Predicting and Explaining Human Semantic Search in a Cognitive Model

Filip Miscevic
Cognitive Science Program,
Complex Networks & Systems
Indiana University Bloomington
fmiscevi@iu.edu

Aida Nematzadeh
Department of Psychology
University of California
Berkeley
nematzadeh@berkeley.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

Recent work has attempted to characterize the structure of semantic memory and the search algorithms which, together, best approximate human patterns of search revealed in a semantic fluency task. There are a number of models that seek to capture semantic search processes over networks, but they vary in the cognitive plausibility of their implementation. Existing work has also neglected to consider the constraints that the incremental process of language acquisition must place on the structure of semantic memory. Here we present a model that incrementally updates a semantic network, with limited computational steps, and replicates many patterns found in human semantic fluency using a simple random walk. We also perform thorough analyses showing that a combination of both structural and semantic features are correlated with human performance patterns.

1 Human Semantic Processing

The study of human semantic memory—word meanings, their relations, and their storage—is challenging due to the complexity of factors involved. Finding (1) the right representation for word meanings and their relations, (2) the mechanism responsible for learning the representation, (3) the appropriate search algorithm to efficiently retrieve information from semantic memory, and (4) the suitable empirical data to evaluate the proposed representations and algorithms is a difficult task. Previous research has extensively explored each of these areas (*e.g.*, Collins and Loftus, 1975; Steyvers and Tenenbaum, 2005; Griffiths et al., 2007).

Psychologists frequently use a task known as semantic fluency (or verbal fluency) to examine human semantic representation and processing (Troyer et al., 1997; Ardila et al., 2006). Participants are asked to produce as many words as they can from a given category (*e.g.*, animal) in a fixed amount of time (*e.g.*, three minutes). The resulting data—which words people recall and in what order—can shed light on how people represent word meanings and their relationships, and how they search such semantic information. For example, Hills et al. (2012) found that participants tend to reply in semantically-related bursts of words—*e.g.*, they recall words from the pet subcategory of animals (*dog*, *cat*) then switch to a different subcategory, such as African animals (*lion*, *zebra*), etc.—indicating that people tend to follow a strategy of *exploiting* a semantically-related patch of words, then *exploring* to find a new patch, much like animals foraging for patches of food in their environment.

Recent work has investigated the properties of semantic representations and processing algorithms that can account for this type of behavior in the semantic fluency task. Different researchers have found that a match to human behavior can be achieved in either of two ways: (a) using a simple (vector-based) semantic representation in combination with an informed, two-stage algorithm to exploit and explore the space (Hills et al., 2012); or (b) creating a richer representation—structured as a semantic network—and using a simple random walk to access it (Abbott et al., 2015; Nematzadeh et al., 2016). These findings suggest that the choice of representation and search algorithm are interdependent, such that the same empirical data can be replicated through different combinations of representation and algorithm that make different trade-offs on the locus of complexity (Abbott et al., 2015).

However, if both combinations account for the human data considered thus far, the question of which model more plausibly captures what occurs in a search in human semantic memory remains open. As Abbott et al. (2015) suggest, further experiments, such as those performed by Hills et al. (2015), can help elucidate the differences between these approaches to modelling human semantic memory. In particular, if there are key aspects of human semantic search that can be explained by one model and not the other, then this goes towards disconfirming the latter. One of the goals of the current paper is to show that a random walk over a semantic network reproduces even the additional empirical patterns of human semantic fluency task examined by Hills et al. (2015).

In addition to these experimental approaches, other findings and theoretical considerations may come to bear on resolving the question of which model most

aptly reflects human semantic search.

For example, people appear to have a structured semantic memory that encodes many kinds of relational knowledge (Miller and Fellbaum, 1991). In this way, complexity costs are incurred during learning (while creating the structured representation) rather than every time the representations are accessed. As such, accessing the knowledge later becomes a more efficient process. Hence, it may be reasonable to suggest that a simple search algorithm operating over a structured semantic network is a preferable model.

Another open issue is precisely what kind of semantic representations realistically capture word relations, especially semantic similarity, which typically form the basic structure of a semantic network (*e.g.*, Miller and Fellbaum, 1991). Work modeling human semantic fluency behavior using a simple random walk over a semantic network has drawn on several different kinds of semantic word representations. Abbott et al. (2015) constructed their semantic network using human association norms (Nelson et al., 1998), so that weighted edges between words directly capture the similarities between them that are relevant to the fluency task (Jones et al., 2015). Nematzadeh et al. (2016) built two networks based on different semantic representations learned from text corpora: a simple vector-based representation model, called BEAGLE, learned from Wikipedia (Jones and Mewhort, 2007, previously used by Hills et al. (2012)), and probability distributions learned from child-directed corpora (Fazly et al., 2010). Given that a random walk over semantic networks from each of these sources—human association norms, vector-space representations, and probability distributions—all model human fluency behavior, how do we choose between them?

An important set of considerations that we explore here involves the cognitive plausibility of how a semantic representation could be learned. While the human association norms used by Abbott et al. (2015) accurately reflect human judgments of word relatedness, it is not clear how the similarity assessments captured in such norms can be learned through exposure to language. The BEAGLE vector-space representations, on the other hand, are learned from instances of natural language. However, acquisition is a batch process over Wikipedia data, which is arguably not a good proxy for the linguistic input from which individuals acquire their semantic lexicon. The probability distributions used by Fazly et al. (2010), however, are learned by a cognitive model from a corpus of child-directed speech. These representations thus meet important criteria for cognitive plausibility, in that they are learned from naturalistic linguistic input.

One final crucial issue that has remained unaddressed to date is the incremental learnability of the semantic network structure itself. Children simultaneously learn word meanings as well as the relations between them (Jones et al., 1991). Thus, it is important to model the simultaneous incremental learning of both semantic word representations and their structure in a semantic network. This has been neglected by previous work discussed so far. Even in the work where semantic representations are learned, only the word representations and not their relations are learned. Instead, the semantic network is created by exhaustively comparing all the word representations after training—a process that is too computationally demanding to be cognitively plausible.

Our contributions in this paper are threefold: First, we show that a semantic network created incrementally within an online word learning model—from naturalistic child language acquisition data—can yield human performance in semantic search using a simple random walk. Our work here confirms that a semantic network created and updated incrementally—while the model is learning words—has the appropriate structure to yield patterns observed in the semantic fluency task, despite having noisy and incomplete connections as a result of being generated from partial knowledge acquired at each time step. Second, as mentioned, we show that the new approach to creating the semantic network produces a structure that also mimics other aspects of human behavior in semantic fluency, going beyond earlier models in the scope of empirical data accounted for (Abbott et al., 2015; Nematzadeh et al., 2016).

Finally, we extend previous analyses of semantic organization to determine more precisely which network properties are correlated with the observed human performance patterns. While other work has focused on the importance of structural properties of the network in determining human behavior (Goñi et al., 2010; Steyvers and Tenenbaum, 2005), we find that both structural *and* semantic properties are necessary to generate patterns observed in human semantic fluency data.

2 Incremental Network Creation

We use the approach of Nematzadeh et al. (2014) to incrementally build a semantic network, which draws on the probabilistic cross-situational word learning model developed by Fazly et al. (2010).

2.1 Incremental Word Learning Model

The semantic network is generated from word meanings (representations) learned by the model of Fazly et al. (2010), trained on the Manchester corpus (Theakston et al., 2001) of the CHILDES database (MacWhinney, 2000). Each input to the model consists of an *utterance* from the corpus, labelled with a *scene* consisting of semantic features for each word. For example, consider the following utterance (U) and selected features from its accompanying scene (S):

U: {*look*, *at*, *the*, *monkey*, *eat*, *a*, *banana*}
S: { . . . , VERTEBRATE, MAMMAL, . . . , FRUIT, . . . }

Just as a child must learn the referent of each word

in a sentence, the learner must infer which features in the scene are associated—or *aligned*—with each word. The model captures this association as the probability of a feature f given a word w, $P(f|w)$, which it incrementally updates from the co-occurrence of f with w across all observed utterance–scene pairs. The meaning of each word w is then represented as the probability distribution $P(\cdot|w)$ over all semantic features, which is estimated through latent variables that model the possible alignments of words and features in an utterance–scene pair. An incremental Expectation Maximization algorithm is used to update $P(\cdot|w)$ (Neal and Hinton, 1998). Hence, as in children, word meanings are gradually learned after many exposures to utterances and scenes.

In particular, for a single utterance–scene pair processed at time t, the alignment (a) probability of each feature (f_i) in the scene and word in the utterance (w_j) is calculated by:

$$P_t(a_{ij}|f_i) = \frac{P_{t-1}(f_i|w_j)}{\sum_{w' \in u} P_{t-1}(f_i|w')}$$

$P_{t=0}(f_i|w_j)$ is initially randomly uniformly distributed. Once the alignment probabilities are calculated, the word meanings are updated:

$$P_t(f_i|w_j) = \frac{\sum_{u \in U_t} P_t(a_{ij}|u, f_i)}{\sum_{f' \in M_t} \sum_{u \in U_t} P_t(a_{ij}|u, f')}$$

Here, U_t represents the set of utterances processed up to and including time t, and M_t is the set of features observed up to and including time t. Note that the summations do not have to be calculated anew each time; the terms from the first $t-1$ utterances can be stored and updated with the contributions from the tth utterance–scene pair.

The learned representation for a word, $P(\cdot|w)$, can be treated as a vector representation of the word over all semantic features. In the present study, we focus on animal nouns, as they are the target of the semantic fluency task in humans. The semantic features of noun meanings used are derived from WordNet hypernyms (Fellbaum, 1998, `http://wordnet.princeton.edu`), and embed hierarchical conceptual knowledge of nouns.

The more features (hypernyms, in this case) two animal words (*e.g.*, "CAT","DOG" vs. "CAT","FROG") have in common, the more similar their learned representations will be. The model learns not only the features associated with that particular word, however, but also features that often occur in the same context as the word. For example, in the above utterance–scene pair, the model may come to associate a non-zero probability with the feature FRUIT and the word *monkey*. Hence, the learned meanings of words capture not only a conceptual hierarchy for that word but also information learned from the context of their usage.

2.2 Incremental Learning of Semantic Networks

Children do not just learn the meanings of words, they also learn the relations between them at the same time (Jones et al., 1991). We use the approach taken by Nematzadeh et al. (2014) to enable the model to learn word meanings and the relationships between them simultaneously, without exhaustively considering all possible relationships between the words.

Since the probability distribution $P(\cdot|w)$ for a given word w is stored as a vector over all semantic features, the cosine of the angle between them can be computed as a measure of their similarity. A semantic network can thus be constructed by representing each word as a node in the network, with an edge between them if the cosine similarity between two words is greater than a threshold ρ.

Whenever a new utterance–scene pair U–S is processed, the probabilities $P(\cdot|w_u)$ of all $w_u \in U$ are updated, affecting the cosine similarities between words w_u and all other words. The semantic network must be updated to reflect these changes in cosine similarities—*i.e.*, some edges may be added, some removed, some changed in weight. However, rather than calculating the (new) cosine similarities between each w_u and *all* other words, the model use a limited set of calculations. It first updates the current edges connecting w_u to its neighbors. Then it selects a small set of new words w_i that *potentially* have a high probability of being similar to w_u. This is accomplished by incrementally forming semantic clusters over word meanings that are adjusted when a word's meaning is updated (Anderson and Matessa, 1992). Each newly updated word meaning w_u is compared to an average (*i.e.*, prototype) representation of each cluster to determine its probability of belonging to that cluster. Finally, n words are selected from each cluster and their cosine similarity to w_u updated, where n is proportional to the probability of w_u belonging to that cluster. The number of computations is limited as w_u is only compared to the cluster prototypes and a restricted number of words from each cluster.

By limiting the number of computations at each step of learning, the model is more cognitively plausible than exhaustively updating the semantic network after each utterance. However, it also means that the resulting semantic network will be noisy—it may have missing, superfluous or incorrectly-weighted edges.

3 Experimental Data and Approach

In this section, we explain the details of the semantic fluency experiment as well as the semantic representation and search algorithm used in our simulations. All of the code and data necessary to reproduce our experiments are available at `https://github.com/FilipMiscevic/random_walk`.

3.1 Evaluation: Semantic Fluency Data

We evaluate our simulations using data from a semantic fluency experiment in which participants were tasked with naming as many animals as they can in three minutes (Hills et al., 2012, 2015). Hills et al. (2012) inferred that the recalled words (*e.g.*, *dog*, *cat*, *lion*, *zebra*) form semantically-related categories or "patches", based on their inter-item retrieval times (IRT)—the time elapsed between the naming of two sequential items that have not previously been recalled. They find that the IRT increases as search within a semantically-related category progresses. A switch into a different semantic category occurs when the IRT exceeds the participant's average IRT across the entire trial. The IRT then decreases and the pattern begins again (see Figure 2a). This result shows that participants exhibit different behavior when recalling words from within a semantic category compared to switching into a new semantic category. Hills et al. (2012) argue that this pattern is a consequence of an informed two-stage search process: local cues, such as similarity to the most recent response, are used to search within patches, and global cues, such as the overall frequency of a word, are used to switch into new patches. Here we replicate previous results that demonstrate that the IRT pattern (Figure 2a) can be predicted by a simple search given structured representations (Abbott et al., 2015; Nematzadeh et al., 2016). In addition, we show that this process matches other patterns observed in the semantic fluency experiment (Hills et al., 2015).

3.2 Representation: A Semantic Network

We assume words and their relations are structured as a semantic network—a graph whose nodes are words, and edges reflect the similarity between the word meanings. We compare two sets of semantic networks, one set created *after training* the word learner explained in Section 2.1, while the other is built *incrementally during the training*, as described in Section 2.2. While the model learns many words, we only consider animal words, as we can evaluate those against the semantic fluency experiment of Hills et al. (2012). We also include the word *animal* itself in the semantic networks, as this is the cue word used in the experiment.

Two words w_i and w_j are connected in the semantic network if the cosine similarity between their feature vectors, $P(\cdot|w_i)$ and $P(\cdot|w_j)$, is above the threshold, $\rho = 0.8$. An exception is made for words connected to the word *animal*: because *animal* is a hypernym of the other animals, its cosine similarity will be less than the cosine between animals of the same subcategory. As such, to ensure that *animal* remains connected to some words in the network, edges radiating from it are kept if the similarity is at least $\rho_{animal} = 0.4$. Both models learn the representations of all 93 animal words present in the corpus; however, not all nodes are guaranteed to be connected to the rest of the network due to this thresholding. These thresholds were determined by a grid search over the possible values of ρ and ρ_{animal} (*i.e.*, $(0,1]$). The model predicts the human data over a notable range of parameter values; nonetheless, there are still more networks in that parameter space that do not predict the data. In Section 5, we will explore what characteristics of the networks are responsible for their successful prediction of data.

Batch Network. The word learner was trained on 120k utterance–scene pairs, with the meaning representation of a word, $P(\cdot|w)$, calculated as described in Section 2.1. After training has concluded, a semantic network is constructed using the final learned representations. A total of 70 words is present in this network.

Incremental Network. The learner is trained on 28k utterance–scene pairs.[1] After each utterance–scene pair is processed, the connections in the semantic network are updated as described in Section 2.2. A total of 75 words is present in this network.

Note that although the word representations of each model are learned by the same learning algorithm, they produce very different semantic networks. In the Batch Network, the edges are created only after training is completed, and is accomplished by exhaustively computing the cosine similarity between all word-pair combinations. The Incremental Network, on the other hand, uses a more cognitively plausible approximation of this process whereby edges are incrementally created by comparing only a small percentage of the word pairs.[2] This means that relations captured by the edges of the Incremental Network are noisier and incomplete.

The Incremental Network still only approximates the process of semantic acquisition in people, albeit more plausibly compared with previous work. As described above, however, we empirically set two thresholds that determine whether words are connected or not: one for the word *animal* and another one for all other animal words. Future work will need to explore whether this distinction can be learned while the network is incrementally created.

3.3 Search Algorithm: A Random Walk

We model the search process as a random walk in which semantic information is retrieved by randomly visiting nodes in the semantic network. Recall that in the semantic fluency experiment, the participants were cued by the word *animal* and were asked to name as many animals they can in three minutes. Following Abbott et al. (2015), we simulate this experiment by performing a weighted random walk on each network, beginning with the word *animal*. At each step in the random walk, a neighboring node is visited with a probability proportional to the edge weight connecting

[1] Even with the smaller corpus (28k as opposed to 120k input pairs), the model predicts the semantic fluency data; thus, we used the smaller corpus to speed up our simulations.

[2] This ends up being only 8% of all $\frac{n(n-1)}{2}$ possible comparisons at each time step, where n is the total number of words seen by the learner at each time step.

them, and the visited word is stored. Just as repeated words are not considered in the human recall data, we assume the output of a random walk to be the sequence of unique words encountered—*i.e.*, each word is counted in the output only when retrieved for the first time. The number of steps taken before the walk terminates (including steps to already-visited nodes) is 70, which produces about the same number of words on the networks as human participants on average do (*i.e.*, 37 ± 5). The results we report are averages over 300 such walks.

3.4 Analyzing Random Walks

In the semantic fluency task, the human response patterns are reflected in changes in the inter-item retrieval time (IRT) over the list of responses. In the empirical data, IRT is the time elapsed from one word until the next word is recalled, and increases and decreases are observed as people switch from one semantic patch of words to another, as noted above. Thus, to evaluate the random walks in our semantic networks against this IRT pattern, we must define a measure of time in the simulated walks (since actual model speed is not an appropriate proxy). We also must determine what constitutes a patch and a switch between two patches.

3.4.1 Measuring Time and Semantic Distance

We follow Abbott et al. (2015) in defining the IRT in a random walk on a semantic network as the number of steps taken (i.e., number of edges crossed) between two words. More specifically, we define IRTs for our walks as follows: for each word that has not previously been visited by the random walk, the IRT is the number of steps taken in the random walk since the last word that was seen for the first time. For example, if the model visits the sequence of nodes "CAT,DOG,CAT,RAT", the random walk output is "CAT,DOG,RAT", and the IRT between CAT and DOG is 1, whereas the IRT between DOG and RAT is 2.

The IRT is considered a proxy for semantic distance between the words. Hills et al. (2015) also looked directly at semantic distances in the sequences generated in the human fluency task: They used vector-space representations (of the BEAGLE model) to calculate cosine similarity between consecutive words. As such, in addition to using IRT in assessing our walks, we also draw on the cosine similarities between words.

3.4.2 Identifying Patch Switches

Each word in a random walk is labeled by the category/categories it belongs to, as defined by Troyer et al. (1997). Words (*e.g.*, DOG) can belong to more than one category (*e.g.*, PETS, CANINE). As a result, there are different possibilities for defining what constitutes a patch and where the patch switches occur. We explore two different ways of defining patches over Troyer's categories, following Hills et al. (2015), as summarized in Figure 1.

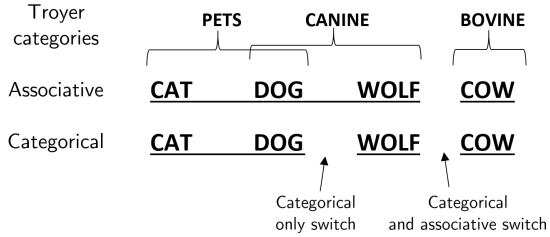


Figure 1: The difference between categorical and associative patch switches, based on Hills et al. (2015).

Categorical patch switch. A patch switch occurs when a word in the sequence has no category in common with *all* of the words in the current patch. In the sequence "CAT,DOG,WOLF", "DOG,WOLF" is a patch switch because WOLF is not in the same category as CAT (is not a PET).

Associative patch switch. A patch switch occurs when a word in the sequence has no category in common with the *last* word in the patch. For example, "DOG,WOLF" is not a patch switch because both words share the Troyer category CANINE, but "WOLF,COW" is a patch switch because they have no categories in common.

From this definition it follows that all associative patch switches are also categorical patch switches. However, a categorical patch switch may not be associative; one such "categorical only" patch switch is illustrated in Figure 1. Hills et al. (2015) argue that human search through memory is more like an associative search, and that the associative patch switch model better explains human IRT patterns. We use the associative patch switch model except where explicitly comparing the differences between the alternatives.

4 Predicting Semantic Fluency Data

Here we compare the results of random walks over the Batch and Incremental Networks in mimicking human semantic fluency data. First, we focus on predicting the pattern of recall observed in human data, then we examine the properties of each patch switch model.

4.1 Recall Patterns

In the human semantic fluency data (Figure 2a), the longest IRTs tend to occur between successive words that do not share a semantic category, presumably reflecting their greater distance in semantic memory (Hills et al., 2012, 2015). This is referred to as a patch switch. In the figure, a patch entry position of 1 indicates the average IRT between the first item in a patch and the item retrieved before it. Similarly, a patch entry position of -1 is the average IRT between the two items preceding a patch switch. Human IRTs in patch entry position 1 (patch switch) are higher than the average IRT, as people take longer to switch to a new patch, then dip below the average IRT at patch position 2 as people recall words within a patch.

As Hills et al. (2012) point out, this behavior is consistent with the marginal value theorem (MVT) of optimal foraging for patches of food in physical space

(Charnov, 1976). In particular, MVT demonstrates that to maximize foraging gains, the optimal moment to leave a current patch is when the instantaneous reward drops below the average reward. In the human semantic search task, since participants are asked to retrieve as many words as they can, shorter IRTs lead to a bigger 'reward', as more words can thus be retrieved within the time limit. Indeed, Hills et al. (2012) demonstrated that those subjects whose search patterns conformed with MVT retrieved the most words. We evaluate whether the IRT patterns of our models also conform to the predictions of MVT as observed in the human data. As such, the first patch-entry position IRT must be significantly greater than the mean IRT (e.g., the ratio between the two is greater than 1) and all other patch entry positions must be no greater than the mean IRT. Finally, successive IRTs within the same patch should be non-decreasing.

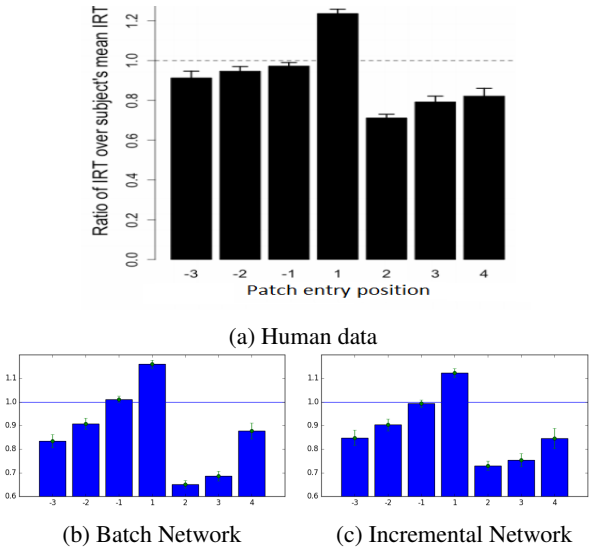
As shown in Figure 2, we observe a similar pattern to the human IRT data in both the Batch and Incremental Networks: the IRT drops between the first and second items in a patch, then steadily increases until the IRT exceeds the long-term average IRT, reflecting a patch switch. A single-sided t-test confirms that the first patch entry IRT is greater than the average IRT ($p \ll 0.001$). We accept the null hypothesis that the patch entry IRT at position -1 is no greater than the average IRT ($0.08 \leq p \leq 0.20$). The other IRTs are significantly less than the average IRT ($p < 0.02$) and successive IRTs within a patch are indeed non-decreasing.

This demonstrates, for the first time, that the combination of a simple search and structured representation that is incrementally created—simultaneously, as words are learned—can predict basic patterns observed in human semantic fluency. Next, we model additional aspects of the human data that have not been considered in previous work (Abbott et al., 2015; Nematzadeh et al., 2016).

A roughly analogous pattern with respect to patch entry positions is found with the average cosine similarities, although here, because cosine represents similarity rather than distance, the direction is reversed, as seen in Figure 3. Words at a patch switch are the least similar to one another. Again, the first patch entry position cosine similarity is significantly less than the average cosine simimlarity ($p < 0.05$). The other patch entry position cosines are on average no smaller than the average ($p \geq 0.05$). This supports the notion that words within patches are more similar (and hence, closer in semantic memory) to each other than words between patches.

4.2 Patch Switch Type Proportion and Duration

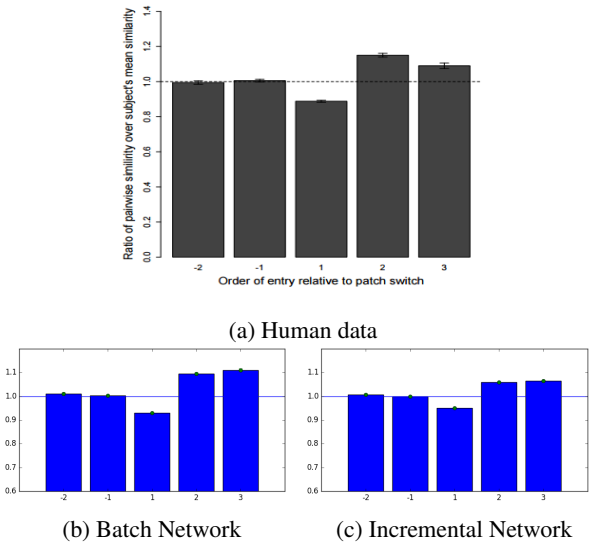
Hills et al. (2015) categorize patch switches on the human data by whether they are associative or categorical-only (see Figure 1). Two observations are made from this data. Firstly, as in Figure 4a, the proportion of associative patch switches steadily in-



(a) Human data

(b) Batch Network (c) Incremental Network

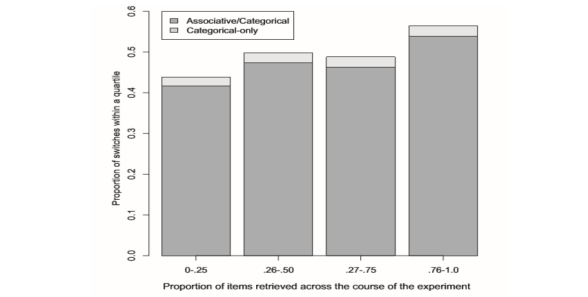
Figure 2: (a) Human IRTs reproduced from Hills et al. (2012). (b,c) IRTs from random walks generated from the simulated semantic networks. Bars are SEM.



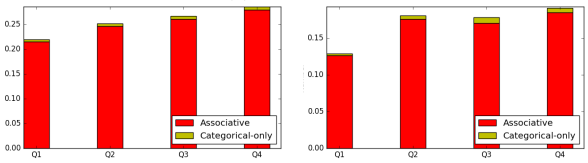
(a) Human data

(b) Batch Network (c) Incremental Network

Figure 3: Cosine similarities between words in successive patch positions normalized by the average long-term cosine similarity in (a) BEAGLE vectors for items retrieved by humans (Hills et al., 2012), (b,c) our semantic networks.

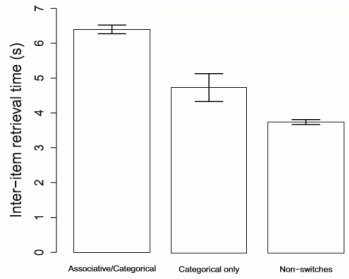


(a) Human data

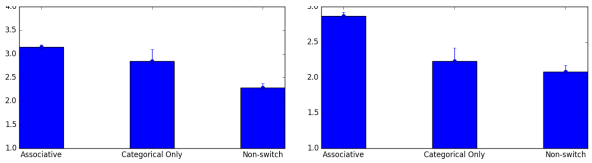


(b) Batch Network (c) Incremental Network

Figure 4: Average proportion of patch switch type on each quartile of the random walk for (a) human data (Hills et al., 2012), (b,c) our semantic networks.



(a) Human data



(b) Batch Network (c) Incremental Network

Figure 5: Average IRTs based on patch switch types for (a) human data (Hills et al., 2012), (b,c) our semantic networks.

creases throughout the four quartiles of the walk, but the number of categorical-only patch switches stays the same. This suggests that as more words are retrieved and semantic patches are depleted, new semantic patches must be explored. However, the categorical-only switches do not change in frequency. We speculate this may either be because they do not contribute to the need to explore different patches, or that they are so uncommon to begin with.

Secondly, as in Figure 5a, associative and categorical-only switches take longer than non-switches, which is expected, as non-switches search within a patch of semantically-related words. Associative switches take the longest, as they delineate the boundaries between the most semantically-different categories (compared to categorical-only switches).

Model Predictions. When we subject the random walks on our networks to these analyses, we observe the same pattern (Figures 4,5). This is the first work to confirm that a random walk on semantic network is consistent with the observed pattern on the duration and proportion of different types of switches.

Hills et al. (2015) point out the associative patch switch model has a *Markov property*, insofar as that only the proceeding word's category affects the existence of a patch switch with the next word. This is an interesting observation because it suggests that the associative switches may simply be easier to make, as only the previous word's categories affect the transition to the current word. In contrast, a categorical-only switch demands higher memory overhead as the next word is affected by the overall category/categories shared by members in the current patch. Our results show that a random walk on a structured semantic network can predict the timing and proportion of these different types of switches.

5 Explaining Semantic Fluency Data

While our results confirm that a simple search on an incrementally-created semantic network mimics many aspects of semantic fluency behavior, not all the semantic networks predict aspects of the human data, such as adherence to MVT. Adding edges to the semantic network depends on the similarity between words reaching a certain threshold. We experimented with a wide range of thresholds on similarity of word pairs (see Section 3.2) and observed that patterns consistent with MVT, as in the human IRT data (Figure 2a), appear only within a certain parameter range. Since the choice of threshold affects the overall structure of the semantic network, we explore the features that distinguish those semantic networks that reproduce human semantic fluency patterns from those that do not.

Previous research has emphasized that semantic networks representing human knowledge have particular structural properties; namely, a small-world structure, as explained below (Steyvers and Tenenbaum, 2005). However, Nematzadeh et al. (2016) observe that having a small-world structure is not a sufficient condition to guarantee a match to the observed human behavior in semantic search. A factor that has remained unexplored is how the quality of a network's semantic connections—whether semantically similar words are connected through a path or not—affects a network's ability to replicate findings in human semantic search. We hypothesize that this semantic quality is also important in predicting the semantic fluency data, because even two networks identical in every way except for their node labels would produce very different behavior as the relationships between the words they represent would be completely different.

Here we perform an extensive analysis considering both structural and semantic properties of the networks to assess which features contribute to the model's adherence to MVT, a major pattern in the human data.

By identifying these features, we can better understand the salient aspects of semantic memory that give rise to patterns in human semantic search. We first explain how we measure the structural and semantic features of the networks. Then we discuss how we build a regression model to determine which features are responsible in predicting the semantic fluency data.

5.1 Measuring Structure and Semantics

A network exhibits small-world structure if it is sparse and highly connected at the same time—there are not a lot of edges in the network, but most nodes are connected through a set of high-degree nodes. As a result, the network consists of a set of highly-connected components that are connected through the high-degree nodes. Small-worldness is often quantified by σ:

$$\gamma = \frac{C}{C_{random}}, \qquad \lambda = \frac{L}{L_{random}}, \qquad \sigma = \frac{\gamma}{\lambda}$$

where C is the average local clustering coefficient and L is the average path length, and the subscript *random* refers to the metric of an equivalent Erdős-Renyi network. A network is considered to be small-world when $\sigma > 1$ (or more strictly, $\gamma \gg 1$, $\lambda \approx 1$) (Watts and Strogatz, 1998). Intuitively, $\gamma \gg 1$ reflects a structure of tightly connected components in the network, and $\lambda \approx 1$ reflects relatively short path distances between nodes compared to a random network.

We observe that all of the semantic networks capable of reproducing the human patterns are small-world, but not all small-world networks generate these patterns, which is consistent with the findings of Nematzadeh et al. (2016). As a result, we consider other structural and semantic features. The structural features include the number of vertices ($|V|$), number of edges ($|E|$), and the sparsity of the network (average nodal degree).

Quality of semantic connections. In addition to the structure of a network, we examine the quality of its semantic connections. We explore this by first identifying the semantic clusters formed in each network using the HDBSCAN algorithm (Campello et al., 2013), and then evaluating these clusters using Troyer's categories as our gold-standard data Troyer et al. (1997). We assume that each cluster in the network can have exactly one category (e.g., pets). To determine the category label of a cluster, we examine the Troyer category memberships of each of its words, and assign the category label based on which category is shared by the most words of the cluster.

We use the standard measures of precision, recall, and F-score to assess the quality of each cluster, and average these across all clusters, weighted by cluster size, to obtain weighted precision, weighted recall, and weighted F-score for a network. We also consider the number of clusters in each network as a feature, $|H|$.

5.2 Analyzing the Contribution of Features

We characterize which structural and semantic features of a network are most important (in predicting human data) by fitting logistic regression models on all the possible combinations of features.

Prior to training, feature values were transformed into z-scores (i.e., for a given feature x for a given network i, the standardized value is $(x_i - \bar{x})/\hat{s}$; $\bar{x}$ is the sample mean of the feature for all networks and $\hat{s}$ is sample standard deviation). This permits the coefficients of regression to be compared directly in terms of their contribution in predicting the data.[3]

5.2.1 Experimental Set-Up

Logistic classifier models were trained on a set of Batch and Incremental networks. During training, we ensure an equal representation of networks that adhere to and do not adhere to MVT. This is a binary condition satisfied according to the criteria explained in Section 4.1. Networks were first generated across the entire parameter space of the similarity thresholds (i.e., all combinations of ρ and ρ_{animal} ranging from 0 to 1, in increments of 0.1). We excluded networks where the number of nodes reachable by the starting word 'animal' was smaller than 30, as they would not be able to produce as many words as human participants did (37 ± 5) (Hills et al., 2012). Since the number of non-IRT producing networks outnumbered the IRT producing networks, we uniformly sampled the parameter space in which IRT pattern-producing networks occurred so that the number of each would be equal. Using this procedure, 42 Batch and 56 Incremental networks were generated. In each case, exactly half of the networks produce the IRT pattern consistent with MVT.

Model selection. For each set of Batch and Incremental networks, we examine which features best predict the human data by building and evaluating logistic regression models for all possible combinations of the features. Model selection was performed in two steps. First, the models with the highest stratified-3-fold (SKF) cross-validation score were taken. From these, the model with the fewest number of features was selected.

5.2.2 Results of Logistic Regression

Table 1 shows the features that appeared in the logistic regression model that achieved the best SKF cross-validation score for each of the types of networks. Since each feature was standardized (with $mean = 0$ and $variance = 1$), the magnitude of the coefficients can be interpreted directly. We note that small-worldness (σ) and weighted F-score are influential predictors for both Batch and Incremental networks. In both models, weighted F-score is the most influential predictor. Although σ is the least influential predictor, we find it significant that it is a shared

[3] Although some of these features are dependent (*e.g.*, $|E|$ and *sparsity*), we do not include their interactions in our regression analysis. We focus on understanding whether a subset of individual features can explain the human data and thus examine all possible combinations of features.

Networks	Acc.	Features and Coefficients				
Batch	93%	**σ**	λ	C	sparsity	**weighted F-score**
		0.58	0.74	-1.92	0.94	**0.94**
Incremental	90%		**σ**	γ	$\|E\|$	**weighted F-score**
			0.65	0.71	-1.64	**1.07**

Table 1: Features used to train the logistic regression models for predicting IRT pattern production with the highest stratified 3-fold cross-validation accuracy (Acc.). Shared features are bolded.

predictor for both networks. Structural properties relating to the number of edges ($|E|, sparsity$) as well as clustering coefficient (C, γ), are structural properties that have been previously characterized in semantic networks (Steyvers and Tenenbaum, 2005; Goñi et al., 2010). Hence, we conclude that both topological features—namely, small worldness (high clustering coefficient and short average path length)—and semantic features—high weighted F-score (good precision and recall in clusters)—are jointly associated with reproducing the IRT pattern.

6 Conclusions

Learning word meanings and representing them in semantic memory are processes that often occur simultaneously, notably in early language acquisition. A cognitive model capable of integrating these two processes will therefore more realistically capture language acquisition and usage. It is noteworthy that both the Batch and Incremental Networks perform comparably on all of the data examined here. We consider this strong support for the hypothesis that semantic networks learned incrementally on a naturalistic language corpus can replicate search patterns in the free recall task, a claim that is neither obvious nor trivial to demonstrate. Furthermore, some of the performance characteristics we use in measuring the fit of the model to the human data—namely, whether the IRT patterns produced by the model are consistent with MVT or not—are binary conditions: either the behavior is replicated or it is not, so, barring additional criteria, a graded scale by which to score performance is not possible. Future work will seek to better characterize the performance differences between the two models.

We deploy a model that can generate semantic networks incrementally from naturalistic language use, *i.e.* child-directed speech, while it gradually learns the word meanings, lending it plausibility as a cognitive model. We show this model replicates human performance on semantic fluency tasks; namely, with regards to patch entry IRT, patch entry cosine similarity patterns, patch switch type proportions, and patch switch type IRTs. We show, furthermore, that the Markov property of the random walk does indeed align with the associative nature of search in the human semantic fluency task (Hills et al., 2015).

By investigating the structural and semantic features of these and other networks, we show that small-worldness alone does not explain the ability of a network to replicate the human patterns. Having highly connected components, *and* ones that reflect the semantic categories of words, are both properties that may be necessary in predicting semantic search behavior observed in humans.

References

Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. 2015. Random walks on semantic networks can resemble optimal foraging. *Psyc. Rev.* 122(3).

John R. Anderson and Michael Matessa. 1992. Explorations of an incremental Bayesian algorithm for categorization. *Machine Learning* 9(4):275–308.

Alfredo Ardila, Feggy Ostrosky-Sols, and Byron Bernal. 2006. Cognitive testing toward the future: The example of semantic verbal fluency (animals). *International Journal of Psychology* 41(5):324–332. https://doi.org/10.1080/00207590500345542.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. *PAKKD 2013* .

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

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psyc. Rev.* 82(6):407.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cog. Sci.* 34(6):1017–1063.

Christiane Fellbaum, editor. 1998. *WordNet, An Electronic Lexical Database*. MIT Press.

Joaquín Goñi, Gonzalo Arrondo, Jorge Sepulcre, Iñigo Martincorena, Nieves Vélez de Mendizábal, Bernat Corominas-Murtra, Bartolomé Bejarano, Sergio Ardanza-Trevijano, Herminia Peraita, Dennis P. Wall, and Pablo Villoslada. 2010. The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cognitive Processing* 12:183–196.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psyc. Rev.* 114(2):211.

Thomas T Hills, Michael N Jones, and Peter M Todd. 2012. Optimal foraging in semantic memory. *Psyc. Rev.* 119(2):431.

Thomas T Hills, Peter M Todd, and Michael N Jones. 2015. Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science* 7:513–534.

Michael N Jones, Thomas T Hills, and Peter M Todd. 2015. Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psyc. Rev.* 122(3).

Michael N Jones and Douglas JK Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psyc. Rev.* 114(1):1.

Susan S. Jones, Linda B. Smith, and Barbara Landau. 1991. Object properties and knowledge in early lexical learning. *Child Development* 62(3):499–516.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, volume 2: The Database. Erlbaum, 3rd edition.

G. A. Miller and C. Fellbaum. 1991. Semantic networks of English. *Cognition* 41(1–3):197–229.

Radford M. Neal and Geoffrey E. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, Springer, pages 355–368.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 1998. The University of South Florida free association, rhyme, and word fragment norms .

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. 2014. A cognitive model of semantic network learning. In *Proceed. Conf. on Empirical Methods in Natural Lang. Processing*.

Aida Nematzadeh, Filip Miscevic, and Suzanne Stevenson. 2016. Simple search algorithms on semantic networks learned from language use. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cog. Sci.* 29(1):41–78.

Anna L. Theakston, Elena V. Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language* 28:127–152.

Angela K. Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology* 11(1):138–146.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442.