Exploring the random genesis of co-occurrence graphs

Jens Gustedt, Pedro Schimit, Hari K. Raghavan

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Exploring the random genesis of co-occurrence graphs

Jens Gustedt — Pedro Schimit — Hari K. Raghavan

N° 7186

January 2010

Domaine 3

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7186--FR+ENG

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche NANCY - GRAND EST

Exploring the random genesis of co-occurrence graphs

Jens Gustedt, Pedro Schimit, Hari K. Raghavan

Domaine : Réseaux, systèmes et services, calcul distribué
Équipe-Projet AlGorille

Rapport de recherche n° 7186 — January 2010 — 22 pages

Abstract: Using the network random generation models from Gustedt [2009], we simulate and analyze several characteristics (such as the number of components, the degree sequence and the clustering coefficient) of the generated networks. This is done for a variety of distributions (fixed value, Bernoulli, Poisson, binomial) that are used to control the parameters of generation process. These parameters are in particular the size of newly appearing sets of objects, the number of contexts in which new elements appear initially, the number of objects that are shared with 'parent' contexts, and, the time period inside which a context may serve as a parent context (*aging*). The results show that these models allow to fine-tune the generation process such that the graphs adopt properties as they can be found in real world graphs.

Key-words: clustering coefficient graph network random generation

Centre de recherche INRIA Nancy – Grand Est
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Explorer la génèse randomisée de graphes de co-apartenance

Résumé : En utilisant le modèle de génération de réseaux de Gustedt [2009], nous simulons et analysons plusieurs caractéristiques des réseaux (comme le nombre de composantes connexes, la séquence des degrés et le coëfficient de clustering). Nous faisons ceci pour une variété de lois de probabilité (valeur fixe, Bernoulli, Poisson, binomiale) qui sont utilisés pour contrôler les paramètres du processus de génération. Ces paramètres sont en particulier la taille des ensembles d'objets où apparaissent initialement, le nombre d'objets qui sont partagés avec les contextes 'parents', et, la période durant laquelle un contexte peut servir en tant que context parent (*vieillissement*). Nos résultats prouvent que ces modèles permettent de finement calibrer le processus de génération pour que les graphes adoptent des propriétés similaires à ceux qui sont observés pour des graphes réels.

Mots-clés : coefficient regroupement graphique réseaux génération aléatoire

1 Introduction and Overview

Application graphs that follow a vein of 'hidden cliques covers defining a graph' have been identified by Guillaume and Latapy [2004]: protein-protein interaction networks, the core network of the Internet, web connections (http links), the co-starring relation among film actors and the co-occurrence relation of words in sentences.

In Gustedt [2009] a family of random models that covers the inter-relationship between those cliques (called *contexts* in the sequel) has been proposed: each newly introduced context depends on a previously one from which part of the objects and their connections are copied. For the graph of co-authorships, e.g, a new paper often emerges from a previous one by slightly modifying the list of authors, some people cease contributing for the new one, others, such as experts of a particular subdomain or new PhD students join in.

Such a dependency had been completely neglected in classical random graph models, such as promoted by Erdős and Rényi [1960], where all edges occur independently from each other. Driven by observations in application domains models have been investigated that take the bias of dependency of choices into account. They have in particular been boosted by the work of Barabási and Albert [1999] for so-called *preferential attachment models*, see also Latapy [2007] for an overview, or Dorogovtsev and Mendes [2003] for a textbook. Another important model that takes such an historic bias into account is the *node duplication model*, see Bhan et al. [2002]. Many variants of these schemes have been proposed in the literature, an overview of these can be found in de Silva and Stumpf [2005].

In this paper, we use our new family of models for the random generation of large sparse graphs that are induced by the co-occurrence of 'objects' in 'contexts', see Section 2 for the terms. In particular, this family of models contains classical random graphs, the preferential attachment model and the duplication model as special cases. As the later two, the generated graphs depend on an initial graph, often called seed, that is assumed to pre-exist. Although such a seed graph substantially adds to the modeling capacity of the generation process, see Hormozdiari et al. [2007], we will only explore the simplest case for such a seed, namely a clique of an appropriate size.

In Gustedt [2009] it had been shown that this new generic generation process allows to produce families of graphs with non-vanishing clustering coefficient ($\overline{cc}$). The objectives of this paper are to make this families of graphs more tangible, to check their properties experimentally and to show how different types of distributions for parameters of the genesis influence the expected structural properties. In particular, we will observe $\overline{cc}$ as some

parameters to generate the graphs are changed, and as the size of the generated graphs grows. In addition to $\overline{cc}$ we will investigate other characteristics of the resulting graphs, in particular the degree distribution.

In Section 2, we present the co-occurrence model, the random distributions that are used for the construction and the construction process itself. In addition, we present the algorithm that is used to compute a faithful estimation of $\overline{cc}$. It is a derivation of a know algorithm of Schank and Wagner [2005] that improves the handling of border cases with very few triangles.

In Section 3, all the results of our simulations are presented, analyzed and discussed. Furthermore, a graphical view is used to have another idea of what happens with the graph for the different used parameters. In Section 4, our conclusions are presented. All figures are given after the references.

2 Model

The model of reference, for details see Gustedt [2009], for our genesis of graphs can be described basically with two types of entities: *contexts* and *objects* which are considered a level up of organization for nodes and edges in a graph. Examples for these two type of entities are authors (objects) of scientific papers (contexts), proteins (objects) in metabolic reactions (contexts), people (objects) in different social contexts, or computer network interfaces (objects) connected to a network switch (context).

The objects given by an application form the vertices of a graph for which the edges correspond to the co-occurrence's of objects in a context. In graph theoretic terms an equivalent description is to investigate a graph together with a *clique cover*, where the cliques of the cover correspond to the application contexts.

For the genesis of a co-occurrence graph G we assume that the contexts appear one after another on a discrete time scale, usually denoted by τ. Thus such a graph can be seen as one instance of a potentially unbounded family of graphs $G_0 \subseteq G_1 \subseteq \ldots \subseteq G = G_\tau \subseteq \ldots$ adding a new context co_τ at each step. G presents just one observation of the application domain at a particular point in time.

Generally, we may not assume that the genesis of a newly appearing context is independent of all previously existing ones. In the contrary, many applications have strong dependencies among their contexts: social contexts evolve by incrementally adding or removing members, metabolic reactions evolve by a mutation of one of the participating components etc. To model this dependency on the past, we assume that a new context co_τ (with the exception of some initial ones) has a *paragon* context $co_{\rho(\tau)}$, with $\rho(\tau) < \tau$,

of which it inherits some of the objects (a set called old_τ) and to which it adds some new ones (new_τ).

In the sequel of this paper we will assume that such a set new_τ may be identical during a period of time i.e

$$new_{start(\tau)} = \cdots = new_\tau = \cdots = new_{start(\tau)+\mathcal{L}(\tau)-1}.$$

The contexts

$$co_{start(\tau)}, \ldots, co_{start(\tau)+\mathcal{L}(\tau)-1}$$

with same *new* are then called a *stable sequence* and the parameter $\mathcal{L}$ is the *length* of this sequence.

As a metaphor for this property think of a new member of a social network (such as a PhD student) that is initially introduced into a bunch of contexts (e.g connections provided by the PhD adviser) and whose connection only will evolve if she or he will be able to recruit students by her- or himself. Clearly such an assumption is reductive and will not cover all desired applications. Gustedt [2009] also treated a more general case where so-called *sporadic* objects may appear, but in the purpose of not over-charging the presentation we will not follow this vein, here. Also, for simplicity in many cases we will assume that *new* is of cardinality 1, i.e that only one object is introduced at a time.

In summary, if we view the genesis of a co-occurrence graph as the described process of successively adding contexts we have a dependency from four different choices that have to be made:

$|\boldsymbol{new}(\tau)|$ regulates the number of initially redundant objects in the new context co_τ.

$\mathcal{L}(\tau)$ The length of the current subsequence models the *head-start* of the *new* contexts.

$\rho(\tau)$, $\boldsymbol{old}(\tau)$ The paragon relation together with the choice of the intersection between new context and paragon model the dependency from previous choices, the *historical bias*.

Observe that since objects in *new* are supposed to have an empty history the cardinality $|new(\tau)|$ is sufficient to describe such a process up to isomorphism. Also, observe that among these four choices the last two are dependent, namely the choice of *old* is dependent on ρ.

2.1 Models of random sampling

In a real world sequence of contexts these choices will depend on a multitude of conditions and properties of the objects. If we want to model the genesis synthetically in a simulation we have to replace these "*good reasons*" by some estimations of the statistical properties of these choices.

The first restriction that we impose for this work here is that in addition to *new* also for *old* we will just model its cardinality $|old|$ by some prescribed statistical distribution. According to a sampled value x for $|old|$ the set $old(\tau) \subseteq co_{\rho(\tau)}$ is then chosen uniformly at random.

At first, $\rho(\tau)$ also is prescribed to simple distribution: existing contexts with time $< \tau$ are chosen uniformly for the paragon relation, too. But since this model is probably too over-simplistic for many applications we will also present other sampling functions, see "aging" below.

For the remaining three numerical quantities $|new|$, $|old|$ and $\mathcal{L}$ we experiment different types of distributions, see e.g Knight [1999] for some of the terminology:

fixed valued, Dirac, $\delta(\mu)$: The simplest type of genesis that we handle in the current work is the case that a choices is fixed to an integer $\mu \geq 0$. In particular fixing all three quantities, leads to some border cases of well known families of random graphs, see Gustedt [2009], among them Erdős and Rényi [1960] and Barabási and Albert [1999] distributed graphs and also chordal graphs, respectively k-trees. Besides its theoretical interest, also some application domains may be close to such a model, e.g data control flow graphs of certain types of programs, see Nishizeki et al. [1976], Thorup [1998], or, growth of crystals.

two valued, Bernoulli, Bern(μ): For a probability (and mean) value $0 \leq \mu \leq 1$ we have an outcome of 1 with probability μ and 0 otherwise. As an extension of the usual Bernoulli distribution for a mean value $\mu \geq 1$ the outcome is chosen from the two values $\lfloor \mu \rfloor$ and $\lceil \mu \rceil$ with probabilities $p = \lceil \mu \rceil - \mu$ and $1 - p = \mu - \lfloor \mu \rfloor$. In the special case that μ is an integer this distribution coincides with the previous.

Poisson, Pois(μ): A mean value μ is prescribed and the quantity is chosen with a Poisson distribution with that mean. From a application perspective such a model can be justified if a range of application properties must be covered and we are interested in all objects that occur in such a range. (e.g expertise for a number of scientific domains)

binomial, Bin(μ, p): A mean value μ and a probability $0 < p < 1$ are prescribed and the quantity is chosen with a binomial distribution with these parameters.

shifted, $D(\mu, ...)^{\geq \nu}$ A minimum value ν is prescribed and then D, one of the distributions above, is applied for the excess over the minimum value. E.g $\mathrm{Bin}(\mu, p)^{\geq \nu}$ is obtained by adding ν to a sample distributed with $\mathrm{Bin}(\mu - \nu, p)$.

Such shifted distributions can be used to model situations where minimum requirements on the quantities have to be fulfilled. In particular this is interesting for *new* where generally we want at least one new element in each context co_τ. Otherwise $co_\tau \subseteq co_{\rho(\tau)}$ and it would be redundant for the generation of the co-occurrence graph.

2.2 About the simulations

As was stated above the sample algorithm has to choose several quantities at random. These quantities can be separated into three different groups that are independent of each other: For a set of (i) $|new_{\tau_0}|$ elements we create (ii) $\mathcal{L}_{\tau_0}$ contexts. For each of these contexts at $\tau_0, \tau_0 + 1, \ldots$ we choose (3) a paragon at time $\rho(\tau)$ and dependent from this paragon a number of $|old_\tau|$ inherited elements. Since these three are chosen independently of each other, the sampling algorithm can be parallelized. A discussion of this parallel algorithm will be given in a different paper; for our purpose here is important that this parallelization helps to scale the process to very large samples.

To be able to observe a real evolution of the graph invariants for growing sizes of graphs, we want to generate graphs randomly over several orders of magnitudes. It would not be appropriate to draw the size of the graphs uniformly in a given range, since then graphs from the smaller end would occur to rarely. Therefore, in all cases the total number of nodes in the graph is determined as follows: we chose a minimum I_{min} and the maximum I_{max} number and then chose a value α uniformly at random as $\ln(I_{min}) \leq \alpha \leq \ln(I_{max})$. The total amount of nodes is then set to $N = \lfloor e^\alpha \rfloor$.

2.3 The estimation of the clustering coefficient

With one exception, the graph invariants that we want to compute for our samples have straight forward sequential and parallel algorithms that lead to fast implementations. This holds in particular for the number of connected components and for the degree sequence and an estimate of its slope.

Unfortunately, an exact computation of the clustering coefficients of as many graphs as we handle in our experiments would not be feasible. Remember that the clustering coefficient $\overline{cc}$ of a graph is the average over the local densities of its vertex neighborhoods. Here, different variants of $\overline{cc}$ are possible if we set the local density of vertices with degree ≤ 1 to 0 or 1. For this paper we choose the case of 1, but for most of the graphs this differentiation is only cosmetic.

We use a refinement of the approach of Schank and Wagner [2005] to estimate this coefficient. The main idea is here to choose a node with degree greater or equal to 2 at random, then 2 of its neighbors and to verify if these neighbors are connected. This is iterated r times, and if c connected neighbors have been found, c/r is taken as an estimate of the clustering coefficient. Clearly if $r \to \infty$ this value is expected to tend to the exact value of the clustering coefficient.

The pseudo-code of the algorithm, adapted from the reference for our case (non-weighted graph), is presented in Algorithm 1. The original work

Algorithm 1: approximation for CC of Schank and Wagner [2005]

Input: integer r, vector $A_{1,\ldots,|V'|}$ of nodes $V' = \{v \in V : d(v) \geq 2\}$
Output: Approximation of CC
Data: Nodes variables: u, v, j, integer variable: $c \Leftarrow 0$

```
begin
    for i ∈ (1,...,r) do
        j ⇐ RandomNode(A)
        u ⇐ RandomAdjacentNode(A_j)
        repeat
            w ⇐ RandomAdjacentNode(A_j)
        until u ≠ w ;
        if ExistEdge(u, w) then
            c ⇐ c + 1
    return c/r
end
```

of Schank and Wagner [2005] didn't say much about the value for r for which we obtain an accurate approximation of $\overline{cc}$. Fixing r to a predefined value, eventually depending on the size of the graph, doesn't lead to a satisfactory approximation. A problem arises when only a small number of edges contributes to the clustering coefficient. The estimated values then only fall into a discrete set of values and give a false impression of the distribution.

This happens mainly for graph families that are close to the Barabási-Albert ($|co_\tau| = 2$) and Erdős-Rényi models, where the $\overline{cc}$ tends to zero when the graphs are growing.

To avoid this problem, we fix some value for r beforehand but then repeat the process if the amount of neighborhood edges that we found was too low to be statistically significant. By that, we obtain a sequence $c_1, c_2, \ldots$ of estimations of $\overline{cc}$. We stop the process when this sequence of values stabilizes.

3 Simulations results

In this section, all the results, graphs and analysis are presented. Firstly, a graphic view of the cases simulated are shown, to have an idea about the general behavior of the graph as some parameters changes. Such a visual approach must have its limits in the size of the graphs that are represented. Therefore, we then analyze degree sequences for larger graphs. To see the behavior not only for individual randomly generated graphs but for whole families, we then switch to a presentation of graphs that show just one parameter as a function of the size of the generated graphs. Particularly interesting here is the estimation of the clustering coefficient since up to our knowledge an experimental study large families and sizes of graphs have not been presented in the literature, yet. Also by that we will experimentally verify the claims of Gustedt [2009] about existence respectively non-existence of lower bounds for the clustering coefficient.

The figures that are presented here consist only of a very small part of the experiments that have been done. They are merely chosen as samples or examples to show typical situations. The interested reader is invited to look into a technical report that shows much more examples, see Gustedt and Schimit [2008], or run the program itself, see `http://parxxl.gforge.inria.fr/`.

3.1 A graphic view of the network

Using the software package Graphviz, see `http://www.graphviz.org`, some pictures of the graphs with the range of parameters as given above have been generated, see Figures 1 to 8. In the graphs all vertices are colored with the same random color for all vertices of the same connected component. Only the lowest numbered (=oldest) vertex in each connected component is left white.

Figure 1 show simple choices for the generation. All cardinalities are fixed to a value of 1. Thus the graphs are necessarily trees and the only random choice concerns the connection of a new vertex to existing ones.

Figure 2 shows the change of the choice function for the stable sequence (the amount of connections a new vertex gets initially) from a fixed value to a random distribution.

Now in Figure 3 we see what happens if we don't enforce that a new context necessarily has an intersection with an old context. The value for the intersection here is distributed with Poisson, and thus empty intersections may appear: the graphs become disconnected. If the distribution for the stable sequence then is fixed to 1, the graph is chordal, *i.e* has no cordless cycles, Figure 3(a). If the distribution for the stable sequence is Poisson some cordless cycles start to appear, Figure 3(b).

Figure 4 gives examples for which the distribution of the intersection size and the amount of new vertices are also is Poisson distributed.

Figures 5 to 8 show examples with aging, *i.e* where the probability of a new context connecting to an old context is regulated by means of the 'age' of the old context. Here, the age of a context co_i is simply taken as the difference $t - i$ from the actual index t.

We use two different random models for aging. The first, shown in Figures 5 and 6 follows a Poisson distribution with a mean value at 10% of the age. The second, Figures 7 and 8, have the same mean value, but follows a Binomial distribution, instead,

Again, we see that the distribution of the number of new vertices has an influence on the appearance of cordless cycles: Figures 5 and 7 have the value fixed to 1 (no cordless cycles) whereas Figures 6 and 8 have a Poisson distribution and show several long cycles.

3.2 Degree Sequences

For much larger graphs, it is unfortunately impossible to show them graphically but we have to move to the visualization of the degree sequence. Here we made the arbitrary choice to present these for some graphs with 100,000 nodes. As usual, the degree sequence is plotted with the attained degrees on x-axis and the number of nodes that have this degree on the y-axis. Both axis obey a log-scale. The y-axis shows a relative scale on the left (the share of vertices with a certain degree) whereas the scale on the left show absolute numbers.

Additionally to the degree sequence the plots show a 'slope' for the mid-range of the values (thin green line). Plots for different numbers of nodes give very similar pictures so they are omitted for brevity.

Figures 9 and 10 show cases without aging. They clearly show a typical "power law" behavior: the straight line approximates the behavior very well. Only for large degrees, whence the occurrence count drops below 10 or so the plot fans out and stripes corresponding to the discrete set of y-values become visible.

Figures 11 and 12 show similar plots, but this time for processes with aging. We see clearly that the behavior for larger degree values is substantially different. The plots are curved, *i.e* for large degree values the number of vertices that have them falls much faster. The fan out is narrowed.

3.3 The Clustering Coefficient

To get an idea on how $\overline{cc}$ is evolving with different sets of parameters we have to compute it for a lot of graphs that are on different orders of scale. Therefore for each graph we have sampled an exponent x uniformly in the interval $[3, 6]$ and then chosen the number of graph nodes to be 10^x. Thereby we obtain a sample space of graphs with 100 to about 2,000,000 – 10,000,000 nodes that covers the different orders of magnitude uniformly. Also for this section the choice function ρ for the paragons contexts has no aging and simply has a uniform distribution over the pre-existing contexts.

Figures 13 to 15 show $\overline{cc}$ as a function of the number of nodes for cases for which all other sample parameters are fixed values. As predicted in Gustedt [2009] we see that $\overline{cc}$ behaves quite differently whether or not the context size is 2 (Figure 13) or more (Figure 14): in the first case it decreases exponentially, in the other it is bounded from below.

We also see that varying the length of the stable sequences doesn't change the picture qualitatively but only quantitatively. This quantitative difference is quite marked for the bounded case (Figure 14), whereas the point clouds have large intersections in the decreasing case (Figure 13).

Figure 15 presents families that are parametrized by the context intersection, instead. All graphs here have stable sequences length fixed to 2 but the context sizes are fixed to different values, namely 3, 4 and 5 and the intersection to 2, 3 and 4, respectively. Again we see that there is no qualitative difference in the plots but a quantitative one. Observe though, that this quantitative difference only shows for graphs that are relatively large, more than about 1000 nodes.

Our last plot then shows results for an additional randomized parameter: Figure 16 plots $\overline{cc}$ in function of the average size of the contexts in the construction, with an average value that is uniformly sampled between 3 and 6. Graphs with non-integer average z are chosen to have an extended Bernoulli distribution (see Sec 2.1) to sample contexts with $\lfloor z \rfloor$ and $\lceil z \rceil$ in

the appropriate proportion. The length of the stable sequence here is fixed per plot, it varied between 1 and 5 in the experiment. In the figures we only show values 1 and 2. For values greater than 2 the plots look very much the same as for 2.

In all cases we observe an increase in $\overline{cc}$ as the average size grows. with growing values in function of the sequence length.

4 Conclusions

In an effort to study a new network generation model, we have presented a large series of simulations. The model presented in Section 2 is more complex than other well knows network models, but we think that it is more appropriate to better understand the generation large graphs as they appear in applications.

One of the positive properties of this model is its capacity to maintain the clustering coefficient near a specific range of values even for larger graphs, and thus approximating natural networks. This property had been proved in previous work and is here shown in large scale experiments. As predicted the clustering coefficient vanishes for context size 2. This generalizes well the known effects for the Barabási-Albert and Erdős-Rényi models. On the other hand whence the context size is larger than 2 the clustering coefficient is clearly bounded away from 0.

Shown for the first time, another property of the model occurred, namely the ability to '*shape*' the falling slope of the degree distribution of the resulting graphs. In case the attachment of new contexts to existing ones is done unbiased, we observe conventional power laws. In cases we also introduce a bias, here, referred to as '*aging*', we are able to generate graphs with '*super*' power laws, for which the decrease shows an inclined curve in logscale.

The properties and parameters of the generated graphs have also been shown to be '*continuous*' for cases where we don't fix the major parameters of the model but just prescribe discrete distribution for them.

5 Acknowledgments

In this document, we report the work which has been done during internships of Pedro Schimit and Hari K. Raghavan that were financed via the international internship program INRIA, under the supervision of Jens Gustedt.

Some of the experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr).

Références

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

A. Bhan, D. J. Galas, and T. G. Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18:1486–1493, 2002.

Eric de Silva and Michael P.H Stumpf. Complex networks and simple models in biology. *J. R. Soc. Interface*, 2:419–430, 2005. URL http://rsif.royalsocietypublishing.org/cgi/crossref-forward-links/2/5/419.

S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003. URL http://sweet.ua.pt/~f2358/.

P. Erdős and A. Rényi. On the evolution of random graphs. *Madyar Tnd. Akad. Mat. Kut. Int. Kőzl.*, 6:17–61, 1960.

Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of *all* complex networks. *Information Processing Letters*, 90(5):215–221, 2004.

Jens Gustedt. Generalized attachment models for the genesis of graphs with high clustering coefficient. In Santo Fortunato, Giuseppe Mangioni, Ronaldo Menezes, and Vincenzo Nicosia, editors, *Complex Networks - Results of the 2009 International Workshop on Complex Networks (CompleNet 2009)*, volume 207, pages 99–113. Springer Berlin / Heidelberg, 2009. URL http://hal.inria.fr/inria-00312059/en/. RR-6622.

Jens Gustedt and Pedro Schimit. Numerical results for generalized attachment models for the genesis of graphs. Rapport technique, INRIA, 2008. URL http://hal.inria.fr/inria-00349461/en/. RT-0361.

Fereydoun Hormozdiari, Petra Berenbrink, Nataša Pržulj, and S. Cenk Sahinalp. Not all scale-free networks are born equal: The role of the seed graph in PPI network evolution. *PLoS Comput Biol*, 3(7):e118, 07 2007. doi: 10.1371/journal.pcbi.0030118. URL http://dx.plos.org/10.1371%2Fjournal.pcbi.0030118.

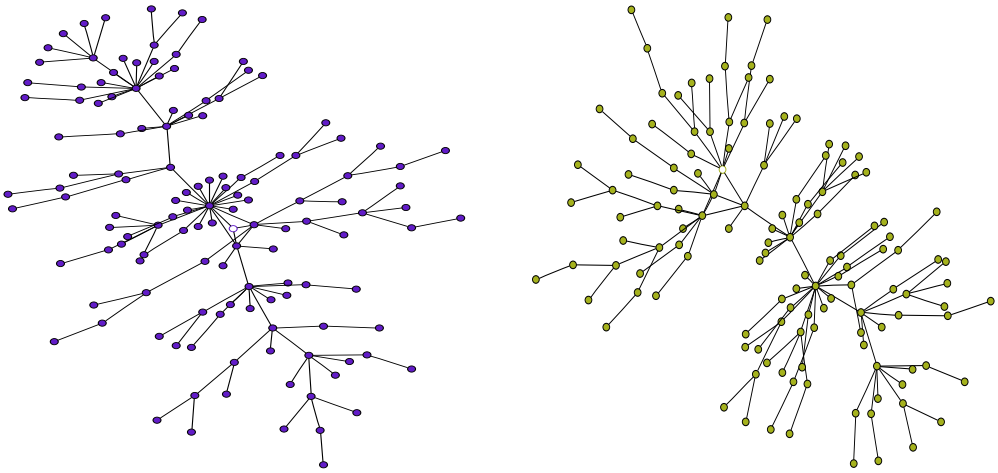
Keith Knight. *Mathematical Statistics*. Chapman & Hall/CRC, 1999.

Matthieu Latapy. *Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique*. Habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris, France, 2007.

T. Nishizeki, K. Takamizawa, and N. Saito. Algorithms for detecting series-parallel graphs and D-charts. *Trans. Inst. Elect. Commun. Eng. Japan*, 59(3):259–260, 1976.

Thomas Schank and Dorothea Wagner. Approximating clustering coefficient and transitivity. *J. Graph Algorithms Appl.*, 9(2):265–275, 2005.

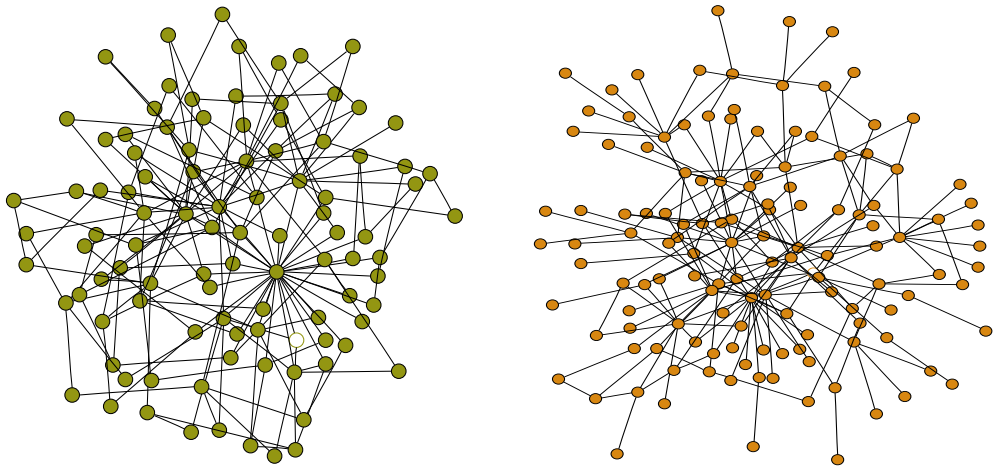
Mikkel Thorup. All structured programs have small tree-width and good register allocation. *Inf. Comput.*, 142(2):159–181, 1998.



(a) stable sequence Bern(1), new vertices Bern(1)

(b) stable sequence Pois(1)$^{\geq 1}$, new vertices Bern(1)

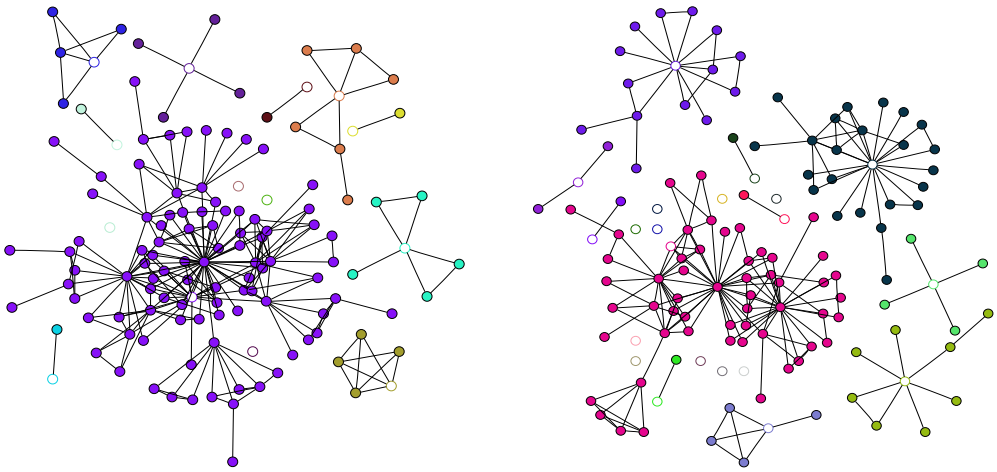
FIG. 1 – 128 vertices, no aging, intersection Bern(1)



(a) stable sequence Bern(1.5), new vertices Bern(1)

(b) stable sequence Pois(1.5)$^{\geq 1}$, new vertices Bern(1)

FIG. 2 – 128 vertices, no aging, intersection Bern(1)



(a) stable sequence Bern(1), new vertices Bern(1)

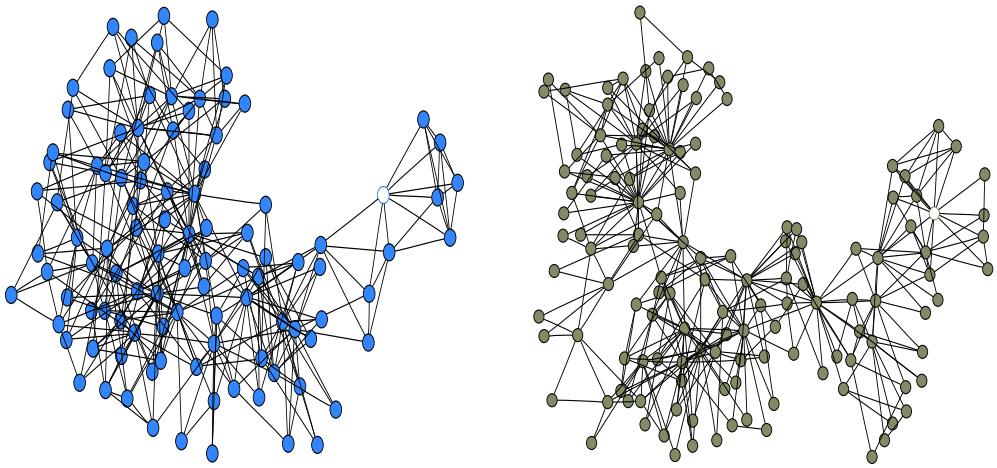
(b) stable sequence Pois(1)$^{\geq 1}$, new vertices Bern(1)

FIG. 3 – 128 vertices, no aging, intersection Pois(2)

(a) stable sequence Bern(2), new vertices Pois(1)

(b) stable sequence Pois(2)$^{\geq 1}$, new vertices Pois(1)

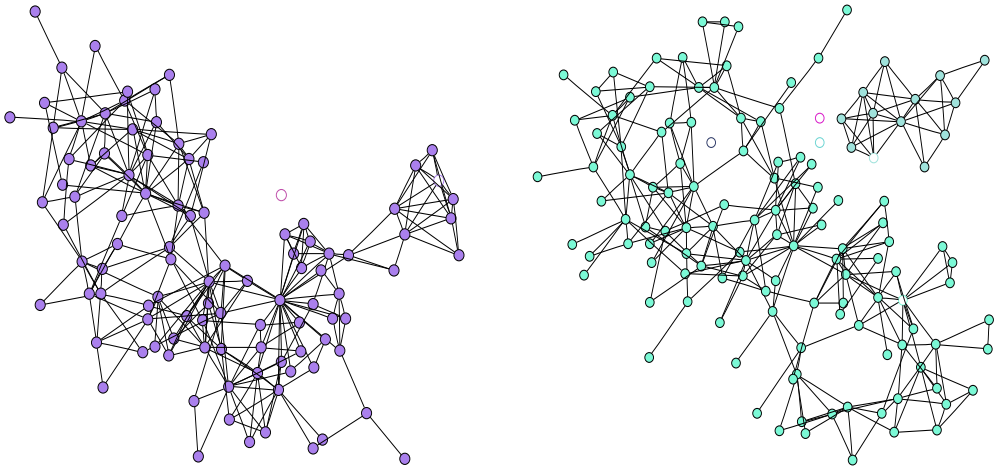
FIG. 4 – 128 vertices, no aging, intersection Pois(2.5)



(a) stable sequence Bern(1.5), new vertices Pois(1)

(b) stable sequence $\mathrm{Pois}(1.5)^{\geq 1}$, new vertices Pois(1)

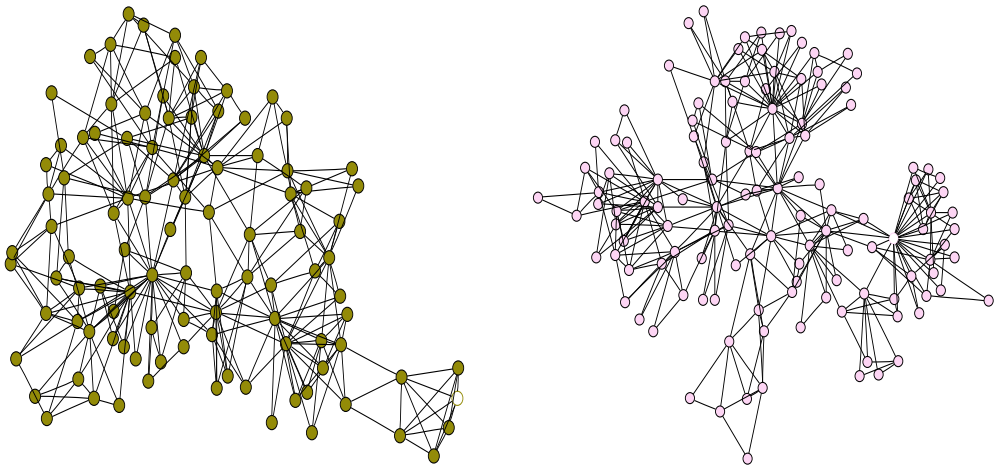
FIG. 5 – 128 vertices, aging $\mathrm{Pois}(0.1 \cdot t)$, intersection Bern(2)



(a) stable sequence Bern(1.5), new vertices Pois(1)

(b) stable sequence $\mathrm{Pois}(1.5)^{\geq 1}$, new vertices Pois(1)

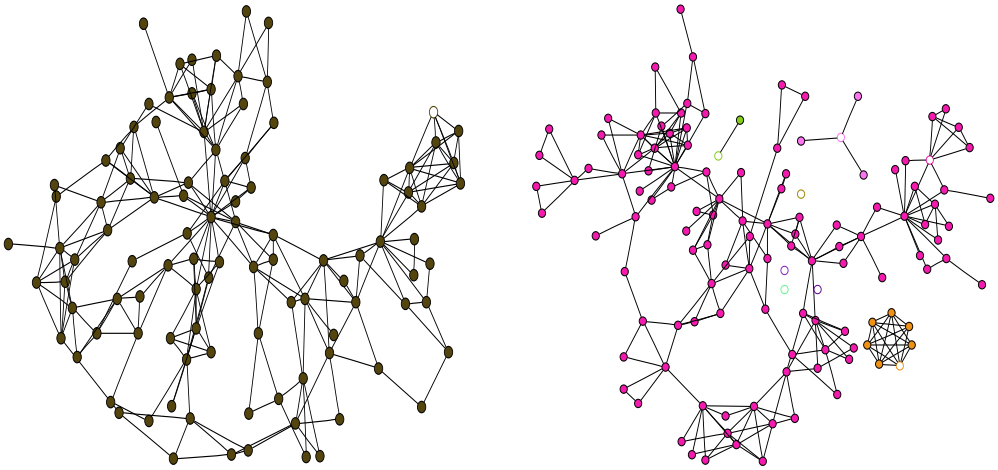
FIG. 6 – 128 vertices, aging $\mathrm{Pois}(0.1 \cdot t)$, intersection Pois(2)



(a) stable sequence Bern(1.5), new vertices Pois(1)

(b) stable sequence $\text{Pois}(1.5)^{\geq 1}$, new vertices Pois(1)

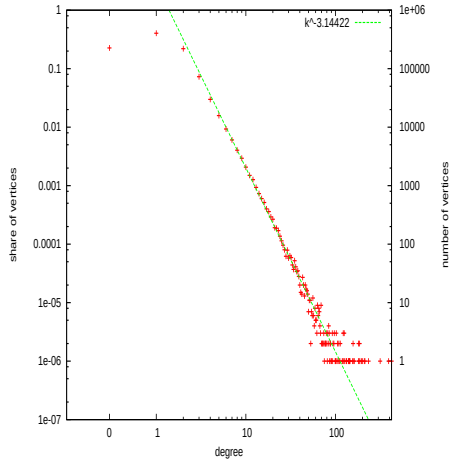
FIG. 7 – 128 vertices, aging $\text{Bin}(0.1 \cdot t, p)$, intersection Bern(2)



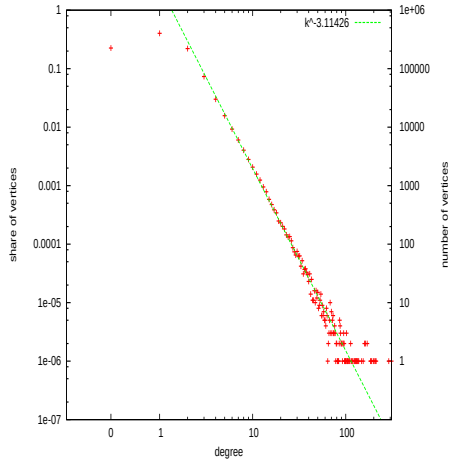
(a) stable sequence Bern(1.5), new vertices Bern(1)

(b) stable sequence $\text{Pois}(1.5)^{\geq 1}$, new vertices Bern(1)

FIG. 8 – 128 vertices, aging $\text{Bin}(0.1 \cdot t, p)$, intersection Pois(2)

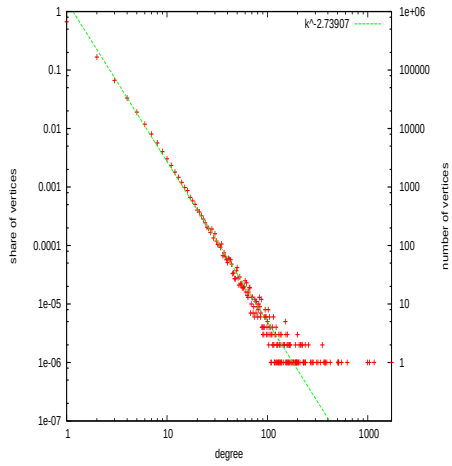


(a) stable sequence Bern(1), new vertices Bern(1)

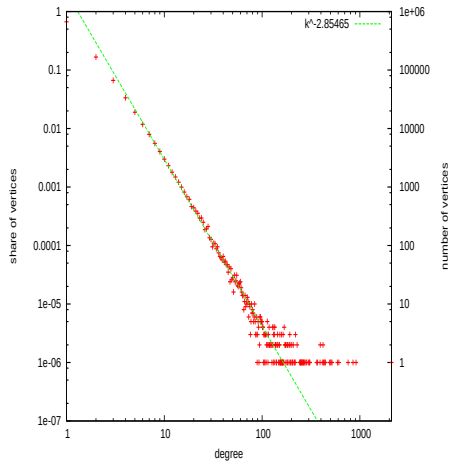


(b) stable sequence $\mathrm{Pois}(1)^{\geq 1}$, new vertices Bern(1)

FIG. 9 – 1000000 vertices, no aging, intersection Pois(1)

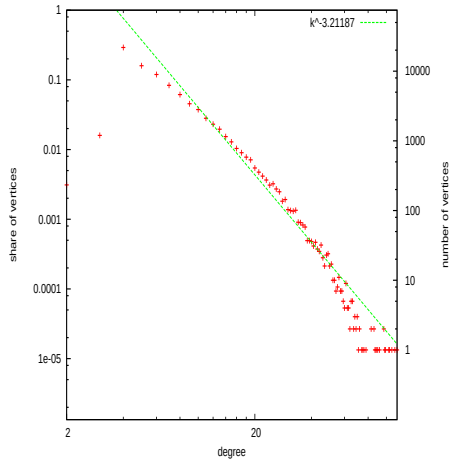


(a) stable sequence Bern(1), new vertices Bern(1)

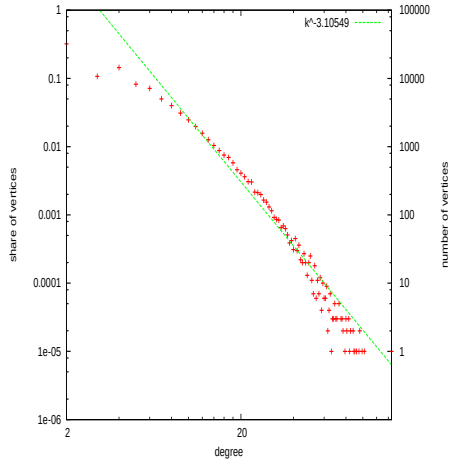


(b) stable sequence $\mathrm{Pois}(1)^{\geq 1}$, new vertices Bern(1)

FIG. 10 – 1000000 vertices, no aging, intersection $\mathrm{Pois}(1)^{\geq 1}$

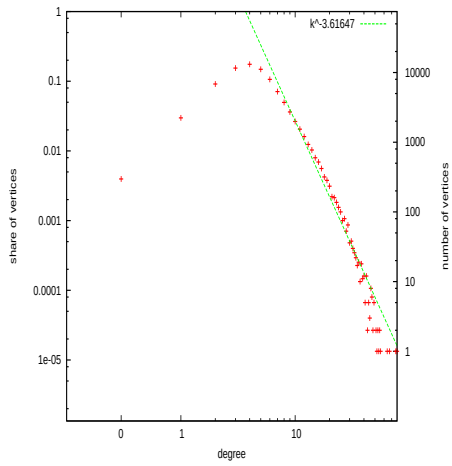


(a) stable sequence Bern(1.5), new vertices Bern(1)

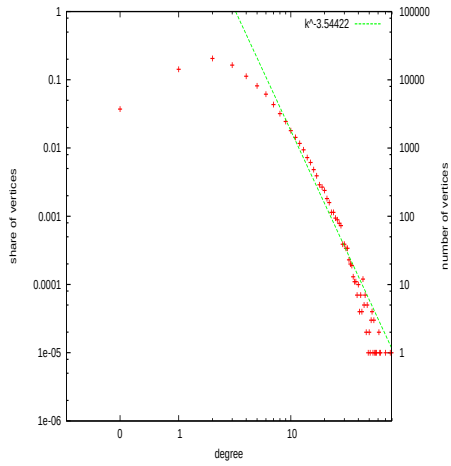


(b) stable sequence $\text{Pois}(1.5)^{\geq 1}$, new vertices Bern(1)

FIG. 11 – 100000 vertices, aging $\text{Bin}(0.1 \cdot t, p)$, intersection Bern(2)



(a) stable sequence Bern(1.5), new vertices Bern(1)



(b) stable sequence $\text{Pois}(1.5)^{\geq 1}$, new vertices Bern(1)

FIG. 12 – 100000 vertices, aging $\text{Bin}(0.1 \cdot t, p)$, intersection Pois(2)

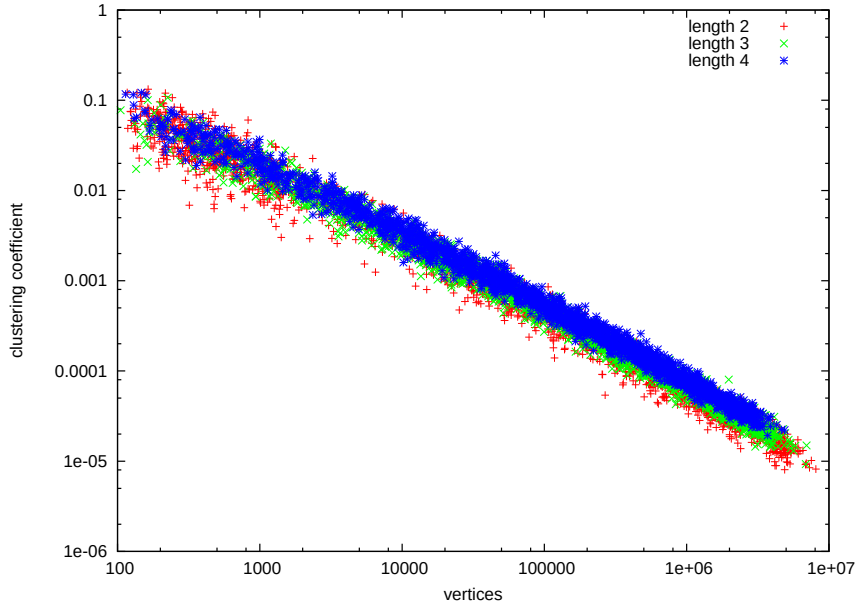


FIG. 13 – Clustering coefficient of families of graphs with fixed context size of 2 and varying stable sequence length

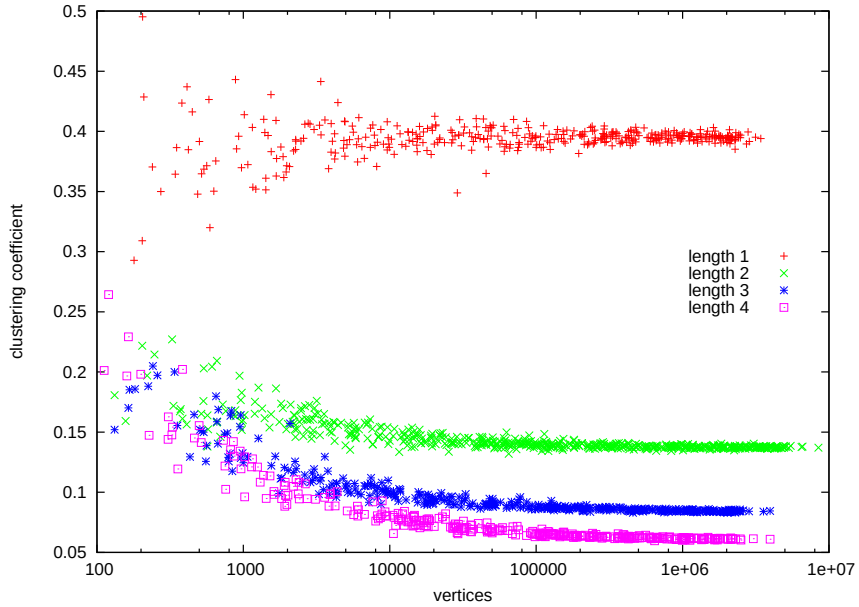


FIG. 14 – Clustering coefficient of families of graphs with fixed context size 3, fixed intersection 2, and varying stable sequence length

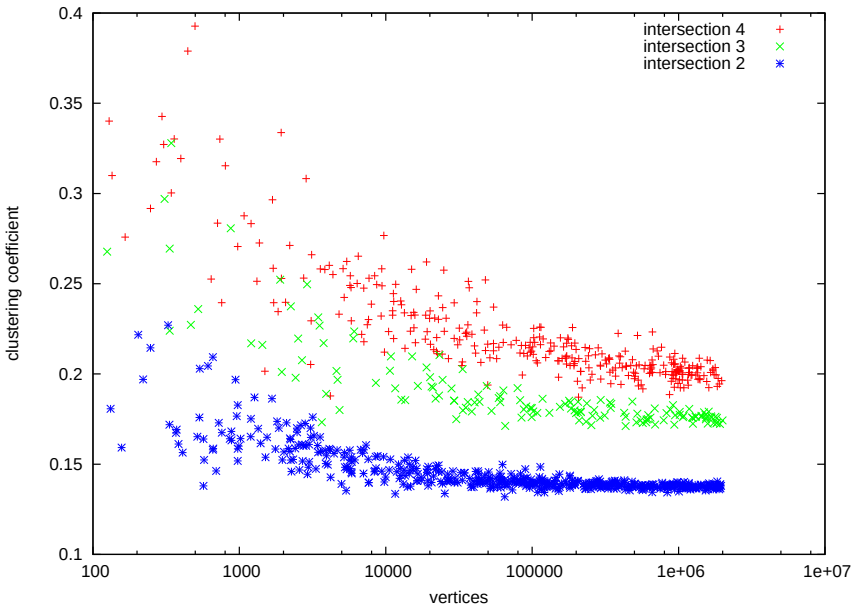


FIG. 15 – Clustering coefficient of families of graphs, contexts with fixed stable sequence length 2 and varying context size and intersection

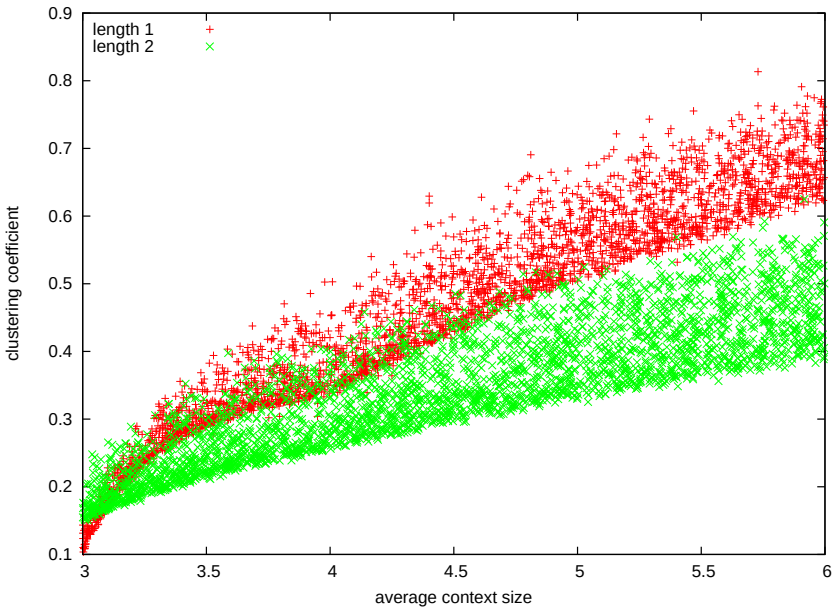


FIG. 16 – Clustering coefficient in function of the average size of the contexts for random vertex numbers.

Centre de recherche INRIA Nancy – Grand Est
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399