to simulate the task of free association (see Fig. 3).

4. Discussion

The purpose of this paper is to provide an existence proof that spatial models can produce the structure observed in free association data provided that they have a plausible process model to simulate the association task. It is premature to reject spatial models of lexical semantic representation simply because the raw spaces must respect metric axioms but human behavior does not (e.g., Griffiths et al., 2007; See also; Ji, Lemaire, & Choo, H., Ploux, 2008). Human semantic memory may also respect metric axioms, but the behavior produced when a choice mechanism is applied to this memorial representation can produce violations of asymmetry, the triangle inequality, and can produce association networks that are small-world and scale-free (cf. Utsumi, 2015).

As an existence proof, these results should not be taken as evidence against any particular model. Even with the Luce choice rule, LSA had difficulties with network structure and the violations of metric axioms. However, this may be due to our assumptions when fixing parameters of the choice model. Fitting the sensitivity and bias parameters to the data may well have produced a model that performed very well when applied to LSA. Nonetheless, the performance of the simpler BEAGLE-LC and Jaccard-LC models make it clear that spatial representations of semantics are still viable models.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*(4), 249.
Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.
Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In *Cognitive dynamics: Conceptual and representational change in humans and machines* (Vol. 13, pp. 17–56).
Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology, 12*, 263–282.
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review, 114*, 211–244.
Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2015). Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science, 40*(6), 1460–1495.
Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition, 111*(2), 151–167.
Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*(2), 185–207.
Holman, E. W. (1979). Monotonic models for asymmetric proximities. *Journal of Mathematical Psychology, 20*, 1–15.
Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.
Jaccard, P. (1912). The distribution of the flora in the Alpine Zone. *New Phytologist, 11*, 37–50.
Ji, H., Lemaire, B., Choo, H., & Ploux, S. (2008). Testing the cognitive relevance of a geometric model on a word-association task: A comparison of humans, ACOM, and LSA. *Behavior Research Methods, 40*(4), 926–934.
Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, & Griffiths. *Psychological Review, 122*, 570–574.
Jones, M. N., & Kintsch, W. (2006). Asymmetric similarity in a self-organizing lexicon. In *Paper presented at the 47th meeting of the psychonomic society*.
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. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, & J. T. Townsend (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254).
Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representations with high-dimensional random vectors. *Cognitive Computation, 1*, 139–159.
Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*, 450–463.
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.
Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review, 15*(4), 838–844.
Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psyc Bulletin & Review, 15*, 598–603.
Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.
McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language, 27*(5), 545–559.

Please cite this article in press as: Jones, M. N., et al., In defense of spatial models of semantic representation, New Ideas in Psychology (2017), http://dx.doi.org/10.1016/j.newideapsych.2017.08.001

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *JEP: General, 115*, 39–57.
Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*, 94–140.
Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49*, 197–237.
Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*, 657–663.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.
Tversky, A., & Gati, I. (1982). Similarity, separability and the triangle inequality. *Psychological Review, 89*, 123–154. Utsumi, A. (2015). A complex Network approach to distributional semantic models. *PloS one*, *10*(8), e0136277. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small world networks. Nature, 393(6684), 440.
Utsumi, A. (2015). A complex Network approach to distributional semantic models. *PloS One, 10*(8), e0136277.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small world networks. *Nature, 393*(6684), 440.

Please cite this article in press as: Jones, M. N., et al., In defense of spatial models of semantic representation, New Ideas in Psychology (2017), http://dx.doi.org/10.1016/j.newideapsych.2017.08.001