Evaluating the psychological plausibility of word2vec and GloVe distributional semantic models

Ivana Kajić, Chris Eliasmith
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada N2L 3G1
{i2kajic, celiasmith}@uwaterloo.ca

Abstract

The representation of semantic knowledge poses a central modelling decision in many models of cognitive phenomena. However, not all such representations reflect properties observed in human semantic networks. Here, we evaluate the psychological plausibility of two distributional semantic models widely used in natural language processing: word2vec and GloVe. We use these models to construct directed and undirected semantic networks and compare them to networks of human association norms using a set of graph-theoretic analyses. Our results show that all such networks display small-world characteristics, while only undirected networks show similar degree distributions to those in the human semantic network. Directed networks also exhibit a hierarchical organization that is reminiscent of the human semantic network.

Keywords: semantic spaces, distributional semantic models, free association norms, network analysis

Introduction

The representation of semantic knowledge is instrumental to many models of linguistic processing in cognitive modelling and machine learning. In particular, the decision of how to represent such knowledge entails the selection of a vocabulary and a computational representation of vocabulary items.

Many computational models studying human semantic memory and related processes have been relying on The University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 2004, USF Norms). Because it is a psychologically plausible representation of a semantic network, the USF Norms have been successfully used to reproduce human-level performance on tasks such as verbal semantic search (Abbott, Austerweil, & Griffiths, 2015; Kajić et al., 2017), and recognition memory and recall (Steyvers, Shiffrin, & Nelson, 2004).

Another common choice for the representation of semantic knowledge is models derived from co-occurrence and word frequency data. Such models learn vector representations of words from large linguistic corpora and are often referred to as spatial or distributed semantic models (DSMs).

In the domain of natural language processing, word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014) have been two widely used DSMs. They are shown to achieve high accuracy on a variety of lexical semantic tasks such as word analogy and named entity recognition. Their capacity to perform well on such tasks also makes them attractive candidates for semantic representations in cognitive models. Yet, it remains unclear which, if any, aspects of such representations are psychologically plausible.

This study evaluates psychological plausibility of GloVe and word2vec models by analyzing semantic networks constructed from those models and comparing them to semantic networks constructed from the USF Norms.

In particular, we evaluate networks in terms of their small-world characteristics, degree distributions and hierarchical organization. We characterize networks that capture properties of human association networks, and identify differences that might have important implications for modelling of human semantic memory.

Semantic Spaces

Vector-based word representations are generated to capture statistical regularities observed in natural language. Often, a high-dimensional co-occurrence matrix is created by counting word occurrences in a set of texts or contexts. Then, by applying a dimensionality reduction method such a matrix is factorized into components that can be used to reconstruct low-dimensional vectors representing individual words. DSMs using this approach have been known as *count-based* models, with Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990, LSA) being one such prominent example. GloVe vectors used in this work are derived from a count-based model (Pennington et al., 2014) that is based on methods similar to LSA.

Although networks created from LSA vectors have been criticized as unable to reproduce connectivity patterns between words as observed in the USF Norms (Steyvers & Tenenbaum, 2005), this has been challenged by more recent work demonstrating that some DSMs indeed produce degree distributions that resemble that of human association norms (Utsumi, 2015).

In contrast to count-based models such as LSA, more recently developed *predictive* models use iterative training procedures in complex neural networks to learn word vectors based on the contexts in which those words occur. Word2vec (Mikolov et al., 2013) is a popular predictive model that computes vectors from a large corpus of text by maximizing the probability of a target, which can either be a single word or a set of context words.

In this work, we use pre-trained GloVe and word2vec vectors. GloVe vectors were trained using the Common Crawl dataset containing approximately 840 billion word tokens. Word2vec vectors were trained on the Google News dataset containing about 100 billion words. In both cases, the resulting vectors used here contain 300 dimensions. To allow for

a comparable analyses, we restrict the size of vocabularies of these two datasets to that of the USF Norms, which contains 5018 words.

Constructing Semantic Networks

We generate undirected and directed semantic networks from word2vec and GloVe semantic models, and compare them to corresponding networks constructed from the USF Norms.[1] The cosine angle is used as a measure of similarity between two word vectors. We have also tested inner angle as a similarity measure, but found no major differences between the two and therefore report results of analyses based on cosine similarity. The data processing and analysis code is available online at `https://github.com/ctn-archive/kajic-cogsci2018`.

Undirected Networks

We create three undirected networks: one from the human free association norms (*USF Norms*), one from the word2vec vectors (*word2vec*) and one from the GloVe vectors (*glove*). In every network, a node represents a word and an edge between two nodes corresponds to an associative relationship between two words. To construct the undirected version of the USF network, we place an edge between two nodes representing words w_1 and w_2 if an association pair (w_1, w_2) or (w_2, w_1) occurs in the USF word association database.

To construct undirected networks from DSMs, we compute the similarity between all word pairs in the corresponding vocabulary. An edge is placed between two nodes in a network if the similarity between words represented by those nodes exceeds a certain threshold τ. To select a threshold for each network, we first test a sequence of uniformly distributed thresholds (with two decimal places) in different ranges. Then, we select the threshold that produces a network with the average node degree $< k >$ that is closest to that of the USF network. Overall, increasing the threshold had the effect of shifting the degree distribution from the "right" (the regime where many nodes have many connections) to the "left" (very sparse connectivity with most nodes having only few connections). For the *word2vec* network the threshold is $\tau = 0.38$ and for the *glove* network it is $\tau = 0.53$.

The degree distribution of the *word2vec* network was strongly correlated ($r = 0.74, p < 0.001$) with the degree distribution of the USF network, while a moderate correlation ($r = 0.49, p < 0.001$) was observed with the *glove* network.

Directed Networks

A directed network is a more accurate representation of human word associations, as it captures the directionality of associative relationship between a cue word and a target word. Only 26.4% of all association pairs in the USF Norms are reciprocal.[2] To construct the directed version of the USF network, a directed edge is placed between two nodes w_1 and w_2 only if w_2 was an associate of a cue word w_1.

We adopt two different methods to construct directed networks from DSMs: the *k*-nn method (Steyvers & Tenenbaum, 2005) and the *cs*-method (Utsumi, 2015). In both cases, the local neighborhood of a node i is determined by placing outgoing edges to k other nodes that represent words most similar to the word represented by the node i. The *k*-nn method determines the number k for each node as the number of associates of that word in the USF Norms, resulting in a network that has the same out-degree distribution as the directed USF network. The *cs*-method finds the smallest number k for which a certain empirical threshold R is exceeded that produces the same average degree connectivity $< k >$ as in the directed USF network. We refer to the networks constructed by the *k*-nn method as *word2vec-knn* and *glove-knn*, and those constructed with the *cs*-method as *word2vec-cs* and *glove-cs*.

We observe strong and significant correlations between degree distributions of the directed networks *word2vec-knn* ($r = 0.69$), *glove-knn* ($r = 0.83$), *word2vec-cs* ($r = 0.75$), *glove-cs* ($r = 0.88, p < 0.001$ in all cases), and the directed USF network.

Network Statistics

Previous research (Steyvers et al., 2004; Utsumi, 2015; Morais, Olsson, & Schooler, 2013) has identified that human association networks can be characterized in terms of their small-world properties: although the networks are very sparse (i.e., a node in the network is on average connected to only a small subset of all nodes in the network), they have a small average shortest path length L, $L = \frac{1}{n(n-1)} \sum_{i,j \in G} d(i,j)$, where n is the number of nodes in the network, i and j are two different nodes and $d(i,j)$ is the shortest distance between the two nodes measured as the number of edges between them.

In addition, the average clustering coefficients of such networks are higher than clustering coefficients of random networks of the same size that have the same probability of a connection between any two nodes. The average clustering coefficient C for a network with n nodes is computed as $C = \frac{1}{n} \sum_{i \in G} c_i$ where c_i is a local clustering coefficient of a node i, given by: $c_i = \frac{2t_i}{k_i(k_i-1)}$. t_i is the number of triangles in the neighborhood of the node i. A triangle is a connectivity pattern where a node i is connected to two other nodes j and k, and at the same time the nodes j and k are also connected. The denominator in the equation for c_i is the number of possible connections in a neighborhood of a node with the degree k_i: $k_i(k_i-1)/2$. In the context of semantic networks, such small-world structure is important for the efficient search and retrieval of items from memory.

To test whether networks constructed from word2vec and GloVe models exhibit small-world characteristics, we run different graph-theoretic analyses. The sparsity s of a network is

[1] We will refer to all networks constructed from word2vec and GloVe models as synthetic semantic networks, to differentiate them from the experimentally derived USF network.

[2] A reciprocal association is one where the word w_1 is an associate of a word w_2 and vice versa.

Table 1: Graph-theoretic statistics of networks derived from USF Norms, word2vec and GloVe vectors. The results of undirected networks are presented in the first three rows. Abbreviations: L = the average shortest path length, k = average node degree, C = the average clustering coefficient, C_k = connectivity, the number of nodes in the largest connected component (expressed in %), D = the network diameter, m = the number of edges, n = the number of nodes, L_{rnd} = the average shortest path of a randomly connected network of similar size, C_{rnd} = the average clustering coefficient of a random network, s = the sparsity of the network (expressed in %).

	L	$< k >$	C	C_k	D	m	n	L_{rnd}	C_{rnd}	s
USF undirected	3.04	22.0	0.186	100.00	5	55,236	5,018	3.03	0.004	0.44
word2vec	4.24	21.3	0.325	99.84	12	52,317	4,902	3.04	0.004	0.44
glove	4.61	22.1	0.373	98.88	12	51,244	4,632	2.99	0.005	0.48
USF directed	4.26	12.7	0.187	96.51	10	63,619	5,018	3.62	0.005	0.25
word2vec-cs	4.81	12.5	0.237	99.28	11	62,328	4,977	3.64	0.005	0.25
word2vec-knn	4.77	12.7	0.232	99.32	12	63,165	4,977	3.64	0.005	0.26
glove-cs	5.06	12.3	0.266	97.21	12	61,470	4,988	3.65	0.005	0.25
glove-knn	5.03	12.7	0.259	97.91	13	63,262	4,988	3.62	0.005	0.25

computed by dividing the average node degree $< k >$ with the total number of edges in the network. Other measures such as the clustering coefficient C, the average shortest path length L and the diameter D have been performed on the largest connected component of each network.

The results of analyses are summarized in Table 1. Our results for the two USF networks are consistent with the previous reports (Steyvers et al., 2004; Utsumi, 2015). Due to the methods used to construct the networks, all synthetic networks have sparsity that is comparable to the sparsity of the human association network. Also, their average shortest path lengths are consistently higher, but still comparable to those of the human association networks. However, in undirected networks, the diameter of synthetic networks is more than twice as long as that of the USF network, meaning that the distance between the two farthest words is longer in the synthetic networks than it is in the association network. Furthermore, all synthetic networks also exhibit a degree of clustering that is higher than that of the USF network. This effect is more pronounced in the undirected versions of the network.

Degree Distributions

To obtain the distribution of degrees in a network, we count the number of nodes with k degrees, where k ranges from one to k_{max}. The k_{max} value denotes the highest node degree and it is different for different networks. The distribution of in-degrees of the directed association network is known to follow a truncated power-law distribution $P(k) \sim e^{\lambda} k^{-\alpha}$, or, in some cases, a pure power-law $P(k) \sim k^{-\alpha}$ (Utsumi, 2015; Morais et al., 2013). The power-law predicts that most nodes in the network have a few connections, while a small number of nodes, regarded as *hubs*, have a rich local neighborhood.

In our analyses of degree distributions, we first test the plausibility of a power-law behavior using the goodness-of-fit test. Then we test whether other heavy-tailed distributions provide a better fit using the loglikelihood-ratio (LR) test. To fit and evaluate different models, we use the Python powerlaw package (Alstott, Bullmore, & Plenz, 2014).

First, we fit the empirical degree distribution to a power-law model using the maximum likelihood estimation for the parameter α. The fit is performed for values of $k > k_{min}$, where k_{min} was determined such that the Kolmogorov-Smirnov (KS) distance between the empirical distribution and the model distribution for values greater than k_{min} is minimized. Given a model, the goodness-of-fit test uses the KS distance between the model and the empirical distribution, as well as the model and thousands of distributions sampled from the model, to evaluate its plausibility. It produces a p-value that is a fraction of sampled distributions that have a greater KS distance than the empirical distribution. Large p-values denote that sampled distributions are more distant than the empirical distribution, in which case the model is regarded as a plausible fit to the empirical data.

The LR-test is a comparative test that evaluates which of the two distributions is more likely to generate samples from the empirical data based on maximum likelihood functions of each distribution. The resulting R value is positive if the first distribution is more likely, and negative otherwise. The alternative heavy-tailed distributions we tested are: truncated power-law, (discretized) lognormal and exponential.

Results of our analyses are summarized in Table 2. As consistent with previous research, we find that the truncated power-law, rather than a pure power-law, is a better description for the distribution of degrees for the direct USF network (Morais et al., 2013; Utsumi, 2015). In addition, our results indicate that the lognormal distribution is a plausible model for the directed USF network, as it is not possible to distinguish between the lognormal and truncated power law distributions ($R = -0.43, p = 0.67$). Pure power-law is excluded as a plausible model for all our networks as p-values for the goodness-of-fit test are all close to 0.

We also find that the truncated power-law is a plausible model for the undirected USF network and both undirected versions of the synthetic networks. It is important to notice

Table 2: Goodness-of-fit test for the power-law distribution and loglikelihood ratio tests evaluating plausibility of the power-law versus other heavy-tailed distributions. The results of undirected networks are presented in the first three rows. Abbreviations: KS = Kolmogorov-Smirnov statistic, LR = loglikelihood ratio.

	Power Law		Power Law vs. Truncated Power Law		Power Law vs. Lognormal		Power Law vs. Exponential	
	KS	p	LR	p	LR	p	LR	p
USF undirected	0.014	0.01	-1.80	0.02	-0.91	0.37	7.46	0.00
glove	0.035	0.00	-7.29	0.00	-5.03	0.00	7.82	0.00
word2vec	0.064	0.00	-7.72	0.00	-5.31	0.00	-5.89	0.00
USF directed	0.055	0.00	-2.54	0.00	-2.36	0.02	0.27	0.79
glove-cs	0.016	0.00	-1.14	0.17	-0.67	0.50	3.52	0.00
glove-knn	0.020	0.00	-1.05	0.16	-0.82	0.41	1.08	0.28
word2vec-cs	0.032	0.00	-0.57	0.40	-0.51	0.61	0.29	0.77
word2vec-knn	0.028	0.00	-0.13	0.82	-0.12	0.90	0.69	0.49

that the exponential distribution, as well as the lognormal distribution cannot be ruled out for the undirected *word2vec* network. However, due to the high p-values of the LR tests it is not possible to reach similar conclusions for the degree distributions of the directed synthetic networks.

To better understand these numerical results, we plot the empirical data and model fits on a semi-log scale in Figure 1. Degree distributions are expressed as complementary cumulative distribution functions, and fits for the power-law, truncated power-law, lognormal, and exponential models are shown. The scarcity of nodes with high degrees (>80) in certain variants of synthetic graphs such as *glove-knn, word2vec-cs* and *word2vec-knn* are likely to contribute to large p-values in LR tests in Table 2. While the distribution of degrees of USF networks is bounded from above by the power-law distribution and from below by the exponential distribution, this is only somewhat the case for the *glove* networks and less so for the *word2vec* networks, indicating differences between degree distributions of human and synthetic word networks.

Hierarchical Topology

Human association networks have been shown to exhibit a hierarchical organization resulting from the high modularity of the network (Utsumi, 2015). Such modules, or clusters, are highly interconnected groups of nodes that form only a few connections to nodes that are not part of the group. The presence of such clusters indicates that there are features shared among nodes in the network, such as semantic or lexical relatedness.

While the average clustering coefficients are reported in Table 1, to investigate the presence of hierarchical structure, we consider the relationship between a node degree and local clustering coefficients c_i (Ravasz & Barabási, 2003). In networks that exhibit hierarchical organization, the local clustering coefficient is dependent on the node degree and has been observed to follow a scaling law of the form $C(k) \sim k^{-\gamma}$ (Ravasz & Barabási, 2003). While many hierarchical networks have been observed to have $\gamma = 1$, hierarchical structure has also been observed in networks with $\gamma < 1$.

To investigate whether the tendency for clustering is dependent on the node degree, we implement methods proposed by Utsumi (2015). We first compute local clustering coefficients for all nodes in the largest connected component in each network. Then, we compute the average clustering coefficient for each neighborhood size k and connect those values to form a line. Finally, we use linear regression in the logarithmic space to determine the slope of the regression line and the correlation coefficient.

The results are shown in Figure 2. First, we confirm that the directed USF network exhibits hierarchical organization with $\gamma = 0.75$ ($r = -0.97$). We also found strong negative correlation between the size of a neighborhood and the average clustering coefficient for the undirected USF network ($\gamma = 0.76, r = -0.97$). For undirected versions of synthetic networks, we find a small positive slope $\gamma = -0.05$ and a positive correlation ($r = 0.27$) for the *glove* network, and similarly $\gamma = -0.10$ ($r = 0.49$) for the *word2vec* network. Therefore, there is no dependency between the local clustering coefficients and the node degree in the undirected versions of synthetic semantic networks.

In contrast, some of the directed semantic networks exhibit higher levels of hierarchical organization. Directed networks constructed with the k-nn method have negative slopes with strong correlations: $\gamma = 0.49$ ($r = -0.96$) for *glove-knn* and $\gamma = 0.39$ ($r = -0.91$) for *word2vec-knn*. The hierarchical relationship is less apparent in networks constructed with the *cs*-method (*glove-cs*: $\gamma = 0.32, r = -0.88$, *word2vec-cs*: $\gamma = 0.17, r = -0.54$).

Discussion

The goal of the present study is to evaluate the psychological plausibility of semantic networks constructed from the widely used word2vec and GloVe distributional semantic models. To this end, a number of graph-theoretic analyses were per-

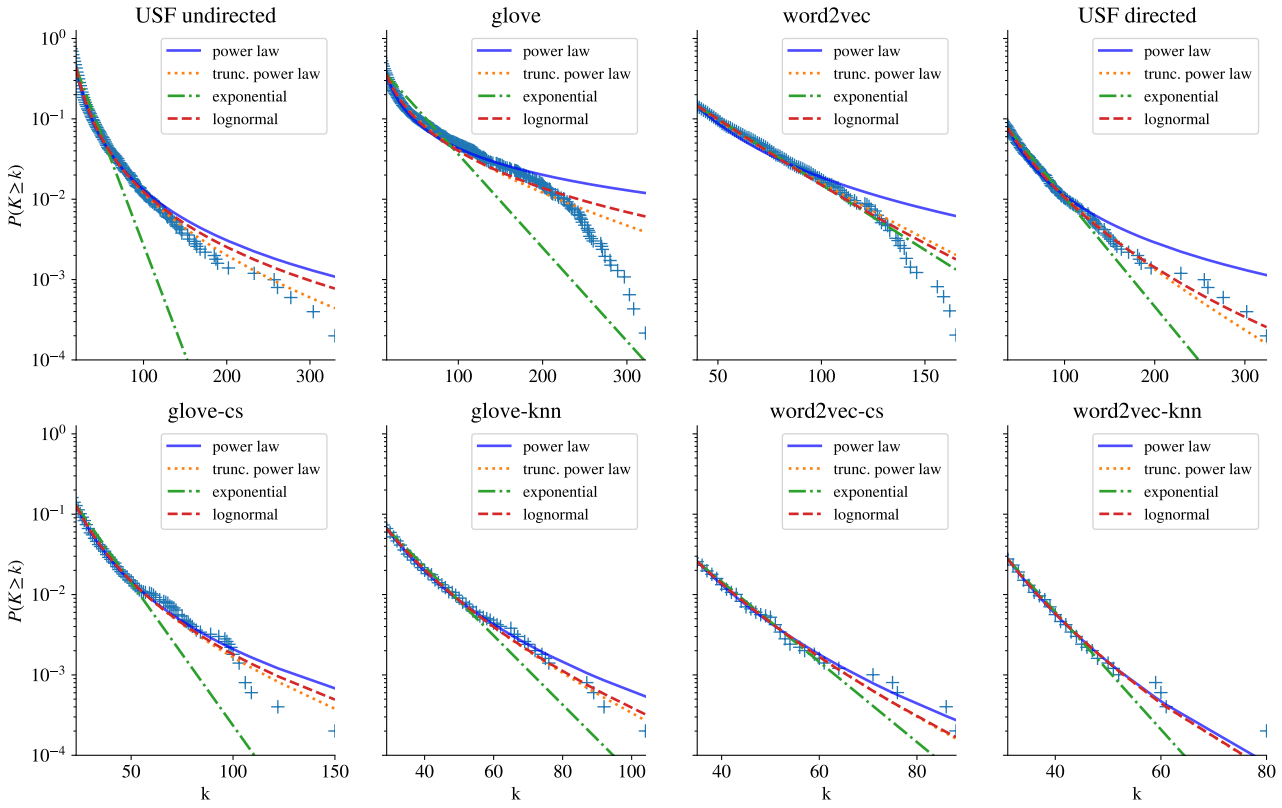

Figure 1: Complementary cumulative distributions and model fits for the *USF Norms*, *glove* and *word2vec* networks. Distributions and fits for undirected networks are shown in the first three plots in the first row. The x-axis for each network is bounded by k_{min} and k_{max} that are unique for each network (see text for details).

formed that compared undirected and directed versions of networks with semantic networks constructed from human association norms.

We found that all networks exhibit the small-world property, characterized by short path lengths and high clustering coefficients. In other words, it is possible to efficiently search in such networks as any two words in a network are only a few words apart. These results are consistent with previous studies that demonstrated small-world structure in different DSMs (Steyvers & Tenenbaum, 2005; Utsumi, 2015).

Degree distribution analyses based on a goodness-of-fit test revealed that the power-law is not a plausible model in any of the networks. This finding may not be as surprising considering that some semantic networks have distributions that can be described well with alternative heavy-tailed distributions (Morais et al., 2013; Utsumi, 2015). We contribute to the existing research by adding that the truncated power-law is a plausible explanation of the degree distribution for the undirected USF network. The synthetic undirected networks were also explained best by the truncated power-law. However, the lognormal and the exponential distribution are also a plausible fit for the undirected *word2vec* network.

Truncated power-law behavior could not be inferred for the directed *word2vec* and *glove* networks. Analyzing the tails of distributions in Fig. 1 provides more insight as to why it is difficult to obtain a clear fit in those cases. Directed networks have only very few nodes with a high number of connections. For example, there are only four nodes with $k > 80$ for *glove-knn* and less than ten nodes with $k > 55$ for both *word2vec* networks. What distinguishes different heavy-tailed distributions are nodes "contained" in the tail of a distribution, and in this case it is possible that the LR test did not have enough data to reliably discriminate between different distributions.

We found that directed networks, more specifically those constructed with the *k*-nn method, exhibit a moderate level of hierarchical organization that is reminiscent of clustering observed in the human association network. The directed *glove-cs* and *glove-knn* networks contain nodes with high connectivity that exhibit fluctuations in clustering coefficients, as observed with the network created from association norms. Such nodes are hubs, some of which have higher clustering coefficients since they are embedded in clusters. Hubs with lower clustering coefficients act as intermediaries in the network by connecting different modules of the network that have less connections.

Overall, these results indicate that different semantic networks constructed from word2vec and GloVe models are capable of capturing some aspects of human association networks. However, for most synthetic datasets there are clear differences with the empirical networks. The *glove-knn* net-

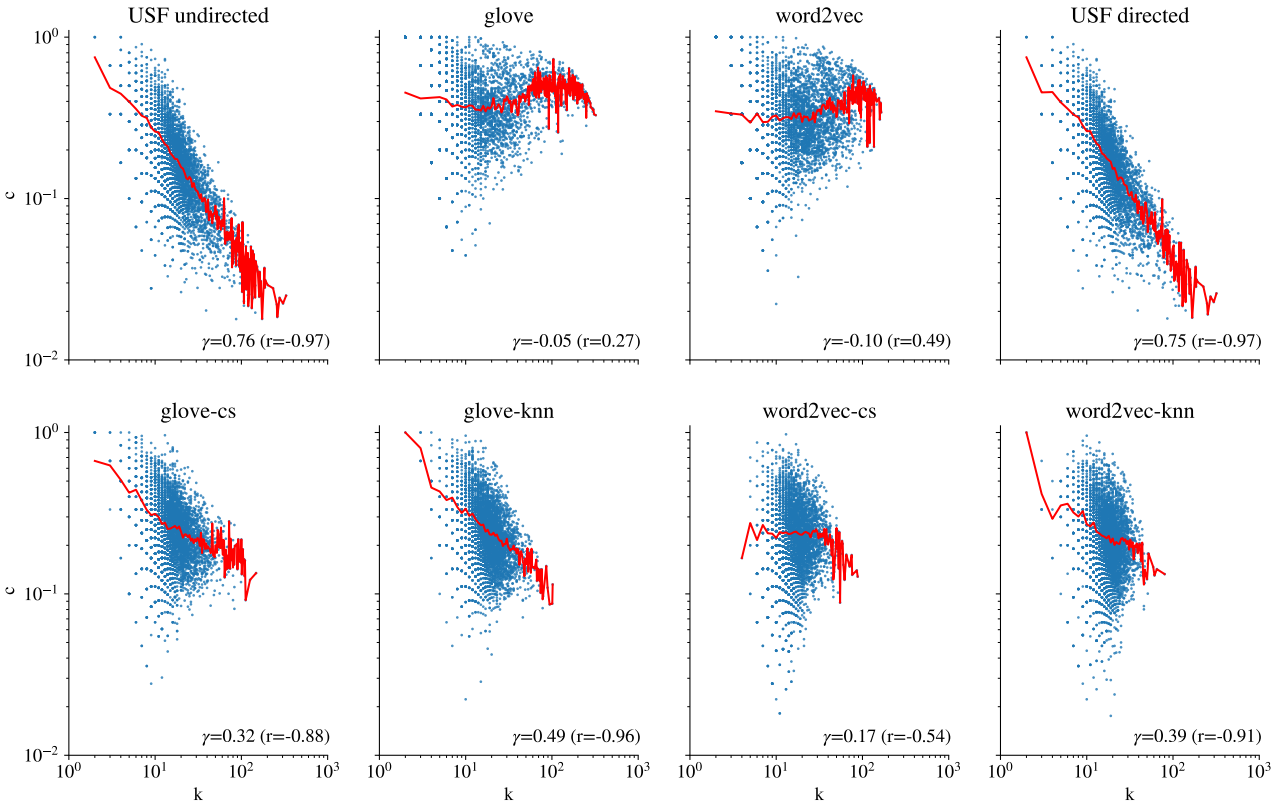

Figure 2: Scatter plots of local clustering coefficients. Every blue dot represents a local clustering coefficient of a node with the degree k. The red line connects averages. The first three plots in the first row are obtained from undirected networks.

work exhibits properties that are most similar to those of the USF Norms, but future work should address methods that yield a greater number of nodes with rich neighborhoods.

Acknowledgments

The authors would like to thank Terry Stewart for useful comments and discussions. This work has been supported by AFOSR, grant number FA9550-17-1-002.

References

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

Alstott, J., Bullmore, E., & Plenz, D. (2014). powerlaw: a python package for analysis of heavy-tailed distributions. *PloS One*, *9*(1), e85777.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Am. soc.: for inf. sci.*, *41*(6), 391.

Kajić, I., Gosmann, J., Komer, B., Orr, R. W., Stewart, T. C., & Eliasmith, C. (2017). A biologically constrained model of semantic memory search. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 631–636). Austin, TX: Cognitive Science Society.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cog. Sci.*, *37*(1), 125–145.

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.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Ravasz, E., & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Phys. Rev. E*, *67*(2), 026112.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In *Experimental Cognitive Psychology and its Applications* (pp. 237–249). American Psychological Association.

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

Utsumi, A. (2015). A complex network approach to distributional semantic models. *PloS One*, *10*(8), e0136277.