Topological structure of dictionary graphs

Henryk Fukś and Mark Krzemiński

Department of Mathematics, Brock University
St. Catharines, Ontario L2S 3A1, Canada

E-mail: hfuks@brocku.ca

Abstract. We investigate the topological structure of subgraphs of dictionary graphs constructed form WordNet and Moby Thesaurus data. In the process of learning a foreign language, the learner knows only a subset of all words of the language, corresponding to a subgraph of a dictionary graph. When this subgraph grows with time, its topological properties change. We introduce the notion of pseudocore and argue that the growth of the vocabulary roughly follows decreasing pseudocore numbers – that is, one first learns words with high pseudocore number followed by smaller pseudocores. We also propose alternative strategy for vocabulary growth, involving decreasing core numbers as opposed to pseudocore numbers. We find that as the core or pseudocore grows in size, the clustering coefficient first decreases, then reaches a minimum, and starts increasing again. The minimum occurs when the vocabulary reaches a size between 10^3 and 10^4. A simple model exhibiting similar behavior is proposed. The model is based on a generalized geometric random graph. Possible implications for language learning are discussed.

PACS numbers: 05.90.+m, 02.50.-r

Submitted to: *J. Phys. A: Math. Gen.*

1. Introduction

"If you were asked to name the trait which most decisively distinguishes human beings from all other creatures on the planet, what would you choose? Love? Warfare? Art and music? Technology? Perhaps. But most people who have considered this question at length have come up with a single answer: language". This statement, taken from the introduction of the book by R. L. Task [1] needs no further justification. In addition to being the single most remarkable characteristic of humans, language presents itself as an immensely complex structure. Like many other complex systems, language can be viewed as a collection of discrete components interacting with each other in various ways. Depending on the "magnification factor", once can consider these components to be phonemes or letters, syllables, words, phrases, or even entire sentences.

In this article, we will mainly consider words and their interaction within a language. Although lower or higher level language features are important and interesting, one cannot underestimate importance of vocabulary in learning and using a language. Contemporary language acquisition specialists, for example, recognize the central importance of the vocabulary, and in the last two decades a lot of research effort went into the study of vocabulary learning strategies, determining what it means to "know a word", and methods of testing vocabulary knowledge and use [2].

One of the first questions that is encountered when one learns a new language is "how much vocabulary do I need to know?". Of course, the most ambitious goal would be to know all words of the language. This, however, is usually impossible to achieve. For example, although comprehensive dictionaries of English can easily contain over 10^5 headwords, it has been demonstrated that educated native speakers of English know only a fraction of this lexicon – about 20000 word families [3].

When one learns a new (second) language, the set of known words is steadily increasing with time. Many language scholars agree that the significant threshold in the language learning process occurs around 3000-5000 word families. It turns out that once this threshold is reached, learners can understand well above 90% of the running words in a typical text [4], and such high text coverage appears to be a necessary condition for transferring reading skills from the first to the second language [5].

In [6], we investigated a simplistic model of vocabulary growth by considering a graph G_{Web} obtained from Webster dictionary. Dictionary headwords were vertices of the graph, and two words were connected by an edge if one words appeared in the definition of the other word. We then assumed that the learner learns consecutive words in the order of decreasing frequency, that is, the most frequent words first. When the learner knows n top ranking words, he essentially "knows" a subgraph of G_{Web} spanned by these n words. With time, this subgraph grows in size, since n steadily increases. We found that many properties of the subgraph, such as, for example, its diameter or density, change monotonically with n, but the clustering coefficient exhibits rather curious behavior: first it decreases with growing n, yet around $n = 4000$ it starts increasing again. This indicates that some change in topology of the subgraph

takes place when the size of the vocabulary reaches 4000. While it remains unclear if this phenomenon has anything to do with the threshold described in the previous paragraph, it is nevertheless should be investigated in greater detail.

In this paper, we will first show that the idea of a "growing subgraph" can be replaced by a purely static model, and that the minimum reached by the clustering coefficient can be understood in static terms as well, as a certain property of the so-called clustering spectrum of the entire graph, to be introduced in subsequent sections. We will also show that this minimum occurs in graphs constructed from very different data than those used in [6], confirming that it a robust phenomenon.

2. Interaction graphs

We used two very different data sets: the WordNet database [7] and the Moby lexicon project [8]. From WordNet database, we extracted all nouns, verbs, adjectives and adverbs together with their definitions. All headwords and words occurring in definitions were stemmatized using the Porter stemming algorithm [9] to remove common morphological and inflectional endings. All compound terms and their definitions were removed. This means that, for example, definitions of terms such as *fall off* were discarded, but *fall* itself remained. This resulted in 45204 unique headword stems, which were assigned to separate vertices of a graph. Two vertices A and B (headword stems) were connected with an edge if an only if A occurred in the definition of B or conversely. The resulting graph has 551940 edges, and in what follows it will be referred to as $\mathcal{G}_{\mathcal{W}}$.

From the Moby lexicon we used the Moby thesaurus to construct another graph. This time, we used as vertices all words (including compounds) occurring in the thesaurus, both as headwords and as synonyms. It should be noted here that the term "synonym" is used in Moby thesaurus in a very broad sense. Under a given headword, one finds a large number of words which are not only strict synonyms, but also hypernyms and hyponyms, and other words with a meaning similar to the headword. Similarly as before, two vertices A and B are joined with an edge if the thesaurus lists A as one of the synonyms of B or conversely. The resulting graph, to be called $\mathcal{G}_{\mathcal{M}}$, has 103306 vertices and 1783351 edges. Note that stemmatization was not performed for $\mathcal{G}_{\mathcal{M}}$, since all headwords and synonyms in the thesaurus are listed in their canonical form.

It should be noted at this point that the thesaurus graph has been studied extensively in recent years [10, 11, 12], and graphs based on the WordNet database have been investigated too [13]. These studies revealed small-world type structure in thesaurus graphs, and semi-empirical description of their degree distribution and other properties has been proposed.

In Figure 1 degree distributions of $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$ are shown. Even though both graphs seem to exhibit power-law like behaviour for large degree values of degree n, they behave quite differently for small n. In order to describe these degree distribution

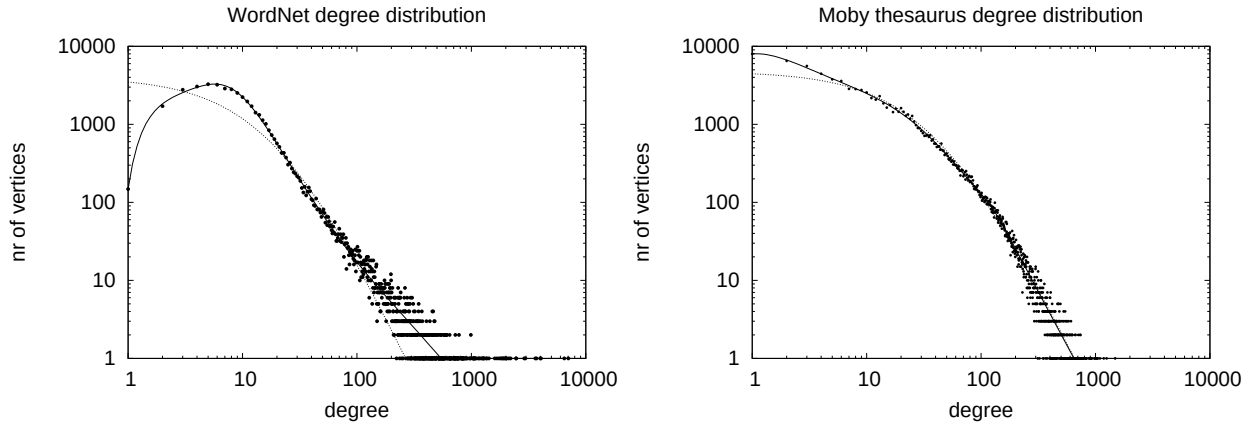

Figure 1. Degree distribution of $\mathcal{G}_{\mathcal{W}}$ (left) and $\mathcal{G}_{\mathcal{M}}$ (right). The dashed lines represent lines of best fit for eq. (1), while the solid lines for eq. (2-3).

curves empirically, we fitted a number of known distribution functions. One of them was the distribution proposed by C. Tsallis and M. P. de Albuquerque for citations of scientific papers [14], which was also used in earlier studies of the thesaurus,

$$f(n) = \frac{N_0}{[1 + (q-1)\lambda n]^q/(q-1)}, \tag{1}$$

where N_0, λ, and q are parameters. Here n denotes degree, and $f(n)$ number of vertices having degree n. The dashed lines in Figure 1 represents the lines of best fit for this equation. One can see that while the Moby thesaurus degree distribution is reasonably well described by eq. (1), the fit for WordNet graph is quite bad. In fact, in both cases the fit is less than ideal.

Since for subsequent investigations a more accurate description of degree distributions was needed, we produced a very precise fit using functions with a much larger number of parameters. For large degrees n, the distribution appears to follow a power law, hence it is reasonable to write

$$f(n) = R(n)n^{\alpha}, \tag{2}$$

where $R(n)$ is a function representing deviation from the power law such that $R(n) \to 1$ as $n \to \infty$. Unfortunately, in order to obtain very accurate fit, $R(n)$ must be rather complicated. We found that the following function works quite well for both $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$:

$$R(n) = A \exp\left(\frac{P_4(\ln n)}{P_5(\ln n)}\right), \tag{3}$$

where P_4 and P_5 are, respectively, polynomials of fourth and fifth degree (in the log-log plot this becomes a rational function). Solid lines in Figure 1 represent lines of best fit for the above function. We obtained $\alpha \approx -2.74$ for Moby thesaurus and $\alpha \approx -1.66$ for the WordNet dictionary. Obviously, the choice of (3) this rather arbitrary, and due to large number of parameters, it is not suprising that it is possible to obtain a good

fit. For that reason, we do not attach any particular meaning to this choice. The fitted function effectively provides a "smoothing" of data, and not much more.

3. Core decomposition

Having the interaction graphs defined, we now turn our attention to the process of language acquisition. It is well known that the frequency of occurrence of a word in a large corpus (textual or spoken) roughly follows a power law known as the Zipf law. As a result, relatively small number of high-frequency words suffices to cover a significant proportion of text, as we already remarked in the introduction. A natural consequence of this is the recommendation of language specialists to learn the most frequent words first, and then proceed to less frequently encountered words [2].

Since a dictionary such as WordNet dictionary is also an English text (although of a special type), the rank-frequency distribution of words in that dictionary follows the Zipf law. Obviously, when we construct the graph $\mathcal{G}_{\mathcal{W}}$, the degree of a given vertex (stem) will be closely related to the frequency of occurrence of that stem in the dictionary. Vertices of high degree will generally correspond to high-frequency stems and conversely. One can say, therefore, that when one learns the language, one should start with high-degree stems and proceed toward stems of lower degree.

As one learns new vocabulary following the strategy outlined above, then at any given moment, the set of know words (stems) forms a subgraph of $\mathcal{G}_{\mathcal{W}}$ or $\mathcal{G}_{\mathcal{M}}$. The notion of pseudo-core will be convenient to describe these subgraphs.

Definition 1 *For a non-negative integer k, k-pseudocore of a graph is the maximal subgraph such that its vertices have degree greater or equal to k, where by the "degree" in this definition we mean the degree of the vertex in the original graph, not in the subgraph. If G is a given graph, we define $G_{[k]}$ to be the k-pseudocore of G.*

Using this definition, we may say that if one starts learning vocabulary by following the rank-frequency list, the known vocabulary will initially consists of vertices of $G_{[k_{max}]}$, where k_{max} is the largest degree in G, then one expands the vocabulary to $G_{[k_{max}-1]}$, followed by $G_{[k_{max}-1]}$, etc.

The reason why we used the prefix "pseudo" in the above definition is because the notion of k-core is much more often used in graph theory. The definition of $k-core$ is similar.

Definition 2 *For a non-negative integer k, the k-core of a graph is the maximal subgraph such that its vertices have degree greater or equal to k. By the "degree" in this definition we mean the degree of the vertex in the subgraph. If G is a given graph, we define $G_{\{k\}}$ to be the k-core of G.*

For $k = 0, 1, 2, \ldots$, subgraphs $G_{\{k\}}$ form a nested sequence of graphs where $G_{\{k+1\}} \subset G_{\{k\}}$. Construction of the sequence of k-cores is known as k-core decomposition [16]. There exists an algorithm for k-core decomposition [15, 16] with time complexity of $O(n + e)$, where n is the number of vertices in G and e is the number of edges of G.

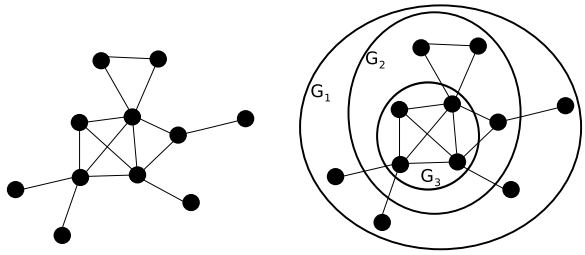

Figure 2. Example of k-core decomposition of a graph (after [15]).

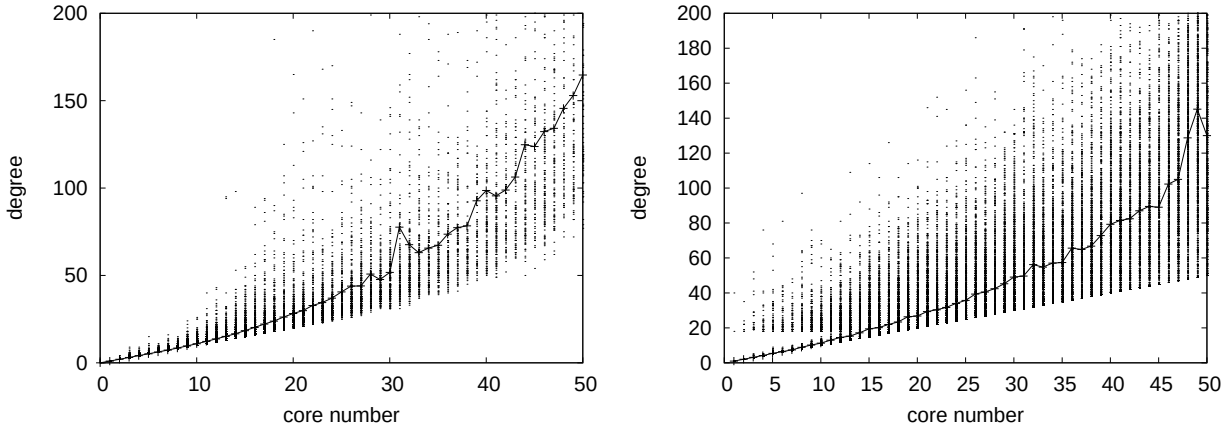

Figure 3. Degree of vertices vs. core number for $\mathcal{G}_{\mathcal{W}}$ (left) and $\mathcal{G}_{\mathcal{M}}$ (right). Solid line represents average degree of vertices with a given core number.

This means that even for very large graphs, k-cores can be computed in an efficient way. Figure 2 shows an example of a graph and its k-core decomposition.

The core (pseudocore) number of a given vertex is the number of the highest core (pseudocore) to which the vertex belongs. Set of all vertices with core (pseudocore) number k will be called the k-layer (k-pseudolayer). Note that the pseudocure number of a vertex of graph G is equal to its degree in G.

How different are cores and pseudocores? In Figure 3 we plot degrees of vertices versus core numbers for all vertices of $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$, together with average degree for a given core number, represented by the solid line. One can observe that the average degree is mostly an increasing function of the core number. This means that on average, vertices of high degree have high core number, and vertices of smaller degrees – smaller core numbers. Obviously, high-degree vertices are those which represent high-frequency words. This suggests that instead of considering "growth of vocabulary graphs" following pseudocores as described earlier, one could consider alternate strategy, learning first words belonging to the k-layer with the highest k, the words learned afterwords belong to $(k-1)$-layer, then to $(k-2)$-layer, etc. We will consider both ways (following cores and pseudocores), but in both cases the notion of the core or the pseudocore effectively replaces a dynamical process ("growing graph") by a static property of the graph, which can be investigated as a purely topological feature of large

graphs, without any reference to time.

4. Clustering spectra

As mentioned in the introduction, among quantities which are typically used to describe topology of large graphs, the clustering coefficient is particularly interesting. The clustering coefficient, originally introduced in [17], represents the average probability that two neighbours of a given vertex are also a neighbour of one another. More formally, given a vertex v of a graph G, let use denote by $N(v)$ the number of edges between vertices of v.

Definition 3 *The local clustering coefficient $C_v(G)$ is defined as*

$$C_v(G) = \frac{N(v)}{\binom{\deg(v)}{2}}$$

where $\deg(v)$ *is the degree of v, that is, the number of edges connected to v. The clustering coefficient of the entire graph G is then defined as $C(G) = |G|^{-1} \sum_v C_v(G)$, where $|G|$ denotes the number of vertices in G and the sum runs over all vertices of G.*

Obviously, the local clustering coefficient varies widely from node to node. In [18, 19] it has been argued that for networks exhibiting hierarchical organization, the local clustering coefficient $C(m)$ of a node with m links follows the scaling law

$$C(m) \sim m^{\gamma}, \tag{4}$$

where $\gamma = -1$. In order to check wheather this applies to $\mathcal{G}_{\mathcal{M}}$ and $\mathcal{G}_{\mathcal{M}}$ we plotted the local clustering coefficient of a vertex versus degree of that vertex for all vertices of $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$, as shown in Figure 4. The distribution of data points is rather wide, and one can only say that the upper boundary of the dataset seems to follow (rather roughly) $m^{-0.8}$ for $\mathcal{G}_{\mathcal{W}}$ and m^{-1} for $\mathcal{G}_{\mathcal{M}}$. While the correlation between $C(m)$ and m is not as strong as in some graphs reported in [18], nevertheless some signs of the behavior similar to eq. (4) are present. This may indicate that elements of hierarchical organization do exists in $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$.

Let us also remark that for a vertex with a given degree, the value of the clustering coefficient can vary quite significantly. To illustrate this, consider two vertices of $\mathcal{G}_{\mathcal{M}}$, corresponding to words *anxiously* and *tractor*. These two words have the same degree equal to 8, yet their local clustering coefficients are, respectively, 1 and 0.142857. The headword *anxiously* is connected to eight headwords: *impatiently, keenly, avidly, promptly, quickly, readily, with open arms, eagerly.* Each of the words from this set is connected to all others in the set, and therefore the local clustering coefficient of *anxiously* is equal to 1. The headword *tractor* is also connected to eight words, which are *machinery, automobile, pusher, tank, truck, creeper, duck, amphibian.* Among these eight words, however, there are only four pairs of direct neighbours, namely *machinery – pusher, automobile – machinery, tank – truck* and *duck – truck.* As a result, the local clustering coefficient of *tractor* is equal to $\frac{4}{28} = 0.142857$.

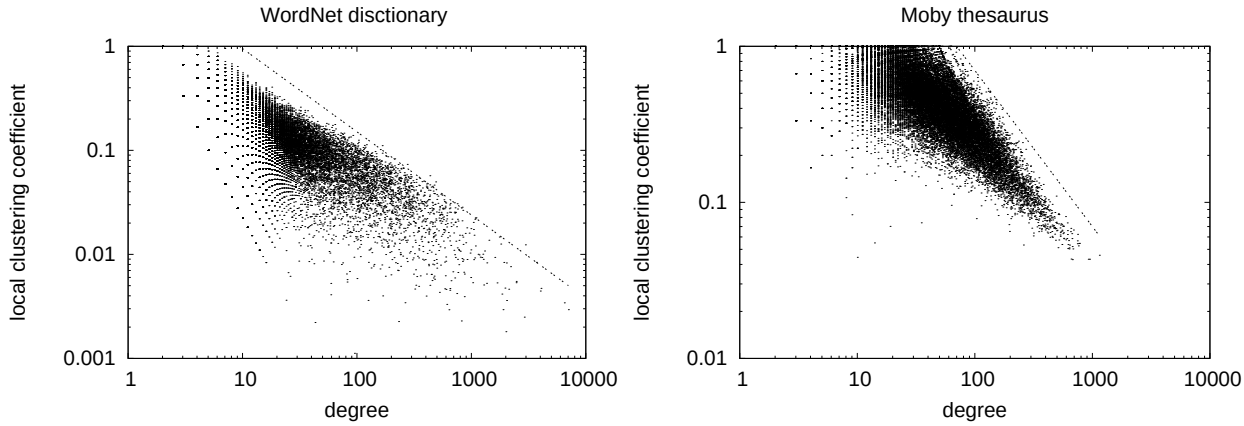

Figure 4. Plot of the local clustering coefficient as a function of the degree of a vertex for all vertices of $\mathcal{G}_{\mathcal{W}}$ (left) and $\mathcal{G}_{\mathcal{M}}$ (right). Dashed line in the case of $\mathcal{G}_{\mathcal{W}}$ (left) has slope -0.8, and for $\mathcal{G}_{\mathcal{M}}$ (right) -1.0.

The local clustering coefficient can be understood as the ratio of the number of edges that exist in the neighbourhood of v to the maximum number of edges that could potentially exist in that neighbourhood of v, which happens to be $\binom{\deg(v)}{2}$. The clustering coefficient $C(G)$ of the whole graph G is obtained by averaging $C_v(G)$ over all vertices v belonging to G. Clustering coefficient of the graph is a measure of the "cliquishness" of the graph. One can say that $C_v(G)$ characterizes local "cliquishness" at vertex v, while $C(G)$ characterizes global "cliquishness" of the entire graph. In practice, however, the local clustering is too detailed to be useful, simply because we have as many $C_v(G)$ numbers as vertices in the graph. The global clustering, on the other hand, is too coarse, being just one scalar value for the entire graph. We will now show how to construct an intermediate characterization of clustering, which lies (in terms of usefulness) somewhere between "microscopic" $C_v(G)$ and "macroscopic" $C(G)$. It will be called "core/pseudocore clustering spectrum".

Definition 4 *A set of pairs* $(|G_{\{k\}}|, C(G_{\{k\}}))$*, where* $|G|$ *denotes the number of vertices of* G*, will be called core clustering spectrum of* G*. Similarly, a set of pairs* $(|G_{[k]}|, C(G_{[k]}))$ *will be called pseudocore clustering spectrum. The value of* k *in ranges from* 1 *to* k_{max}*, where* k_{max} *is the largest* k *for which, respectively,* $G_{\{k\}}$ *or* $G_{[k]}$ *is non-empty.*

We will visualize the core clustering spectrum by plotting points $(|G_k|, C(G_k))$ on a plane. The value of k will range from 1 to k_{max}, where k_{max} to the largest k for which G_k is non-empty.

For some graphs, the core clustering spectrum is very narrow, meaning that the number of points in the spectrum is small. This is the case, for example, for classical Erdös-Rényi random graphs. In other cases, the spectum may be quite wide, as we will see in subsequent sections.

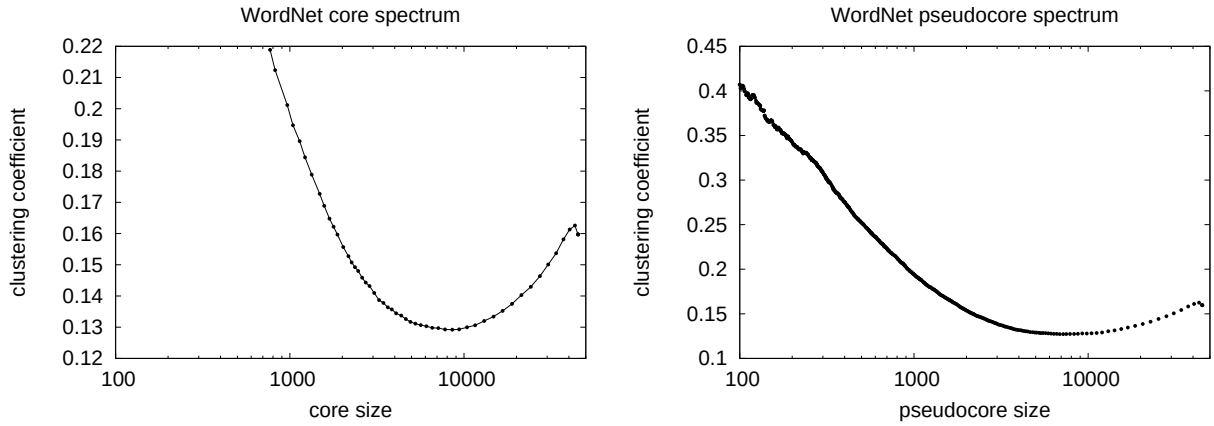

Figure 5. Core and pseudocore clustering spectrum of $\mathcal{G}_{\mathcal{W}}$.

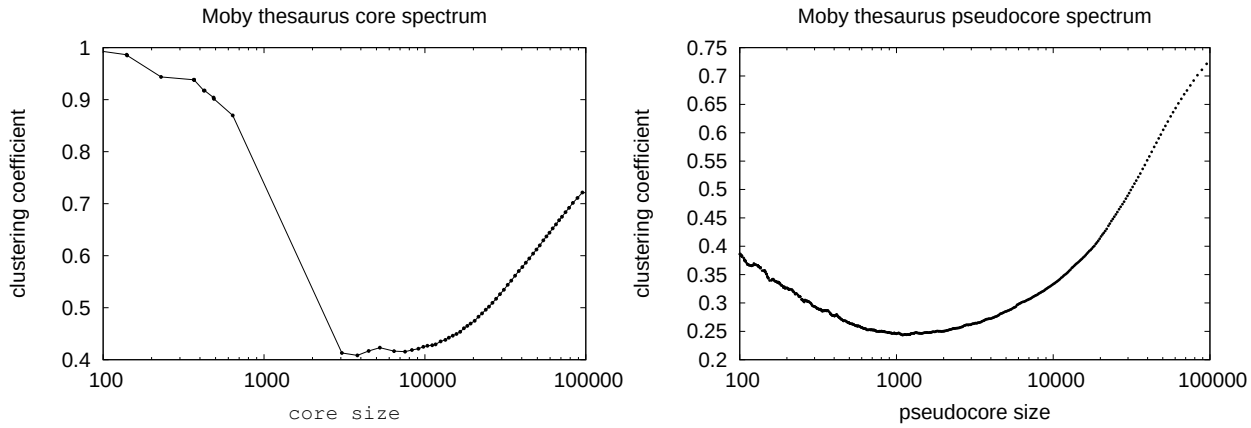

Figure 6. Core and pseudocore clustering spectrum of $\mathcal{G}_{\mathcal{M}}$.

5. Characterization of the structure

Both core and pseudocore clustering spectra have been computed for $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$. Results are shown in Figures 5 and 6. It is rather remarkable that all four graphs exhibit well-defined minima occurring somewhere between the core size 10^3 and 10^4. An immediate question which presents itself after inspection of these spectra is: are there any known random graph models which would exhibit similar U-shaped spectra? We computed spectra of classical random graphs, Barabasi-Albert random graphs with a variety of parameters, "power law cluster graph", GNP graph, and several others. None of them exhibits a minimum in the spectrum, and most of the time their spectra are monotonic functions of the core size.

We also generated random graphs with the same degree distribution as $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$ using the so-called configuration model [20] as well as using Havel-Hakimi algorithm [21]. To be precise, we used fitted functions given by eq. (2) as degree distributions.

Using the first of these methods, one initially creates a degree sequence drawn from the distribution (2). Vertices are created with stubs for attaching edges, such that the

number of stubs is equal to the degree of the vertex. Two randomly selected available stubs are connected with an edge, and this procedure is repeated until all stubs are exhausted.

Another method for constructing a random graph with a given degree sequence is known as Havel-Hakimi algorithm [21]. The algorithm creates the desired graph by successively connecting the node of highest degree to other nodes of highest degree, resorting remaining nodes by degree, and repeating the process.

We used both methods to create random graphs of the same size as $\mathcal{G}_{\mathcal{W}}$ and $\mathcal{G}_{\mathcal{M}}$ using the fitted eq. (2) as the degree distribution. We found that in spite of the "right" degree distribution, core spectra of these graphs do not resemble the dictionary graph spectrum at all. In both cases, clustering coefficient decreases with the growing core size, and no minimum is present.

6. Toward the model

In [11], it has been demonstrated that thesaurus dictionaries share some of the statistical properties of low-dimensional (d=2) Euclidean (geometric) graphs. For this reason, in the search for a simple model of dictionary graphs, we next turned our attention to geometric graphs.

Geometric random graph [22] is a type of random graph which is constructed by placing vertices at random uniformly and independently on the unit square. Vertices u, v are connected if and only if the distance between them is less or equal than a given threshold r, that is, when $d(u, v) \leq r$. The distance $d(u, v)$ is often computed assuming periodic boundary condition, in which case the unit square effectively becomes a torus.

Clustering spectrum of a geometric graph defined above does not, unfortunately, exhibit any minimum, so the normal geometric random graph cannot serve as a model of the dictionary graph.

Let us, however, consider a natural generalization of the geometric random graph, in which the parameter r is vertex-dependent. To be precise, we place vertices at random uniformly and independently on the unit torus. Vertices are numbered by an index i ranging from 1 to n. Each vertex has its own "range parameter" $r(i)$. Two vertices labelled i and j are connected if and only if $d(i, j) \leq r(i)$ or $d(i, j) \leq r(j)$, that is, when one of them is within the range of the other.

Suppose now that $r(i)$ is an increasing function of i. This would mean that vertices with large i have large range, and are likely to be connected to a larger number of other vertices than those with small i. This is precisely what we would want if vertices represented words of the language, and i was the reversed order in which the words are learned. The words one learns first are the high-frequency words, and in the dictionary graph they should be linked to large number of other words. One would therefore expect that a generalized geometric random graph with increasing $r(i)$ might have properties similar to the dictionary graph.

We considered a simple form of $r(i)$, chosen rather arbitrarily. Let n be the desired

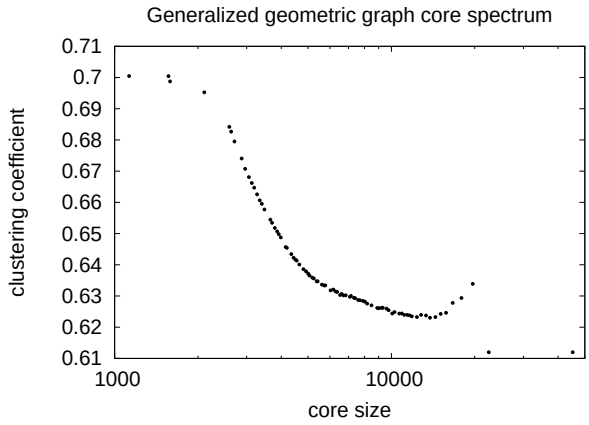

Figure 7. Core clustering spectrum of generalized geometric random graph with $\gamma = 5$ and the same number of edges and vertices as $\mathcal{G}_{\mathcal{W}}$.

number of vertices in the generalized geometric random graph, and m be the desired number of edges. We take

$$r(i) = \lambda \left(\frac{i}{n}\right)^{\gamma}, \tag{5}$$

where $\gamma > 0$. The constant λ is determined by the requirement that the total number of edges should be equal to m, meaning that

$$\frac{1}{2} n\pi \sum_{i=1}^{n} r(i)^2 = m. \tag{6}$$

The factor $1/2$ appears in front of the sum since all edges are counted twice. This leads to

$$\lambda = \sqrt{\frac{2m}{n\pi n^{2\gamma}}} \left(\sum_{i=1}^{n} i^{2\gamma}\right)^{-1/2}. \tag{7}$$

Approximating the sum by integral, after integration we obtain

$$\lambda \approx \sqrt{\frac{2m(1 - n^{-1-2\gamma})}{(1+2\gamma)\pi}}. \tag{8}$$

Using this form of $r(i)$, we generated a number of graphs, to be called generalized geometric random graphs, using different values of parameter γ, and we computed their core clustering spectra. A typical core clustering spectrum of a generalized geometric random graph is shown in Figure 7, for the graph with $\gamma = 5$ and the same number of edges and vertices as $\mathcal{G}_{\mathcal{W}}$. Once can see that the spectrum indeed exhibits well-defined minimum, occurring roughly around 15000. Obviously, it is still far from the spectrum of $\mathcal{G}_{\mathcal{W}}$ shown in Figure 5, yet the general shape is quite similar. One clear difference is that in Figure 5 there is only a slight drop in the clustering coefficient at the right end of the spectrum, while this drop is much larger in Figure 7 (last two points of the spectrum). This is due to a large number of isolated components, which, in the case of generalized geometric graph, are simply isolated vertices, which belong to the outermost

core and significantly contribute to the drop in the average clustering observed at the right end of the spectrum.

7. Interpretation and conclusions

We demonstrated that core clustering spectra of dictionary graphs posses some features which are not present in commonly studied models of random graphs, with he exception of generalized geometric random graphs. This indicates that the topological structure of dictionary graphs may be mathematically very interesting, and the connection with geometric graphs needs to be further explored, as it may shed some more light on dictionary graph topology.

Are there, however, any practical implications of these findings? There might be, since our results suggest two strategies for learning vocabulary of a foreign language:

(i) we start with words which are most frequent, thus have the highest pseudocore number, progressing toward less frequent words.

(ii) we start with words of the highest core number, progressing toward smaller core numbers.

Is there any advantage of the second strategy? We argue that there might be some. Consider again Figure 6. Once can see that while the shape of both core and pseudocore spectrum is similar in the sense that they both fist decrease, and then start increasing with the growing subgraph size. Nevertheless, for a given subgraph size, the clustering coefficient is higher for the core than for the pseudocore. This means that the strategy (ii) allows one to maintain higher clustering coefficient than (i), yet on average, more frequent words are still learned at the beginning, and less frequent words later. Maintaining high clustering coefficient means that newly learned words are placed in areas of semantic space which are connected to cliques of known words, so that one ventures into new territories without straying too far from well-known areas. This may result in slightly smaller text coverage, but, on the other hand, higher "coherence" of the vocabulary, and perhaps better grasp of the meaning of newly learned words.

An equally interesting problem arises from the possible connection of dictionary graphs with geometric graphs. On one hand it should be possible to tune the construction of generalized geometric graphs to find a model with clustering spectrum (and other properties) more closely resembling $\mathcal{G}_{\mathcal{M}}$ or $\mathcal{G}_{\mathcal{W}}$, thus producing a more accurate model of these graphs, by choosing, for example, a different form of $r(i)$ or by changing the way the graph is generated.

But on the other hand one can also go in the opposite direction. If the topological structure of dictionary graphs somewhat resembles low-dimensional generalized geometric graphs, it may be possible to embed $\mathcal{G}_{\mathcal{M}}$ or $\mathcal{G}_{\mathcal{W}}$ in low-dimensional Euclidean space in such a way that the distribution of edge lengths is similar to what is implied by (5). Coordinates of a given word in this space could then have some

interesting linguistic meaning. Work in this direction is ongoing and will be reported elsewhere.

Acknowledgments

One of the authors (H.F.) acknowledges financial support from NSERC (Natural Sciences and Engineering Research Council of Canada) in the form of the Discovery Grant. Calculations of clustering spectra and other graph-related quantities were performed using Igraph library [23] available under GNU General Public Licence.

References

[1] R. L. Trask. *Language: the basis*. Routledge, New York, 1999.

[2] I. S. P. Nation. *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge, 2001.

[3] R. Goulden, P. Nation, and J. Read. How large can a receptive vocabulary be? *Applied Linguistics*, 11:341–363, 1990.

[4] J. B. Carrol, P. Davies, and B. Richman. *The American Heritage Word Frequency Book*. Houghton Mifflin, New York, 1971.

[5] B. Laufer. How much lexis is necessary for reading comprehension? In P. J. L. Arnaud and H. Béjoint, editors, *Vocabulary and Applied Linguistics*, pages 126–132. Macmillian, London, 1992.

[6] H. Fukś and C. Phipps. Toward a model of language acquisition threshold. In E. Wamkeue, editor, *Proceedings of the 17th IASTED International Conference on Modelling and Simulation*, page 263. Acta Press, 2006.

[7] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998. http://wordnet.princeton.edu.

[8] Grady Ward. The moby lexicon project. http://icon.shef.ac.uk/Moby.

[9] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[10] A. E. Motter, A. P. S. de Moura, Y. C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Phys. Rev. E*, 65, 2002.

[11] O. Kinouchi, A. S. Martinez, G. F. Lima, G. M. Lourenco, and S. Risau-Gusman. Deterministic walks in random networks: an application to thesaurus graphs. *Physica A*, 315:665–676, 2002.

[12] A. D. Holanda, I. T. Pisa, O. Kinouchi, A. S. Martinez, and E. E. S. Ruiz. Thesaurus as a complex network. *Physica A-Statistical Mechanics And Its Applications*, 344:530–536, 2004.

[13] Mariano Sigman and Guillermo A. Cecchi. Global organization of the wordnet lexicon. *PNAS*, 99(3):1742–1747, February 2002.

[14] C. Tsallis and M. P. de Albuquerque. Are citations of scientic papers a case of nonextensivity? *Eur. Phys. J. B*, (13):777–780, 2000.

[15] J. I. Alvarez-Hamelin, L. DallAsta, A. Barrat, and A. Vespignani. K -core decomposition : a tool for the visualization of large scale networks. *Advances in Neural Information Processing Systems*, 18:41, 2006. arxiv.org, cs.NI/0504107.

[16] V. Batagelj and M. Zaversnik. Generalized cores. CoRR, cs.DS/0202039, 2002.

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

[18] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67(2):026112, Feb 2003.

[19] S. N. Dorogovtsev and J. F. F. Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1485):2603–2606, 2001.

[20] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[21] G. Chartrand and L. Lesniak. *Graphs and Digraphs*. Chapman and Hall/CRC, 1996.

[22] Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

[23] Gábor Csárdi. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, 2006.