# CORE DECOMPOSITION SPECTRA OF LARGE GRAPHS AND THEIR APPLICATIONS IN MODELLING

Henryk Fukś and Mark Krzeminski
Department of Mathematics
Brock University
St. Catharines, ON, Canada
email: hfuks@brocku.ca

**ABSTRACT**
Systems with large number of interacting components are often modelled by random graphs and networks. In models of this type, one frequently needs to characterize graph clustering at both local and global level. We propose a method of characterization of clustering in large graphs and networks using the concept of $k$-core decomposition. The plot of clustering coefficient of $k$-core versus size of $k$-core will be called *the spectrum of clustering coefficients*. We show that $k$-core spectrum may play an important role in language graphs, such as graphs constructed from language dictionaries, where it can be used to describe some dynamical phenomena by purely static, topological quantities. In the last part of the paper, we propose a random graph model of a dictionary graph for which the $k$-core spectrum has similar features as in real dictionary graphs. The model is based on generalization of geometric random graphs in which the range parameter varies from vertex to vertex.

**KEY WORDS**
Mathematical modelling, complex networks, core decomposition, language modelling

## 1 Introduction

In recent years, large graphs and networks with complex topological structure became one of the leading paradigms of mathematical modelling [5]. A large variety of natural and technological phenomena can be described and modelled using the frameworks of graph theory and network theory [10].

In many cases, if we want to investigate a complex system with large number of interacting components, we need a good "null model", that is, a model which is as simple as possible, yet it captures the desired features of the real system as closely as possible. Obviously, the notion of "as simple as possible" need to be defined first. In the case of large graphs and networks, we often need to construct a model which has somewhat similar topology as the system under consideration. Various quantities characterizing structure of large networks have been introduced, including shortest path related functions, functions characterizing graph components, centrality measures, spectral properties, and many others. Among these quantities, one of the most important and most frequently used is the clustering coefficient or transitivity [5]. Clustering coefficient can be defined either locally (for each vertex) or globally (as average of local clustering coefficients). The local clustering is not very convenient to use, as it is typically a very large vector. Averaging it over the entire graph discards too much information, and proposed different averaging schemes, such as weighted averaging, are still not very satisfactory. In what follows, we will introduce a measure of clustering which lies somewhere between the local and the global level. We will then show an application of this concept in a model of language dictionary graph.

## 2 Core decomposition and clustering spectra

The clustering coefficient was originally introduced in [12]. It 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 $G$, the local clustering coefficient is defined as

$$C_v(G) = \frac{\text{number of edges between neighbours of } v}{\binom{\deg(v)}{2}},$$

where $\deg(v)$ is the degree of $v$, that is, the number of edges connected to to $v$. Clustering coefficient can thus 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 then defined as the average of $C_v(G)$ 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". 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 usability) somewhere between "microscopic" $C_c(G)$ and "macroscopic" $C(G)$.

Before we do this, we will first introduce the notion of $k$-core. For a non-negative integer $k$, $k$-core of a graph

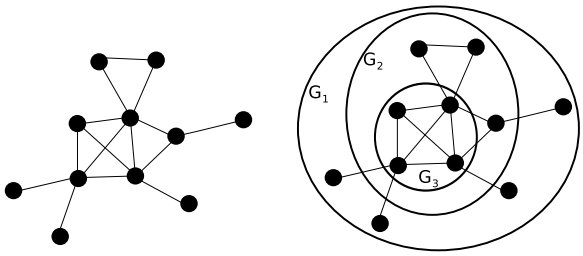


Figure 1. Example of $k$-core decomposition of a graph (after [3]).

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 [4]. There exists an algorithm for $k$-core decomposition [3, 4] with time complexity of $O(n + e)$, where $n$ is the number of vertices in $G$ and $e$ is the number of edges of $G$. This means that even for very large graphs, $k$-cores can be computed in an efficient way. Figure 1 shows an example of a graph and its $k$-core decomposition.

We are now ready to define the promissed alternative tool for characterization of clustering, to be called "$k$-core spectrum of clustering coefficients". It will be defined as a set of pairs $(|G_k|, C(G_k))$, where $|G|$ denotes the number of vertices of $G$. We will visualize the $k$-core spectrum of clustering coefficients 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 spectrum is very narrow, meaning that $k_{max}$ is rather small, and the number of points in the $k$-core 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.

## 3 Clustering spectrum of a dictionary graph

In [8], the authors studied properties of a graph constructed from a large dictionary of English language available from Project Gutenberg web site, also known as *The Gutenberg Webster's Unabridged Dictionary* [9]. The dictionary has been converted into a large graph $\mathcal{G}$ with vertices representing individual words. If one word occurs in the definition of another word, then these two words (vertices) are linked with an edge. The resulting graph has about $10^5$ vertices and $10^6$ edges, and rather complicated topology, which is not fully understood yet. It has been observed that the degree distribution of the dictionary graph, as shown in Figure 2, is well approximated by the curve

$$P(d) = d^{-\alpha} \exp(A - B\, d^{-\beta}) \qquad (1)$$

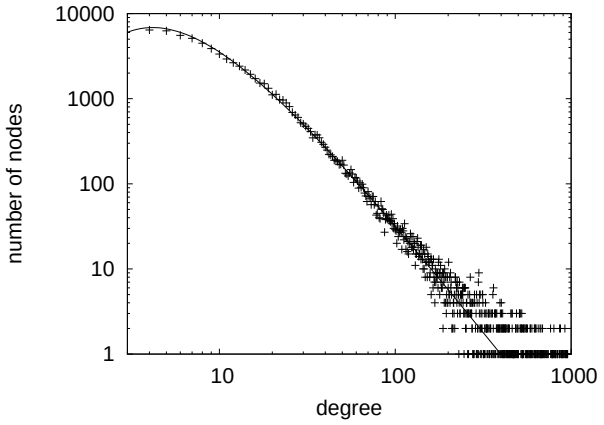


Figure 2. Degree distribution of the dictionary graph.

where $P(d)$ is the number of vertices with degree $d$, $\alpha = 2.61075$, $A = 15.6987$, $B = 10.2904$, and $\beta = 0.83179$. Note that this is just a convenient empirical curve, which we will use to describe the degree distribution of the dictionary graph, and that it has no other special meaning or theoretical justification.

Using the dictionary graph, one can build a simplistic model of second language acquisition, as proposed in [8]. The person learning English as a second language knows at a given moment only a subset of all words represented by vertices of $\mathcal{G}$. Let $W$ denote the set of known words, and let $\mathcal{G}/W$ be the subgraph of $G$ generated by $W$. Obviously, as $W$ grows, $\mathcal{G}/W$ grows as well. An important feature of the learning process is that the words which are more frequently encountered are learned first, while the more specialized and rare words are learned later. It is possible to rank words of English (or other) language by frequency of their occurrence in a large text corpus, and then assume that the words are roughly speaking learned in order of their appearance on the list. This assumption, although very crude, reflects the basic mechanism of second language learning.

The learning process, therefore, can be modelled by a growing graph $\mathcal{G}/W$, such that vertices are being added to it in order given by the rank-frequency list. It has been discovered that the clustering coefficient of graph $\mathcal{G}/W$ initially decreases as the graph grows, reaches a minimum when $W$ consists of approximately 3000 words, and then increases again. The location of this minimum appears to be in roughly the same place as other special points in language learning, as discussed in [8]. The exact reason for this is not currently known.

The growing graph model of language acquisition is dynamical in the sense that it involves time. The existence of a minimum of the clustering coefficient has been originally formulated as a dynamical question too. We will now show that one can use the notion of $k$-core spectrum to reformulate the model in such a way that it won't involve time any more. The existence of the minimum of the clustering coefficient will then become a topological property

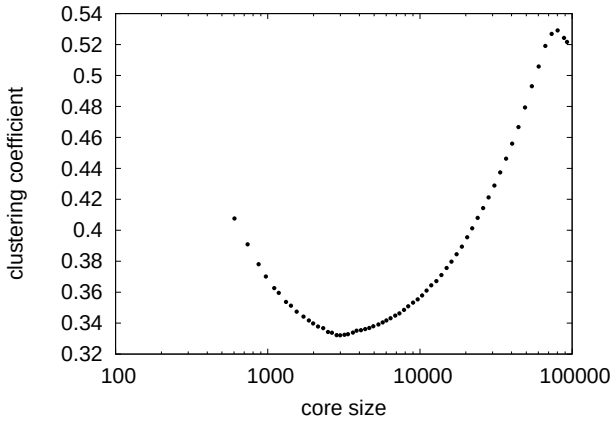


Figure 3. $k$-core spectrum of clustering coefficients for the dictionary graph.

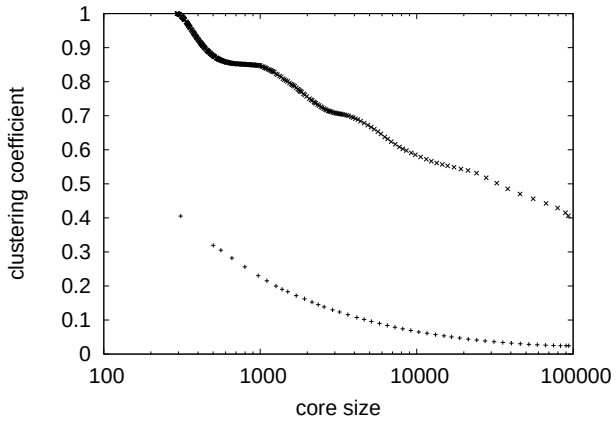


Figure 4. $k$-core spectrum of clustering coefficients for configuration model graph ($+$) and the graph obtained with Havel-Hakimi algorithm ($\times$).

of the dictionary graph instead of a dynamical property.

We found that the position of the word in the rank-frequency list is very highly correlated with its coreness. By coreness we mean the number of the highest $k$-core to which the word belongs. The most frequent words have generally high coreness, and the rare words have low coreness. We can, therefore, approximately assume that the learner first learns words belonging to $k_{max}$-core of $\mathcal{G}$, then words of $(k_{max}-1)$-core, and so on. Words central to the language are learned first, and consecutive vocabulary layers are being added as the learning progresses.

The $k$-core spectrum of clustering coefficients for the dictionary graph is shown in Figure 3. We immediately notice a rather remarkable feature, namely, it exhibits a very well-defined minimum when the core size reaches approximately 3000, precisely in the same place as mentioned earlier. However, now this minimum is just a property of the dictionary graph, and we do need to refer to any dynamical process to describe this phenomenon.

Eventhough the dictionary graph has somewhat "random" appearance when one tries to visualize it, and its degree distribution exhibits a power law characteristic to some random graph models, its $k$-core spectrum is unlike the spectrum of any other random graph. We computed spectra of classical random graphs, Barabasi-Albert random graphs with 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.

## 4 Random graph models

Since standard models of random graphs do not have the desired minimum in their $k$-core spectrum, the next step is to generate random graphs with exactly the same degree distribution as the dictionary graphs. It is rather straightforward to produce a random graph with the degree distribution given by eq. (1). One could reasonably suspect that such random graph could serve as "zero-th order" model of the dictionary graph. The simplest method to create a random graph with a given degree distribution is so-called configuration model [10]. Using this method, one first creates a degree sequence of length $N$ drawn from the distribution (1). Then $N$ vertices are created with stubs for attaching edges, such that the number of stubs equals to the degree of the vertex. We connect two randomly selected available stubs with an edge, and repeat this procedure until all stubs are exhausted.

Another method for constructing a random graph with a given degree sequence is known as Havel-Hakimi algorithm [6]. 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 the dictionary graph and having the degree distribution given by eq. (1). This was done using NetworkX package [1]. We then computed $k$-core spectrum of clustering coefficients for the resulting graphs using igraph library [7]. The results are shown in Figure 4. 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.

## 5 Generalized geometric random graphs

Failure to produce the desired $k$-core spectrum using methods described in previous sections suggests that a different approach is needed. We will now describe a new random graph model, based on the idea of geometric graphs.

Geometric random graph [11] is a type of a random graph which is constructed as follows. We first place vertices at random uniformly and independently on the unit square. Then we connect two vertices, $u$, $v$, if and only

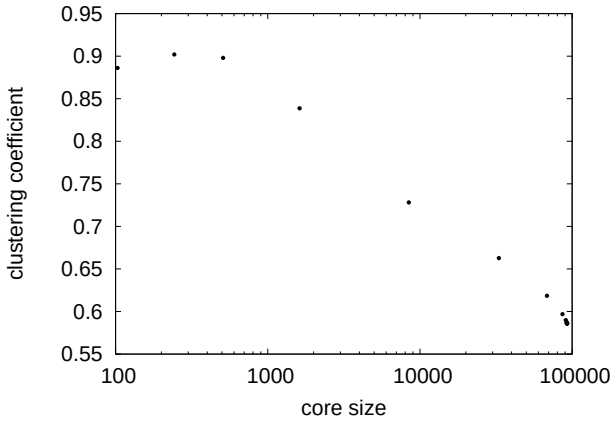


Figure 5. $k$-core spectrum of clustering coefficients for the geometric graph with $r = 0.00914$.

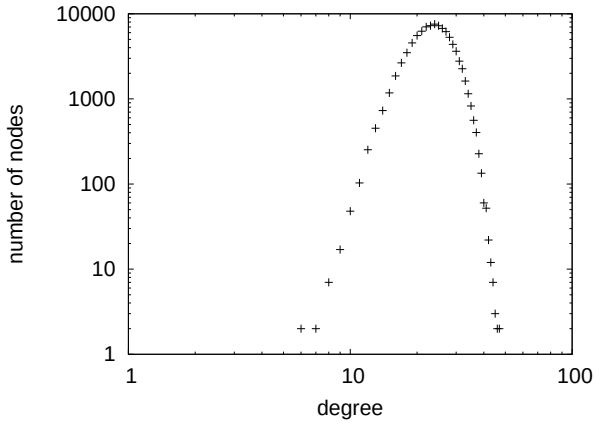


Figure 6. Degree distribution for the geometric graph with $r = 0.00914$.

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. Figures 5 and 6 show respectively $k$-core spectrum and degree distribution of a geometric graph with the same number of vertices and edges as the dictionary graph, corresponding to $r = 0.00914$. As we can see, the degree distribution is far from the power law, and the spectrum does not exhibit any minimum. Clearly, the normal geometric random graph cannot serve as a model of the dictionary graph.

We will now propose a natural generalization of the geometric random graph, in which the parameter $r$ is not constant, but varies from vertex to vertex. To be precise, we place vertices at random uniformly and independently on the unit torus. Vertices are numbered by 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.

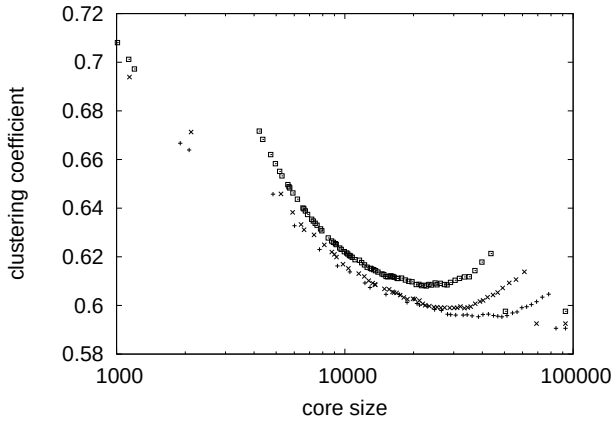


Figure 7. $k$-core spectrum of clustering coefficients for the generalized geometric graph with $r(i)$ defined by eq. (2) with $\gamma = 1$ (+), $\gamma = 2$ (×), and $\gamma = 4$ (□)

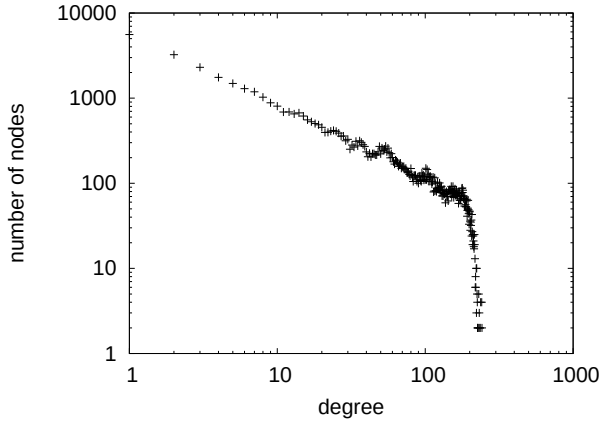


Figure 8. Degree distribution for the generalized geometric graph with $r(i)$ defined by eq. (2) with $\gamma = 4$.

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.

To test this hypothesis, we considered a simple form of $r(i)$. Let $n$ be the desired number of vertices in the generalized geometric random graph, and $m$ be the desired number of edges. We take

$$r(i) = \lambda \left(\frac{i}{n}\right)^{\gamma}, \tag{2}$$

where $\gamma > 0$. The constant $\lambda$ is determined by the require-

ment that the total number of edges should be equal to $m$, meaning that

$$\frac{1}{2}n\pi\sum_{i=1}^{n} r(i)^2 = m. \tag{3}$$

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

$$\lambda = \sqrt{\frac{2m}{n\pi n^{2\gamma}}}\left(\sum_{i=1}^{n} i^{2\gamma}\right)^{-1/2}. \tag{4}$$

Approximating the sum by integral, after integration we obtain

$$\lambda \approx \sqrt{\frac{2m(1-n^{-1-2\gamma})}{(1+2\gamma)\pi}}. \tag{5}$$

Figure 7 shows $k$-core spectrum of clustering coefficients for the generalized geometric graph with $r(i)$ defined by eq. (2) for several values of the exponent $\gamma$. We can see that even in the linear case, that is, for $\gamma = 1$, the spectrum exhibits clear and well defined minimum. When $\gamma = 4$, the minimum occurs roughly when the size of $k$-core is approximately equal to 20000. This is well beyond the minimum in the dictionary graph, which occurs at 3000. Nevertheless, this appears to be the first random graph model known to us which possesses a minimum in the spectrum. The degree distribution of the generalized geometric graph with $\gamma = 4$ is shown in Figure 8, and as one can see, it shows features of a power law, similarly to the dictionary graph.

## 6 Why geometric?

Eventhough the spectrum of the generalized geometric graph shown in Figure 7 is not identical with the spectrum of the dictionary graph, we believe that is should be possible to find another function $r(i)$ which would produce a graph with closely matching spectrum. In fact, when in (2) we set $\gamma = 20$, $k$-core spectrum of the resulting generalized geometric graph has a local minimum located around $6000$, as shown in Figure 9. Preliminary results indicate that further one may need to increase the dimensionality of space to further improve the match.

The intriguing question is this: why geometric graphs seem to be the best models of dictionary graphs? Where does the "geometric" part come from? A possible, although quite speculative explanation can be formulated if one assumes that words occupy regions of some abstract space, to be tentatively called "semantic space". Words which are frequently used are likely to occupy large volume of this space, as they have wide meaning and can be used in many different contexts. Highly specialized words, on the other hand, are less frequently used and more narrowly defined, so one can assume that they occupy smaller volumes of the semantic space. When the volume occupied by one word overlaps with another word (such as when one of them is needed to define the other), then we connect them with an edge, obtaining a graph with a topology similar to the dictionary graph.

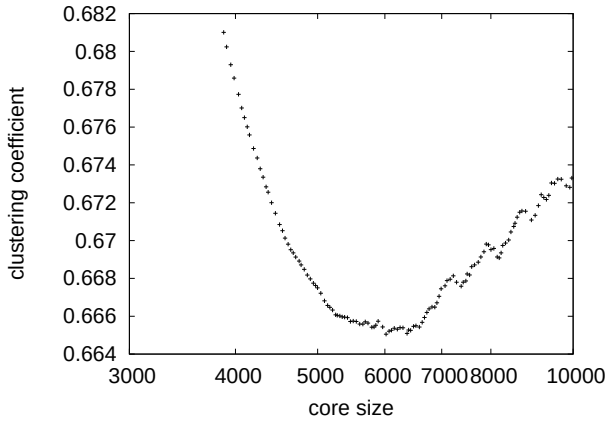


Figure 9. Part of $k$-core spectrum of clustering coefficients for $r(i)$ defined by eq. (2) with $\gamma = 20$.

## 7 Conclusions and further work

The concept of $k$-core spectrum of clustering coefficients is a useful way of characterizing clustering of random graphs, allowing to describe the the "middle ground" between the local and the global clustering. We demonstrated that in the case of the dictionary graph, $k$-core spectrum of clustering coefficients has a natural interpretation as series of clustering coefficients of a growing vocabulary graph. We were also able to devise a model, based on generalized geometric random graph, for which the spectrum behaves in a similar way as the spectrum of the dictionary graph.

We should point out that the shape of the $k$-core spectum for dictionary graphs appears to be independent of the dictionary itself. In fact, it appears to be independent of the language as well. Figure 10 shows the spectrum for a of French language dictionary graph, constructed from *Dictionnaire de l'Académie Française, 6th Edition* [2]. This graph is smaller that the Webster graph, having $28238$ vertices and $790730$ edges, but the shape of its spectrum still closely resembles Figure 3.

More work needs to be done to improve the model. We are currently performing a systematic search of functions $r(i)$, trying to find a function which would produce $k$-core spectrum closer to the spectrum of the dictionary graph. We are also attempting to calculate the $k$-core spectrum for various random graph models rigorously, without resorting to numerical computations. Among other questions, a question of particular interest is what conditions must $r(i)$ satisfy for the generalized geometric random graph to have a minimum in its spectrum. This issue is currently under investigation and will be reported elsewhere.

## 8 Acknowledgements

The first author (H.F.) acknowledges partial financial support in the form of Discovery Grant from the Natural Sciences and Engineering Research Council of Canada

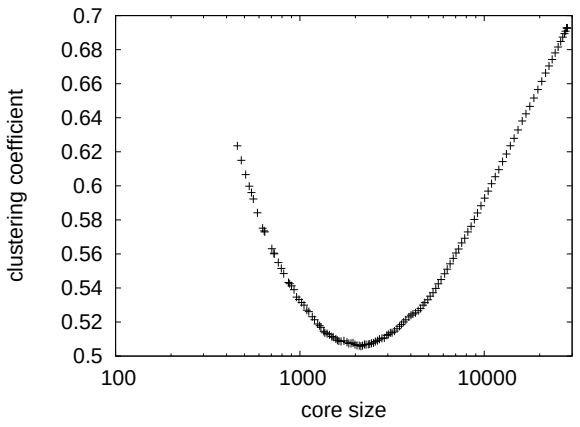


Figure 10. $k$-core spectrum of clustering coefficients for the French dictionary graph.

(NSERC). Authors benefitted from work of Bryan Penfound (supported by BUSRA award) and Jeff Haroutunian (supported by NSERC USRA award) who, respectively, performed numerical investigations of core decomposition and simulations of geometric random graphs.

## References

[1] NetworkX, Python package for analysis of complex networks. https://networkx.lanl.gov.

[2] *Dictionnaire de l'Académie Française, 6th Edition*. 1835. Electronic version courtesy of Mark Olsen, ARTFL Project, University of Chicago.

[3] 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.

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

[5] S. Bornholdt and H. G. Schuster, editors. *Handbook of Graphs and Networks*. Wiley-VCH, Weinheim, 2003.

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

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

[8] 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.

[9] Project Gutenberg. *The Gutenberg Webster's Unabridged Dictionary*. Plainfield, N.J, 1996. http://www.gutenberg.org/etext/673.

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

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

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