Complex Network Analysis of Distributional Semantic Models

Akira Utsumi (utsumi@inf.uec.ac.jp)
Department of Informatics, The University of Electro-Communications
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Abstract

A number of studies on network analysis have found the small-world and scale-free properties in the network of free word association, which reflects human semantic knowledge. Nevertheless, there have been very few attempts to apply network analysis to distributional semantic models (DSMs), despite the fact that DSMs have been extensively studied as a model of human semantic knowledge. In this paper, therefore, we analyze the small-world and scale-free properties of DSM networks. We demonstrate that DSM networks exhibit the same properties as the word association network. Especially, we show that DSM networks have the distribution of the number of connections that follows the truncated power law, which is also observed in the association network. This result indicates that DSMs provide a plausible model of semantic knowledge. Furthermore, we propose a modified version of Steyvers and Tenenbaum's (2005) growing network model, which involves the processes of semantic differentiation and experiential correlation. This model can better explain different distributions generated by various DSM implementations.

Keywords: Semantic network; Distributional semantic model; Truncated power-law distributions; Network models

Introduction

Recently, network analysis or network science has attracted considerable attention in cognitive science (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). The network properties revealed through the analysis of complex cognitive phenomena tell us about the behavior of the underlying cognitive processes, and even simple network models can provide valuable insights into the cognitive mechanisms. In particular, a large number of network studies have investigated language-related phenomena (Borge-Holthoefer & Arenas, 2010), among which the most studied one is free word association (De Deyne & Storms, 2008; Nelson, McEvoy, & Schreiber, 2004). This is because free word association reflects our lexical knowledge acquired through world experience, and thus reveals the structure of human semantic memory or mental lexicon.

Network studies on word association have demonstrated the small-world and scale-free properties of semantic network (De Deyne & Storms, 2008; Morais, Olsson, & Schooler, 2013; Steyvers & Tenenbaum, 2005). For an association network where each word is represented by a node and an association relation between two words is represented by an edge joining the corresponding nodes, the small-world property indicates that any two word nodes are connected by traversing only a few edges, although the network is highly structured locally. The scale-free property indicates that most word nodes are poorly connected, while a relatively small number of words are highly connected; as a result, the distribution of the number of connections for each node follows a power law. All the existing studies agree on the small-world property of the association network, but some studies (Morais et al., 2013) suggest that the association network is not completely scale-free; rather the network is characterized by a power law truncated by an exponential cutoff, where the most connected words have a smaller connection than would be expected in a purely power-law distributed network.

These network properties are expected to reveal the cognitive mechanism underlying the semantic structure of language (Borge-Holthoefer & Arenas, 2010). For example, the small-world structure sheds light on an efficient search process in semantic memory. Investigating various network models that generate scale-free networks provides valuable insight about psychological processes involved in lexical development.

In contrast to the growing network-analytic interest in word association, a distributional semantic model (henceforth, DSM) has rarely been investigated in network analysis, despite the fact that DSMs have been extensively studied as a valid model of semantic memory (Landauer, McNamara, Dennis, & Kintsch, 2007). In a DSM, the lexical meaning of a word is represented by a high-dimensional vector, and the degree of semantic relatedness between any two words can be easily computed from their vectors. Word vectors are constructed from large bodies of text (i.e., corpus) by observing distributional statistics of word occurrence. Steyvers and Tenenbaum (2005) tested whether latent semantic analysis (LSA), one of the most popular versions of DSMs, exhibits the same network structure as word association, and found that LSA networks were small-world, but not scale-free. They concluded from this result that LSA is limited as a model of human semantic knowledge. However, their finding does not imply that DSMs in general fail to model semantic memory, because a variety of methods for constructing semantic spaces other than LSA are devised in the DSM framework (e.g., Turney & Pantel, 2010). Furthermore, as Morais et al. (2013) pointed out, their analysis of the scale-free property was quite subjective in that their claim of power-law behavior was derived solely from the observation of the behavior of distribution without any statistical tests.

In this paper, we analyze the small-world and scale-free properties of semantic networks constructed from various DSMs in a more systematic way, and examine the conditions under which DSM networks have the same properties as the association network. Through this network analysis, we test whether a DSM can provide a psychologically plausible model of semantic memory. In addition, we discuss a cognitive mechanism for lexical development by proposing a network model that simulates the behavior of DSM networks.

Complex Network Analysis

We first define some terminology from graph theory used in this paper. A network (or graph) G is defined by a pair (V,E) comprising a set V of nodes or vertices and a set E of edges that connect a pair of nodes. The number of nodes $|V|$ is denoted by n, and the number of edges $|E|$ is denoted by m. An edge is directed or undirected, and a graph containing only directed (undirected) edges is said to be directed (undirected). Two nodes are neighbors if they are connected by an edge. The degree k_i of a node v_i is the number of edges that connect to it (i.e., the number of neighbors). The average degree over all nodes is denoted by $\langle k \rangle$. In a directed network, the degree of a node v_i has two types, namely in-degree k_i^{in} and out-degree k_i^{out}, which respectively refer to the numbers of edges incoming to or outgoing from v_i.

A path is a sequence of edges that connects one node to

another. The path length is the number of edges along that path. An undirected graph is connected when there is a path between every pair of nodes. A directed graph is strongly connected when there is a path in both directions between every pair of nodes. A (strongly) connected component of a graph is a subgraph that is (strongly) connected, but no longer connected when any other node in the graph is added. The number of nodes of the largest (strongly) connected component is denoted by n_{CC}. Note that, in this paper, we restrict the following analyses to the largest connected component.

Small-world networks can be characterized by high local clustering and short path lengths between nodes. On the one hand, small-world networks have clusters of nodes that are densely connected to each other. On the other hand, two nodes that belong to different (and distant) clusters are also connected by only a few edges. These two features can be quantitatively measured by the clustering coefficient C and the average shortest path length L. The clustering coefficient C is defined as $C = \sum_{v_i \in V} C(v_i)/|V|$ where $C(v_i) = T_i / \binom{k_i}{2}$. In this definition, T_i denotes the number of edges that exist between neighbors of a node v_i, and thus $C(v_i)$ represents the probability that two neighbors of v_i are connected by an edge. As a result, the clustering coefficient C represents the probability that two neighbors of a randomly chosen node are themselves neighbors. On the other hand, L is defined as the average of shortest path length over all (ordered) pairs of distinct nodes in an undirected (directed) network.

More formally, small-world networks are defined in terms of how they differ from the random networks with the same type of edges, number of nodes, and number of edges. Let C_{random} and L_{random} be the clustering coefficient and the average path length of the corresponding random network. A network G is said to be small-world if $C \gg C_{random}$ and $L \geqslant L_{random}$ (Watts & Strogatz, 1998). It means that small-world networks are highly clustered, unlike random networks, yet they have small path length, like random networks.

The scale-free property can be characterized by a broad, heavy-tailed degree distribution that follows the power law $P(k) \sim k^{-\alpha}$. Many real networks such as WWW, citations of scientific papers, and food web have been found to be scale-free, and the exponent α of these distributions usually ranges between 2 and 3. Amaral et al. (2000) also demonstrated that many real systems that are not scale-free can be grouped into two additional classes: a broad-scale network, characterized by a degree distribution that has a power-law regime followed by a sharp cutoff, i.e., $P(k) \sim k^{-\alpha} \mathrm{e}^{-\lambda k}$, and a single-scale network, characterized by a degree distribution with a fast decaying tail that follows the exponential $P(k) \sim \mathrm{e}^{-\lambda k}$.

The degree distribution of a given network can be examined by constructing a binned histogram or plotting a cumulative degree distribution on a logarithmic scale. If a degree distribution follows the power law, the plotted distribution shows a straight-line behavior, whether cumulative or not. Hence, the easiest way to evaluate whether a given degree distribution follows the power law is to observe the shape of the plotted distribution. However, this method is subjective, especially when we test whether a distribution follows the power law. Therefore, we not only observe the plotted data, but also apply Clauset, Shalizi, and Newman's (2009) statistical framework for testing the goodness-of-fit between the data and the power law, and whether the power law is a more plausible model than alternative distributions. In this framework, the power law is fitted to the data (i.e., plotted distribution) and the scaling parameter α is estimated using maximum likelihood estimation (henceforth, MLE). This fitting procedure assumes a lower bound k_{min} to the power-law behavior, for an empirical reason that most naturally occurring distributions only follow a power-law distribution above some lower bound. The lower bound k_{min} is identified by minimizing the Kolmogorov-Smirnov distance D between the data and the theoretical power-law fit. In this paper, to avoid a biased estimate resulting from throwing away many legitimate data points, we estimate the optimal k_{min} within the range of $k_{min} \leq 50$. The goodness-of-fit test is conducted by empirically estimating the probability p that D for the observed data is smaller than that for the synthetic data randomly drawn from the power-law distribution that best fits the observed data. In other words, p denotes the probability of obtaining the observed data under the null hypothesis that the observed data follows the estimated power-law model. If p is small, the null hypothesis is rejected. Clauset et al. (2009) suggest that the power law is ruled out when $p \leq 0.1$, and we also use this criterion. The estimated power-law model is compared with alternative models using a likelihood-ratio test. As suggested by Clauset et al. (2009), model selection by the likelihood ratio is significant when $p < 0.1$.

Table 1: Statistics for the semantic network ($n = 5,018$) built from the USF association norms

	m	n_{CC}	$\langle k \rangle$	L	L_{random}	C	C_{random}
Directed	63,620	4,845	12.7	4.26	3.64	0.187	0.005
Undirected	55,236	5,018	22.0	3.04	3.03	0.187	0.005

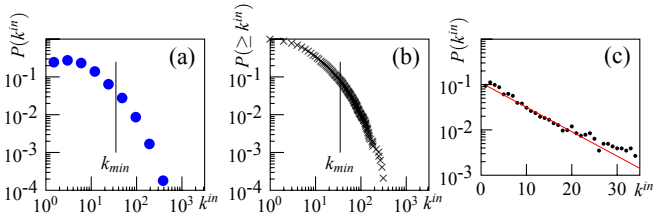


Figure 1: In-degree and cumulative in-degree distributions of the USF association network

Analysis of Word Association Network

Method

We used the English free association norms collected at the University of South Florida (Nelson et al., 2004), which was also used in previous studies on association networks. Following these studies, we constructed a directed network as follows. First, only cue words were represented as nodes (i.e., words that appeared only as an associate were not considered). Second, two word nodes x and y were connected by a directed edge from x to y, if the word y was listed as an associate of the cue x by at least two of the participants of the association experiment. We also generated an undirected network by replacing directed edges with undirected ones.

Result

Table 1 shows the network statistics for the USF association network. The association network has a small-world structure because $C \gg C_{random}$ and $L \geqslant L_{random}$. This result is completely consistent with the existing findings on the analysis of association networks (De Deyne & Storms, 2008; Morais et al., 2013; Steyvers & Tenenbaum, 2005).

Figure 1 (a) and (b) plot the in-degree distribution of the directed association network and its cumulative distribution. [1] These graphs show that the distributions deviate from the pure power law, as argued by Morais et al. (2013). Indeed, the goodness-of-fit test for the best-fit power-law model

($\alpha = 2.91,\ k_{min} = 35$) ruled out the possibility of the pure power law, $D = 0.048$, $p = .01$. Furthermore, likelihood-ratio tests indicate that the truncated power law (i.e., the power law with an exponential cutoff) is a significantly better fit to the observed distribution than the pure power-law form, $\log LR = -8.36$, $p < .001$. (Although the pure power law is favored over the exponential, it is not significant, $\log LR = 2.06$, $p = .80$.) These results are completely consistent with Morais et al.'s (2013) finding. Note that the power-law exponent α was estimated at 2.91 in this paper and 2.92 in Morais et al. (2013), but these values are higher than those of the other studies claiming the pure power-law fit, i.e., $\alpha = 1.79$ (Steyvers & Tenenbaum, 2005), or 2.13 (De Deyne & Storms, 2008). Interestingly, our estimate of α for the truncated power-law distribution is 1.78, which is close to their estimates of pure power law. This also seems to suggest that the truncated power law may better describe the in-degree distribution in the USF association network.

We also applied the fitting procedure to the observed in-degree distribution below $k_{min} (= 35)$. This analysis is motivated by the existing finding that some semantic networks exhibit the power law with the initial exponential decay (Motter, de Moura, Lai, & Dasgupta, 2002). The result is that the exponential distribution with $\lambda = 0.124$ better fits the data than the pure power law ($\log LR = -421.9$, $p < .0001$) and the truncated power law ($\log LR = -389.9$, $p < .0001$). Figure 1 (c) indeed shows that this exponential fit appears to be the case; $P(k)$ decreases roughly linearly with the degree on the semilogarithmic (i.e., log-linear) scale and the slope of the red line equals to $\lambda \log e$.

Analysis of DSM Networks

Method

To compare DSM networks directly with the USF association network, we used only the cue words of the USF association norms when creating DSM networks. As a corpus for DSMs, we used the written and non-fiction parts of the British National Corpus, which contained 491,106 documents, 73,422 distinct words, and 4,702 cue words.

We created a semantic network from a given semantic space by first computing the cosine similarity between any pairs of words and then determining local neighborhoods using the cosine similarity. Local neighborhoods were determined by two methods, namely the *k*-nn method and the *r*-method. The *k*-nn method was used in Steyvers and Tenenbaum (2005), while the *r*-method is devised in this study.[2] Both methods create a directed edge from each word to its nearest neighbors. They differ in the way of determining the number of nearest neighbors for each word. In the *k*-nn method, the number of neighbors for a word w_i is set to the number of associates of that word in the USF association norms. In the *r*-method, the number of neighbors for a word w_i is determined to be the largest $|N|$ such that $\sum_{w_j \in N} cos(w_i, w_j) / \sum_{w_j \in V} cos(w_i, w_j) \leq r$ where N is a set of nearest neighbors of w_i and V is a set of all words except w_i. The threshold r is determined so that the created DSM network has the same $\langle k \rangle$ as in the directed association network.

In the DSM framework, semantic spaces are constructed by the following three steps (Turney & Pantel, 2010).

1. Initial matrix construction: A word-context frequency matrix $\boldsymbol{A}$ with n_w rows for words and n_c columns for contexts are constructed. The *i*-th row corresponds to the word vector for the *i*-th word w_i.

2. Weighting: The elements of the matrix $\boldsymbol{A}$ are weighted.

3. Smoothing: The dimension n_c of the row vectors of $\boldsymbol{A}$ is reduced to n_r.

The notion of context in Step 1 can be generally classified into two types: "documents as contexts" and "words as contexts." For a documents-as-contexts (or word-document) matrix, an element a_{ij} of $\boldsymbol{A}$ is the frequency of a word w_i in a document d_j. For a words-as-contexts (or word-word) matrix, its element a_{ij} is the cooccurrence frequency of two words w_i and w_j within a certain range such as a window of some words. In this paper, we used both types of context, and a context window of size 2 (i.e., two words on either side of the target word) for a word-word matrix. In Step 2, we employed two popular weighting methods: tf-idf and ppmi. In the tf-idf method, the weight is calculated by the product of the local weight based on the term frequency and the global weight based on the inverse document frequency or entropy. In this paper, we used the product of the logarithm of the word frequency and the entropy (Utsumi, 2011). In the ppmi method, the weight is calculated by pointwise mutual information and negative values are replaced with zero (Bullinaria & Levy, 2007). In Step 3, matrix smoothing was conducted using singular value decomposition (SVD). In this paper, we set $n_r = 300$, which is used in typical applications of LSA. Note that LSA corresponds to the combination of a word-document matrix, tf-idf weighting, and SVD smoothing.

As a result, we obtained 24 DSM networks from all possible combinations of two methods for determining neighborhoods (*k*-nn or *r*), two initial matrices (word-document or word-word), three weighting options (tf-idf, ppmi, or unweighted), and two smoothing options (SVD or unsmoothed).

Table 2: Statistics for some representative examples of the DSM networks ($n = 4,702$)

	m	n_{CC}	$\langle k \rangle$	L	L_{random}	C	C_{random}
Word-document matrix, unweighted, unsmoothed, *k*-nn method							
Directed	60,262	4,519	12.8	4.84	3.59	0.222	0.006
Undirected	49,274	4,702	21.0	2.94	3.06	0.228	0.005
Word-document matrix, tf-idf, smoothed, *r*-method							
Directed	59,613	4,156	12.6	5.83	3.58	0.317	0.006
Undirected	48,622	4,702	20.7	3.88	3.07	0.308	0.005
Word-word matrix, unweighted, unsmoothed, *r*-method							
Directed	59,621	3,091	12.8	8.31	3.45	0.366	0.008
Undirected	55,520	4,702	23.6	3.05	2.95	0.335	0.005
Word-word matrix, ppmi, smoothed, *r*-method							
Directed	59,613	4,474	12.7	5.77	3.60	0.251	0.006
Undirected	48,504	4,702	20.6	3.76	3.07	0.242	0.005

Result

Small-world property Table 2 shows the network statistics for some representative examples of DSM networks. These results clearly indicate that the DSM networks have the small-world structure, i.e., high clustering coefficient, small shortest path length, and high connectivity. Although we do not show

[1] In this paper, we address only the in-degree distribution, because using the out-degree or the degree of the undirected network would introduce a bias, which comes from the task characteristics such as the number of associations (De Deyne & Storms, 2008).

[2] Steyvers and Tenenbaum (2005) also used the ε-method, in which local neighborhoods are computed by thresholding the cosine similarity; any pair of words whose cosine value is equal to or higher than a threshold ε is connected by an undirected edge. However, the symmetric nature of this method is not appropriate for modeling the semantic knowledge underlying word association. In human word association, a word x is an associate of a cue word y does not imply that y is an associate of x, but the ε-method cannot capture such the difference. Therefore, we did not use the ε-method in this paper.

Table 3: Summary of statistical testing for power-law behavior in the in-degree distributions of directed DSM networks

	Unsmoothed			Smoothed		
matrix / method	raw	tf-idf	ppmi	raw	tf-idf	ppmi
word-document / *k*-nn	+∗+	+∗+	+∗+	+∗∗	+∗∗	+∗∗
word-document / *r*	+∗+	+∗+	+∗+	+∗∗	+∗∗	+−∗
word-word / *k*-nn	+−+	−−∗	−−∗	+−∗	+∗∗	+−∗
word-word / *r*	+−+	−−∗	+∗∗	+−∗	+∗∗	+∗∗

Note. A three-symbol code used in this table is defined as follows. The left symbol denotes the result of the goodness-of-fit test for the power law, the middle one denotes the result of the likelihood ratio test for the truncated power law over the power law, and the left one denotes the result of the likelihood ratio test for the exponential over the power law. The symbol + denotes that the power law fits to the data (left) or the power law is favored over the alternative (right). The symbol − denotes that the power law is ruled out (left) or the alternative is favored over the power law (middle or right). The symbol ∗ denotes no significant preference between the power law and the alternative. Red and green cells respectively denote that the in-degree distribution follows the pure power law and the truncated power law.

the statistics of all the 24 DSM networks, other networks also have the same small-world structure.

Scale-free property In this section, we first discuss the result of Clauset et al.'s (2009) statistical tests for all the DSM networks, and then examine the in-degree and cumulative in-degree distributions when necessary.

Table 3 shows the results of the goodness-of-fit test for the best-fit power-law model and the likelihood-ratio tests for the truncated power law and the exponential over the power law. The code +∗+ definitely indicates that a pure power-law distribution is most appropriate, and +∗∗ also indicates that a pure power law is likely to be most appropriate. On the other hand, the codes −−∗, −−+, +−+ and +−∗ indicate that a truncated power-law degree distribution is most appropriate, while the codes −∗− and +∗− indicate that a exponential distribution is most appropriate. Other possible codes show that the most appropriate distribution cannot be determined by this test. Note that the test result for the USF association network is coded as −−∗, which favors a truncated power law.

Overall, Table 3 shows that all the DSM networks exhibit a power-law distribution (denoted by red cells) or a truncated power-law distribution (denoted by green cells). This result provides direct evidence against Steyvers and Tenenbaum's (2005) argument that LSA networks do not produce the power-law distribution and thus LSA cannot provide a plausible model of semantic memory. A DSM has an ability to reproduce semantic networks with the same properties as the association network, and therefore it can provide a psychologically plausible framework for modeling human semantic memory.

A more detailed analysis of the result of the statistical tests and the in-degree distributions reveals how the parameters for constructing semantic spaces affect the properties of DSM networks, which are summarized as follows.

1. Unsmoothed word-word matrix: The DSM networks derived from the unsmoothed word-word matrices typically exhibit the truncated power-law degree distribution, which is similar to the distribution of the USF association network.

2. Smoothed word-word matrix: Some of the smoothed word-word-based DSMs yield a scale-free network whose degree distribution follows the pure power law, while other DSMs still yield the truncated power-law distribution.

3. Unsmoothed word-document matrix: The DSM networks constructed from the unsmoothed word-document matrices exhibit neither a power-law distribution nor its variants.

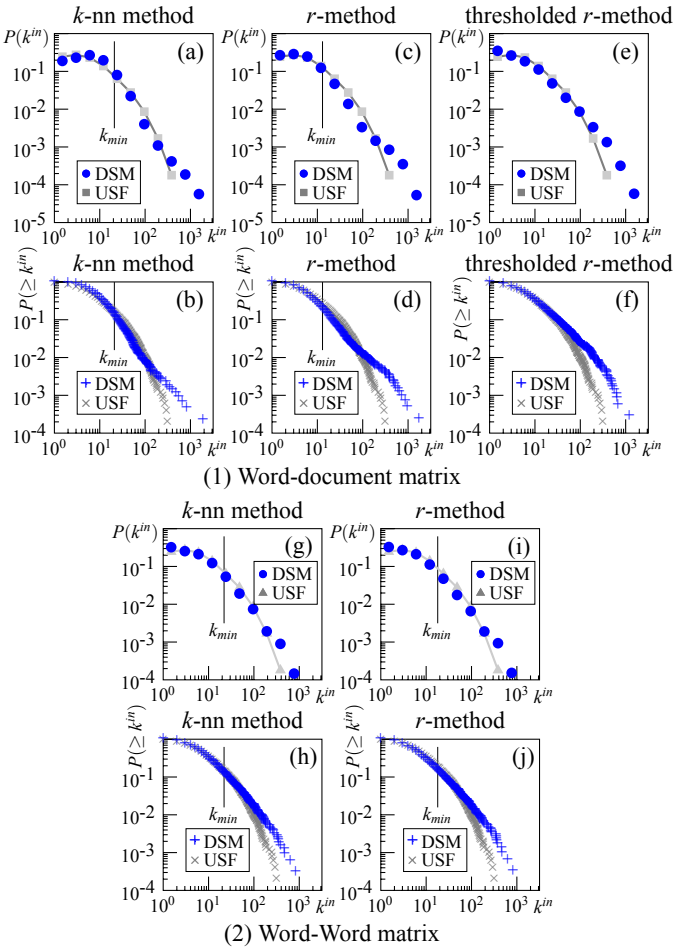


Figure 2: In-degree and cumulative in-degree distributions of the DSM networks generated from the initial matrix

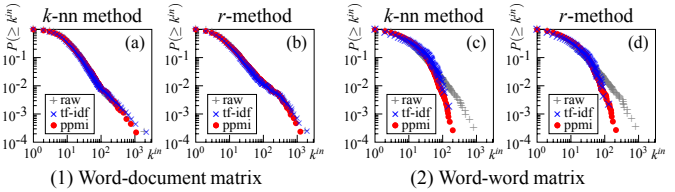


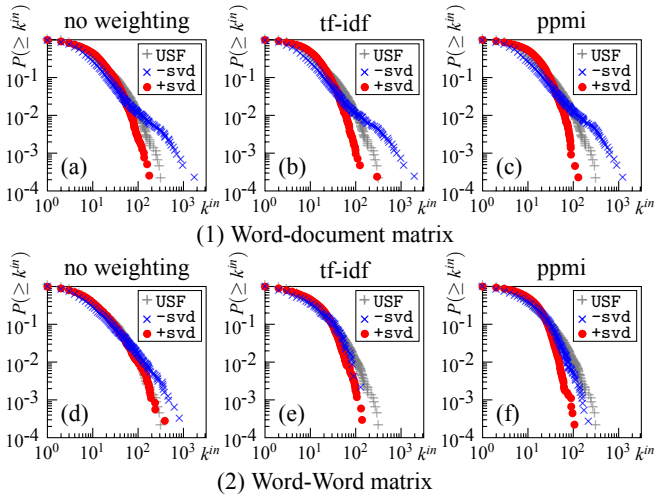
Figure 3: Cumulative degree distributions of the DSM networks generated from different weighting schemes

4. Smoothed word-document matrix: The smoothed word-document-based DSMs yield a scale-free network whose degree distribution follows the pure power law.

Concerning Result 1, the initial unweighted matrix generates networks whose degree distribution follows the truncated power law, which is similar to the distribution of the association network, as shown in Figure 2 (g)-(j) and in the second column (+−+) of Table 3. Both tf-idf and ppmi weighting schemes change the degree distribution into a more truncated form (i.e., with a sharper cutoff), as shown in Figure 3 (c) and (d), but the distribution still follows the truncated power law.

Concerning Result 2, SVD smoothing indeed affects some DSM networks (i.e., networks with tf-idf weighting) based on the word-word matrix so that their degree distribution follows the pure power law, as shown in the fifth through the last columns (+∗∗) of Table 3 and in Figure 4 (d)-(f).

On the other hand, Result 3 can be confirmed by Figure 2



(1) Word-document matrix

(2) Word-Word matrix

Figure 4: Cumulative in-degree distribution of the DSM networks generated after and before SVD smoothing: The case of the *r*-method

(a)-(d). Although the second column of Table 3 indicates that the pure power law is most appropriate, the degree distributions for the unweighted word-document matrix, in fact, take a different shape from three distributions addressed in this paper; the distributions decay exponentially for small k, but the decay is suddenly slower in a linear fashion, which is rarely observed in real-world systems. Weighting does not change these unnatural distributions, as shown in Figure 3 (a) and (b).

These unnatural distributions would result from an unsuccessful construction of word-document-based networks, in which word pairs with very low cosine similarity are connected by edges. In general, a word-document matrix is more sparse than a word-word matrix generated from the same corpus; the percentage of zero elements was 99.36% for the word-document matrix used in this paper, while 78.84% for the word-word matrix. The higher sparseness of the word-document matrix leads to very low cosine similarity. Indeed, in the word-document-based network by the *r*-method, about an half of the word pairs connected by an edge had the cosine of 0.05 or less. These low-cosine word pairs are likely to include highly frequent words, because they are likely to have more non-zero elements in their vector representation. The fatter tail observed in the degree distribution for the word-document matrix may be a consequence of this frequency effect; some frequent words are connected by a large number of edges. Hence, if a network is created by thresholding the cosine similarity of word pairs, it is expected that its degree distribution becomes close to the (truncated) power law. Indeed, when we created a network using the *r*-method by limiting word pairs to be joined by an edge to those with the cosine of 0.05 or more, its degree distribution followed the truncated power law, as shown in Figure 2 (e)-(f). The same result was obtained in the networks created by the *k*-nn method.

Concerning Result 4, SVD smoothing also leads to the word-document-based networks to follow the pure power law, as shown in Figure 4 (a)-(c) and in the fifth through the last columns of Table 3. This is because SVD compensates the data sparseness of the original DSMs.

Network Model and Semantic Relations

One question that naturally arises is what structure underlying various semantic networks governs their different behaviors. In this section, we provide one possible answer to the question in terms of a network model that explains a distinction of semantic relations between connected word nodes.

Barabási and Albert (1999) proposed a simple model for a scale-free network with the mechanism of network growth and *preferential attachment*, which leads to the pure power-law degree distribution with the exponent of 3. In this model, a small fully connected network of M nodes is constructed first, and then a new node is added to the network successively (i.e., network growth), by connecting it to M existing nodes selected with probabilities proportional to their degrees (i.e., preferential attachment). In order to provide a psychologically plausible explanation of semantic growth, Steyvers and Tenenbaum (2005) extended the Barabási-Albert model by introducing the process of semantic differentiation into the simple mechanism of preferential attachment. In the Steyvers-Tenenbaum model, after an existing node v_i is chosen for differentiation with probability proportional to their degrees just as the Barabási-Albert model does, a new node is connected to M randomly chosen nodes in the neighbors of the node v_i. However, their model cannot explain the observed difference among distributions of DSM networks, because it does not have free scaling parameters enough to generate a variety of pure and truncated power-law distributions.

One solution to the limitation lies in the reasonable assumption that semantic growth cannot be explained solely by the process of semantic differentiation. Two word nodes connected by preferential attachment can be regarded as semantically or taxonomically similar, because a new word added to the network by semantic differentiation corresponds to more specific variations on existing words. However, a new word can be associated with other words by another relation, namely an attributive or collocational relation. This process does not require preferential attachment because there is no reason to assume that highly complex concepts (i.e., those with many connections) are likely to be an attribute of a new concept. For example, in the USF association norms, the four most listed associates of the cue *cherry* are *red*, *pie*, *fruit*, and *apple*. When we consider the situation where a new word *cherry* is added to the network, adding edges from *cherry* to *fruit* and *apple* means that *cherry* differentiates the concepts of *fruit* and its neighbor *apple* by introducing their new subcategories. On the other hand, adding edges to *red* and *pie* can be interpreted differently; these edges may be added because *cherry* has the attributes of "being red" and "being a material of pie". [3] We refer to this process as *experiential correlation*.

In lexical semantics, this distinction of semantic relations is known as *syntagmatic-paradigmatic distinction*. Two words are syntagmatically related if they cooccur more often than would be expected by chance. Syntagmatically related words tend to cooccur in a noun phrase (e.g., *red cherry*) or a verb phrase (*eat cherries*). Because these phrases represent a relation between a concept and its attribute, syntagmatic relations are caused primarily by experiential correlation. On the other hand, two words are paradigmatically related if they do not cooccur but they can substitute for one another, in other words, they cooccur with similar words. Paradigmatic relations tend to be taxonomically similar by virtue of synonym, antonym or other coordinates, and thus they are caused by semantic differentiation.

In order to integrate the process of experiential correlation into the Steyvers-Tenenbaum model, we consider *random attachment*, which is introduced into the Barabási-Albert model by Liu, Lai, Ye, and Dasgupta (2002). In Liu et al.'s (2002) model, a new node is attached to the existing nodes preferen-

[3] The case of *pie* is controversial; it can also be interpreted as due to the semantic differentiation in which *cherry* subcategorizes a pie.

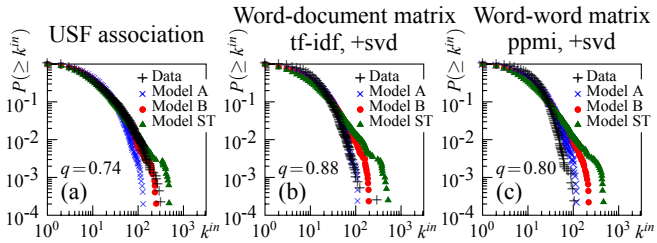


Figure 5: Cumulative in-degree distributions of the real semantic networks and simulated networks

tially with probability $1-p$ or randomly with probability p. The resulting network has the degree distribution that follows a mixture of power-law and exponential behaviors. Clearly, the distribution completely follows the power law if $p = 0$, while it follows the exponential if $p = 1$. When $0 < p < 1$, the distribution exhibits an approximately exponential behavior for small k, and a power-law-like behavior for large k. Note that, as mentioned earlier, the degree distribution of the USF association network follows the exponential below k_{min} and the truncated power law above k_{min}. This suggests that both preferential and random attachments are required for appropriately simulating the behavior of semantic networks.

Following Liu et al.'s (2002) model, we propose a modified version of the Steyvers-Tenenbaum model with both preferential and random attachments. A new node is attached to M existing nodes preferentially by semantic differentiation with probability $1-p$, and randomly by experiential correlation with probability p. For preferential attachment, nodes to be connected to a new node are chosen from only the neighbors of a node v_i that were previously added by preferential attachment. In random attachment, these nodes are chosen from all the existing nodes. To simulate the observed distributions of the DSM networks by this modified model, we must determine p. In this paper, we determine p using the fraction q of edges that connect a pair of syntagmatically related words in the target network as follows: $p = q$ (Model A) and $p = q/2$ (Model B). Model A assumes that all syntagmatic relations are caused by random attachment, while Model B assumes that syntagmatic relations are caused equally by random and preferential attachments (for this possibility, see footnote 3).

Figure 5 shows some simulation results by the modified Steyvers-Tenenbaum model. In these simulations, the direction of each edge was chosen randomly, pointing toward the old node with the probability γ. We determined γ so that the generated network had approximately the same connectivity n_{CC} as the real network. In Figure 5, the modified model (Model A or B) better simulates the distributions of the real semantic networks than the original Steyvers-Tenenbaum model (Model ST), thus suggesting that the network model with both preferential and random attachments is more appropriate for explaining the behaviors of semantic networks. Whether Model A or B better fits the observed distribution depends on the semantic network; Model A better reproduces the distribution for the DSM networks, while Model B is more appropriate for the USF association network. This result suggests that syntagmatically related words may be more likely to be connected by random attachment in the DSM networks than in the association network. It is unclear why this difference occurs, an answer to which must await further research.

Concluding Remarks

The complex network analysis reported in this paper demonstrates that DSM networks have the same properties as the association network; they are small-world and their degree distributions follow the truncated power law with an initially exponential decay. In addition, the modified Steyvers-Tenenbaum model with both preferential and random attachments better reproduces these degree distributions. Future research directions include analyzing a more fine-grained structure (e.g., hierarchical structure and semantic field) of semantic networks, developing a more plausible network model for simulating the behavior of semantic networks, and modeling semantic representation as navigation on semantic networks.

References

Amaral, L., Scala, A., Barthélémy, M., & Stanley, H. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, *97*(21), 11149–11152.

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348–360.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy*, *12*, 1264–1302.

Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Liu, Z., Lai, Y.-C., Ye, N., & Dasgupta, P. (2002). Connectivity distribution and attack tolerance of general networks with both preferential and random attachments. *Physics Letters A*, *303*, 337–344.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, *37*, 125–145.

Motter, A. E., de Moura, A. P., Lai, Y.-C., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E*, *65*(6), 065102.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

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

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, *35*(2), 251–296.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.