Comparing models of semantic fluency: Do humans forage optimally, or walk randomly?

Johnathan E. Avery and Michael N. Jones
Cognitive Computing Laboratory
Indiana University, Bloomington
[averjo] [jonesmn]@indiana.edu

Abstract

Hills, Jones, and Todd (2012) observed that response patterns during the semantic fluency task (e.g., "name all the animals you can in a minute") display statistical signatures of memory search that mirror optimal foraging in physical space. They proposed a model of memory search based on exploration-exploitation tradeoffs known to produce optimal foraging patterns when animals search for food resources, applied to a spatial model of semantic memory. However, Abbott, Austerweil, and Griffiths (2015) demonstrated that optimal foraging behavior could also naturally emerge from a random walk applied to a network representation of semantic memory, without reliance on a foraging process. Since then, this has been a very active are of debate in the literature, but core confounds have prevented any clear conclusions between the random walk and cue switching model. We control confounds here by using a fixed training corpus and learning model to create both spatial and network representations, and evaluate the ability of the cue switching model and several variants of the random walk model to produce the behavioral characteristics seen in human data. Further, we use BIC to quantitatively compare the models' ability to fit the human data, an obvious comparison that has never before been undertaken. The results suggest a clear superiority of the Hills et al. cue switching model. The mechanism used to search memory in the fluency task is likely to have been exapted from mechanisms evolved for foraging in spatial environments.

Keywords: Semantic memory; memory search; model.

Introduction

Free recall from memory has been one of the most active areas of Cognitive Science since the field's inception, allowing the study of encoding, memory organization, and the processes with which humans search for and retrieve stored information. One of the most commonly used tasks to study retrieval from semantic memory is the semantic fluency task (SFT). In SFT, the participant is presented with a natural category label (e.g., "animals") and is required to produce as many exemplars of the category as possible within a fixed amount of time (e.g., "dog, cat, llama, …).

The task is commonly used in experimental psychology (Raajmakers & Shiffrin, 1981; Romney, Brewer, & Bachelder, 1993), but is also widely used in neuropsychological batteries, and is known to be sensitive to clinical group diagnoses (e.g., Alzheimer's and Parkinson's; Troyer et al., 1998). Given the ubiquity of the task and the fact that it taps general mechanisms of memory search, a formal model is greatly needed to understand the underlying nature of the mechanism that drives SFT not only for our basic understanding of cognition, but for transfer of models to applied domains that attempt to use SFT data for detection of early stage dementias (Johns et al., 2017).

Responses in SFT typically occur in temporal bursts of related items, with time lags between clusters thought to involve the search time required to navigate to a new cluster (e.g., {farm animals} → {fish}). Hills, Jones, and Todd (2012) made the observation that the temporal pattern of items produced in SFT exhibited statistical signatures that are characteristic of animals foraging for food in physical space (*optimal foraging theory*: Charnov, 1976), suggesting that our memory search mechanisms may have been exapted from primitive mechanisms that evolved to search for food resources in the physical environment.

The argument in favor of optimal foraging in SFT relies on three components. First, there is a decrease in semantic similarity when transitioning between patches. A change in similarity denotes switching between local and global cues. Second, transitions between patches occur when a patch has been sufficiently depleted. Third, cluster transitions can be predicted by the marginal value theorem (MVT; Charnov, 1976). The MVT predicts that patch switching in spatial foraging occurs when the time spent searching locally exceeds the average transition time between foraging events across the environment (the marginal value).

Hills et al. tested a variety of search models applied to a semantic space simulated by a corpus-based distributional model of semantic representation, BEAGLE (Jones & Mewhort, 2007). The specific search model that best explained the human data was a two-stage model that used local similarity to generate items until no other proximal item was found, and then switched to a global frequency cue to select the next item (and search by local similarity resumed). The fact that the global-local switch model was the best explanation of the human data was theoretically significant for two reasons: 1) it produces patterns of optimal foraging, and 2) the process it uses mirrors the best accounts of how animals make exploration-exploitation decisions in when foraging for food. Just as a honey bee must decide when to give up on a local patch of flowers and accept the costs that accompany the search for a new unknown patch, humans show the same pattern in memory search when deciding when to give up on the farm animals and search for a new resource-rich patch to exploit.

However, there is disagreement whether behavioral patterns of optimal foraging truly require a cognitive mechanism based on spatial search for food resources.

The Debate

While the Hills et al. (2012) account of SFT as exapation of evolved mechanisms makes for a good story, we strive for parsimony in science and need to carefully consider issues of model mimicry. Abbott, Austerweil, and Griffiths (2015) demonstrated that the behavioral patterns in the Hills et al. data that appear to indicate an optimal foraging mechanism could also be produced by a simple random walk mechanism applied to a network representation of semantic memory. This observation began a lengthy debate over the interaction between memory representations and the retrieval mechanisms that operate on them.

In a rebuttal, Jones, Hills and Todd (2015) noted that the semantic network used by Abbott et al. (2015) was constructed by human free association norms, a task remarkably similar to the SFT data they were fitting to. The concern was that a simple retrieval process seemed to suffice in the Abbott et al. work because much of the requisite complexity to simulate the data was hidden in the process used to construct the network. Jones et al. referred to this as the *representational Turk* problem in cognitive modeling: A simple account of the process required to explain human behavior could be achieved if human behavior was used to construct the representation, but then we have sacrificed explanation as behavior is used to predict behavior. As a *reductio ad absurdum*, they proposed a perfectly fitting zero-parameter model of SFT could be achieved if we simply made the assumption that SFT data were used to construct the memory space. In contrast, the original Hills et al. (2012) representational space was built by a theoretical model of human semantic learning applied to statistical redundancies in our linguistic data. Jones et al. suggested that to fairly evaluate search models they must be applied to the same memory representation.

In response, Nematzadeh, Miscevic, and Stevenson (2016) attempted to address the representational Turk problem by using the WordLearner algorithm (Faizly, Alishahi, & Stevenson, 2010) to construct a semantic network from a corpus of child-directed speech, the CHILDES corpus (MacWhinney, 2000). They replicated the basic phenomena suggestive of optimal foraging when a random walk algorithm was applied to this constructed network. Importantly, the network was constructed with a learning model applied to natural language rather than human behavioral data, which sidesteps the representational Turk problem. While this is a step in the right direction, it is still impossible to compare the pure processes of random walking to optimal foraging because a different corpus, register, vocabulary, and learning model were used to create the network from those used to create the spatial representation used in the original Hills et al. (2012) work. A random walk must be applied to a network, and the cue switching model may be applied to a spatial representation or a network. But it is important to carefully equate the construction of the space and the network if conclusions about the retrieval mechanisms are to be made, and there remain several confounds in the literature that preclude comparison of the core retrieval mechanisms.

Table 1. There is a lack of consistency between the studies examining models of semantic retrieval. Discrepancies between studies prevent a direct comparison between results and offer no solid ground on which to draw conclusions.

	HJT	**AAG**	**NMS**
Learning Environment	Wikipedia	USF FA	CHILDES
Learning Model	BEAGLE	Human	Word Learner
Representation	Space	Network	Network
Search Model	Foraging	RW	RW

Table 1 summarizes the dimensions along which the studies differ in their attempts to explain behavior on the SFT. It is evident that at no level can a direct comparison be made between the cue switching and the random walk models in any of the studies. In order for such a theoretical comparison to be made (the final row), the factors in the above three rows must be held constant. As noted previously, the cue switching model requires a spatial representation and the random walk requires a network representation, but for a fair comparison these representations must be constructed with the same learning model on the same learning environment.

The goal of this paper is threefold. Firstly, we construct a semantic matrix using the learning environment (Wikipedia) and learning model (BEAGLE) that accompanies the original Hills et al. (2012) data. We use this spatial matrix for the cue switching model, and apply a thresholding algorithm to the same matrix to build an equivalent network representation for the random walk model. Secondly, we perform both qualitative and quantitative model comparisons (Busemeyer & Diederich, 2010). In our *qualitative* model comparison, the models are evaluated on their ability to correctly produce the qualitative phenomena that have been used as evidence of optimal foraging; note that only qualitative comparisons have been used in this debate so far—the novel contribution here being a clear deconfounding of learning model and environment to determine if both the random walk and cue switching model can produce the characteristic phenomena. In addition, we also conduct *quantitative* model comparisons: Using maximum likelihood and BIC, we evaluate which search model most closely approximates the human data from the original Hills et al. study. If all models can duplicate the signature patterns of optimal foraging in the human data, the next step is to evaluate which model gives the most accurate account of the production patterns. Thirdly, we also implement and evaluate newer versions of the random walk model suggested by Zemla and Austerweil (2017).

Method

Equating Spatial and Network Representations

It is an error to think of spatial and network representations as fundamentally different. If both are defined by similarity data, a space and a network are both fundamentally matrices and are isomorphic. A fully connected network with edge weights determined by BEAGLE similarity is the exact same matrix as a BEAGLE similarity space—describing it as a space or network simply adds confusion because they are the same matrix. If a thresholding rule is applied to the spatial similarity matrix, it becomes equivalent to a partially connected network with the same threshold. If the threshold rule applied to the spatial similarity matrix is binarized (nodes are connected or not), then this becomes isomorphic to an unweighted network. If the same learning model and learning environment are used to create the similarity matrix, then it may be described as either a network or space. For simplicity here, we will talk of a network, and will compare both the foraging and random walk processes when applied to the same network representation.

Steyvers and Tenenbaum (2005) generated a semantic network by applying a threshold epsilon to a cosine similarity matrix of terms. Using this technique, we generated 101 networks from the BEAGLE space matrix used in the original Hills et al. (2012) study, using equally spaced epsilon values ranging from $0 \leq \varepsilon \leq 1$. Similarity-derived edge weights were updated at the varying levels of ε as:

$$W_{ij} = \begin{cases} 0 & Wij < \varepsilon \\ W_{ij} & Wij \geq \varepsilon \end{cases} \quad (1)$$

A network generated from an epsilon = 0 is a complete graph, while a network generated from epsilon = 1 is a graph without edges between nodes. Nodes within the network represent the full set of animal words produced by participants in the original Hills et al. (2012) study.

An additional node for the cue 'animal' was added to the network. Both foraging and random walk models require a global cue. Hills et al. (2012) used word frequency as the secondary cue within their cue-switching model, whereas Abbott et al. (2015) used free association norms to connect the node 'animal' to other nodes. Word frequency was used to establish the edge weight between the node 'animal', where

$$w('animal', X) = \frac{f(target)}{\sum_i f(target\_i)}$$

Frequency counts were taken from the Wikipedia corpus used in the original Hills et al. study. This yields weights such that words with higher frequency have stronger weights connecting them to the node 'animal'. The threshold ε was not applied to edge weights derived from word frequency. Removing this step for these edge weights ensures that the network is connected at every value of ε. There are no semantically isolated nodes within the network at a threshold $\varepsilon < 0.38$. When $\varepsilon \geq 0.38$, there are nodes that can only be reached via edges derived from word frequency.

To summarize, the same cosine similarity matrix was used as the base representation for both the foraging and random walk model. The similarity matrix learned from the BEAGLE model trained on the Wikipedia corpus from the original Hills et al. (2012) paper was used to create a network representation for both the cue switching model and the random walk model, and performance across levels of the threshold parameter was evaluated. Hence, the spatial and network representations are based on the same learning environment and the same learning model, holding constant the previous confounds in the literature that were the first two rows of Table 1.

The Cue Switching Model

Foraging behavior entails a strategic tradeoff between local exploitation and global exploration. Optimal foraging in semantic memory instantiates this tradeoff as switching between local and global cues. The cues are independent and attended to separately. Within this model, these are operationalized as a similarity cue and a frequency cue, respectively. The cue switching model is a dynamic model where the cues are attended to differently between the different processes. When searching locally in memory, both the local similarity cue and the global frequency cue are being taken into account, resulting in a strategic behavior where the yield of a local search is weighed against the yield of another semantic patch. When making cluster switches, the global frequency cue alone is taken into account, resulting in a strategic behavior that seeks an underexploited semantic cluster.

We implemented the cue switching model of Hills et al. (2012), which incorporates multiple cues dynamically within a Luce choice rule (cf. architectures from SAM and ACT-R):

$$P(X_{n+1}|Q_1, Q_2, X_n) = \frac{W(X_{n+1}, X_n)^{\beta_l} \times W(X_{n+1}, "animal")^{\beta_g}}{\sum_{k=1}^{N} W(X_k, X_n)^{\beta_l} \times W(X_k, "animal")^{\beta_g}} \quad (2)$$

where W(A,B) is the weight of the edge (based on BEAGLE cosines) between the nodes A and B. The cue switching model of Hills et al. (2012) is a two-parameter model, where β is a free parameter to simulate the saliency of a given cue. Given the word produced in position *n*, equation 2 defines the probability of next producing word *n+1* as a function of local or global cues. To maintain consistency with the random walk, our global cue here is the word "animal" in space.

The Random Walk Model

In contrast to the cue switching model, the random walk model does not strategically switch between cues. Presupposing a network structure of representation, the activation of nodes occurs based on two cues as well. The original model of Abbott et al. (2015) performs a local traversal of the network by randomly visiting nodes based on the edge weights of directed connections. The model also contains a parameter that can perform random jumps back to the global node "animal" and continues to traverse the network randomly. Hence, the random walk can also predict the probability for a sequence of items produced in SFT:

$$P(X_{n+1}|Q_1, Q_2, X_n) =$$

$$\rho P(X_{n+1}|Q_2) + (1-\rho)P(X_{n+1}|Q_1(X_{n+1}, X_n)) \qquad (3)$$

where ρ is a free parameter that captures the degree to which retrieval is prone to jumping to another part of the network. The likelihood of moving to another node is governed by the edge weight between that node and the node 'animal'. This is marked by the strength of the connection weight, given by $W(a, b)$ in

$$P(a|Q) = \frac{W(a,Q)}{\sum_{k=1}^{N} W(a_k,"animal")} \qquad (4)$$

Qualitative Model Comparisons

We first evaluate the ability of the original random walk model from Abbott et al. (2015) to produce the core qualitative phenomena indicative of optimal foraging that were seen in the original Hills et al. (2012) study. Namely, switching between semantic patches must be predicted by the marginal value theorem. We attempted to recreate these results by implementing the different search mechanisms on complete semantic network. Figure 1 illustrates that for both models, patch switches occur at the point predicted by the marginal value theorem for both the random walk and cue switching models. Though the effect is small for both models, the key element is that the qualitative pattern of the marginal value theorem holds for both models.

Both models are able to reproduce behavior that is consistent with the MVT. This demonstrates how particular statistical signatures may be achieved via very different avenues. Two possibilities are left – either both accounts of behavior in recall are given participation points and the matter is dropped, or we dig deeper to determine which model provides a better account of human behavior.

Quantitative Model Comparisons

Under qualitative analysis, neither model produced results that could demonstrate its preeminence as the best account of human behavior. In order to determine which model is a better account, we found the best fitting parameters for each participant and calculated the negative log likelihood of those parameters.

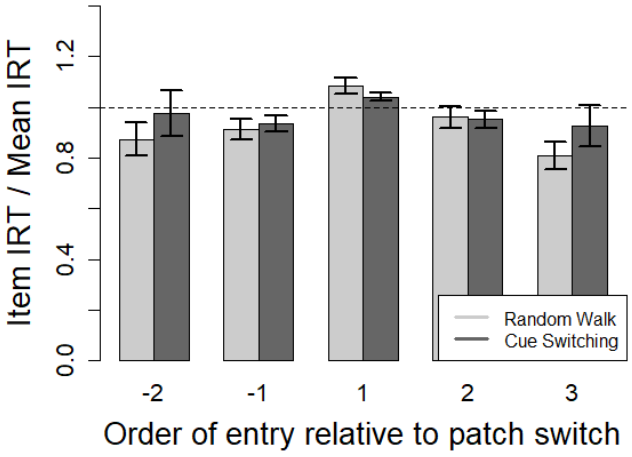

Figure 1. The item IRT relative to position in patch, where '1' corresponds to the first item in a patch, and '-1' corresponds to the last item in the previous patch. As predicted by the marginal value theorem, a patch switch is predicted to occur when the IRT exceeds the mean IRT.

For each participant's best fitting parameters, we derived the Bayesian Information Criterion (BIC) to evaluate the model's performance according to:

$$BIC = 2NLL + p \times log(N)$$

where p is the number of parameters for the given model. The BIC penalizes models for their complexity, where models with more parameters receive a higher BIC. Models with lower BIC are considered to better depict the data, since they produce a better balance of parsimony and accuracy. Performance on a particular network of threshold epsilon was aggregated by taking the median BIC for a participant for a retrieval model. BIC was calculated for both models as well as for how well a base model that included only word frequency fit the data**.**

Weighted Networks

Figure 1 depicts the median BIC for the three models at the varying thresholds. Optimal performance occurs at different levels of epsilon between the two models. The cue switching model performs best on a complete network, whereas the random walk model performs best on a connected but not complete network. More specifically, the random walk model performs best at a threshold where certain nodes become isolated from one another via connections derived semantically. Recall that, in the absence of edge weights derived from frequency, nodes become isolated semantically at $\varepsilon = 0.38$. In the instance of a semantically isolated node, the model may still reach the node via the frequency-derived connections. The random walk model demonstrates a marked

– though brief - increase in performance at $\varepsilon > 0.38$. This performance confirms suspicions raised by Abbott et al. (2015) when considering the effectiveness of a random walk model.

When the network is at $\varepsilon = 1$, the random walk model is relying solely on frequency information. This results in a larger BIC because of the penalty for additional parameters in the model. Notably, there is no crossover interaction between model fit at any epsilon threshold. This indicates the primacy of the cue switching model in accounting for the data.

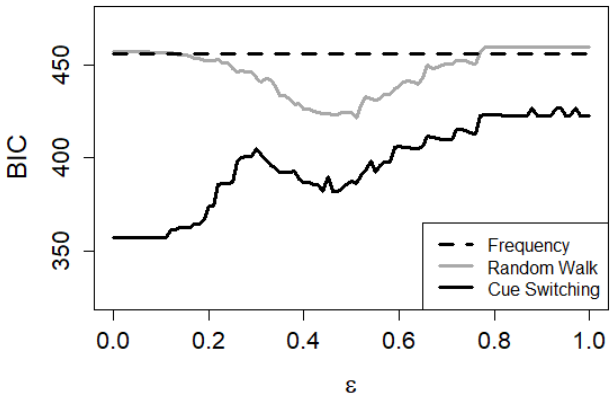

Figure 2. Performance of the models of retrieval on weighted semantic networks at increasing thresholds ε.

Unweighted Networks

Nematzedah et al. (2017) suggested there may be a difference in model performance given an unweighted network. In order to test this, all weighted networks were converted to unweighted networks, such that:

$$W_{ij} = \begin{cases} 0 & Wij = 0 \\ 1 & Wij > 0 \end{cases} \quad (5)$$

Both edge weights derived from semantic similarity as well as frequency were updated. In this way, some semantic information is preserved in the form of an edge, but nuanced information about similarity and frequency in the edge weights was eliminated.

On both the weighted and unweighted networks, the random walk model performs optimally at some intermediary value of ε. Notably, the cue switching model performed best on a complete weighted network, whereas on an unweighted network it performs best on some intermediary threshold of ε.

Both the random walk and cue switching models perform worse on an unweighted network. On the weighted network, the best BIC for the random walk model is at $\varepsilon = 0.51$ with a BIC = 421.16, and $\varepsilon = 0.1$ with a BIC = 357.26 for the cue switching model. On the unweighted network, the best BIC for the random walk model is at $\varepsilon = 0.51$ with a BIC = 434.70, and $\varepsilon = 0.44$ with a BIC = 432.37 for the cue switching model.

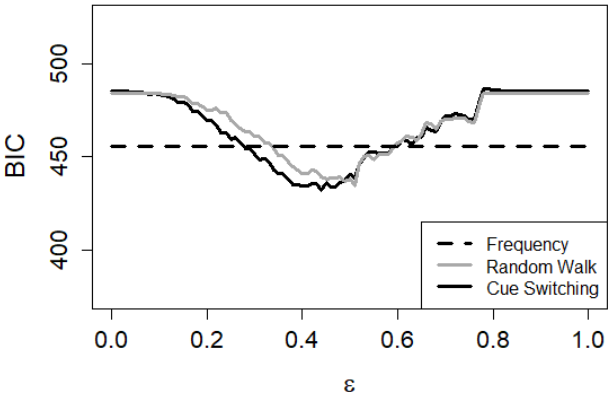

Figure 3. Performance of the models of retrieval on unweighted semantic networks at increasing thresholds ε.

The Extended Random Walk Models of Zemla and Austerweil (2017)

Insofar as the mathematical formulations for the foraging and random walk models reflect the theories of behavior at large, it is evident that the cue switching model more accurately accounts for the behavior of memory retrieval on the semantic fluency task.

In order to test whether the difference in performance is merely due to the mathematical formulation of a random walk model, we investigated different instantiations of random walk models. The alternative random walk models were suggested by Zemla & Austerweil (2017). These models ranged in complexity from zero to two free parameters. They were intended to provide a new take on the debate between random walk models and cue switching models, tending toward being implementations of the random walk models.

Node degree search. This is a zero free parameter model. The probability of subsequent nodes is based on number of connections between nodes. When at a given node, the most probable subsequent node is the node with the most connections. For a node degree search, edge weights are irrelevant.

Cluster depth first search. This is a zero free parameter model. A cluster depth first search searches all connected nodes to a given node before moving to the next node to search through its connected nodes.

Random Walk with Random Jumps. This is a two free parameter model. This model is a modification of the random walk model examined previously. The second parameter is the degree to which the random walk will move to a random node within the network.

Figure 4 depicts the BIC performance of the best fitting parameters for the additional models. Additionally, Figure 4 includes the performance of the random walk model previously displayed in Figure 2 as a baseline of comparison. The cluster depth first search was not included in Figure 4 because its minimum BIC = 611.55 at $\varepsilon = 0.64$. Notably,

none of the other suggested random walk models performed better than the original suggested by AAG. The zero free parameter models performed very poorly. Where other models increase in performance at intermediary thresholds ε, the node degree search and cluster depth first search decrease in performance.

The best fitting model of the alternative implementations is the random walk model with a random jumping parameter. However, adding a parameter that allows the model to randomly jump to a node does not account for the behavior better than assuming that a participant is attending to two cues separately.

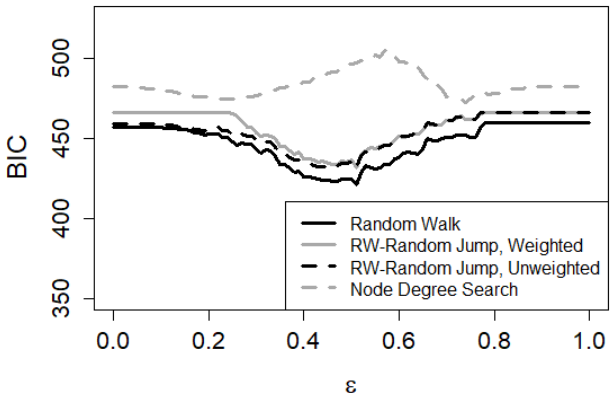

Figure 4. The performance of alternative implementations of a random walk model of retrieval across at increasing thresholds ε.

Discussion

This study offers a comparison between different accounts of human performance on semantic fluency task. A cue switching model describes recall as a tradeoff between local and global cues. A random walk model posits that recall is a random process that occurs on some well-structured semantic representation.

A random walk model offers the simplest explanation of behavior. The model requires fewer free parameters that operate within a less complex mechanism. Conversely, the cue switching model argues that lower level mechanisms involved in spatial foraging were exapted into higher level cognitive processes over the course of evolutionary history. Where the random walk model offers a simple mechanism on its own, the cue switching model argues for fewer behavioral mechanisms overall.

Both models create statistical signatures indicative of foraging behavior. Therefore, the models can't be evaluated on this qualitative aspect alone. We evaluated each model by the amount of error it produced by maximizing fit to the individual set of items produced by participants in the Hills et al. (2012) experiment. Although both models were able to generate the qualitative pattern of behavior that is consistent with MVT, the original cue switching model of Hills et al. (2012) greatly outperforms the random walk model according to this BIC. Other instantiations of a random walk were also evaluated, all of which performed more poorly than the original random walk model. In this respect, the current study supports Hills et al.'s original cue switching model over the random walk model as a better explanation of human behavior in SFT.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological review*, *122*(3), 558.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30*, 149–165.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

Charnov, E. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9*(2), 129–136.

Hills, T. T., Jones, M. N., & Todd, P. T. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*, 431-440.

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., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, & Griffiths. *Psychological Review, 122*, 570-574.

Nematzadeh, A., Miscevic, F., & Stevenson, S. (2016). Simple search algorithms on semantic networks learned from language use. *arXiv preprint arXiv:1602.03265*.

Raaijmakers, J., & Shiffrin, R. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*(1), 28-34.

Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society*, *4*(2), 137-143.

Zemla, J. C., & Austerweil, J. L.(2017). Modeling semantic fluency data as search on a semantic network. *Proceedings of the Cognitive Science Society*.