Searching Semantic Memory as a Scale-Free Network: Evidence from Category Recall and a Wikipedia Model of Semantics

Graham William Thompson (gthompson2@ucmerced.edu)
Christopher T. Kello (ckello@ucmerced.edu)
Cognitive and Information Sciences
University of California, Merced
5200 North Lake Road, Merced, CA 95343 USA

Abstract

How is semantic memory structured and searched? Recalling items from semantic categories is a classic assay of semantic memory, and recall dynamics tend to exhibit semantic and temporal clustering, as if memory items are organized and retrieved in clusters. Recent analyses show this clustering to be approximately scale-free in terms of distributions of inter-response intervals (IRIs). This finding is replicated and extended in the present study by asking participants to type as many animals as they can recall from semantic memory. To begin to explain these results, the organization of semantic memory is modeled as a network based on Wikipedia entries for nearly 6,000 animals. The Wikipedia animal network is found to be scale-free in terms of its degree distribution, and aspects of the network are found to correlate with aspects of recall. Semantic similarity based on Wikipedia entries is found to compare favorably with a measure based on latent semantic analysis. It is concluded that semantic memory processes can be usefully theorized as searches over scale-free networks.

Keywords: semantic memory, scale-free networks, Lévy foraging; category recall; latent semantic analysis; Wikipedia

Introduction

Category recall is a classic approach to investigating semantic memory. Participants produce as many items from a semantic category as possible in a specified period of time (Bousfield & Sedgewick, 1944). Items tend to be recalled in clusters. For the category of "animals", for instance, part of a typical sequence might be "lion, tiger, cougar, leopard… kitten, cat, tabby". This sequence contains two groups of semantically similar items, big wild cats followed by house cats. Such clusters can be of varying kinds and sizes, and they tend to correspond with short IRIs, relative to longer pauses when switching from one cluster to the next (Grunewald, Lockhead & Gregory, 1980).

Clustering seems to be a general feature of semantic memory. Work in this area has a long history, with early experiments showing that, when participants memorize words presented in random order, they tend to recall those words in clusters based on semantic categories (Bousfield & Sedgewick, 1953). Therefore clustering must be related to memory encoding, retrieval, or both. In clinical work, semantic category recall is used as a diagnostic for mental disorders. Schizophrenic patients, people with semantic dementia and people with Alzheimer's all show specific deficits in category generation tasks (see Murphy, Rich & Troyer, 2006).

Previous work has sought to account for clustering in category recall with patch foraging models (see Hills, Jones & Todd, 2012). Patch foraging theorizes semantic memory as a set of patches of similar items. Memory search consists of series of quick retrievals of items from within a patch, interleaved by longer times switching to the next patch when a sufficient number of items in the current patch have been found. Framed this way, optimal foraging can be expressed in terms of the time to leave a patch. It is optimal to switch when the instantaneous rate of recall per unit time drops below the long-term expected rate of recall (Charnov,1976). Category switch times, and times in other human search tasks, have been found to be consistent with patch foraging (Cain, Vul & Mitroff, 2012).

Patch foraging models put little, if any, theoretical weight on the distribution of patch sizes, given that only their mean size is needed to compute the expected rate of recall (along with basic assumptions about retrieval times). That said, recent work on category generation tasks has examined IRIs in more depth (Rhodes & Turvey, 2007). When the category recall task was of sufficient length (e.g. ten to twenty minutes for recalling animals), IRIs were found to be power law distributed, where the frequency of an IRI was an inverse power of its size, $P(IRI) \sim 1/IRI^{\alpha}$. Ideally, this distribution has no characteristic scale, in which case the mean and variance diverge as more samples are drawn from the distribution. The implication is that patches in memory also have no characteristic size, instead coming in a wide range of sizes, with the probability of observing ever larger patches increasing steadily over time.

Power law IRI distributions fall outside the purview of patch foraging models, but they have been studied extensively in animal foraging models (Viswanathan et al., 1996). Unlike patch models, animal foraging models explicitly consider the space in which items are to be found, such as trees and bushes in a meadow where birds are foraging for nuts and berries. Interestingly, the same power law distribution found in category recall is also found in inter-retrieval intervals during foraging for a wide range of species (Sims, Southall & Humphries, 2003). Theorists have related these findings to so-called Lévy walks (Mandelbrot, 1982), which are random walks with path lengths drawn from a power law distribution. While it is unlikely that foraging paths are literally random Lévy walks, they may capture an important property of foraging. The reason is that Lévy walks may be optimal search strategies when their

exponent $\alpha \sim 2$, and items are sparsely distributed (Viswanathan et al., 2000). Consistent with optimal Lévy walks, inter-retrieval intervals in animal foraging, as well as IRIs in category recall, have all been found to be distributed like a power law with the predicted exponent value (Rhodes & Turvey, 2007; Sims, Southall & Humphries, 2003).

What do these findings tell us about the process of searching semantic memory in category recall? They suggest that simple Lévy walks might characterize much about memory search, but they also might tell us about the structure of semantic memory. Items in memory are often theorized in terms of networks, in which case one is led to ask whether these networks are structured in such a way that searching them results in power law IRIs. As it turns out, recent work on semantic networks shows their degree distributions might be scale-free, i.e. follow a power law distribution (Steyvers & Tenebaum, 2005).

A network consists of a interconnected nodes, and the degree of a node is its number of connections. Scale-free networks are those whose distributions of node degrees follow an inverse power law. Many natural and manmade networks are scale-free, such as power grids, brain networks and the World Wide Web (Strogatz, 2001). Steyvers and Tenenbaum (2005) analyzed three different types of data as reflections of semantic networks: word associations, WordNet entries (Miller, 1995) and Roget's Thesaurus. In all three cases, data were used to link items in a network based on similarity or associative relations, and in all cases networks were scale-free.

This evidence for scale-free networks suggests that items fall into clusters with no characteristic size, analogous to the link drawn from power law IRI distributions to patches with no characteristic size. This analogy suggests that a scale-free semantic network might account for power law IRI distributions, as well as the semantic clustering of items in category recall tasks. In the present study, we collect data in a category recall task and test whether observed clusters of recalled items, and associated IRIs, can be explained by a model of semantic memory.

We draw and expand upon previous studies as follows. Category recall data are collected via typed instead of spoken responses, as in previous studies. This difference allows us to test whether previous findings replicate when response dynamics are on the order of seconds (typing) instead of milliseconds (speech). Typing also allows us to test whether the same power law distribution occurs in IRIs, as well *within* responses (typing durations). If so, we would have evidence that recall processes unfold continuously throughout the task, rather than in alternating stages of recall and response execution (Kawamoto, Kello & Jones, 1998; Spivey & Dale, 2006).

We then build a semantic network of animals using over 6000 pages from Wikipedia. We follow an information theoretic method used previously to show that the entirety of Wikipedia can be formalized as a scale-free network (Massuci et al., 2012). We test whether this method replicates when analyzing only one subset domain of Wikipedia, and we test whether the resulting measures of animal similarity and network structure can be used to explain an online behavioral measure, i.e. category recall data in this case. We also compare Wikipedia measures of semantic similarity with those generated from latent semantic analysis (LSA) of linguistic corpora (Landauer & Dumais, 1997). LSA has become a standard co-occurrence method, whereas Wikipedia is new encyclopedic method. We end by discussing the implications of results for Lévy and patch foraging models, semantic memory, and search processes in general.

Experiment

Methods

Participants and Procedure. Nineteen undergraduates at University of California, Merced participated for course credit. Participants were instructed to recall as many members from the category of "animals" as they could in twenty minutes, after first completing three minutes of practice with naming colors. Responses were typed and recorded using a Flash interface that stored the timing of each key press. Key press times were used to calculate the intervals from the end of one response to the start of the next, termed inter-response time (IRT), and the time from start to end of respond, termed typing duration.

Results

The average number of animals produced by each participant was 117 (SD = 38.6). Distributions of IRIs and typing durations were plotted in logarithmic coordinates to gauge whether they resembled power law distributions. As shown in Figure 1, the negative linear relation is indicative of a power law, and multi-model inference tests (Akaike, 1974) confirmed that 4 subjects were best fit by a power law, and the other 15 were best fit by a lognormal (which is akin to a constrained or truncated power law in this case). The deviations from linear at left end of these distributions were due to minima that constrained and thereby distorted the power law relationship. Distortion aside, the slope of these distributions in logarithmic coordinates was near -2, which replicates the category recall findings of Rhodes and Turvey (2007). Thus memory retrieval dynamics followed the same pattern for slower typed responses, relative to faster spoken responses.

Typing durations also followed the same power law relation, suggesting that memory retrieval is ongoing during response execution. To test whether this result may have been due to variations in response length, typing durations were normalized by the number of letters in each response. As shown in Figure 1, normalized distributions had the same overall shape as the others. Thus response length did not factor into the results.

In addition to IRIs, the category recall task also yields series of recalled animals. Visual inspection of these series indicated that, as expected, semantically related animals

tended to be recalled in close proximity compared with less related animals. Next we describe the Wikipedia semantic network model and test whether it can account for the relative positioning and clustering of items in recall sequences.

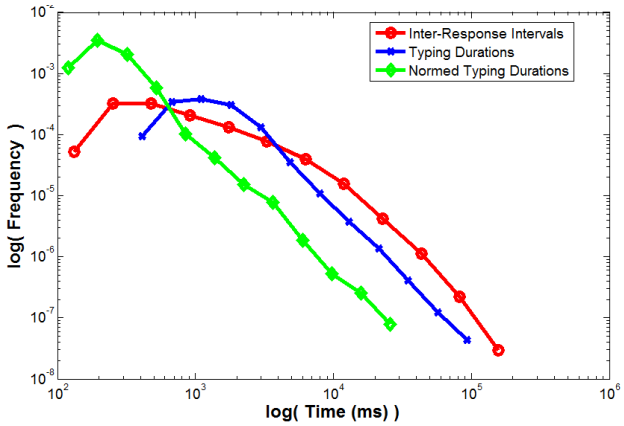

Figure 1. Response histograms in logarithmic coordinates.

Semantic Memory Modeling

The network method developed by Masucci et al. (2011) is based on transforming each given Wiki page into a probability distribution over lemmas, and then using Jenson-Shannon divergence (JSD) to measure the distance between two probability distributions. Animal Wiki pages were found using the Dbpedia ontology (Auer, Bizer & Kobilarov, 2007) which contains a list of all articles in Wikipedia associated with a given tag. A list of 129,027 animal articles in Wikipedia was compiled and all stub articles, redirect pages and articles with under 500 words of main text were removed, leaving 5,701 animal pages. Formatting, references, and function words were removed, and remaining words were lemmatized to collapse across different inflectional forms.

The resulting frequency counts over lemmas on each page were normalized to create probability distributions, and each distribution served as a semantic representation of the corresponding animal. These representations can be used to determine which animals are and are not linked in a semantic network, provided there is a good measure of similarity between probability distributions. Note that two distributions for two given pages may only partially overlap in their corresponding sets of lemmas, which means that a similarity measure must encompass and normalize over varying degrees of overlap.

A well-known measure of similarity between two probability distributions is the Kullback-Liebler (KL) divergence, defined as

$$\mathrm{KL}[\rho_1, \rho_2] = \int_{\Delta} \rho_1(x) \ln \frac{\rho_1(x)}{\rho_2(x)} \, dx.$$

This divergence is asymmetric and non-normalized, whereas JSD is a symmetric extension of KL divergence, normalized between zero and one:

$$\mathrm{JSD}[\rho_1, \rho_2] = \frac{1}{2}\left(\mathrm{KL}\left[\rho_1, \frac{\rho_1 + \rho_2}{2}\right] + \mathrm{KL}\left[\rho_2, \frac{\rho_1 + \rho_2}{2}\right]\right),$$

JSD can be thought of as providing a measure of how much the same lemmas are used with the same frequency between two Wiki pages.

Semantic Network. JSDs were calculated for all pairs of probability distributions, and an undirected semantic network was created by connecting any two animals with a JSD similarity below a given threshold. The threshold was chosen to be just high enough to merge 90% of the animals into a single, interconnected network (every node could be reached from every other node by traversing the network, and unconnected nodes were removed from analysis).

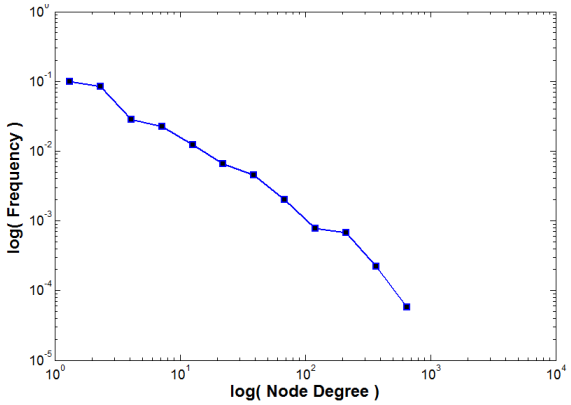

Figure 2. Degree distribution of the Wikipedia semantic network in logarithmic coordinates.

The structure of the resulting network was sparse, small-world and approximately scale-free. Average minimum path length was 3.65, average clustering coefficient was 0.529, the diameter of the network was 14 and the degree distribution followed a power law distribution (Figure 2). This finding replicates Masucci et al. (2011) for a subset of Wikipedia, and it provides convergent evidence with Steyvers & Tenenbaum (2005) that semantic memory can be expressed as a scale-free network.

Accounting for Category Recall Results

Semantic networks are often compared with offline human behaviors that can be expressed as structures. Wikipedia provides crystallized, idealized representations of animals, but it stands to reason that online measures of human behavior would be sensitive to these representations. We examine three such measures from our category recall data: 1) Animal response similarity as a function of distance in recall sequences, 2) First-order transitions in recall sequences, and 3) IRIs as a function of shortest path length in the scale-free semantic network.

To provide a benchmark for these three measures, we also computed semantic similarity using LSA, which has become a standard correlate of lexical semantic representation. As a co-occurrence method, LSA has strengths and weaknesses

compared with our Wikipedia-based method. Its main strength is that LSA can provide a representation for every word in a set of documents, whereas the Wikipedia method can only provide representations for existing entries. However, each entry unambiguously corresponds to a particular semantic item, whereas LSA merges all the different meanings and usages of each given word, like "fish", into a single semantic representation. LSA word vectors cannot be combined to form compound representations that correspond to animals like "flying fish" and "zebra finch".

Responses were corrected for spelling mistakes, multi-word responses like "black bear" were reduced to their superordinate category, i.e. "bear" in this case. LSA vectors were calculated using a term by term comparison from the general reading to first year college corpus with 300 factors from the LSA website. Of the 827 unique items produced in the category recall experiment, we were able to compute 196 animal LSA vectors, and 293 Wiki probability distributions.

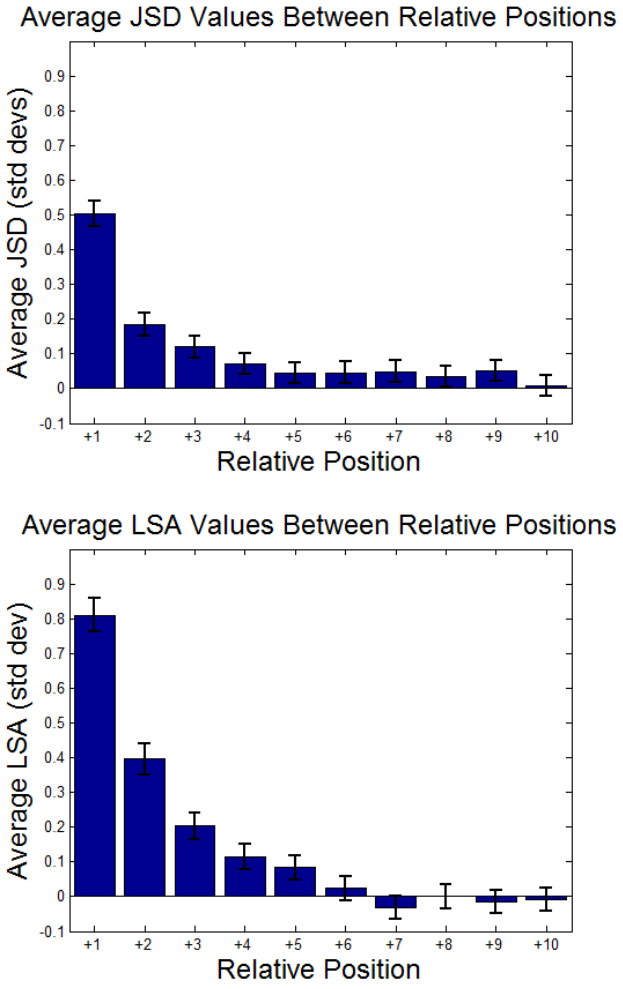

Figure 3. Mean JSD and LSA measures as a function of relative positions, with standard error bars.

Recall Proximity. The first measure we examine is related to evidence marshaled for patch foraging models. Hills et al. (2012) used a co-occurrence method called BEAGLE (Jones & Mewhort, 2007) to show that items produced within a patch are more semantically related as a function of proximity in a sequence. Given the evidence for patches of all sizes, and hierarchical nesting of patches as evidenced by power laws, we tested whether the analysis could be extended to all items in category recall sequences, without setting patch boundaries.

JSD and LSA measures were computed between pairs of animal responses in category recall sequences, as a function of the relative position of items, from 1 (adjacent) to 10 (nine intervening items). JSD and LSA measures of semantic similarity were averaged for each relative position, and then normalized by the mean and standard deviation for all pairwise JSD and LSA similarities, across all relative positions. Results are shown in Figure 3.

The JSD and LSA measures produced comparable results showing that semantic similarity decreased as a function of positional distance in sequences. This result confirms visual inspection of sequences, as well as previous research (Bousfield & Sedgewick, 1953) showing that similar items are recalled in nearby positions. Both measures also compared favorably with the Hills et al. (2012) results, which showed a distinct effect of similarity only for immediately adjacent items in recall sequences. Quantitative differences between LSA and Wikipedia methods were also observed: Compared with Wikipedia, LSA registered relatively higher similarities for immediately adjacent items, but similarity fell off more quickly with increasing distance.

For each recalled item, all animals that could be recalled next were arranged according to the JSD or LSA similarity between them, creating a ranking of possible transitions. Transitions to every ranking were normalized by the sum of all JSD or LSA transitions. Probabilities were then divided by random chance for each analysis (1/293 and 1/196, respectively), to show proportion above or below chance.

Transitional Probabilities. LSA and JSD measures used in the recall proximity analysis can serve as a basis for predicting performance in category recall, but additional mechanisms would be needed to fully simulate recall sequences. As a start, we used LSA and JSD measures to compute first-order transition probabilities as a simple means of predicting each recall item in a sequence, based only on the previous item. This analysis extends the previous one because each transitional probability is computed relative to all possible recall items, which is probably more akin to a model of category recall that simulated dynamics of semantic memory, e.g. by as traversing a scale-free semantic network.

As shown in Figure 4, for both JSD and LSA measures, participants transitioned to the most similar items with a higher probability. JSD transitions to the highest ranked word made up around 7% of total transitions, and for LSA it was around 3%. This effect falls off after the first 30-50 most similar items, and is most pronounced for the JSD

measure. Therefore the JSD measure appears to be better at predicting transitional probabilities compared with the LSA method.

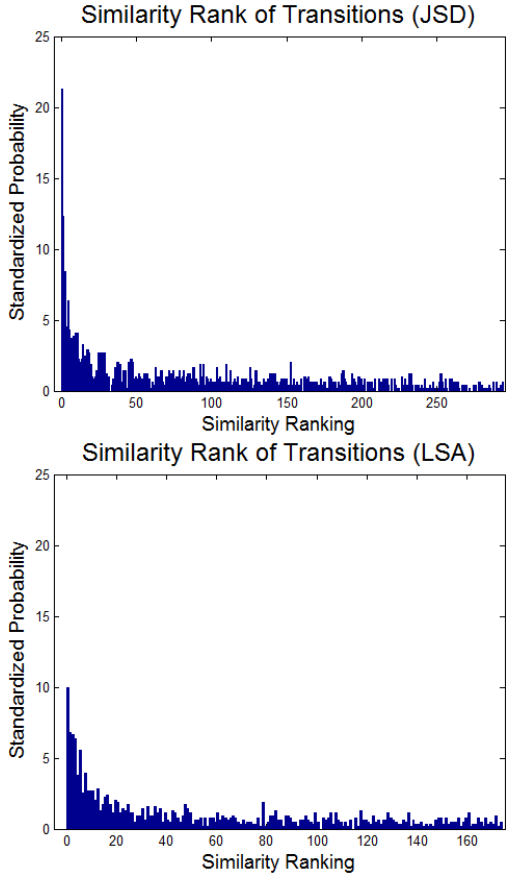

Figure 4. Standardized probability of first-order transitions between items by similarity ranking for JSD and LSA measures.

Accounting for IRIs. The previous two analyses focused on the sequencing of recalled items, but the times between recalls are arguably more at issue in theories of semantic memory. Both patch and Lévy foraging theories aim to explain IRI effects—patch transition times for the former, and IRI distributions for the latter. We tried using JSD and LSA similarities alone, as in the previous two analyses, to account for IRIs. However, they did not correlate reliably with IRIs under any transformation of the data we tried.

Despite the lack of a direct link between semantic similarities and IRIs, a semantic network built from similarities still may account for IRIs by virtue of network structure and dynamics that capture interactions among items in memory. We used the scale-free semantic network reported earlier, based on Wikipedia JSDs, to account for IRIs observed in our category recall experiment. We did not build a network based on LSA values because the lack of semantic representations for word phrases like "tiger shark" prohibited us from creating a sufficiently rich network.

Simulating network dynamics is beyond the present scope, but we accounted for IRIs using a standard measure of network structure that is likely to have a strong influence on network dynamics. Minimum path length is the minimum number of links needed to traverse from one node to another. Minimum path length provides a measure of how disparate two nodes are in the context of an interconnected network, and will directly impact any walker or spreading activation mechanism used to formulate network dynamics.

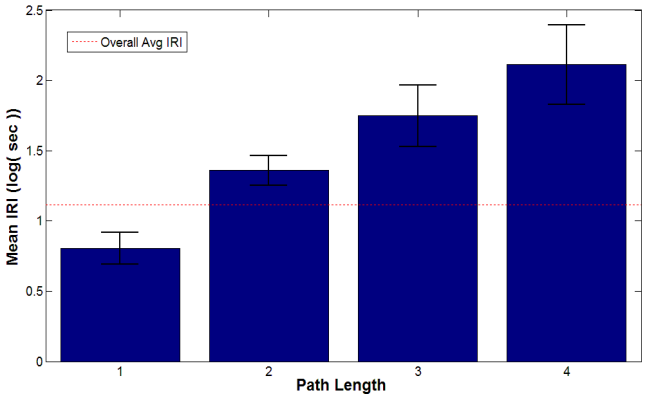

Figure 5. Mean logged IRI between words with different minimum path length separation in the semantic network.

Minimum path lengths were computed between all adjacently recalled items in all sequences from the category recall experiment. Minima ranged from 1 to 6, but we only examined pairs from 1 to 4 because there were too few at 5 and 6 to afford analysis. IRIs were logarithmically transformed due to their heavy tails (as reported earlier), and mean log IRIs were computed for each minimum path length, as shown in Figure 5. There was a significant effect of path length on log IRIs, $\beta = .403$, $t(1067) = 7.44$, $p < .001$, and they accounted for a significant proportion of variance $R^2 = .049$, $F(1, 1067) = 55.36$, $p < .001$. IRIs were shorter for immediately connected items compared to the baseline mean IRI, and IRIs were progressively greater than baseline as path length increased from 2 to 4. This result provides evidence that a scale-free semantic network may account for IRIs in category recall experiments, even when semantic similarities alone are not enough to account for the data.

Discussion

In the present study, we provided initial evidence that category recall performance can be modeled using scale-free semantic networks. The category recall data were from series of typed responses, but IRI distributions had the same shape as in previous experiments using spoken responses. This replication indicates that the present analyses should generalize. Typed responses also indicated that memory dynamics unfold during response execution as well as the pauses between responses, which will be important to account for in future models and simulations.

Previous studies have provided behavioral evidence that semantic memory is organized as a scale-free network (Steyvers & Tenenbaum, 2005), and we showed that a semantic network of animals built from Wikipedia pages is

indeed scale-free. We used a measure of Wikipedia page similarity to account for basic aspects of recall sequences, and we showed that this measure of similarity compared favorably with a more standard LSA measure. We also showed that a basic aspect of scale-free network structure explained some of the variance in category recall IRIs.

The next step in this line of work is to implement network dynamics to test whether scale-free network structure can account for the scale-free, power law distribution of IRIs observed in category recall tasks. This test will bear on Lévy foraging theories that would explain power law IRIs in terms of random or correlated walkers. A parallel step will be to test whether network dynamics can account for evidence suggesting that foragers spend optimal amounts of time foraging within patches, and switch to new patches when current rates of recall fall below the long-run average. While our model does not have clear delineations between patches, because items fall into nested clusters with no characteristic scale, it may still account for patch evidence. Recent modeling work showed that random walks on semantic networks might account for this evidence without reference to patches (Abbott, Austerweil & Griffiths, 2012).

Finally, it will be informative to apply the Wikipedia method of network creation to other phenomena of semantic memory, in other semantic domains. In doing so, it will be important to compare this method with a range of co-occurrence methods, as well as other encyclopedic methods.

Acknowledgments

This work was supported by a grant from the National Science Foundation, BCS 1031903 (PI Kello).

References

Abbott, J., Austerweil, J., & Griffiths, T. (2012). Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems 25* (pp. 3050-3058).

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716-723.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The Semantic Web*, 722-735.

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

Bousfield, WA. 1953. The occurrence of clustering in recall of randomly arranged associates. *The Journal of General Psychology*, *49*, 229–240.

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science*, *23*(9), 1047-1054

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology*, *9*(2), 129-136.

Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225-240.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431-440.

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

Kawamoto, A. H., Kello, C. T., Jones, R., & Bame, K. (1998). Initial phoneme versus whole-word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 862-885.

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*(2), 211-240.

Mandelbrot, B. B. (1982). *The fractal geometry of nature*. Times Books: 232-244.

Masucci, A. P., Kalampokis, A., Eguíluz, V. M., & Hernández-García, E. (2011). Extracting directed information flow networks: an application to genetics and semantics. *Physical Review E*, *83*(2), 1-6.

Masucci, A. P., Kalampokis, A., Eguíluz, V. M., & Hernández-García, E. (2011). Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PloS one*, *6*(2), e17333.

Murphy, K. J., Rich, J. B., & Troyer, A. K. (2006). Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. *Journal of the International Neuropsychological Society*, *12*(4), 570-574.

T. Rhodes & M.T. Turvey. (2007). Human memory retrieval as Levy Foraging. *Physica A*, *385*, 255–260

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., & Metcalfe, J. D. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098-1102

Spivey, M. J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*(6825), 268-276.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413-415.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Havlin, S., Da Luz, M. G. E., Raposo, E. P., & Stanley, H. E. (2000). Lévy flights in random searches. *Physica A: Statistical Mechanics and its Applications*, *282*(1), 1-12.