Spring 2015

Towards an Integrated Model of the Mental Lexicon

Natawut Monaikul
Governors State University

Towards an Integrated Model of the Mental Lexicon

By

Natawut Monaikul
B.S., University of Illinois at Urbana-Champaign, 2013

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Computer Science

Governors State University
University Park, IL 60466

2015

Acknowledgments

I would like to thank my committee: Dr. Wong for encouraging me to do a thesis over a project in the first place, Dr. Tang for being a strong home base I can rely on for help and teaching me to always shoot higher, and Dr. Tweddle for giving me the inspiration for this topic in his amazing Graph Theory class.

I would also like to thank Dr. Cleland-Huang for essentially being my second mother and mentor. Without her, I would not even know where to begin with academic research, and my mind would not be as active as it is today. Her lab and her guidance has helped me become a rigorous researcher always craving to learn more and push onward. I must also thank her for passing on the knowledge of how to recover from failed results and find a diamond amidst a pile of coal.

Last, but not least, I would like to thank my dear family. They are always there to support me in my endeavors, and they can somehow still put up with my countless late-night projects. I could not be more grateful for where they have gotten me today.

Contents

Abstract

Several models have been proposed attempting to describe the mental lexicon—the abstract organization of words in the human mind. Numerous studies have shown that by representing the mental lexicon as a network, where nodes represent words and edges connect similar words using a metric based on some word feature, a *small-world* structure is formed. This property, pervasive in many real-world networks, implies processing efficiency and resiliency to node deletion within the system, explaining the need for such a robust network as the mental lexicon. However, each model considered a single word feature at a time, such as semantic or phonological information. Moreover, these studies modeled the mental lexicon as an unweighted graph. In this thesis, I expand upon these works by proposing a model that incorporates several word features into a weighted network. Analyses on this model applied to the English lexicon show that while this model does not exhibit the same small-world characteristics as a weighted graph, by setting a minimum threshold on the weights (reminiscent of action potential thresholds in neural networks), the resulting unweighted counterpart is a small-world network. These results suggest that a more integrated model of the mental lexicon can be adopted while affording the same computational benefits of a small-world network. An increased understanding of the structure of the mental lexicon can provide a stronger foundation for more accurate computational models of speech and text processing and word-learning.

Keywords: small world; mental lexicon; lexical network; graph theory

Chapter 1

Introduction

When we read the word "bug," we almost immediately recognize the word. Given that this thesis is computer science-related, memories of programming *errors* may have been evoked. General readers may have formed a mental image of a bug or *insect*, perhaps an *ant* or a *spider*, noting it has several *legs* or *antennae*. For a Danish speaker, this may have prompted visuals of a *belly*. The cursory reader may have accidentally read the word as *bog* and have begun envisioning a *mire* or *moss*. Regardless of the interpretation, the fact is that an isolated word has been accessed instantaneously and has activated so many related concepts.

How do we recognize and retrieve words so quickly? Of course, an answer to this would require a vocabulary "filing system" of some sort through which lexical access takes place, which poses a more fundamental question: how do we store words and concepts in our minds? This question has interested philosophers, linguists, and computer scientists alike for decades and has brought about the concept of the *mental lexicon*–the abstract organization of words in the mind. A deeper understanding of the mental lexicon can lead to more plausible and thorough explanations of linguistic activities and phenomena, such as reading comprehension, speech processing, and speech errors, as well as to more accurate and efficient computational models for artificial intelligence systems mimicking natural language understanding and production.

Recent technological advances and psycholinguistic experiments have made the abstract concept a reality by allowing researchers to physically model the mental lexicon as a complex network, leading to graph-theoretic analyses that reveal underlying properties exhibited by the network. These studies have effectively brought us closer to a more complete model of the mental lexicon, but current models are still far from complete. In this thesis, I seek to build upon these models as a nod towards a more integrated model, incorporating features of each model, as well as features previously explored only in psycholinguistic studies. I argue that such an integrated model has a structure similar to previous models, thereby affording the same computational benefits of those models, while more exhaustively including findings from the literature pertaining to the mental lexicon.

The remaining chapters are laid out as follows: Chapter 2 discusses theoretical models of the mental lexicon through the years and recent models analyzed using graph-theoretic techniques (with a brief introduction to the relevant concepts from graph theory), Chapter 3 introduces the integrated model I propose, as well as justifications for each aspect of the model, Chapter 4 presents the main findings of analyses over this proposed model, and Chapter 5 takes a closer look at the model and the findings, along with its implications.

Chapter 2

Literature Review

2.1 Historical Development

2.1.1 The Hierarchical Network Model

An early model of the mental lexicon was presented by Collins and Quillian in conjunction with developing a computational model for language comprehension, known as the Teachable Language Comprehender (TLC) (Quillian, 1969; Collins and Quillian, 1969). In this Hierarchical Network Model (HNM), words and concepts are represented as nodes in a taxonomy with subclass and superclass relationships. Each node also has *properties* corresponding to defining characteristics of the word. However, each node does not contain all of its properties; rather, nodes contain only the properties which distinguish them from their parents. This is to ensure minimal storage of information. A simplified version of the HNM is given in Figure 1.

We see that items are organized in a tree structure, and each node has minimal properties separating it from its parent. For example, while a canary *can fly*, this follows from the fact that a canary *is a* bird; on the other hand, a canary must *be yellow*, but a bird does not necessarily have to be. Collins and Quillian note that while this scheme is space-efficient, it increases retrieval time for properties of a word. For example, for a person to determine if a canary *has ears*, the person must move up two levels in the hierarchy to access the *has ears* property of an animal.

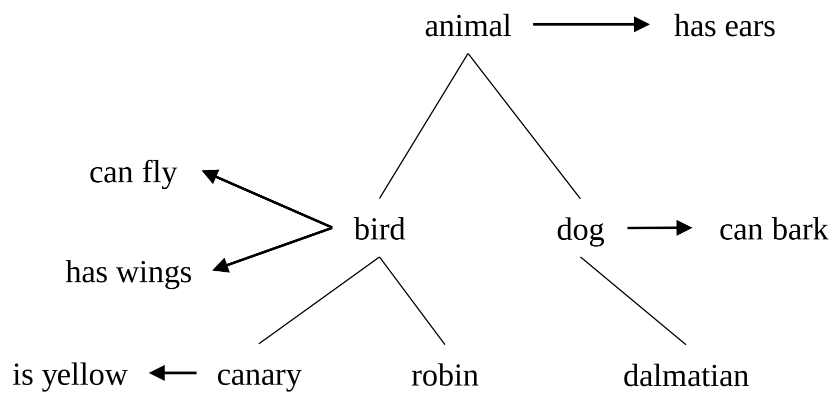


Figure 1: A sample hierarchy in the HNM
adapted from Collins and Quillian (1969)

To justify this decision, they tested human subjects, exploring the psychological reality of the proposed differences in retrieval time. Collins and Quillian (1969) set up a *semantic verification task*, wherein participants must decide whether or not a statement such as “A canary is a bird” is true as quickly as possible (in fact, psycholinguistic tasks such as this have remained a typical method of probing the structure of the mental lexicon on the assumption that linguistic performance is directly influenced by linguistic representation in the mind). According to the structure of the HNM, a sentence such as “A canary is a bird” should be judged correct more quickly than “A canary is an animal,” since the former requires accessing words only one edge apart, whereas the latter requires accessing words two edges apart. Similarly, the model predicts that listeners should react to “A canary is yellow” more quickly than to “A canary has ears,” as the former property is readily accessible when the word “canary” is retrieved. Indeed, their findings are consistent with the model’s prediction—on average, participants deemed sentences relating words closer together in the hierarchy as correct more quickly than sentences relating words farther apart.

While the HNM accounts for the experimental evidence Collins and Quillian provided, the model’s predictions fail to account for results from several subsequent studies. Loftus and Scheff (1971) carried out a study in which participants were asked to assign three superclasses to each of fifty words. Given a word such as "collie," the HNM would predict that the word would be more frequently assigned to the "dog" superclass than the "animal" superclass, since "collie" is closer to "dog" in the hierarchy than "animal." Results were consistent with this prediction; however, participants also more frequently listed "cantaloupe" as an instance of a "fruit" than as an instance of a "melon," even though "cantaloupe" would be closer to the latter than the former. This inconsistency was found in categorizing several other words, such as "chimpanzee" ("animal" over "primate") and "drum" ("musical instrument" over "percussion instrument").

In light of these findings, Smith et al. (1973; 1974) gave semantic verification tasks similar to those of Collins and Quillian to participants to see if these inconsistencies are reflected in reaction times. Indeed, they found that, on average, sentences such as "A collie is a dog" prompted faster judgment than "A collie is an animal," while sentences such as "A cantaloupe is a melon" elicited slower reaction times than "A cantaloupe is a fruit." The same general pattern was found for most of the other sentences involving sets of words in Loftus and Scheff's study that were inconsistent with the HNM.

Even more counterevidence was provided by Rosch (1975). In one of several experiments, participants were again presented with semantic verification tasks. However, instead of testing different superclasses of a word for changes in average reaction time, Rosch varied the *instances* of a superclass. In the pair of sentences "A cat is an animal" and "A dog is an animal," Collins and Quillian's model would predict that roughly equal reaction times would be elicited, since both a "dog" and a "cat" are presumably on the same level of the hierarchy compared to "animal." In this case, results showed that the model's prediction was correct. However, consider now the pair of sentences "A canary is a bird" and "A penguin is a bird." Again, the HNM predicts equal reaction times, but Rosch found a significant difference in average reaction times—participants judged a "canary" to be a "bird" more quickly than a "penguin" to be a "bird." Rosch proposes that since a "penguin" is a less typical "bird" compared to a "canary," the connection between "bird" and "penguin" is weaker than the connection between "bird" and "canary," thereby increasing processing time. Slower reaction times were also elicited for more atypical instances of a category, suggesting that lexical access is sensitive to *typicality effects*—the phenomenon in which more typical members of a category (or subclasses of a word) are more readily accessible than less typical members.

2.1.2 The Spreading Activation Model

To address these shortcomings, Collins and Loftus (1975) proposed a Spreading Activation Model (SAM). An extension to the original HNM, this model assumes a complex network of concepts connected by various types of relationships with varying degrees of strength. A sample subnetwork of the SAM is depicted in Figure 2, where the strength of the relationship between two concepts is reflected in the length of the edge connecting the concepts (shorter edges indicate stronger relationships). The strength of a relationship is determined by the number of shared properties between two concepts, where each property is itself a concept and therefore a node. Furthermore, the more frequently a concept is used by a speaker, the stronger the connections become between the concept and its properties.

In addition to a conceptual *semantic* network, Collins and Loftus also incorporated into the SAM a *lexical* network where the phonological and orthographic properties of words are stored. The nodes in the semantic network are abstract in the sense that they are nameless concepts; the actual names that speakers assign to concepts as they are pronounced and written lie in the lexical component of the SAM. The nodes in the lexical network are connected with strengths according to phonological and orthographic similarity, i.e., by how similar two words sound or look.

The name of the model derives from the proposed manner of processing over the network. When a word is "activated"—read, heard, or thought about—activation spreads to neighboring nodes, which in turn spreads to nodes neighboring those nodes, and so on. The activation is likened to a signal whose strength attenuates as it travels outward from the source node, where the decrease in strength is inversely proportional to the strength of each connection, i.e., the weaker the connection

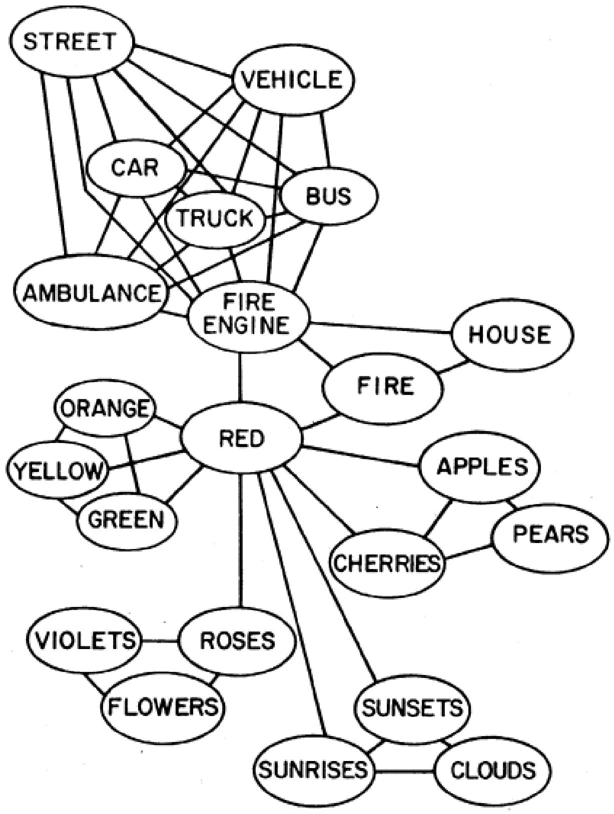


Figure 2: A sample subnetwork of the SAM reproduced from Collins and Loftus (1975)

between two nodes is, the more the signal strength weakens as it passes from the first node to the second. Thus, when a concept is retrieved, properties and other concepts that are strongly connected to the activated concept are more easily accessed.

This SAM therefore accommodates the counterevidence for the HNM discussed in the previous section. Words that are used more frequently have stronger connections, so neighboring concepts or properties (which do not necessarily have to follow a strict hierarchy) can be accessed more quickly than others. Thus, since "canary" is a more typical example of a "bird" than "penguin," the SAM now predicts that "A canary is a bird" should be responded to more quickly than "A penguin is a bird."

In addition to addressing the aforementioned issues, Collins and Loftus also use the SAM to attempt to account for the familiar "tip of the tongue" (TOT) phenomenon. Brown and McNeill (1966) first studied the effect by asking participants to recall low-frequency words given their definitions, and it was found that when the participants were not able to completely retrieve a word, they were still able to recall the number of syllables, the initial letter, and the location of the primary stress of the target word with high accuracy. This suggests that words in the mental lexicon are also partially organized by their phonological structures, an organization reflected in the lexical network component of the SAM.

Another psychological effect the SAM can provide an explanation for is *lexical priming*, in which exposure to one word influences the response towards a related word. Meyer and Schvaneveldt (1971) first studied priming effects through a *lexical decision task*, which requires a participant to determine whether or not a word is indeed a word. In their study, subjects were visually presented two words (each of which was either a real English word or a non-word which resembled an English word) and were asked to judge the validity of the word as quickly as possible. It was found that when the two words had a semantic association, participants deemed the words as real English words more quickly than when the two words had no obvious association. For example, participants displayed a significantly quicker reaction time to the pair "nurse" and "doctor" than to the pair "nurse" and "butter." This was evidence of *semantic priming*, since exposing the subject to one word decreased reaction times in responding to a semantically-related word. Meyer et al. (1974) also provided evidence for *phonological priming* in a similar lexical decision task, in which participants responded more quickly to similar-sounding word pairs such as "bribe" and "tribe" than to arbitrary word pairs such as "bribe" and "hence." In terms of the SAM, these priming effects are explained by the fact that when the participant reads the first word, the word is activated in the mental lexicon. This word presumably has strong connections with semantically- and phonologically-related words, so these related words immediately become activated, allowing the participant to more quickly access and judge the second word presented if the second word is related to the first word.

Despite the amount of psycholinguistic evidence the SAM can now account for, the model is still incomplete as it does not consider the *morphological* and *syn-*

tactic properties of words. A *morpheme* is generally defined as the smallest unit of meaning, ranging from words to prefixes and suffixes. For example, the word "submit," the prefix "re-," and the suffix "-ing" are single morphemes, and the word "resubmitting" is composed of three morphemes. Syntactic properties of a word can include its part of speech (e.g., noun, verb, or adjective) or, in other languages such as French, its grammatical gender. Fromkin (1973) argued that speech errors provide evidence for the storage of morphological and syntactic (as well as phonological and semantic) properties of words. One type of speech error is a *word exchange*, in which two words in the intended utterance are switched. An example of this is saying "Seymour sliced the *knife* with a *salami*" instead of the intended "Seymore sliced the *salami* with a *knife*." Fromkin found that when these word exchanges occurred, the switched words were generally the same parts of speech. In the previous example, the exchanged words are both nouns. This shows that the mental lexicon incorporates syntactic features to some extent. Furthermore, when the word exchange error occurred between words of different parts of speech, the words were transformed to ensure the grammatically of the utterance. An example of this is saying "I think it is *careful* to measure with *reason*" instead of the intended "I think it is *reasonable* to measure with *care*." Although the words that are exchanged are "care" and "reason," the appropriate suffixes were added and removed to produce a grammatical English sentence, demonstrating a morphological component in the mental lexicon. Fromkin also provided evidence for the storage of phonological properties through *phoneme exchanges* such as "a *d*arn *b*ore" instead of "a *b*arn *d*oor," also known as *spoonerisms*, and semantic properties through *word substitutions* such as "blond *eyes*" instead of "blond *hair*," where the substituted word is generally semantically-related to the intended word. The SAM can account for these errors, but it provides no explanation for the word exchange errors.

2.1.3 The Revised Spreading Activation Model

In response, Bock and Levelt (1994) developed the more complete Revised Spreading Activation Model (RSAM). The RSAM features three levels of nodes: the *conceptual* level, which contains words as nodes connected in a similar fashion to the SAM, the *lemma* level, which contains the syntactic information of words, and the *lexeme* level, which contains the phonological and morphological information of words. Bock and Levelt specify that in the lemma level, syntactic information can include the gender of a word, the part of speech of a word, or the *frame* of a verb—for example, the verb "put" requires a subject, a direct object, and a prepositional object, e.g., "*He* [subject] put the *money* [direct object] on the *table* [prepositional object]." A subnetwork of the RSAM is given in Figure 3.

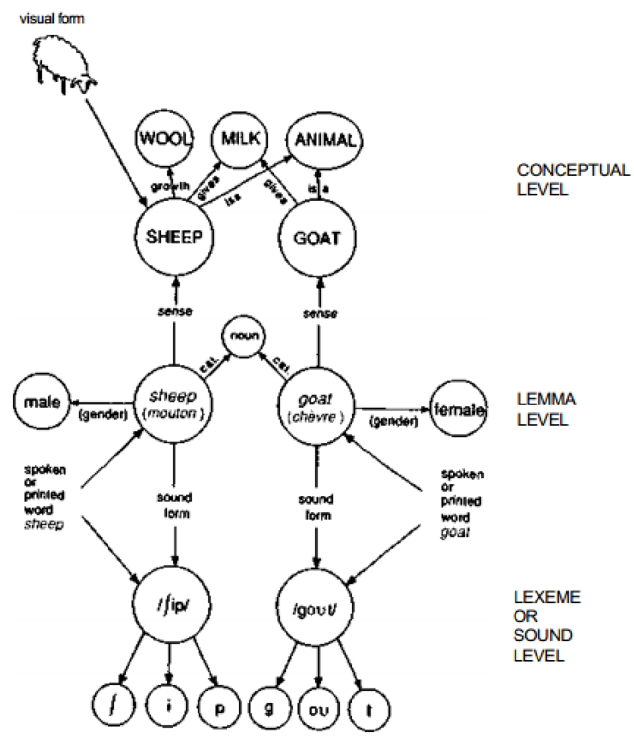


Figure 3: A sample multi-level subnetwork of the RSAM reproduced from Bock and Levelt (1994)

As the name implies, processing over the network is still through spreading activation, but the model provides three interactive levels to account for as many word features as possible. In Section 2.3, I discuss how researchers have separately explored each of these levels using graph-theoretic analyses.

2.2 Graph Theory 101

Before further discussing models of the mental lexicon as complex networks, I introduce some concepts and techniques from the field of graph theory relevant to analyzing the structures of the models.

2.2.1 Basic Definitions

A *graph* (or *network*) G is the pair $(V(G), E(G))$, where $V(G)$ is the set of *vertices* (or *nodes*), and $E(G)$ is the set of *edges* of the graph. Instead of $V(G)$ and $E(G)$, V and G will sometimes be used when the context makes clear which graph V and E refer to. The *order* of the graph G is $|V|$, the number of vertices. For two vertices u and v, we write the edge connecting u and v as (u, v), and if $(u, v) \in E$, then we say that u and v are *adjacent* and are *neighbors*. Note that, despite the ordered-pair notation, (u, v) refers to the same edge as (v, u). The *neighborhood* of v, denoted $N(v)$, is the set of all neighbors of v, and the number of elements in the neighborhood of v is the *degree* of v, $\deg(v)$. As an example, consider the graph G

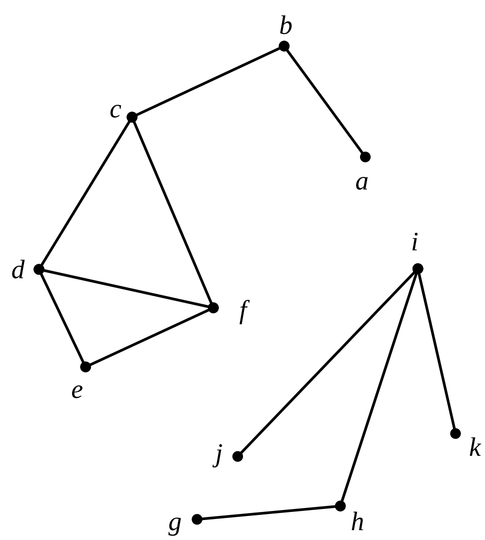


Figure 4: A sample graph G

in Figure 4. Vertices of G include a and h, while edges include (a, b) and (h, i). The vertices c and d are adjacent and are neighbors, and $N(c) = \{b, d, f\}$, so $\deg(c) = 3$.

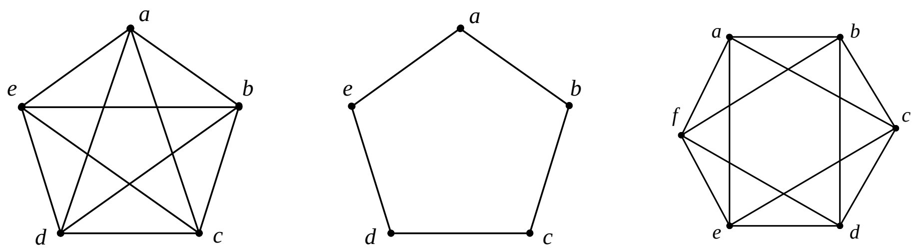
We define a *path* in a graph G as a finite sequence of distinct vertices $v_1, v_2, \ldots, v_n$

such that each $(v_i, v_{i+1}) \in E$ for $i = 1, 2, \ldots, n-1$. Note that we exclude infinite paths from the definition for the sake of simplicity. We also define the *length* of a path P as the number of edges P is composed of. A *closed path*, or a *cycle*, is a path in which the starting and ending vertices of the path are the same. The *shortest path* between two vertices u and v is then the path of minimal length with u and v as endpoints. We denote the length of the shortest path between vertices u and v as $d(u, v)$. For the sake of completeness, we let $d(u, u) = 0$ for all vertices $u \in V$. Furthermore, if there exists no path between u and v, then we say that $d(u, v) = \infty$. Referring back to Figure 4, a path between vertices f and a is f, c, b, a, containing edges $(f, c), (c, b)$, and (b, a). The length of this path is then 3, and we see that this path is also the shortest path between f and a (as opposed to the path f, d, c, b, a). The path c, d, e, f, c is a closed path of length 4.

A graph is said to be *connected* if there exists a path between every pair of vertices in the graph; otherwise, the graph is *disconnected*. A subgraph of a graph G is a graph H such that $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. A *(connected) component* of a graph is a maximally-connected subgraph of the graph, i.e., the addition of any other vertex in the graph to the component will not retain the component's connectedness. In Figure 4, G is disconnected with two components. The vertices $\{c, d, f\}$ and the edges $\{(c, d), (d, f), (c, f)\}$ make up a subgraph of G. The vertices $\{a, b, c, d, e, f\}$ together with the edges joining them form a component of G.

Some families of graphs arise so often in graph theory that they are given special names. A *complete graph (of order n)*, denoted K_n, is a graph in which every pair of vertices is adjacent. A *regular graph (of order n and common degree k)*, denoted $R_{n,k}$, is a graph in which every vertex has the same degree. A *cycle graph (of order n)*, denoted C_n, is a graph in which the n vertices are connected in a cycle. Note that $R_{n,n-1} = K_n$ and $C_n = R_{n,2}$. Examples of each of these types of graphs is given in Figure 5.

A *weighted graph* (or *weighted network*) is a graph $G = (V(G), E(G))$ in which every edge of the graph has a weight (some real number) associated with it. In this case, we write the edge between two vertices $u, v \in V$ with weight $w \in \mathbb{R}$ as



(a) K_5 (b) C_5 (c) $R_{6,4}$

Figure 5

$(u, v, w) \in E$. For a weighted graph, the length of a path between u and v can be defined as the sum of all weights associated with each edge of the path. In real-world networks, these weights can represent the distance between two cities, the cost to build a connection between two objects, the strength of a connection, etc. Figure 6 gives an example of a weighted graph G. Examples of edges in G are $(b, c, 1.24)$ and $(i, k, 1)$. Some definitions for unweighted graphs, such as the degree of a vertex or the shortest path length, have no exact parallel for weighted graphs, so alternate formulations that have been proposed in the literature will be presented as needed in the next few sections.

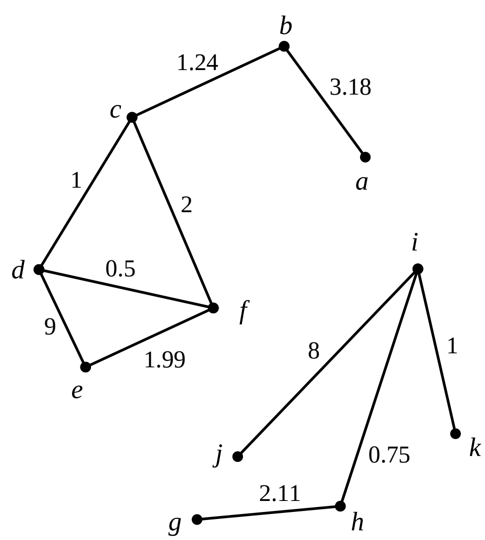


Figure 6: A sample weighted graph G

2.2.2 Graph-Theoretic Measures

There are two main measures for (unweighted) graphs from graph theory used in the discussion of the mental lexicon (and in many real-world networks): the *average shortest path length* and the *clustering coefficient*. There is also the notion of the *average degree* of a graph, denoted $\langle k \rangle$, which is simply the arithmetic mean of the degrees of every vertex, or

$$\langle k \rangle = \frac{1}{|V|} \sum_{v \in V} \deg(v).$$

However, this measure is used more as a means for calculating other measures rather than a comparative measure (discussed in the next section).

The average shortest path length L of a graph G of order n is the arithmetic mean of the lengths of the shortest paths between every ordered pair of distinct vertices in V, or

$$L = \frac{1}{n(n-1)} \sum_{u,v \in V} d(u,v).$$

Note that the average shortest path length is only reasonable for connected graphs, since if there exists no path between vertices u and v, $d(u,v) = \infty$, so for disconnected graphs, we sometimes only consider the largest connected component of the graph, i.e., the component of greatest order in the graph. Intuitively, the average shortest path length of a graph characterizes how "easy" it is to travel from one vertex to another. In Figure 7, the length of the shortest path between a and b is 1, the length of the shortest path between a and c is 2, and so on. The sum of the lengths of the shortest paths between each pair of distinct vertices is 56. Since there are $6 \cdot 5 = 30$ possible ordered pairs of distinct vertices, the average shortest path length of G is $56/30 \approx 1.87$. This means that, on average, we must travel roughly 2 edges to get from one vertex to any other vertex in G.

The *local clustering coefficient* C_v of a vertex v in a graph G is defined as the ratio of the number of edges that exist among the neighborhood of v to the number

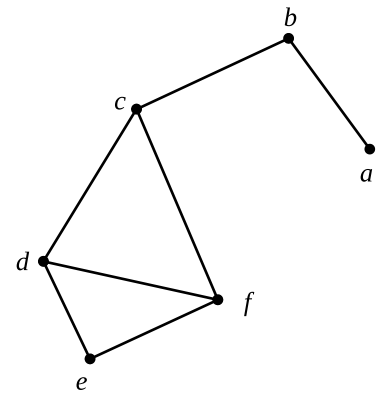


Figure 7: A sample connected graph G

of possible edges that could exist among the neighborhood of v, or

$$C_v = \frac{|\{(w,x) \in E \mid w,x \in N(v)\}|}{\binom{\deg(v)}{2}}.$$

If $\deg(v) < 2$, then we assign $C_v = 0$. This measure was first introduced by Watts and Strogatz (1998). An equivalent method of computing the local clustering coefficient of a vertex is to replace the numerator with the number of *triangles* through the vertex, since a triangle through the vertex exists when two neighbors of the vertex are adjacent. The *(average) clustering coefficient* C is simply the arithmetic mean of the local clustering coefficients for each vertex in a graph G of order n, or

$$C = \frac{1}{n} \sum_{v \in V(G)} C_v.$$

Since the denominator of C_v is the total number of possible edges, $0 \leq C \leq 1$. Generally, the clustering coefficient is only calculated for connected graphs; however, it may be extended to any graph regardless of connectivity, since isolated vertices (vertices of degree 0) simply have a local clustering coefficient of 0. Intuitively, the clustering coefficient characterizes how "tightly-knit" a graph is. In the context of social networks, this may also be thought of as answering the question, "How likely is it that a friend of a friend is also my friend?" In Figure 7, vertex d has three neighbors: c, f, and e. Among these neighbors, only the two edges (c, f) and (e, f)

exist. However, in a neighborhood of three vertices, there could potentially be up to $\binom{3}{2} = 3$ edges, so $C_d = 2/3$. We could also see that there are two triangles through d, one with vertices c, d, f and one with vertices d, e, f, so we again have $C_d = 2/3$. Performing this same calculation for each vertex in G and taking the average, we find that $C = 4/9$.

For comparison, consider K_5, the complete graph of order 5. For any vertex in K_5, every vertex in its neighborhood is adjacent to each other, so the local clustering coefficient for that vertex would be 1. Taking the average over every vertex, the clustering coefficient of K_5 is 1. On the other end of the spectrum, consider C_5, the cycle graph of order 5. Each vertex has two neighbors that are not adjacent to each other, so the local clustering coefficient of each vertex is 0, making the average clustering coefficient 0.

Several counterparts of these two measures for weighted graphs have been proposed. For the following definitions, let w_{uv} denote the weight of the edge joining vertices u and v. If u and v are not adjacent or if $u = v$, let $w_{uv} = 0$. When considering the shortest path between two vertices in weighted graphs, it is necessary to specify the significance of the "shortest" path. For example, in a weighted network where nodes represent cities and weights on edges represent distances between cities, it is natural to define the shortest path between two vertices to be the path whose edge weight sum is minimal. However, in a weighted network where the nodes represent scientists and weights on edges represent the number of papers of which both scientists are co-authors, we might imagine that the more papers they have collaborated on, the stronger the connection between the scientists. An "optimal" path between two scientists might then be one in which the sum of the edge weights is maximal—in some sense, the "strongest" path between the two scientists. In these cases when edge weights represent strengths rather than physical distances, Newman (2001a) defined the *distance* between two adjacent vertices as the (multiplicative) inverse of the weight of the edge joining the two vertices. The *weighted shortest path* between two vertices is then the path in which the sum of the *distances* is minimal. Then, if the weighted shortest path between vertices v_1

and v_n is $v_1, v_2, \ldots, v_n$, the *weighted shortest path length* $d'(v_1, v_n)$ is defined as

$$d'(v_1, v_n) = \sum_{i=1}^{n-1} \frac{1}{w_{v_i v_{i+1}}}.$$

Again, we let $d'(v, v) = 0$ for every vertex v, and $d'(u, v) = \infty$ if there does not exist a path between vertices u and v. The *weighted average shortest path length* L' in a (connected) weighted graph of order n then follows as before:

$$L' = \frac{1}{n(n-1)} \sum_{u,v \in V} d'(u, v).$$

This definition of the weighted average shortest path length has also been supported and utilized by Li et al. (2007) and a slight variant called *average efficiency* by Latora and Marchiori (2003).

The local clustering coefficient has a less natural extension to weighted graphs, so many different formulae for many different applications have been proposed. Barrat et al. (2004) were the first to propose a *weighted local clustering coefficient* (WLCC)—in the context of analyzing a scientific collaboration network and a world-wide air-transportation network—that involves the notion of the *strength* of a node, a counterpart to the degree of a node. For a weighted graph G, the strength $s(v)$ of a node $v \in V$ is defined as the sum of the weights of the edges from that node, or

$$s(v) = \sum_{u \in N(v)} w_{uv}.$$

Then the WLCC for vertex v proposed by Barrat et al., denoted $C'_{v,B}$, is given by

$$C'_{v,B} = \frac{1}{s(v)(\deg(v) - 1)} \sum_{u,x \in N(v)} \frac{w_{uv} + w_{xv}}{2} a_{ux},$$

where $a_{ux} = 1$ if vertices u and x are adjacent and 0 otherwise. This definition looks at adjacency among the neighbors of v while taking into account the weights on the edges with v as an endpoint, scaled against the total strength of v. Another version of the WLCC for a vertex comes from Onnela et al. (2005), which considers *all* of the weights in the triangles through the vertex, rather than just the weights on the edges with the vertex as an endpoint. Let $\hat{w}_{uv}$ denote the weight w_{uv} normalized by the maximum weight among all edges of the graph, i.e.,

$$\hat{w}_{uv} = \frac{w_{uv}}{\max(\{w \in \mathbb{R} \mid (u, v, w) \in E\})}.$$

Then the WLCC for a vertex v proposed by Onnela et al., denoted $C'_{v,O}$, is defined as

$$C'_{v,O} = \frac{1}{\deg(v)(\deg(v) - 1)} \sum_{u,x \in N(v)} \sqrt[3]{\hat{w}_{uv}\hat{w}_{xv}\hat{w}_{ux}}.$$

Note that while $C'_{v,B}$ uses the *arithmetic* mean, $C'_{v,O}$ uses the *geometric* mean, which, as Fleming and Wallace (1986) have shown, is an appropriate mean for normalized measurements. Onnela et al.'s WLCC was developed for the analysis of financial and metabolic networks. Zhang and Horvath (2005) formulated another WLCC for a vertex v, denoted $C'_{v,Z}$, using weights normalized by the greatest weight of the graph:

$$C'_{v,Z} = \frac{\sum\limits_{u,x \in N(v)} \hat{w}_{uv}\hat{w}_{xv}\hat{w}_{ux}}{\sum\limits_{\substack{u,x \in N(v)\\ u \neq x}} \hat{w}_{uv}\hat{w}_{xv}}.$$

The application of this WLCC was to gene co-expression networks. Finally, Holme et al. (2007) proposed their definition of a WLCC similar to $C'_{v,Z}$, but in the context of a Korean university affiliation network. Instead of normalizing each of the weights by the maximal weight of the graph, $C'_{v,H}$, the WLCC of a vertex v according to Holme et al., is defined as

$$C'_{v,H} = \frac{1}{M} \cdot \frac{\sum\limits_{u,x \in N(v)} w_{uv}w_{xv}w_{ux}}{\sum\limits_{u,x \in N(v)} w_{uv}w_{xv}},$$

where

$$M = \max(\{w \in \mathbb{R} \mid (u, v, w) \in E\}).$$

For all of these definitions, the various *(average) weighted clustering coefficients* (C'_B, C'_O, C'_Z, C'_H) of a graph would simply be the arithmetic mean of the respective WLCC of each vertex. Each of these WLCC proposals has been developed in different contexts, and no objective method for determining the "best" definition has been created. However, Saramäki et al. (2007) have summarized the different WLCCs by comparing several features of each WLCC. This comparison is reproduced in Table 8.

Feature	C'_B	C'_O	C'_Z	C'_H
Is the same as C when weights become binary	✓	✓	✓	
Is a value between 0 and 1, inclusive	✓	✓	✓	
Uses maximal weight of graph in normalization		✓	✓	✓
Considers weights of all edges in triangles		✓		✓
Invariant to weight permutations within a triangle		✓		
Considers weights of edges not in any triangles	✓		✓	✓

Table 8: Comparison of different WLCCs
reproduced from Saramäki et al. (2007)

2.2.3 Random Graphs

When studying networks, it is sometimes necessary to have benchmark calculations against which to compare measures in the network. Many researchers have turned to *random graphs* for these benchmarks, comparing their networks against a random network with the same number of vertices and approximately the same number of edges.

The birth of the study of random graphs is attributed to Erdős and Rényi (1959). They defined a random (unweighted) graph of order n to be a graph in which each of the $\binom{n}{2}$ possible edges is added with probability p and proved several properties of these graphs. In this scheme, if $p = 0$, then the graph is empty, while if $p = 1$, then a complete graph is formed. Moreover, the expected number of edges is $\binom{n}{2}p$, and the expected average degree is $(n-1)p$. As far as the average shortest path length L and clustering coefficient C of random graphs, Chung and Lu (2001) have formally shown that for most values of p, L can be approximated by $\ln(n)/\ln(np)$.

Albert and Barabási (2002) gave an informal argument for this approximation of L for random graphs, and Fernholz and Ramachandran (2007) provided a more exact approximation for random graphs with certain properties. Albert and Barabási (2002) also showed that C for a random graph in which each edge was added with probability p can be approximated by p, since the probability that two neighbors of a vertex are adjacent is exactly the probability that any two vertices are adjacent.

Watts and Strogatz (1998) presented another method of creating random (unweighted) graphs. Starting with a regular graph of order n, each edge of each vertex is "rewired," i.e., a different neighbor for the vertex is chosen, with probability p if the rewiring would not create a repeated edge. Note that p is the probability of a rewiring occurring; the new neighbor is chosen uniformly from the $n-1$ possible vertices (a vertex cannot be adjacent to itself). A reproduced example is given in Figure 9, depicting an $R_{20,4}$ graph ($p=0$) on the left, a completely rewired graph ($p=1$) on the right, and a graph generated with an intermediate value of p in the center. Watts and Strogatz used this technique to be able to study random graphs with different amounts of "randomness," where values of p closer to 0 would correspond to nearly regular graphs, while values of p closer to 1 would correspond to a nearly completely random graphs. It was then shown that as p approaches 1, the shortest average path length becomes better approximated by $\ln(n)/\ln(\langle k\rangle)$, and the clustering coefficient becomes better approximated by $\langle k\rangle/n$.

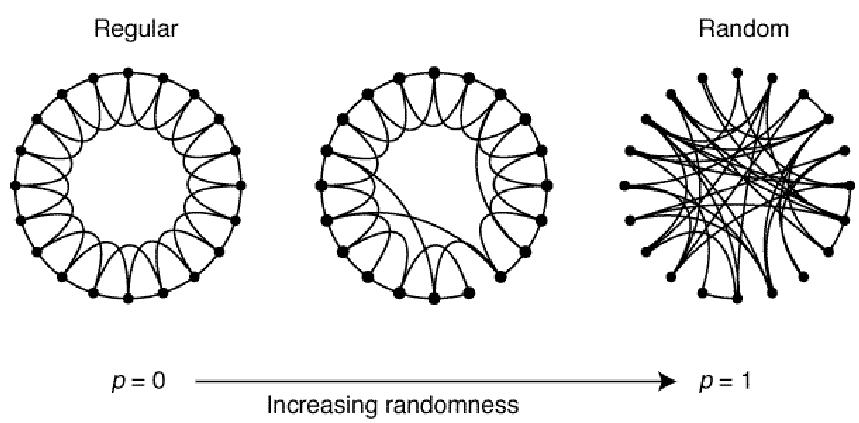


Figure 9: Examples of graphs generated using Watts and Strogatz's (1998) technique, reproduced from the same source

In Erdős and Rényi's (ER) random graphs, the expected average degree is $(n-1)p$, which, for large values of n, can be approximated to np. Then for ER random graphs, the average path length is approximately $\ln(n)/\ln(np) = \ln(n)/\ln(\langle k\rangle)$, and the clustering coefficient is approximately $p = \langle k\rangle/n$. Thus, roughly the same results have been found for these different definitions of random graphs. These formulae for L and C that do not involve p become important when comparing real-world networks to random graphs. An appropriate comparison would require the real-world network and the random graph to have approximately the same number of edges, which is roughly proportional to the average degree of a graph. The formulae allow for calculations of L and C for random graphs when the value of p is unknown but when the value of n and $\langle k\rangle$ are known. Moreover, since the value of p has different meanings in each formulation of a random graph, using n and $\langle k\rangle$ to approximate L and C keep the results consistent across models. The need for these formulae will become apparent in the next section.

While much research has been done on unweighted random graphs, little work exists on their weighted counterparts. Garlaschelli (2009) proposed a weighted random graph model in which each edge is added and assigned a weight w with probability $p^w(1-p)$. Since two non-adjacent vertices essentially have a weight of 0, the probability that any two vertices are non-adjacent is $p^0(1-p) = 1-p$, which implies that the probability that any two vertices are adjacent is $1-(1-p) = p$. Thus, Garlaschelli's weighted random graphs parallel ER random graphs. Garlaschelli also showed that when using this weighted random graph model as a benchmark against which to compare a real-world weighted network of order n and total edge weight sum W, choosing

$$p = \frac{2W}{n(n-1)+2W}$$

creates a weighted random graph that serves as an appropriate reference for the real-world network. If the *average weight* $\langle w\rangle$ of a graph—the arithmetic mean of all weights in the graph—is considered instead of W, this choice of p may be approximated as $\langle w\rangle/(\langle w\rangle + n)$, which is reminiscent of the appropriate choice of p

for ER random graphs.

Garlaschelli's model only considered integer-valued weights, and no estimation of the weighted average shortest path length or the weighted clustering coefficient for these weighted random graphs was given. Li et al. (2007) utilized an *ad hoc* method of generating a weighted random graph for comparison against real-world weighted networks. By randomly redistributing the edge weights of a real-world network, a corresponding weighted random graph is formed with the same number of vertices and average weight. No formal model was proposed, so no derivations of L' and C' were given, but L' and C' were empirically found by generating a random graph with this method and using Newman's (2001a) definition of L' and Onnela et al.'s (2005) version of C'. Although Garlaschelli's model has received more attention due to its formal description, there is no comparative source which can determine the objectively better definition of a weighted random graph.

2.2.4 Small-World Graphs

We have all more than likely experienced the *small-world effect* at one point or another when we meet someone new who happens to be a friend of a friend or when we run into someone we know at an unexpected place. Milgram (1967) was one of the first to investigate this phenomenon and set out to answer the question, "Given any two people in the world, how many intermediate acquaintance links are needed before the two people are connected?" Milgram's experiment involved giving a group of people documents asking if they knew a certain target person (from a group of target people living outside of the state). If a participant knew the target person personally, Milgram was informed; otherwise, the participant was asked to pass the message along to someone who would be more likely to know the target person, and the chain would continue until the target person was reached. From this, Milgram was able to construct a "who knows who" graph and found that, on average, it required roughly six people to reach the target person. This lends credence to the popular expression "six degrees of separation."

Watts and Strogatz (1998) extended the "small-world" concept to real-world (unweighted) networks and coined the term *small-world networks*. They analyzed three networks: the collaboration graph of actors in popular movies, the electrical power grid of the western United States, and the neural network of the nematode worm *C. elegans*. By calculating L and C for each of these graphs, it was found that these networks had high clustering and short average path lengths. As previously discussed, random graphs provide a reasonable reference against which real-world networks can be compared to find the significance of different properties. Watts and Strogatz used the approximations of L and C for random graphs to define a small-world network as one in which the average shortest path length is roughly the same as that of a random graph, but the clustering coefficient is much greater than that of a random graph. There is no standard measure for "much greater," but many researchers have used this definition of small-world networks to show its appearance in many real-world networks. Furthermore, to characterize "roughly the same" for average shortest path lengths, Watts and Strogatz also note that for regular graphs (which, in their method of random graph generation, is on the opposite end of the spectrum from a completely random graph), $L = n/(2\langle k \rangle)$. The average shortest path length of a small-world network should then be significantly closer to that of a random graph than that of a regular graph. The results of Watts and Strogatz's analysis are given in Table 10; for each network, the number of vertices, the average degree, the average shortest path length, and the clustering coefficient, alongside the estimated average shortest path length and clustering coefficient for random graphs given the values of n and $\langle k \rangle$, denoted L_{rand} and C_{rand}, respectively, as well as the average shortest path length of a regular graph with the same order and average degree, denoted L_{reg}. It can be seen, then, that each of the networks has the defined properties of a small-world network. Watts and Strogatz note that this small-world characteristic is important to a network because in dynamical systems, this structure affords enhanced signal-propagation speed and computational power.

Also listed in Table 10 are the findings of several other researchers that demonstrate small-world networks in various real-world networks. Adamic (1999) examined the World Wide Web at the site level as a network in which nodes represented

Network	**Order**	$\langle k \rangle$	L	L_{rand}	L_{reg}	C	C_{rand}
Actor collaboration (Watts and Strogatz, 1998)	225,226	61	3.65	2.99	1,846	0.79	2.7×10^{-4}
Power grid (Watts and Strogatz, 1998)	4941	2.67	18.7	12.4	925.3	0.080	0.005
C. elegans (Watts and Strogatz, 1998)	282	14	2.65	2.25	10.07	0.28	0.05
WWW (Adamic, 1999)	153,127	35.21	3.1	3.35	2,174	0.1078	2.3×10^{-4}
LANL co-authorship (Newman, 2001b)	52,909	9.7	5.9	4.79	2,727	0.43	1.8×10^{-4}
MEDLINE co-authorship (Newman, 2001b)	1,520,251	18.1	4.6	4.91	41,996	0.066	1.1×10^{-5}
SPIRES co-authorship (Newman, 2001b)	56,627	173	4.0	2.12	163.7	0.726	0.003
NCSTRL co-authorship (Newman, 2001b)	11,994	3.59	9.7	7.34	1,670	0.496	3.0×10^{-4}
Ythan Estuary food web (Montoya and Solé, 2002)	134	8.7	2.43	2.26	7.70	0.22	0.06
Silwood Park food web (Montoya and Solé, 2002)	154	4.75	3.40	3.23	16.21	0.15	0.03

Table 10: The shortest average path length L and the clustering coefficient C of various real-world networks compared to the same measures on corresponding random graphs

sites, and two nodes were adjacent if some page in one site pointed to some page in the other site. Newman (2001b) analyzed co-authorship graphs in several different scientific databases: LANL (preprints in theoretical physics), MEDLINE (published papers in biomedical research), SPIRES (published papers and preprints in high-energy physics), and NCSTRL (preprints in computer science). In these networks, nodes represented authors, and two nodes were adjacent if the two authors appeared as co-authors on a paper. Montoya and Solé (2002) studied food webs of the Ythan Estuary in Scotland and Silwood Park in England and represented them as networks in which nodes signified species and edges indicated a predator-prey relationship. In their analysis, Montoya and Solé calculated L_{rand} and C_{rand} by generating over 200 random graphs with the same average degree as the food webs and taking the average values of the average shortest path length and the clustering coefficient of those random graphs. The results from each of these studies show that small-world networks exist in a variety of contexts, from neurological to ecological to sociological. Small-world networks have also been found in models of the mental lexicon, which will be discussed in Section 2.3.

Because more attention has been given to studying small-world networks in unweighted graphs, there exist few results for weighted networks. The definition of a small-world network can easily be extended to weighted networks: a weighted small-world network should have approximately the same weighted average shortest path length and a significantly larger weighted clustering coefficient as those of a comparable weighted random graph. However, this definition encounters problems in practice: since there is no consensus on a "comparable" weighted random graph or an appropriate measure of the weighted clustering coefficient, there is currently no standard method of finding small-world networks in real-world weighted networks.

Li et al. (2007) analyzed two collaboration networks from the LANL database—subnetworks of the LANL network in Newman's (2001b) study that include only papers in astrophysics and condensed matter physics. Again, the nodes in these networks represent authors, and edges indicate that two authors have collaborated on a paper. Weights were assigned based on a method developed by Newman (2001a); the weight between two scientists i and j is given by

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1},$$

where n_k is the number of authors of paper k, and δ_i^k is 1 if scientist i was a coauthor of paper k and 0 otherwise. In this definition, the fewer authors there are on a paper, the stronger the weight between the authors, since, presumably, collaborators have a stronger relationship with each other and spend more time together when there are not as many other collaborators on a paper. As discussed in Section 2.2.3, Li et al. calculated the weighted average shortest path length L' using Newman's (2001a) method and Onnela et al.'s (2005) weighted clustering coefficient C'. They generated weighted random graphs by redistributing weights on a corresponding network of equal order and calculated the same measures using the same methods on these random graphs, respectively L'_{rand} and C'_{rand}. Table 11 gives the results of the analysis of each network, showing that both have high clustering and short average path length by the chosen definitions.

Network	**Order**	L'	L'_{rand}	C'	C'_{rand}
Astrophysics	16,706	247.84	182.54	0.016	1.0×10^{-5}
Condensed matter physics	16,726	265.12	276.98	0.017	8.6×10^{-6}

Table 11: Results from Li et al.'s (2007) analysis of weighted collaboration networks

Latora and Marchiori (2003) have also analyzed weighted neural, social, communication, and transportation networks; however, they proposed different measures for analysis over the traditional average shortest path length and clustering coefficient. Latora and Marchiori formally defined the *efficiency*, divided into *global* and *local*, and the *cost* of a network to show that these networks exhibited an *economic small-world behavior*—the networks had high global and local efficiency and low cost, whereas random networks (generated in a manner similar to Watts and Strogatz's (1998) rewiring procedure) could not achieve the same efficiency without a high cost.

2.3 Small-World Structures in the Mental Lexicon

Since small-world networks have been observed in several different fields, and since the mental lexicon has generally been accepted to be best represented by a network (as in the SAM or the RSAM), researchers have naturally attempted to find similar structures in models of the mental lexicon. Recall that the RSAM includes three inter-connected levels, where each level contains nodes connected by different types of similarity, i.e., the conceptual level by semantic similarity, the lemma level by shared syntactic properties, and the lexeme level by phonological similarity. Researchers have successfully utilized these different kinds of similarity to probe each level separately in search of small-world structures. In this section, I discuss various models of the mental lexicon that have been analyzed in the literature. Exact results for these studies are presented in Section 2.3.5.

2.3.1 Phonological Similarity

Vitevitch (2008) built and analyzed an unweighted *phonological network*; nodes represented English words (taken from the 1964 Merriam-Webster Pocket Dictionary), and two nodes were connected if they formed a *minimal pair*, i.e., the two words differed by exactly one *sound* (also called a *phoneme*). Note that this does not imply a difference of one letter between two words, as English letters do not correspond to strictly one sound, and a combination of letters can form one sound, e.g., "ch." Examples of minimal pairs are "bat" and "cat," "bat" and "chat," and "bat" and "beat"; a counterexample is "bat" and "bar." A sample phonological subnetwork is given in Figure 12. By taking the largest connected component of the phonological network, which encompassed 6,508 nodes, Vitevitch found a small-world structure with high clustering and a short average shortest path length.

Arbesman et al. (2010) also modeled the mental lexicon as a phonological network, but considered languages other than English in diverse language families: Spanish, Mandarin, Hawaiian, and Basque. The analysis of these networks differed from that of English in that, while the average shortest path length was computed only for the largest connected component, the clustering coefficient was computed for the entire network. An appropriate random network for comparing clustering coefficients then had the same order as the entire network, while an appropriate random network for comparing average shortest path lengths had the same order as the largest connected component. Furthermore, the average shortest path length was computed by taking a random sample of 1000 nodes out of the largest connected component rather than over the entire component. Through this analysis, each phonological network also exhibited small-world characteristics, despite the diversity of the languages.

2.3.2 Semantic Similarity

While minimal pairs can be determined directly given the pronunciations of words, semantic similarity is less straightforward. In the HNM, semantic similarity was

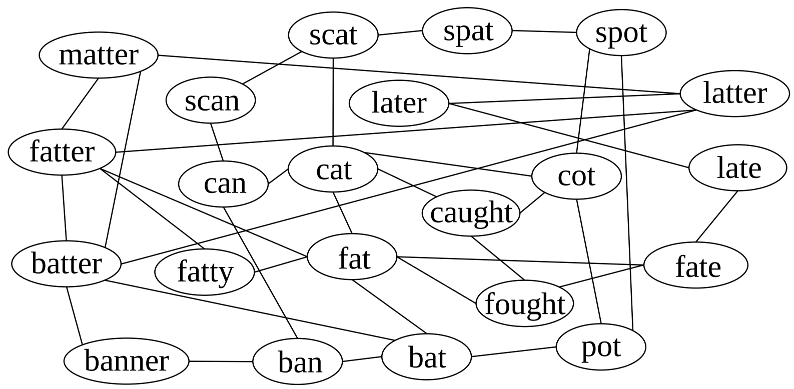


Figure 12: A sample phonological network based on Vitevitch (2008)

determined by the existence of a subset or superset relationship, while in the SAM and the RSAM, semantic similarity was determined by the number of shared properties (see Sowa (1992) for a comprehensive history of the philosophical development of semantic similarity as it applies to categorization). Miller (1995) and Fellbaum (1998) began the (English) WordNet project with the goal of building a comprehensive *semantic network*. WordNet is made of *synsets*—sets of synonymous words—as nodes. A synset generally contains several words, and a word can belong to several synsets ("bank" has several meanings). The synsets are connected by several types of relationships, but most notably: *hypernymy* and *hyponymy* (superset and subset relationships, respectively), *meronomy* and *holonymy* ("is part of" and "has part" relationships, respectively), and *antonymy* (opposite meanings).

Because of the completeness of WordNet as a semantic network, Steyvers and Tenenbaum (2005) performed their analysis on WordNet. Nodes represented words rather than synsets to provide a network of larger order, and each type of relationship (including synonymy) provided an unweighted edge between words. In their study, a comparable random graph was generated by rewiring connections in WordNet randomly, as in Watts and Strogatz (1998). The measures of L_{rand} and C_{rand} were then computed on this random graph instead of using the approximations discussed in Section 2.2.3. As with the phonological networks, an English semantic network based on WordNet was shown to be a small-world network.

In an unpublished paper (discussed in Albert and Barabási (2002)), Yook et al.

(2001) also examined a semantic network in which the edges only indicated synonymous words according to the Merriam-Webster Dictionary. The largest connected component contained only 22,311 words, compared to Steyvers and Tenenbaum's 122,005, but the network was still found to exhibit small-world characteristics. In some sense, this shows that the mental lexicon represented as a semantic network is a small-world network both locally (in Yook et al.'s subnetwork) and globally (in Steyver and Tenenbaum's whole network).

2.3.3 Similarity by Co-Occurrence

Recall that syntactic information can include verb frames, which specify the requirements for words being used with the verb to make a grammatical utterance. In some sense, the syntactic properties of a word influence how they appear together, or *co-occur* (Mel'čuk, 1988). Thus, one way to model the lemma level of the RSAM, in which syntactic properties are stored, is as a *co-occurrence network*. Cancho and Solé (2001) created an unweighted co-occurrence network by connecting words that were separated by at most one word in the British National Corpus. As an example, from the following four sentences, a small co-occurrence network can be created (Figure 13):

(a) John is tall.
(b) John drinks water.
(c) Mary is blonde.
(d) Mary drinks wine.

The final co-occurrence network of over 460,000 words was analyzed and again found to be a small-world network. Since most studies in analyzing the structure of the mental lexicon focus on models that emphasize semantic and phonological similarity, little research exists in models utilizing syntactic information.

2.3.4 Similarity by Association

In the conceptual level of the RSAM, concepts are connected by several types of relationships. While WordNet accounts for many of those relationships, it does

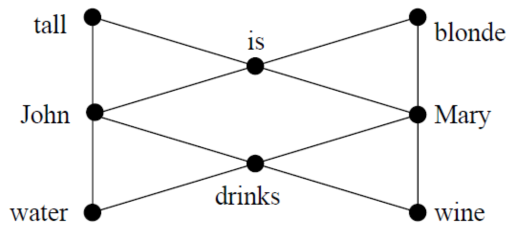


Figure 13: A simple co-occurrence network created from four sentences reproduced from Cancho and Solé (2001)

not include every type. For example, "cat" and "meow" presumably have some sort of relationship (only cats meow), but the relationship does not fall into any of WordNet's relationship types. Since the word "meow" often evokes the word "cat" in English speakers, this can be thought of as an *association.* Nelson et al. (1998) have created a database of associations between words by collecting human-produced associations for several years (amassing over 6,000 participants). Participants were presented with English words and were asked to write the first word that came to their mind—a *discrete association task.* For each presented word, words that were produced by at least two participants were recorded (with their frequencies and other statistical information) in the database as an association. Table 14 lists some sample entries from the database with the most relevant columns: *word* refers to the presented word, *association* refers to the produced word, *group size* is the number of participants presented the given word, *frequency* is the number of times the association appeared as a response, and *strength* is calculated as the frequency divided by the group size.

Word	**Association**	**Group Size**	**Frequency**	**Strength**
absence	class	152	4	0.026
blow	candle	146	7	0.048
lab	experiment	138	20	0.145
notebook	paper	147	40	0.272
skin	clear	149	2	0.013
stranger	unknown	144	21	0.146

Table 14: Sample (condensed) entries from the database of free association norms collected by Nelson et al. (1998)

Steyvers and Tenenbaum (2005) used this database to create an unweighted *associative network*, where nodes (words) were connected if one evoked the other as a response by at least two participants, i.e., if the pair of words appeared in the database as an entry. Unsurprisingly, a small-world structure was also found in the associative network. Note that similarity by association may represent almost any type of relationship. One word may evoke another freely due to various similarities. Thus, the results from this associative network may be dependent on those from the phonological, semantic, and co-occurrence networks. However, it also follows that this associative network encompasses more complex relationships that cannot be accounted for by those networks.

2.3.5 Summary

The final results of each study can be seen in Table 15. As the results indicate, a small-world structure was found in each of the networks modeling the mental lexicon using various types of similarity. Due to the availability of phonological information of words, and given the straightforward metric used to determine phonological similarity, there are more results for phonological networks in several languages. While each network displayed small-world characteristics, no study integrated the different types of similarity into one model for analysis. Furthermore, most of the studies focused on the English lexicon, though the methods used could easily be applied to a lexicon of another language, as long as there is enough information to create the network. Although weighted networks have been analyzed extensively, there is still little consensus on how to determine the existence of a weighted small-world network, so most, if not all, studies on analyzing the structure of the mental lexicon have focused on creating unweighted networks. In the present study, I attempt to fill these gaps in the literature by creating and probing the structure of such an integrated model.

Network	**Order**	$\langle k \rangle$	L	L_{rand}	L_{reg}	C	C_{rand}
English phonological (Vitevitch, 2008)	6,508	9.105	6.05	3.98	357.39	0.126	0.0014
Spanish phonological (Arbesman et al., 2010)	44,833	2.95	10.3	9.9	7,599	0.191	1.17×10^{-5}
Mandarin phonological (Arbesman et al., 2010)	19,712	3.88	10.1	7.3	2,540	0.383	8.55×10^{-5}
Hawaiian phonological (Arbesman et al., 2010)	1,406	3.44	5.5	5.8	204.4	0.241	7.40×10^{-4}
Basque phonological (Arbesman et al., 2010)	35,173	2.50	10.4	11.4	7,034	0.206	1.21×10^{-5}
English semantic (Steyvers and Tenenbaum, 2005)	122,005	1.6	10.56	10.61	38,127	0.0265	1.29×10^{-4}
English semantic (Yook et al., 2001)	22,311	13.48	4.5	3.84	827.6	0.7	6.0×10^{-4}
English association (Steyvers and Tenenbaum, 2005)	5,018	22.0	3.04	3.03	114.0	0.186	0.00435
English co-occurrence (Cancho and Solé, 2001)	460,902	70.13	2.67	3.06	3,286	0.437	1.55×10^{-4}

Table 15: The shortest average path length L and the clustering coefficient C of various models of the mental lexicon compared to the same measures on corresponding random graphs

Chapter 3

Methods

In this section, I will describe the model of the mental lexicon used in the present study and techniques used to uncover its structure in search of a small-world network.

3.1 Model Requirements

As discussed in the previous section, recent models of the mental lexicon have taken several types of similarity into account, but in separate networks. A more integrated model of the mental lexicon should incorporate several features at once (as in the RSAM). The following are important characteristics that were taken into account in the present model and that should be considered in any integrated model of the mental lexicon (for the provided reasons).

1. **The network should be weighted.** Since words can be similar to varying degrees, the mental lexicon should be represented as a weighted network, as with many real-world networks (discussed in Section 2.2.1). This is evidenced in Nelson et al.'s (1998) free association experiment, in which some words evoked a word with greater frequency among participants than other words. This suggests an association strength between words that caused some responses to appear more frequently than others. Collins and Loftus (1975) called for weighted edges in their SAM to represent the number of shared

properties between concepts. Furthermore, two previously discussed studies have noted the importance of a weighted network representation of the mental lexicon in future work (Steyvers and Tenenbaum, 2005; Vitevitch, 2008).

2. **The network should incorporate phonological similarity.** As suggested by Collins and Loftus (1975) in the SAM, Bock and Levelt (1994) in the RSAM, and Fromkin (1973), the network representing the mental lexicon should consider the phonological component of words. Indeed, many studies have shown that in behavioral experiments, participants can more easily recall or produce words that are phonologically similar or have large *phonological neighborhoods*, i.e., many phonologically-similar neighbors (Meyer et al., 1974; Luce and Pisoni, 1998; Rouibah et al., 1999; Copeland and Radvansky, 2001; Marian et al., 2008). Other studies have shown the opposite effect as well under different conditions with potential interactions with word frequency and semantic similarity (Conrad and Hull, 1964; Luce and Pisoni, 1998; Ziegler et al., 2003). This is known as the *lexical competition principle*, where listeners have more trouble recognizing words with large phonological neighborhoods because many more words are activated at once compared to words with small phonological neighborhoods. Whatever the particular conditions of the experiments that lead to facilitated or inhibited lexical access, phonological similarity is playing a role in retrieval, suggesting that encoding phonological similarity in a model of the mental lexicon is necessary.

3. **The network should incorporate semantic similarity.** Semantic similarity as a condition for edges in the mental lexicon has essentially been supported since Collins and Quillian's (1969) HNM, as they, Collins and Loftus (1975), and Bock and Levelt (1994) have all proposed connections based on semantic similarity in their respective models. Meyer and Schvaneveldt (1971) have observed semantic priming effects, where participants who were exposed to one word responded to semantically-related words more quickly than to non-related words (presented immediately after the priming word). Furthermore,

as phonological neighborhood size has been shown to influence lexical access, Buchanan et al. (2001) have also demonstrated the influence of semantic neighborhood size on retrieval through a lexical decision task—namely, that the more semantically-related neighbors a word has, the more quickly the participant can respond to the word (and can therefore access the word). In their study, semantic similarity was determined through word co-occurrence, based on the popular notion proposed by Harris (1954) known as the *distributional hypothesis*, stating that words with similar distributions (say, in a corpus) tend to have similar meanings.

4. **The network should incorporate similarity by co-occurrence.** The distributional hypothesis suggests that co-occurrence is indicative of semantic similarity, and Bock and Levelt (1994) proposed a lemma level containing the syntactic information of words (including verb frames), so co-occurrence must be incorporated into a complete model of the mental lexicon. Furthermore, Morgan (2014) has shown that expressions composed of frequently co-occurring words elicited faster responses from participants, demonstrating the existence of similarity by co-occurrence in the mental lexicon. Because co-occurrence can appear as a result of syntactic and semantic information, this requirement is not entirely orthogonal to the previous requirement. However, it is included to account for co-occurrences that do not necessarily imply a semantic relationship. For example, the expression "strong tea" is a more frequent co-occurrence than "powerful tea," even though "strong" and "powerful" are synonymous, suggesting that the relationship between "strong" and "tea" is not necessarily a semantic one.

5. **The network should incorporate similarity by association.** In some sense, this requirement can be fulfilled by satisfying requirements 2, 3, and 4, since associations can result from several factors (hence, the use of the general word "association"). However, if the measure of semantic similarity used to fulfill requirement 3 cannot account for *all* types of semantic similarity,

similarity by association can accommodate the shortcomings. For example, as discussed in Section 2.3.4, using WordNet to determine semantic similarity cannot explain the association between "cat" and "meow," as the relationship does fall into one of WordNet's categories. Furthermore, since association data has been gathered empirically through psycholinguistic experiments, similarity by association can provide a more accurate reflection of the mental lexicon.

6. **The network should incorporate orthographic similarity.** Not discussed in previous graph-theoretic analyses of the mental lexicon, *orthographic similarity*, i.e., similarity in written forms, is also an important aspect of the mental lexicon. Meyer et al. (1974) have shown in a lexical decision task that when a participant judged a word, if the participant had been exposed to an orthographically-similar word (differed by one letter in its written form) that was not phonologically similar, the participant responded more slowly and made more errors than if the prime had not been related. For example, participants had more trouble reacting to the word "break" when it was preceded by "freak" than when it was preceded by "couch." The orthographic neighborhood of a word was suggested to have an influence on lexical access because of this experiment. Van Heuven and Dijkstra (1998) have also used a lexical decision task to demonstrate that Dutch-English bilingual speakers' responses to English words with many orthographic neighbors in Dutch are inhibited. On the other hand, Ziegler et al. (2003) have found a facilitatory effect—words with more orthographic neighbors were recognized more quickly. As with the studies on phonological neighborhoods, there is competing evidence on whether orthographic neighborhood size facilitates or inhibits lexical access; in either case, it is clear that orthographic similarity affects lexical access in some way and must be included in a model of the mental lexicon.

7. **The network should incorporate word frequency information.** The frequency of a word in a language is generally considered to be its frequency in a sizable corpus of the language. Zipf (1949) posited that the phenomenon of

some words being used significantly more frequently than other words arises from the *principle of least effort*: humans are inclined to use few words frequently to efficiently reach understanding, thereby minimizing effort in processing and production. The question of whether or not humans are sensitive to these word frequencies has been studied extensively. As discussed in Section 2.1, Smith et al. (1973; 1974) and Rosch (1975) have demonstrated typicality effects in semantic verification and categorization tasks (typicality can be thought of as frequently-used words). The effects of word frequency on lexical access have also been pervasive in other psycholinguistic experiments. Many studies have shown that more frequent words are more easily recognized or accessed than less frequent words (Howes, 1957; Rosenzweig and Postman, 1957; Segui et al., 1982; Balota and Chumbley, 1984; Luce and Pisoni, 1998). Therefore, word frequency must be encoded in a complete model of the mental lexicon.

8. **The network should incorporate word concreteness information.** When children are taught new words, the words are often accompanied by visuals because imagery is thought to aid in word retention. Indeed, many studies have found that concrete words are retrieved and responded to more quickly than abstract words (Schwanenflugel and Shoben, 1983; Kroll and Merves, 1986; Bleasdale, 1987; de Groot, 1989; Paivio et al., 1994). Similar effects have been found in whole sentence processing, where concrete sentences were judged more quickly than abstract sentences (Holmes and Langford, 1976; Schwanenflugel and Shoben, 1983). For example, participants were able to deem the sentence "Armed soldiers surrounded the barracks" (many concrete words) as plausible more quickly than the sentence "Mutual distrust dominated the sessions" (many abstract words). Two popular, competing theories attempting to explain the processing advantage of concrete words over abstract words are the *dual-coding theory* proposed by Paivio (1986; 1991), which claims that concrete words are stored with a visual component in addition to linguis-

tic information (whereas abstract words lack the visual component), and the *context-availability model* developed by Bransford and McCarrell (1974) and Kieras (1978), which claims that concrete words simply have more associative links in the mental lexicon than abstract words. However, as with many competing theories, a comprehensive theory of concreteness effects likely combines elements of both theories (Jessen et al., 2000). It is therefore important to encode concreteness information into a model of the mental lexicon. While the level of concreteness or abstractness of a word can be determined by eliciting human judgments, the absence of a formal definition of concreteness as it applies to words renders the information subjective; however, human responses still provide an empirical reflection of word storage in the mind.

3.2 Creating the Model

The requirements presented in the previous section provided the basis for the present model, which was created for the English lexicon. As with every other model of the mental lexicon discussed in Section 2.3, the integrated model I propose represented words as vertices. Edge weights between two nodes were determined by taking the arithmetic mean of the following five measures:

- **Phonological Similarity.** Following the scheme Vitevitch (2008) and Arbesman et al. (2010) used in creating their phonological networks, the phonological similarity between two vertices was 1 if the two words formed a minimal pair and 0 otherwise. This definition of phonological similarity has also been used by Ziegler et al. (2003), and Luce and Large (2001) have shown that when participants were asked to produce the first real word they could think of after hearing a non-real word, a majority responded with a word that was one phoneme different from the non-real word. The pronunciation of words was found using the Carnegie Mellon University (CMU) Pronunciation Dictionary, in which English words are transcribed into phonemes. For words with more than one pronunciation (e.g., “comparable,” where the main stress can lie on "com" or on "par" depending on the speaker), if any of the pronunciations formed a minimal pair with the word being compared, the words were considered phonologically similar.

- **Semantic Similarity.** Several measures of semantic similarity have been proposed, many of which are based on the structure of WordNet (explained, summarized, and compared in Budanitsky and Hirst (2006) and Meng et al. (2013)). In this model, the Wu-Palmer (WUP) similarity metric proposed by Wu and Palmer (1994) was chosen. The motivation for the measure is the idea that more semantically-similar words are closer together in a hierarchy of concepts, as this implies that more features are shared between the words (in the HNM, features of a word are stored with the highest ancestor of the word containing those features). To formally define WUP similarity for any two concepts c_1 and c_2 in a hierarchy, denoted $\mathrm{sim}(c_1, c_2)$, let $\mathrm{dep}(c_1)$ denote the depth of c_1 in the hierarchy, and let $\mathrm{lcs}(c_1, c_2)$ denote the *least common subsumer* of the two concepts, i.e., the deepest concept in the hierarchy that is an ancestor to both concepts. Recall that $d(c_1, c_2)$ denotes the shortest path length between c_1 and c_2, which is applicable here, since a hierarchy is itself a graph. Then WUP similarity is defined as

$$\mathrm{sim}(c_1, c_2) = \frac{2 \cdot \mathrm{dep}(\mathrm{lcs}(c_1, c_2))}{d(c_1, c_2) + 2 \cdot \mathrm{dep}(\mathrm{lcs}(c_1, c_2))}.$$

  The WUP similarity metric essentially considers the distance between two concepts relative to their depths in the entire hierarchy. Since WordNet includes superclass and subclass relationships, it can be used as a hierarchy for calculating WUP similarity. As an example, consider the words "dog" and "cat." The placement of these words in WordNet can be seen in Figure 16, where there are nine concepts between "entity" (the root node) and "placental." Recall that WordNet uses nodes to represent synsets, so each word in Figure 16 is actually the representative word of a synset. The least common subsumer of "dog" and "cat" in this diagram is "carnivore," which has a depth of 12.

Furthermore, the shortest path length from "dog" to "cat" is 4 (from "dog" to "canine" to "carnivore" to "feline" to "cat"). Thus, the WUP similarity score is $2 \cdot 12/(4 + 2 \cdot 12) \approx 0.857$. In WordNet, nouns and verbs are stored in

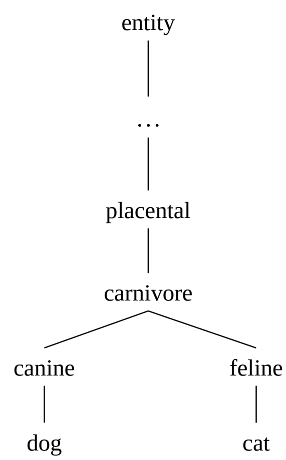


Figure 16: A subnetwork of WordNet containing "dog" and "cat"

separate hierarchies, while adjectives and adverbs are not stored hierarchically. Thus, WUP similarity is only meaningfully calculated between two nouns or two verbs. All other pairs received a semantic similarity score of 0. Future work seeks to improve the measure of semantic similarity so that all parts of speech can be considered. Furthermore, since words can belong to several synsets, the semantic similarity between two words was assigned the greatest WUP similarity score between all pairs of synsets to which each word belongs. Therefore, when considering two words with several meanings, the meanings for which the WUP similarity score is maximal are used to determine the semantic similarity.

- **Similarity by Co-Occurrence.** Reasonable co-occurrence information can be discovered from a sizable corpus in the target language. For this model, co-occurrence frequencies were computed using the Brown Corpus (Francis and Kučera, 1964), a collection of 500 texts from 15 genres, totaling to over one million words. First, all *stop words* were removed, i.e., words with little to no semantic value, such as "the" or "and," and words were *lemmatized*, i.e., converted to their base form with no inflectional endings (e.g., "words" lemmatized is "word," and "running" lemmatized is "run"). Then, frequencies of *bigrams*, ordered pairs of adjacent words, were computed. In this model, the order of each bigram does not matter, as the similarity between word v and word w is the same as the similarity between word w and word v. Therefore, for each distinct pair of words, the greater frequency was taken. For example, the frequency of the bigram "united" and "state" was greater than the frequency of the bigram "state" and "united" (since the expression "United States" is more common than the expression "states united"), so the frequency of the bigram "united" and "state" was used as the frequency count of the unordered pair "united" and "state." The frequencies were then normalized by the greatest frequency count to obtain a co-occurrence similarity score between the word pairs (so that the most frequent pair had a score of 1). Thus, for each pair of vertices, the co-occurrence similarity score was that obtained through this method if the pair indeed co-occurred in the corpus and 0 otherwise.

- **Similarity by Association.** The University of South Florida Free Association Norm Database (Nelson et al., 1998) was used to obtain association similarity scores. The strength of a word pair (as described in Section 2.3.4) was taken as the association similarity score between the pair of words. Again, in this model, words are taken as unordered pairs, so if word v evoked word w in more participants than the number of participants who responded to word w with word v, the association similarity score was assigned the strength of v eliciting w. For example, 140 participants were presented with the word "above," and 79 among them responded with "below," a strength of $79/140 \approx 0.564$. On the other hand, 73 out of 145 participants responded to "below" with "above," a strength of $73/145 \approx 0.503$. Then the association similarity score between the nodes "above" and "below" would be the greater of the two, 0.564.

- **Orthographic Similarity.** As with other types of similarity, many measures have been proposed to determine orthographic similarity, many of which have been collected and compared in Christen (2006). It was found that the *Jaro similarity* measure is most effective in a name-matching task (determining if two strings represent the same name). To some extent, the Jaro similarity measure, originally formulated by Jaro (1989), determines orthographic similarity by the number of matching characters and the number of transposed characters between two strings. Let $s_1 = c_1 \cdots c_n$ and $s_2 = d_1 \cdots d_m$ be two strings of letters $c_1, \ldots, c_n, d_1 \ldots, d_m$. Two letters c_i and d_j are considered *semi-matching* if $c_i = d_j$ and $|i - j| \leq \left\lfloor \frac{\max(n,m)}{2} \right\rfloor - 1$, i.e., the corresponding letters are within a reasonable distance from each other in the words. Two letters are considered *matching* if $i = j$ in the definition of semi-matching, i.e., the matching letters are in the same position in the words. Note that matching letters are also semi-matching letters. Let m_S be the number of semi-matching letters, and let m_M be the number of matching letters. Then the Jaro similarity between the two strings, denoted $\text{Jaro}(s_1, s_2)$, is defined as

$$\text{Jaro}(s_1, s_2) = \begin{cases} \frac{1}{3}\left(\frac{m_S}{n} + \frac{m_S}{m} + \frac{m_S + m_M}{2m_S}\right), & m_S \neq 0 \\ 0, & m_S = 0 \end{cases}.$$

Rawlinson (1976) has found that initial and final letters of words were more important for word recognition than medial letters (hence, the popular phenomenon where "tihs can siltl be raed by Egilsnh spkeaers" despite the rearrangement of medial letters). Thus, in the present model, for two words v and w (represented in their written forms), the orthographic similarity score between v and w was given by $\text{Jaro}(v, w)$ if neither the initial letter of v and the initial letter of w matched nor the final letter of v and the final letter of w matched, $\sqrt{\text{Jaro}(v, w)}$ if the corresponding initial or final letters matched (but not both), and $\sqrt[4]{\text{Jaro}(v, w)}$ if the corresponding initial and final letters matched. Since the Jaro similarity score is a value strictly between 0 and 1, taking square roots increases the score. In this scheme, "brat" and "bond" have a higher orthographic similarity score than "brat" and "from" because of the shared initial letters, despite the fact that both pairs have exactly one matching letter in the same position.

Each of the similarity measures between two words produces a value between 0 and 1, so taking the average of the similarity scores gives a value between 0 and 1 for the weight of the edge connecting the two words. While Bock and Levelt (1994) proposed a multi-level network for the mental lexicon, taking the average value of each similarity measure effectively "flattens" the network into one level, where the contribution from each level is weighted equally. In future work, it would be necessary to create and analyze a multi-level model; however, current literature lacks formalizations of small-world structures in multi-level networks.

The edge weight was further multiplied by the *concreteness values* and the *frequency values* of both words. Brysbaert et al. (2014) have collected concreteness ratings from over 4,000 participants for over 37,000 English words. Participants were asked to rate words on a 5-point scale for their concreteness, which the participants were instructed to interpret as how directly a word can be experienced by one of the senses rather than only be understood through the use of other words. The ratings were averaged across the participants for each word, and these averaged ratings provided concreteness values for words. The value was normalized to be a number between 0 and 1 by dividing the value by 5. To determine the frequency value of a word, a corpus of subtitles was used. Dave (2011) has compiled word lists in several languages with word frequencies from a large database of subtitles. The frequency value of a word was then taken to be the frequency of the word normalized by the maximal frequency among all words in the English word list (so that the most frequent word in the list had a frequency value of 1).

Concreteness values and frequency values are properties of nodes themselves rather than edges. A complete model would use these values to create node weights in addition to edge weights. In the current model, the concreteness value and frequency value of a word were instead used to strengthen the edges incident with the word to simulate node-weighting. This technique is similar to that used by

Wiedermann et al. (2013) in their analysis of multi-level networks.

To summarize, consider two words u and v in the lexical network. Let $\text{sim}_p(u,v)$ denote the phonological similarity score between v and w, $\text{sim}_s(u,v)$ their semantic similarity score, $\text{sim}_c(u,v)$ their co-occurrence similarity score, $\text{sim}_a(u,v)$ their association similarity score, and $\text{sim}_o(u,v)$ their orthographic similarity score. Furthermore, let $c(v)$ denote the concreteness value of v and $f(v)$ the frequency value of v. Then the weight of the edge joining u and v is given by

$$w_{uv} = \frac{\text{sim}_p(u,v)+\text{sim}_s(u,v)+\text{sim}_c(u,v)+\text{sim}_a(u,v)+\text{sim}_o(u,v)}{5} \cdot c(u)c(v)f(u)f(v).$$

By this definition, edge weights take values between 0 and 1, where values close to 0 indicate a weak relationship, and values close to 1 indicate a strong relationship.

3.3 Analyzing the Structure of the Model

The main hypothesis of this work is that a more integrated model of the mental lexicon exhibits small-world characteristics similar to those of previous models. To test this hypothesis on the model described in the previous section, two experiments analyzing the structure of the network were conducted. The following sections describe these experiments.

3.3.1 Experiment 1: The Weighted Lexicon

As presented in Section 2.2.4, unweighted small-world networks are characterized by a short average shortest path length L and a large clustering coefficient C. Weighted small-world networks presumably have the same properties given an appropriate measure of the weighted clustering coefficient. In this experiment, the weighted average shortest path length L' and the weighted clustering coefficient C' were computed for the proposed integrated model of the mental lexicon. Words were taken from WordNet (as *lemmas*, i.e., as uninflected words) and were removed if phonological, concreteness, or frequency information was unavailable. This resulted in a total vocabulary of 22,489 words. 1,000 words were randomly chosen as nodes (due to computational limitations), and edges were created for each pair of nodes following the method described in the previous section. This will be referred to as the *adjusted lexical network*. The value of L' was computed for the resulting largest connected component in a manner similar to Newman's (2001a): the distance between two adjacent vertices u and v was defined as $1 - w_{uv}$ rather than $1/w_{uv}$, and these values were used to find weighted shortest paths. The choice of definition for distance was made to constrain distances between 0 and 1 while retaining the property that distance increases as strength (edge weight) decreases. The value of C' was computed for the resulting largest connected component using Onnela et al.'s (2005) definition of the WLCC. To examine the impact of the frequency and concreteness values on the edge weights, a second weighted network was created from the same 1,000 nodes, where edge weights were determined without the multiplicative factor of $c(u)c(v)f(u)f(v)$ for each pair of nodes u and v. This will be referred to as the *unadjusted lexical network*.

To determine whether or not a small-world structure existed in the network, these computed values were compared against those of weighted random graphs. Two weighted random graphs of order 1,000, one against which to compare the unadjusted lexical network (no concreteness or frequency values) and one against which to compare the adjusted lexical network, were generated by instilling randomness into each feature of the integrated model. Each component of an edge weight was randomly picked for each pair of vertices. Since each similarity score took on values between 0 and 1, real numbers between 0 and 1 were randomly chosen (uniformly) for each type of similarity. The exception to this was for phonological similarity: since the score in this case was either 0 or 1, either 0 or 1 was randomly assigned to the phonological similarity score. Random values between 0 and 1 were also chosen to represent the concreteness and frequency parameters. The edge weight between two vertices was then computed using the same formula for the lexical network, but with these randomized values. The values of L'_{rand} and C'_{rand} were then found for the largest connected component of each random graph using the same definitions as in the lexical network.

3.3.2 Experiment 2: The Unweighted Lexicon

Due to the paucity of research on the defining properties of weighted small-world networks, the weighted (adjusted) lexical network from Experiment 1 was transformed into unweighted networks to analyze for small-world properties. Two transformation methods were tested:

- **Method 1.** A minimum threshold on edge weights was set. If the value of the edge weight was at least the value of the threshold, the edge remained in the network, but as an unweighted edge; otherwise, the edge was discarded. This binary decision process generated an unweighted graph as a function of the minimum threshold.

- **Method 2.** The weakest edges were incrementally removed to simulate more fine-grained threshold settings. The 500 weakest edges were removed, and the remaining edges were treated as unweighted edges, resulting in an unweighted graph. This process was repeated until no more edges could be removed or until the largest connected component was composed of a single vertex. This method was also compared against unweighted networks generated by incrementally removing a random set of 500 edges.

For each unweighted network, L and C were computed for the largest connected component. These values were compared against L_{rand} and C_{rand}, which were computed for an unweighted random network of the same order and approximately the same number of edges as the largest connected component of the unweighted lexical network using the approximations developed by Watts and Strogatz (1998) and Albert and Barabási (2002) discussed in Section 2.2.3.

Chapter 4

Results

4.1 Experiment 1 Results

The results of Experiment 1 are reported in Table 17. The integrated model and the random network contained approximately the same number of edges and had roughly the same weighted average shortest path length, but a small-world structure was not found in the lexical network. In fact, the opposite effect was found: the weighted clustering coefficient of the integrated model was significantly smaller than that of the random network. Sample edges that were created in the adjusted lexical network are given in Table 18. The first two columns list the word pair, the next five columns list the phonological, semantic, association, orthographic, and co-occurence similarity scores, the next four columns list the frequency values and the concreteness values of the two words, and the final column lists the total weight of the edge. Additionally, the 10 strongest edges and the 10 weakest edges of the adjusted and unadjusted lexical networks are listed in Table 19.

Network	**Order**	**Edges**	L'	C'
Lexical Network (Adjusted)	1,000	485,703	1.028	6.190×10^{-5}
Random Network (Adjusted)	1,000	499,500	1.000	0.049
Lexical Network (Unadjusted)	1,000	485,703	0.883	0.219
Random Network (Unadjusted)	1,000	499,500	0.547	0.481

Table 17: Calculated results from Experiment 1

Table 18: A sample of edges in the integrated model

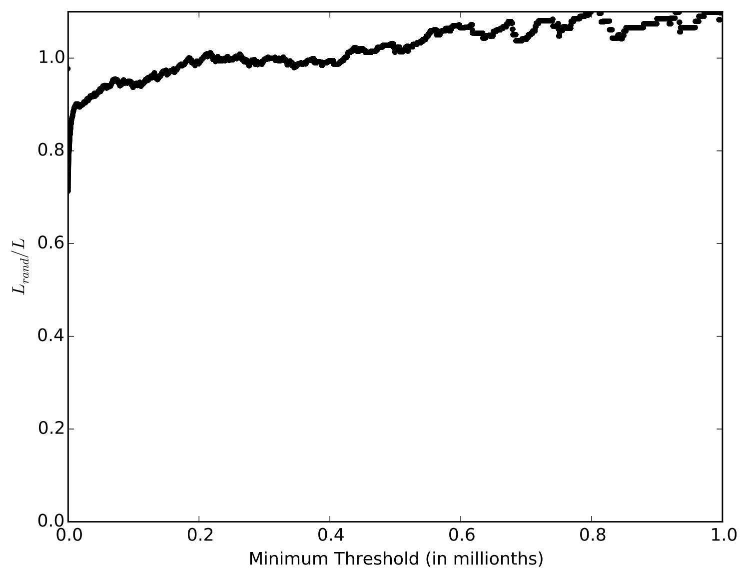
Node1	Node2	Phon	Sem	Assoc	Ortho	Co-oc	Freq1	Freq2	Conc1	Conc2	Weight
care	watch	0	0.737	0	0.483	0	0.024	0.016	0.466	0.922	4.057×10^{-5}
better	boss	0	0.857	0	0.687	0	0.039	0.008	0.382	0.766	2.778×10^{-5}
care	child	0	0.154	0.020	0.695	0.008	0.024	0.011	0.466	0.956	2.081×10^{-5}
teller	bank	0	0.421	0.814	0	0	1.919×10^{-4}	0.005	0.876	0.956	1.824×10^{-7}
flick	flip	1	1	0.026	0.885	0	1.453×10^{-4}	5.629×10^{-4}	0.742	0.794	2.806×10^{-8}
sewage	raw	0	0.267	0	0.667	0.018	6.495×10^{-5}	5.337×10^{-4}	0.904	0.670	3.994×10^{-9}
dogged	doggy	1	0	0	0.907	0	3.051×10^{-5}	2.198×10^{-4}	0.452	0.852	9.848×10^{-10}
unenforced	vacillation	0	0	0	0.397	0	3.281×10^{-7}	3.281×10^{-7}	0.35	0.484	1.447×10^{-15}

Table 19: Strongest and weakest edges in the resulting networks

Adjusted Lexical Network			**Unadjusted Lexical Network**		
Node1	**Node2**	**Weight**	**Node1**	**Node2**	**Weight**
better	watch	4.543×10^{-5}	flick	flip	0.582
care	watch	4.0572×10^{-5}	bank	tank	0.547
better	care	2.945×10^{-5}	mouse	moose	0.545
better	boss	2.778×10^{-5}	poker	poke	0.527
watch	touch	2.624×10^{-5}	gin	pin	0.524
better	child	2.236×10^{-5}	hanger	hacker	0.521
care	child	2.081×10^{-5}	filer	fighter	0.517
watch	child	1.943×10^{-5}	talker	stalker	0.515
better	news	1.782×10^{-5}	mill	male	0.509
better	touch	1.761×10^{-5}	solution	pollution	0.508
...			...		
preferentially	nonstandard	1.402×10^{-15}	bobcat	shuffle	0.0143
aggrieve	indestructibility	1.358×10^{-15}	hijacking	moose	0.0138
draftsmanship	unenforced	1.323×10^{-15}	dopey	bashful	0.0066
preferentially	draftsmanship	1.258×10^{-15}	curse	shaky	5.089×10^{-4}
nonstandard	aggrieve	1.184×10^{-15}	dealer	optimistic	5.089×10^{-4}
aggrieve	unenforced	1.152×10^{-15}	highly	fat	5.089×10^{-4}
aggrieve	vacillation	1.107×10^{-15}	vital	touch	5.089×10^{-4}
preferentially	unenforced	1.090×10^{-15}	frigid	except	5.089×10^{-4}
draftsmanship	aggrieve	9.758×10^{-16}	expect	affair	5.089×10^{-4}
preferentially	aggrieve	8.076×10^{-16}	whatever	fun	5.089×10^{-4}

4.2 Experiment 2 Results

For Method 1, based on the results from Experiment 1, 50,000 evenly-spaced minimum threshold values between 1.0×10^{-16} and 1.0×10^{-5} were chosen to generate unweighted networks from the adjusted lexical network. The values of C_{rand}/C and L_{rand}/L were plotted against the threshold values for each of the 50,000 resulting unweighted networks. Note that in an unweighted small-world network, we would expect to find C_{rand}/C close to 0 (since C should be significantly greater than C_{rand}) and L_{rand}/L close to 1 (since L and L_{rand} should be approximately equal). These plots are given in Figure 20.



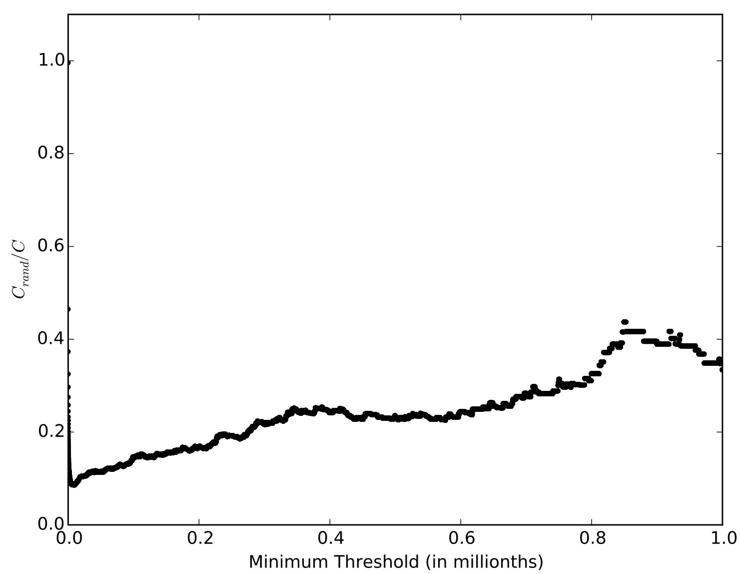
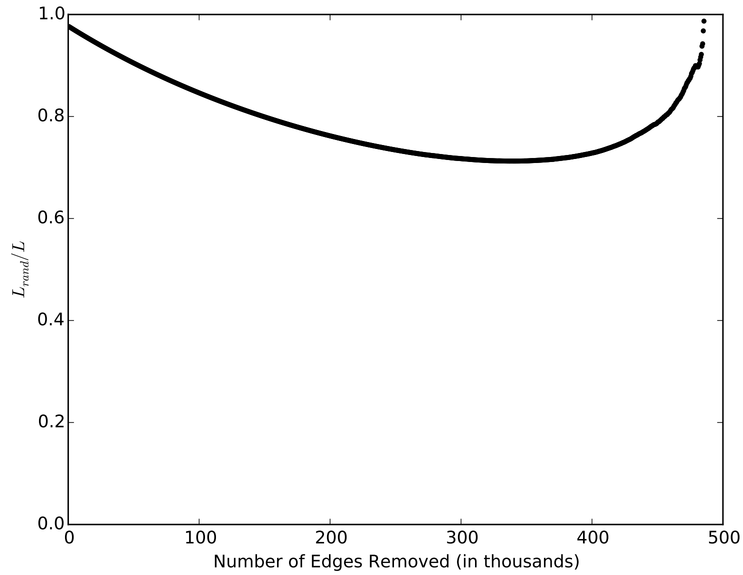


Figure 20: Plots of L_{rand}/L (left) and C_{rand}/C (right) against minimum threshold values using Method 1

For Method 2, the values of C_{rand}/C and L_{rand}/L were again plotted, but against the number of edges removed. Figure 21 shows the plot as each set of 500 weakest edges is removed, and Figure 22 shows the plot as each set of 500 random edges is removed.



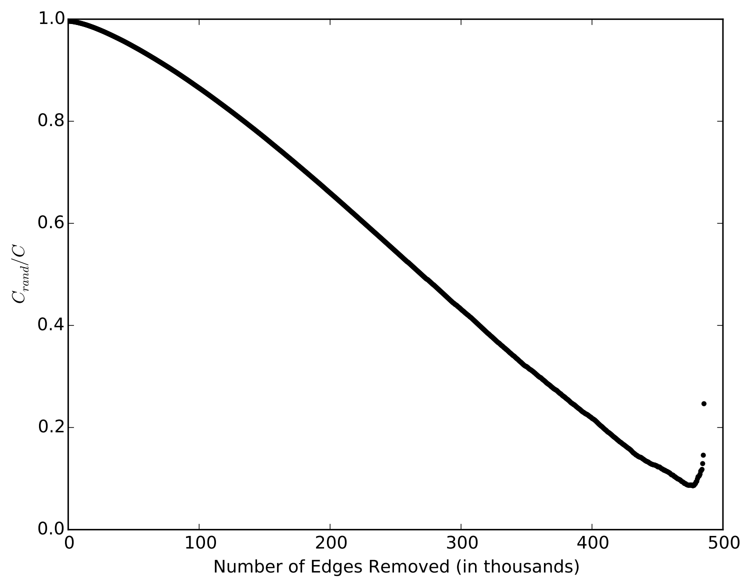
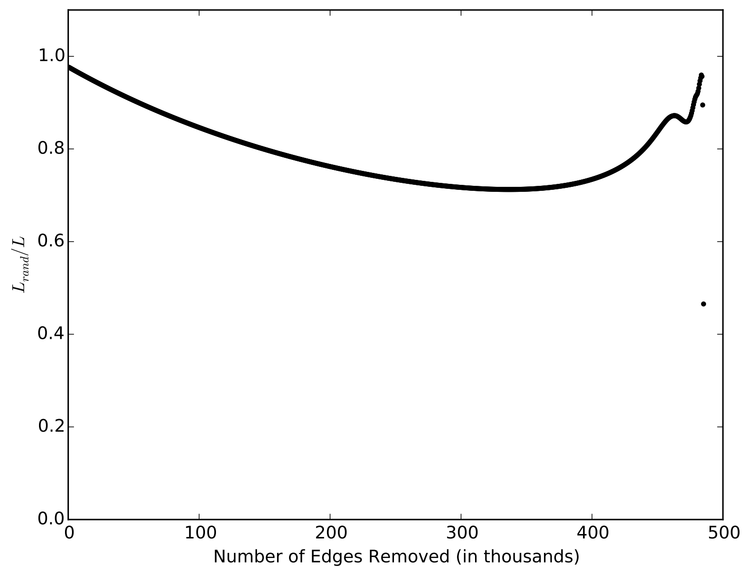


Figure 21: Plots of L_{rand}/L (left) and C_{rand}/C (right) as weakest edges are removed using Method 2



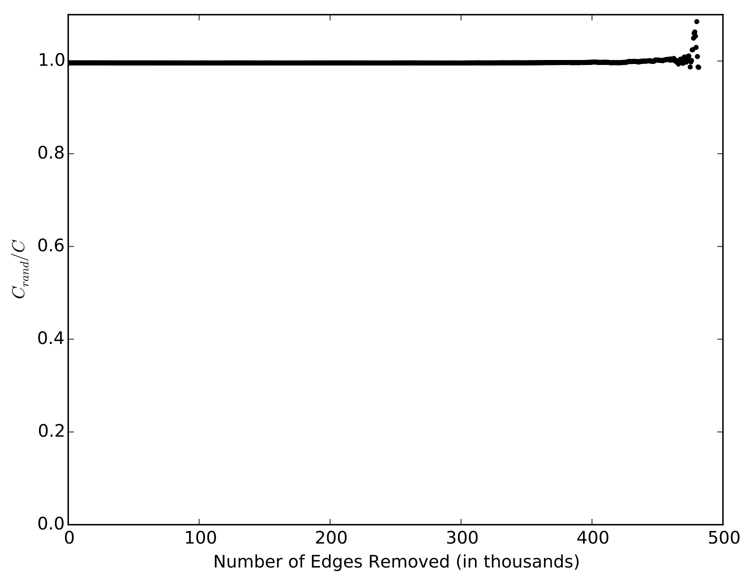


Figure 22: Plots of L_{rand}/L (left) and C_{rand}/C (right) as random edges are removed using Method 2

Chapter 5

Discussion

5.1 Analysis of Results

The results from Experiment 1 show that, using the discussed methods of measuring L' and C', the integrated model of the mental lexicon proposed here does not exhibit small-world behaviors. Several sources of error are possible. While the definition of L' is more intuitive, the definition of C' is less so; the reversed outcome between the weighted clustering coefficients may be due to an inappropriate definition of C' in this context. A different method for calculating C' may produce different results. Another explanation could be the method of weighted random graph generation. As previously mentioned, no standard method of producing a weighted random graph serving as an appropriate comparison for a real-world network exists. Furthermore, the definition of a weighted small-world network was extended directly from the definition of an unweighted small-world network; there may be unforeseen and unstudied nuances introduced by extending the definition to weighted networks.

Other sources of error may be in the integrated model itself. With 7 nodes, there can be a maximum of $\binom{7}{2} = 21$ edges among them. The fact that only 7 words appeared in the 10 strongest edges (nearly half of the possible number of edges between 7 nodes) of the adjusted lexical network suggests that the frequency value or the concreteness value may be exerting too strong of an influence compared to the other measures incorporated in the edge weights. The same can be said about the 10 weakest edges. If this is the case, then the frequency value or the concreteness value is strongly promoting a connection that would otherwise be weak. However, this effect may be desirable, as it has been shown that word frequency and familiarity play a significant role in word recognition and recall (discussed in requirement 7 in Section 3.1). Nonetheless, there is the possibility that the technique of simulating node weight by strengthening its edges by some factor is flawed. This is further evidenced by the word pairs in the 10 strongest edges in the unadjusted lexical network, which contain a variety of words. However, in this network, words which were minimal pairs were pushed to the top, suggesting that the phonological similarity score may be masking the other similarity scores. Taking the arithmetic mean of the various similarity scores is a naïve approach that does not consider the possibility that the different types of similarity may actually be interacting with unbalanced roles, e.g., similarity by association may be more important in lexical storage than phonological similarity, or the possibility that each similarity score, though a value between 0 and 1, has skewed distributions and must be adjusted accordingly. Regardless, it has been shown that different types of similarity indeed interact. Rouibah et al. (1999) demonstrated that in a task requiring semantic judgments, a phonological priming effect can be found, and in a task requiring phonological judgments, a semantic priming effect can be found, suggesting that both phonological and semantic information were accessed, independent of the target task.

The individual methods of determining similarity can also be brought to question, as many measures exist for each type of similarity. While the other measures take a real value between 0 and 1, phonological similarity takes on one of two values. A more fine-grained score is necessary, since two words may sound similar without necessarily being minimal pairs (e.g., “pass” and “bath”). Minimal pairs were used in this work because defining phonological similarity beyond the notion of minimal pairs becomes complex. It may not be enough to consider phoneme deletions, insertions, or mutations; a pair of phonemes may sound more similar than another pair of phonemes (e.g., /b/ and /p/ vs. /b/ and /e/). More information must be collected about phoneme similarity or more complex algorithms must be developed and used (see Kondrak (2000) for such an algorithm) to obtain a more accurate and fine-grained phonological similarity score.

Other methods of determining semantic similarity exist; however, many of the methods based in WordNet use the notion of *information content* of a word, which is a measure dependent on the frequency of the word in a corpus. These measures were not considered here to minimize the amount of overlap between scores—in this case, between the semantic similarity score and the frequency value (and to some extent, the co-occurrence similarity score, which was also based on word frequency in a corpus). Though an effort was made to keep similarity scores orthogonal (so that no score was dependent on another score), the method for determining one type of similarity may not be defined well enough to account for every component of the type of similarity. For example, any semantic similarity measure based in WordNet will not be able to account for every human intuition of semantic similarity—this is where similarity by association becomes important. However, currently, association similarity scores can only be obtained empirically through elicitation tasks, which can be subjective.

Another shortcoming in testing this model is that nodes represented words solely in their written forms. This raises two important questions: What should a node in the mental lexicon represent? And how should multiple meanings and pronunciations be stored? In Bock and Levelt's (1994) RSAM, nodes on each level could represent different units of language, from phonemes to morphemes to whole words. In the present model, the multi-level approach was consolidated by collapsing the levels into a single-level network. By choosing to use a node to represent the written form of a word, the phonological, syntactic, and semantic information of the word were all stored in the single node representing the word. This means that nodes were distinguished only orthographically, so words with multiple meanings or pronunciations did not receive their due distribution in the network; all of the meanings and pronunciations of the word were located within the node itself. For example, in Table 18, the words "better" and "boss" had a high semantic similarity score. This is due to the fact that "better," though generally thought of as meaning "of higher quality," can mean "a superior." Since a "boss" is a "superior," this interpretation of "better" is closely semantically-related to "boss." Ideally, these different meanings of "better" would be stored separately so that the less common meaning is not used in determining the association or co-occurrence similarity scores. However, many of the studies in Section 2.3.5 successfully chose the written forms of words to be nodes in their networks (Cancho and Solé, 2001; Steyvers and Tenenbaum, 2005; Vitevitch, 2008; Arbesman et al., 2010), but none mentioned how words with multiple meanings or pronunciations were handled.

Furthermore, there have been mixed results about whether every possible meaning of a word is activated when the word itself is activated or only the correct meaning of the word in context is activated. For example, when participants heard a sentence that began with "Jack tried the punch . . . ," reaction times to the words "hit" and "drink" were both decreased (Gernsbacher and Faust, 1991). This suggests that all meanings of a word are activated with the word. On the other hand, when participants were presented with a sequence of three words, if the first and third words were related to the same meaning of the second word ("save," "bank," and "money"), the third word was recognized more quickly than when the third word was unrelated to the second (Schvaneveldt and Meyer, 1976). Moreover, if the first and third words were related to different meanings of the second word ("river," "bank," and "money"), the third word was not recognized significantly more quickly or more slowly than to a word unrelated to the second. This suggests that the context of the first word selected the appropriate meaning of the second word without activating the other meanings of the second word. The former case would be supported by a network in which every meaning of a word is stored with the word in a node, so the activation of a node would activate the properties within the node. The latter case would be supported by a network in which different meanings of a word are stored in different nodes, so context can guide the activation of one meaning over another separately. Future studies should consider the implications each model has on the network structure.

In contrast to Experiment 1, Experiment 2 showed promise for the integrated model. Based on the results, we see that setting minimum threshold values generated small-world graphs. In Figure 20, when extremely weak edges have been removed, the resulting unweighted networks have average shortest path lengths close to those of random graphs and clustering coefficients greater than those of random graphs. While a rising trend can be observed in both plots, the trend does no harm to the analysis because minimum thresholds greater than 1.0×10^{-5} led to extremely disconnected graphs in which the largest connected component had very few vertices, if not a single vertex. Furthermore, from Tables 10 and 15, ratios of C_{rand} to C for previously-studied small-world networks ranged from 5.87×10^{-5} to 0.273, so the ratios found in Figure 20 are appropriate indicators that the generated unweighted lexical networks were small-world.

The second method tested in Experiment 2 also suggested small-world behaviors in the resulting unweighted networks. As weakest edges were removed, the unweighted counterparts became more and more small-world, as shown in Figure 21. The average shortest path length remained relatively close to the average shortest path length of a comparable random graph, and the clustering coefficient constantly surpassed that of a random graph as edges disappeared. Figure 22 assures that this trend does not arise even when edges are removed randomly; the clustering coefficient remained nearly equal to that of a random graph as edges were removed. This implies that weak edge removal, whether performed by setting a threshold or incrementally, can uncover small-world properties in the lexical network, suggesting that when a weighted model of the mental lexicon is used, extremely weak edges may need to be disregarded.

Setting a minimum threshold for edge weights in this model of the mental lexicon mimics *threshold potentials* in neurons of a neural network, a threshold that excitation in a neuron must surpass in order to fire (Rushton, 1927; Hodgkin and Huxley, 1952). Removing weak connections can then be likened to activation that fails to spread from one node to another due to poor signal strength. The unweighted counterpart of the integrated model retains connections that have a greater potential for activation, which may be important in applying the model to computational models of language processing.

5.2 Applications

Models of speech perception, speech production, and lexical access are all intimately related to models of the mental lexicon. Indeed, a popular model of speech perception called the TRACE Model is based on an interactive network similar to the RSAM (McClelland and Elman, 1986). An analysis of lexical networks as in the present study can help support the foundation for these models of speech processing and understanding by providing evidence for techniques that can both accommodate previously-studied psycholinguistic phenomena and afford the model efficient computational ability (as discussed by Watts and Strogatz (1998) for small-world networks).

5.3 Future Work

As discussed in the analysis, there are many future directions to be taken. Each level of the model should be analyzed separately to see if one level plays a greater role than another and to see if a multi-level approach would produce similar or better results. A possible method for teasing out the influence exerted by each level is to isolate pairs of words with a single type of similarity and present them to participants to empirically determine the usefulness of each similarity in processing. Another method would be to reverse engineer the weights by again determining empirical similarity between words through controlled experiments.

Future directions also include exploring different methods of calculating similarity at each level. Latent semantic analysis has been used extensively and effectively to uncover underlying relationships between words based on the principle that words used in similar contexts tend to have similar meanings (Landauer and Dumais, 1997). The linear algebra-based technique has been successful in determining synonyms given a large amount of text in the absence of a thesaurus or other source indicating word relations (Landauer et al., 1998a) and several other linguistically-based tasks requiring automated word-similarity detection (Soboroff et al., 1997; Gordon and Dumais, 1998; Landauer et al., 1998b; Gong and Liu, 2001). This method may provide a more accurate measure of semantic similarity than the WordNet-based technique in the present study.

The problem of the lack of weighted small-world network analyses in the literature is also a topic for future work. Testing different formulations of the weighted clustering coefficient and different methods of generating comparable weighted random graphs may produce more enlightening results. Furthermore, the present study calculated the weighted average shortest path length by taking $1-w$ for each weight w instead of Newman's $1/w$. Though the average shortest path length comparisons showed roughly equal results in the experiments, using the multiplicative inverse may demonstrate different trends and should be investigated.

This study examined a subset of the English lexicon. A larger set of words should naturally be tested, but computational complexity increases dramatically as more words are chosen to be included in the network. Moreover, there is the question of the generalizability of these requirements and this integrated model to other languages. A cross-linguistic analysis would be an interesting future direction, as well as applying the model to a bilingual lexicon.

5.4 Conclusion

In this study, I proposed a set of requirements for an integrated model of the mental lexicon as a weighted network and built such a model for analyzing in search of small-world structures, a ubiquitous property in many real-world networks. While this integrated model of the mental lexicon failed to show small-world characteristics as a weighted network, this may have been due to the design choices of the model or the lack of current understanding of weighted small-world networks. Although the tested model fulfilled the requirements in Section 3.1, there are still many other ways to satisfy the requirements. Furthermore, this list of requirements may not be exhaustive, but each requirement is important in a complete model of the mental lexicon, as evidenced by numerous psycholinguistic studies. By transforming the weighted lexical network into an unweighted network by removing weak connections, representing connections that are least likely to be activated in processing over the network, the desired small-world properties were found, suggesting that the underlying structure of this integrated model can provide the same computational benefits as other small-world networks and is indeed a step towards a complete model of the mental lexicon.

Bibliography

[1] Adamic, L. A. The small world web. *Lecture Notes in Computer Science 1696* (1999), 443–452.

[2] Albert, R., and Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics 74* (2002), 47–96.

[3] Arbesman, S., Strogatz, S. H., and Vitevitch, M. S. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos 20*, 3 (2010), 679–685.

[4] Balota, D. A., and Chumbley, J. I. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance 10*, 3 (1984), 340–357.

[5] Barrat, A., Barthélémy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. *Proc. of the National Academy of Sciences of the U.S.A. 101*, 11 (2004), 3747–3752.

[6] Bleasdale, F. A. Concreteness-dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition 13*, 4 (1987), 582–594.

[7] Bock, K., and Levelt, W. Language production: Grammatical encoding. In *Handbook of Psycholinguistics*, M. A. Gernsbacher, Ed. Academic Press, San Diego, CA, 1994, pp. 945–984.

[8] Bransford, J. D., and McCarrell, N. S. A sketch of a cognitive approach to comprehension: Some thoughts on what it means to comprehend. In *Cognition and the Symbolic Processes*, W. Weimer and D. Palermo, Eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 1974, pp. 189–230.

[9] Brown, R., and McNeill, D. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior 101*, 11 (2004), 3747–3752.

[10] Buchanan, L., Westbury, C., and Burgess, C. Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review 8*, 3 (2001), 531–544.

[11] Budanitsky, A., and Hirst, G. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics 32*, 1 (2006), 13–47.

[12] Christen, P. A comparison of personal name matching: Techniques and practical issues. Tech. rep., The Australian National University, 2006.

[13] Chung, F., and Lu, L. The diameter of sparse random graphs. *Advances in Applied Mathematics 26* (2001), 257–279.

[14] Collins, A. M., and Loftus, E. F. A spreading-activation theory of semantic processing. *Psychological Review 82*, 6 (1975), 407–428.

[15] Collins, A. M., and Quillian, M. R. Retrieval times from semantic memory. *Journal of Verbal Learning and Verbal Behavior 8* (1969), 240–247.

[16] Conrad, R., and Hull, A. J. Information, acoustic confusion, and memory span. *British Journal of Psychology 55* (1964), 429–432.

[17] Copeland, D. E., and Radvansky, G. A. Phonological similarity in working memory. *Memory & Cognition 29*, 5 (2001), 774–776.

[18] Dave, H. Frequency word lists, 2011. `https://invokeit.wordpress.com/frequency-word-lists/`.

[19] de Groot, A. M. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition 15*, 5 (1989), 824–845.

[20] Erdős, P., and Rényi, A. On random graphs I. *Publicationes Mathematicae 6* (1959), 290–297.

[21] Fellbaum, C. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[22] Fernholz, D., and Ramachandran, V. The diameter of sparse random graphs. *Random Structures and Algorithms (RSA) 31*, 4 (2007), 482–516.

[23] Fleming, P. J., and Wallace, J. J. How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM 29*, 3 (1986), 218–221.

[24] Francis, W. N., and Kučera, H. A standard corpus of present-day edited American English, for use with digital computers. Brown University, Providence, RI, 1964. `http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/`.

[25] Fromkin, V. A. Slips of the tongue. *Scientific American* (1973), 181–187.

[26] Garlaschelli, D. The weighted random graph model. *New Journal of Physics 11* (2009).

[27] Gernsbacher, M. A., and Faust, M. The role of suppression in sentence comprehension. In *Understanding Word and Sentence*, G. B. Simpson, Ed. Elsevier Science Publishers B. V., North Holland, 1991, ch. 5, pp. 97–128.

[28] Gong, Y., and Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, 2001), ACM, pp. 19–25.

[29] Gordon, M. D., and Dumais, S. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science 49*, 8 (1998), 674–685.

[30] Harris, Z. S. Distributional structure. *Word 10* (1954), 146–162.

[31] Holme, P., Park, S. M., Kim, B. J., and Edling, C. R. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A 373*, 821 (2007).

[32] Holmes, V. M., and Langford, J. Comprehension and recall of abstract and concrete sentences. *Journal of Verbal Learning and Verbal Behavior 15* (1976), 559–566.

[33] Howes, D. H. On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America 29* (1957), 296–305.

[34] i Cancho, R. F., and Solé, R. V. The small world of human language. *Proceedings of the Royal Society of London B 268*, 1482 (2001), 2261–2265.

[35] Jaro, M. A. Advances in record linkage methodology as applied to the 1985 census of Tampa, Florida. *Journal of the American Statistical Association 84*, 406 (1989), 414–420.

[36] Jessen, F., Heun, R., Erb, M., Granath, D.-O., Klose, U., Papassotiropoulos, A., and Grodd, W. The concreteness effect: Evidence for dual coding and context availability. *Brain and Language 74*, 1 (2000), 103–112.

[37] Kieras, D. Beyond pictures and words: Alternative information processing models for imagery effects in verbal memory. *Psychological Bulletin 85* (1978), 532–554.

[38] Kondrak, G. A new algorithm for the alignment of phonetic sequences. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics* (2000), 288–295.

[39] Kroll, J. F., and Merves, J. S. Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition 12*, 1 (1986), 92–107.

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

[41] Landauer, T., Foltz, P., and Laham, D. Introduction to latent semantic analysis. *Discourse Processes 25* (1998a), 259–284.

[42] Landauer, T., Laham, D., and Foltz, P. Computer-based grading of the conceptual content of essays. Unpublished manuscript, 1998b.

[43] Latora, V., and Marchiori, M. Economic small-world behavior in weighted networks. *European Physical Journal B – Condensed Matter 32* (2003), 249–265.

[44] Li, W., Lin, Y., and Liu, Y. The structure of weighted small-world networks. *Physica A: Statistical Mechanics and its Applications 376* (2007), 708–718.

[45] Loftus, E. F., and Scheff, R. W. Categorization norms for 50 representative instances. *Journal of Experimental Psychology 91*, 2 (1971), 355–364.

[46] Luce, P. A., and Large, N. R. Phonotactics, density and entropy in spoken word recognition. *Language and Cognitive Processes 16* (2001), 565–581.

[47] Luce, P. A., and Pisoni, D. B. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing 19*, 1 (1998), 1–36.

[48] Marian, V., Blumenfeld, H. K., and Boukrina, O. V. Sensitivity to phonological similarity within and across languages. *Journal of Psycholinguistic Research 37*, 3 (2008), 141–170.

[49] McClelland, J. L., and Elman, J. L. The TRACE model of speech perception. *Cognitive Psychology 18* (1986), 1–86.

[50] Mel'čuk, I. *Dependency Syntax: Theory and Practice.* State University of New York, Albany, NY, 1988.

[51] Meng, L., Huang, R., and Gu, J. A review of semantic similarity measures in WordNet. *International Journal of Hybrid Information Technology 6*, 1 (2013), 565–581.

[52] Meyer, D. E., and Schvaneveldt, R. W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology 90*, 2 (1971), 227–234.

[53] Meyer, D. E., Schvaneveldt, R. W., and Ruddy, M. G. Functions of graphemic and phonemic codes in visual word-recognition. *Memory & Cognition 2*, 2 (1974), 309–321.

[54] Milgram, S. The small-world problem. *Psychology Today 1* (1967), 61–67.

[55] Miller, G. A. WordNet: A lexical database for English. *Communications of the ACM 38*, 11 (1995), 39–41.

[56] Montoya, J. M., and Solé, R. V. Small world patterns in food webs. *Journal of Theoretical Biology 214*, 3 (2002), 405–412.

[57] Morgan, J. A. Exploring the psycholinguistic validity of extended collocations. Master's thesis, Portland State University, 2014.

[58] Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. The University of South Florida word association, rhyme, and word fragment norms, 1998. `http://www.usf.edu/FreeAssociation/`.

[59] Newman, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E 64* (2001a).

[60] Newman, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences 98*, 2 (2001b), 404–409.

[61] Onnela, J.-P., Saramäki, J., Kertész, J., and Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Physical Review E 71* (2005).

[62] Paivio, A. *Mental Representations: A Dual Coding Approach.* Oxford University Press, Oxford, 1986.

[63] Paivio, A. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology 45* (1991), 255–287.

[64] Paivio, A., Walsh, M., and Bons, T. Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition 20*, 5 (1994), 1196–1204.

[65] Quillian, M. R. The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM 12*, 8 (1969), 459–476.

[66] Rawlinson, G. E. *The significance of letter position in word recognition.* PhD thesis, University of Nottingham, 1976.

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

[68] Rosch, E. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General 104* (1975), 192–233.

[69] Rosenzweig, M. R., and Postman, L. Intelligibility as a function of frequency of usage. *Journal of Experimental Psychology 54* (1957), 412–422.

[70] Rouibah, A., Tiberghien, G., and Lupker, S. J. Phonological and semantic priming: Evidence for task-independent effects. *Memory & Cognition 27*, 3 (1999), 422–437.

[71] Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K., and Kertész, J. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E 75* (2007).

[72] Schvaneveldt, R. W., and Meyer, D. E. Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance 2*, 2 (1976), 243–256.

[73] Schwanenflugel, P. J., and Shoben, E. J. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition 9*, 1 (1983), 82–102.

[74] Segui, J., Mehler, J., Frauenfelder, U., and Morton, J. The word frequency effect and lexical access. *Neuropsychologia 20*, 6 (1982), 615–627.

[75] Smith, E. E., Shoben, E. J., and Rips, L. J. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review 81*, 3 (1974), 214–241.

[76] Soboroff, I. M., Nicholas, C. K., Kukla, J. M., and Ebert, D. S. Visualizing document authorship using n-grams and latent semantic indexing. In *Proceedings of the 1997 Workshop on New Paradigms in Information Visualization and Manipulation* (New York, NY, 1997), ACM, pp. 43–48.

[77] Sowa, J. F. Semantic networks. In *Encyclopedia of Artificial Intelligence*, S. C. Shapiro, Ed., second ed. Wiley, New York, 1992.

[78] van Heuven, W. J. B., and Dijkstra, T. Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language 39* (1998), 458–483.

[79] Vitevitch, M. S. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research 51* (2008), 408–422.

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

[81] Wiedermann, M., Donges, J. F., Heitzig, J., and Kurths, J. Node-weighted interacting network measures improve the representation of real-world complex systems. *Europhysics Letters 102* (2013).

[82] Wu, Z., and Palmer, M. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (Las Cruces, NM, 1994), pp. 133–138.

[83] Yook, S., Jeong, H., and Barabási, A.-L. Unpublished, 2001.

[84] Zhang, B., and Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology 4*, 17 (2005).

[85] Ziegler, J. C., Muneaux, M., and Grainger, J. Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language 48* (2003), 779–793.

[86] Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, 1949.