Exploring the random genesis of co-occurrence graphs

Jens Gustedt, Hari K. Raghavan, Pedro Schimit

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Exploring the random genesis of co-occurrence graphs

Jens Gustedt — Hari K. Raghavan — Pedro Schimit

N° 7186 — version 2

initial version January 2010 — revised version December 2010

Domaine 3

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*, Hari K. Raghavan†, Pedro Schimit‡

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

Rapport de recherche n° 7186 — version 2 — initial version January 2010 — revised version December 2010 — 28 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 distribution 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

* INRIA Nancy – Grand Est, France
† National Institute of Technology Tiruchirappalli, India
‡ University of São Paulo, Brezil

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 distribution 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

Graphs are used to model the interrelationship between objects of many scientific or engineering domains, called *applications* hereafter. Often they model a co-occurrence relation of these objects in a certain *context* under investigation:

- As the main example of this paper here, we will use the co-authorship relation between authors (objects) of scientific papers (contexts), see Newman [2001b,a].
- In *biology*, the protein-protein interaction network is composed by proteins (objects) and two proteins are related if they occur in the same metabolic reaction (context), Jeong et al. [2000]
- In *sociology*, the co-starring relation among film actors relates to actors (objects) if they appeared in the same movie (context), Watts and Strogatz [1998].
- In *linguistics* and *information retrieval*, the co-occurrence relation between words relates words (objects) that commonly occur in the same sentence (context), Doyle [1962], Patel et al. [1997], Co-occurrence networks.
- In *computer science*, the data dependency graph of a computer program relates variables (objects) that occur in the same statement (context), Banerjee [1976], Ferrante et al. [1987].

Many formalisms have been used to capture the different aspects of such a co-occurrence and there is currently no common vocabulary that is used to describe this kind of modeling:

- *Graphs* or *networks* model the co-occurrence relation of pairs of objects (the vertices of the graph). The underlying contexts are only found implicitly in such a graph as *cliques*, i.e complete subgraphs. The structure of the contexts as a whole can be seen as a *cover* of the graphs with such cliques, i.e a family of cliques of the graph that together represent each of the edges of the graph, see Nishizeki et al. [1976], Thorup [1998], Fenn et al. [2006].
- A similar description arises in so-called *content based networks* in which nodes are connected if they share some "content". This has originally been modeled by the overlap structure of sets of strings (called *genes*), see Balcan and Erzan [2006].
- *Hypergraphs* with the set of objects as a vertex set and the set of contexts as hyperedges model the co-occurrence of several objects in the same context, see e.g Zhang and Liu [2010], Ghoshal et al. [2009], Zlatić et al. [2009].
- *Bipartite graphs* have been used to model the relationship between the set of objects under investigation (one bipartition class) and the set of context

(the other bipartition class), see Guillaume and Latapy [2004], Lambiotte and Ausloos [2005b,a], Choudhury et al. [2010].
- A *concept* or *Galois lattice* where the objects define the *atoms* and the maximal contexts the *co-atoms* can be used to represent the refinement relation between different contexts, so-called *concepts*, see Barbut and Monjardet [1970], Wille [1982].

In Gustedt [2009] a family of random models that covers the inter-relationship between such contexts 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, Wagner [1994]. 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. This family of models contains classical random graphs, the preferential attachment model and the duplication model as special cases, and it also provides random models for well studied classes of graphs such as k-trees, Arnborg and Proskurowski [1989], and more generally chordal graphs, Fulkerson and Gross [1965].

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), Jeong et al. [2000], 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*, see above, where the cliques of the cover correspond to the application contexts.

In general, the *inversion problem* that consists of retrieving the original context (cliques, contents ...) structure from the generated graph is difficult, but can be solved exactly for the special case that the graph is a so-called k-tree, Bodlaender [1996]. In the context of content based networks, Ramasco and Mungan [2008] are able to approximate the "content" structure statistically.

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 stable sequence models the *head-start* of the subset of *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.

The simplest type of genesis that we handle in the current work is the case that a choice 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].

Other distributions that we experimented and that are integrated into the software are Bernoulli, Poisson, Zipf, binomial and some variants of these. In particular it is often convenient to *shift* a distribution by an additional constant: *e.g* to impose that *new* is not empty we may chose it to be distributed as $1 + x$ where x is Poisson distributed.

2.2 Aging

In a first approach at a given point in time τ of the generation process all existing contexts will be considered being equally likely to be chosen as a paragon. This corresponds to a situation where no contexts ever will be

forgotten throughout the process and thus also all objects that have been introduced may receive connections at any future point in time. In such a model, objects and context have no *age*, they never get old. Another probably unrealistic aspect of this simplified model is the inherent assumption that all contexts are immediately productive and available to be derived in new ones, they are mature from the start.

For many applications this does not seem to be realistic. E.g for the co-author graph old papers are unlikely to be the offspring of new collaborations with only a slight modification in the set of authors; authors move to different institutions, change their scientific interests or simply retire.

As a first approximation of such an aging process, we model the probability of a new context connecting to an old context by means of the *age* of the old context. Here, the age of a context co_i is simply taken as the difference $\tau - i$ from the actual index τ. The bias of the aging process is then introduced by choosing contexts with a probability corresponding to their age. As examples we will see one such process that follows a Poisson distribution with a mean value at 10% of the age and another one with the same mean value, but following a Binomial distribution, instead.

2.3 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 (like the degree distribution or the clustering coefficient) 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 $N = \lfloor e^\alpha \rfloor$.

2.4 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 distribution 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: $c \Leftarrow 0$

```
1  begin
2      for i ∈ (1,...,r) do
3          j ⇐ RandomNode(A)
4          u ⇐ RandomAdjacentNode(A_j)
5          repeat
6              w ⇐ RandomAdjacentNode(A_j)
7          until u ≠ w ;
8          if ExistEdge(u, w) then
9              c ⇐ c + 1
10     return c/r
11 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 distributions 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 4. In the graphs all vertices of the same

connected component are colored with the same random color. Only the lowest numbered (=oldest) vertex in each component is left white.

Figure 1 shows the change of the choice function for the stable sequence (the amount of connections a new vertex gets initially) from a fixed value (1(a) and 1(b)) to a random distribution. Figures 1(c) and 1(d) show the impact of the type of the distribution for the stable sequence length $\mathcal{L}$; whereas the first has a more equilibrated shape, the second has a larger degree variation with more pending vertices of degree one and some of middle and high degree.

Observe also for the later two graphs that they have many large cycles but almost no triangles. These so-called *chordless cycles*, i.e cycles of length greater than 3 that don't have a shortcut, are important structural characteristics from a graph theoretic point of view.

Now in Figure 2 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 2(a). If the distribution for the stable sequence is Poisson some cordless cycles start to appear, Figure 2(b).

Figures 2(c) and 2(d) show how the change from a fixed value for the stable sequence length $\mathcal{L}$ to a Poisson distribution with the same mean changes the structural properties of the resulting graph.

Figures 3 to 4 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 Figure 3 follows a Poisson distribution with a mean value at 10% of the age. The second, Figure 4, have the same mean value, but follows a Binomial distribution, instead.

In all these 8 graphs the oldest vertex of the giant component is located towards the right of the picture, and in all of them but 4(b) we see that this vertex does not play a central role in the graph. In the contrary, e.g Figure 3 shows well how the genesis of the graphs advances in time from the initial vertices on the right of the figure to more recently generated ones on the left.

3.2 Degree Distributions

For much larger graphs, it is unfortunately impossible to show them graphically but we have to move to the visualization of the degree distribution. Here

we made the arbitrary choice to present these for some graphs with 100,000 nodes. As usual, the degree distribution is plotted with the attained degrees on x-axis and the number of nodes that have this degree on the y-axis. The y-axis that covers 5 or 6 orders of magnitude obeys a log-scale and shows a relative scale on the left (the probability for a certain degree) whereas the scale on the right show absolute numbers. The x-axis is in log-scale for the first four plots and linear for the four others. Plots for different numbers of nodes give very similar pictures so they are omitted for brevity.

Additionally to the degree distribution the plots show approximations with known distributions, Pareto (power law), exponential or Poisson, where appropriate.

Figure 5 shows cases without aging. They clearly show the expected power law behavior, see Krapivsky and Redner [2005]: 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.

Figure 6 shows similar plots, but this time for processes with aging analogous to the one in Figure 4. We see clearly that the behavior for larger degree values is substantially different. The power law is not such a good approximation and we see that for the upper two (with fixed intersection size between contexts) the distribution seems to be between power law and exponential.

For the lower two examples (with Poisson distribution for the intersection size) we have a case that can be understood with two different ranges for the degrees. The first range of low degree vertices corresponds to vertices that are (still) peripheral for the creation process. The degree of those vertices is dominated by the distribution of the degree in the stable sequences that introduces them. Here the distribution is well fitted with a Poisson distribution.

The second range of higher degree corresponds to vertices that later have received connecting edges when new contexts duplicate their initial contexts. For "not too old" vertices the probabilities in appearing in such a context that is duplicated are about the same, but for "older" vertices to be found in a copied context is then less and less likely. So the tail of the distribution falls faster than any power law and is best approximated with an exponential distribution.

3.3 Using Application Parameters

Figure 7 shows the result of experiments where we have used the statistical parameters of co-author graphs from two different domains as have been

reported by Newman [2001b]. There, it has been shown that the number of authors per paper follow a power law distribution. We have thus sampled this quantity with a Zipf distribution of the appropriate mean value and exponent. The other parameters of the model then where adjusted with the goal to obtain a similar degree distribution as in the original. For example the "Los Alamos" graph in Figure 7(a) has mean values of 1.27, 1.5 and 4 for the size of the intersection between contexts, the number of new authors per context and the length $\mathcal{L}$ of the stable sequence, respectively. For the "Medline" graph on the right these mean values are 2.6, 1.2 and 6.1. For both types of graphs we supposed an aging such that the mean value of the age of the parent context is at 10 % of the already generated contexts.

The two degree distribution plots represent emulated graphs in the same order of magnitude than in the original paper. By comparing to the plots, there, we see that the degree sequences of the emulated graphs are quite similar to the original ones.

3.4 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 8 to 10 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 8) or more (Figure 9): in the first case it decreases exponentially, in the other it is bounded from below. These observations are also consistent with the computations for the clustering coefficient in bipartite networks as given by Newman et al. [2001].

We also see that varying the length $\mathcal{L}$ of the stable sequences doesn't change the picture qualitatively but only quantitatively. This quantitative difference is quite marked for the bounded case (Figure 9), whereas the point clouds have large intersections in the decreasing case (Figure 8).

Figure 10 presents families that are parametrized by the context intersection, instead. All graphs here have stable sequences length $\mathcal{L}$ 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 qual-

itative 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 11 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 have context sizes of $\lfloor z \rfloor$ or $\lceil z \rceil$ chosen randomly in the appropriate proportion. The length $\mathcal{L}$ 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.

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 other distributions. Depending on the setting they are situated somewhere between an exponential and a power law or are combining a Poisson distribution for the low degree vertices with an exponential tail.

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

We like to thank the reviewers of the initial version of this paper for their helpful comments. These largely helped to improve the presentation.

Major part of this work has been done during internships of Pedro Schimit and Hari K. Raghavan at INRIA under the supervision of Jens Gustedt. These internships were financed via the international internship program of INRIA.

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).

References

Stefan Arnborg and Andrzej Proskurowski. Linear time algorithms for NP-hard problems restricted to partial k-trees. *Discrete Appl. Math.*, 23:11–24, 1989.

Duygu Balcan and Ayşe Erzan. Dynamics of content-based networks. In V. N. Alexandrov et al., editors, *ICCS 2006, Part III*, volume 3993 of *LNCS*, pages 1083–1090. Springer-Verlag, 2006.

Utpal Banerjee. Data dependence in ordinary programs. Master's thesis, Univ. Illinois, Dept. Computer Science, November 1976.

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

M. Barbut and B. Monjardet. *Ordre et classification*. Hachette, 1970. 2 volumes.

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

Hans L. Bodlaender. A linear time algorithm for finding tree-decompositions of small treewidth. *SIAM J. Comput.*, 25(1305-1317), 1996.

Monojit Choudhury, Niloy Ganguly, Abyayananda Maiti, Animesh Mukherjee, Lutz Brusch, Andreas Deutsch, and Fernando Peruani. Modeling

discrete combinatorial systems as alphabetic bipartite networks: Theory and applications. *Phys. Rev. E*, 81(3):036103, Mar 2010. doi: 10.1103/PhysRevE.81.036103.

Co-occurrence networks. Wikipedia, 2008-2010. URL https://secure.wikimedia.org/wikipedia/en/wiki/Co-occurrence_networks.

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/.

Lauren B. Doyle. Indexing and abstracting by association. *American Documentation*, 13(4):378–390, 1962. doi: 10.1002/asi.5090130404.

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

Daniel Fenn, Omer Suleman, Janet Efstathiou, and Neil F. Johnson. How does Europe make its mind up? Connections, cliques, and compatibility between countries in the Eurovision Song Contest. *Physica A: Statistical Mechanics and its Applications*, 360(2):576 – 598, 2006. ISSN 0378-4371. doi: DOI:10.1016/j.physa.2005.06.051. URL http://arxiv.org/abs/physics/0505071.

Jeanne Ferrante, Karl J. Ottenstein, and Joe D. Warren. The program dependence graph and its use in optimization. *ACM Transactions on Programming Languages and Systems*, 9(3):319–349, 1987. doi: 10.1145/24039.24041.

D. R. Fulkerson and O. A. Gross. Incidence matrices and interval graphs. *Pacific J. Math*, 15:835–855, 1965. URL http://projecteuclid.org/Dienst/UI/1.0/Summarize/euclid.pjm/1102995572.

Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and M. E. J. Newman. Random hypergraphs and their applications. *Phys. Rev. E*, 79(6):066118, Jun 2009. doi: 10.1103/PhysRevE.79.066118. URL http://arxiv.org/abs/0903.0419.

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.

H. Jeong, B. Tombor, Z. N. Albert, R. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804): 651–654, 2000. URL http://www.nd.edu/~networks/Publication%20Categories/03%20Journal%20Articles/Biology/Large-Scale%20MetabolicNetworks_Nature%20407,%20651-654%20%282000%29.pdf.

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

P. L. Krapivsky and S. Redner. Network growth by copying. *Phys. Rev. E*, 71(3):036118, Mar 2005. doi: 10.1103/PhysRevE.71.036118. URL http://physics.bu.edu/~redner/pubs/pdf/gnc.pdf.

R. Lambiotte and M. Ausloos. n -body decomposition of bipartite author networks. *Phys. Rev. E*, 72(6):066117, Dec 2005a. doi: 10.1103/PhysRevE.72.066117. URL http://arxiv.org/abs/physics/0507154.

R. Lambiotte and M. Ausloos. Uncovering collective listening habits and music genres in bipartite networks. *Phys. Rev. E*, 72(066107), 2005b. URL http://arxiv.org/abs/physics/0508233.

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.

M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132, 2001a. doi: 10.1103/PhysRevE.64.016132. URL http://www-personal.umich.edu/~mejn/papers/016132.pdf. see also Newman [2006].

M. E. J. Newman. Scientific collaboration networks. I. network construction and fundamental results. *Phys. Rev. E*, 64:016131, 2001b.

M. E. J. Newman. Erratum: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [phys. rev. e 64, 016132 (2001)]. *Phys. Rev. E*, 73(3):039906, Mar 2006. doi: 10.1103/PhysRevE.73.039906. URL http://pre.aps.org/pdf/PRE/v73/i3/e039906. see also Newman [2001a].

M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64 (2):026118, Jul 2001. doi: 10.1103/PhysRevE.64.026118. URL http://arxiv.org/abs/cond-mat/0007235.

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.

Malti Patel, John Bullinaria, and Joseph Levy. Extracting semantic representations from large text corpora. In *In Proceedings of the 4th Neural Computation and Psychology Workshop*, pages 199–212, London, 1997. Springer.

José J. Ramasco and Muhittin Mungan. Inversion method for content-based networks. *Phys. Rev. E*, 77(3):036122, Mar 2008. doi: 10.1103/PhysRevE.77.036122.

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.

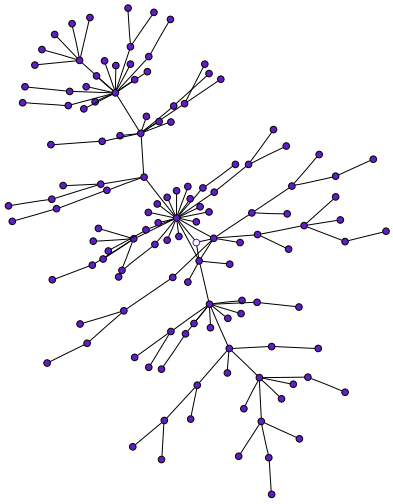
A. Wagner. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA*, 91(10):4387–4391, 1994. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC43790/.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. URL http://tam.cornell.edu/tam/cms/manage/upload/SS_nature_smallworld.pdf.

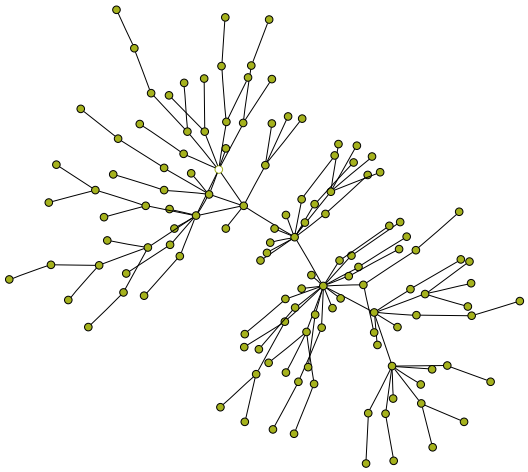
R. Wille. Restructuring the lattice theory: An approach based on hierarchies of concepts. In I. Rival, editor, *Ordered sets*, pages 445–470. Reidel, 1982.

Zi-Ke Zhang and Chuang Liu. Hypergraph model of social tagging networks. Technical Report arXiv:1003.1931, arXiv, 2010.

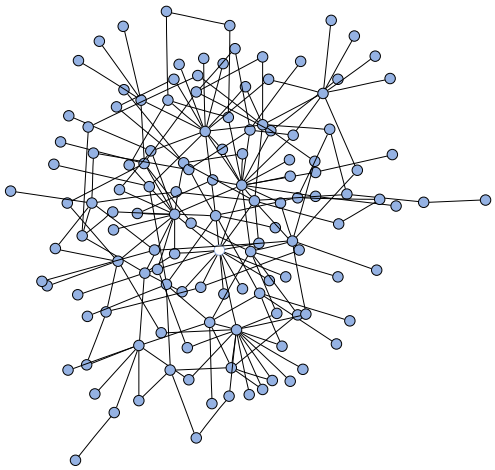
Vinko Zlatić, Gourab Ghoshal, and Guido Caldarelli. Hypergraph topological quantities for tagged social networks. *Phys. Rev. E*, 80(3):036118, Sep 2009. doi: 10.1103/PhysRevE.80.036118. URL http://arxiv.org/abs/0905.0976.



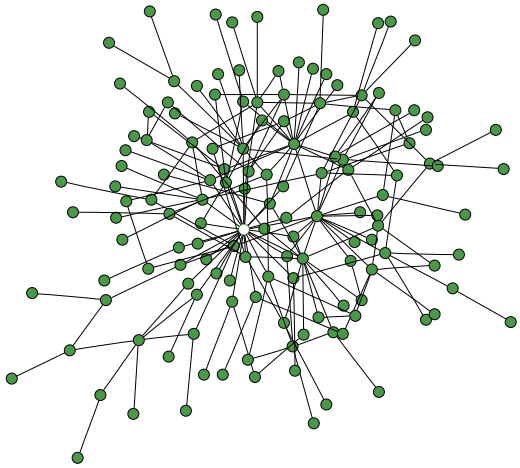
(a) The lengths $\mathcal{L}$ of the stable sequences are fixed to 1 thus the resulting graph is a tree.



(b) Another tree instance with same rule as (a)

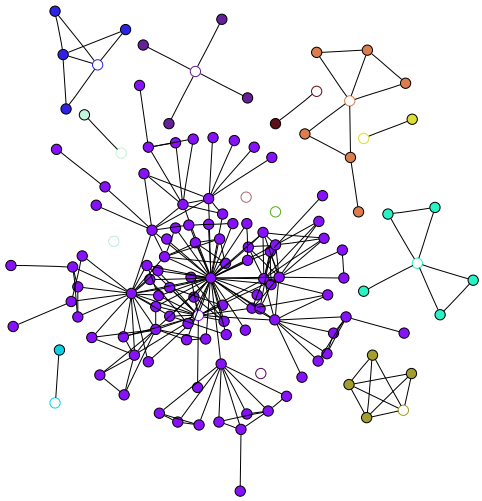


(c) The lengths $\mathcal{L}$ of the stable sequences are chosen to be 1 or 2 with equal probability.

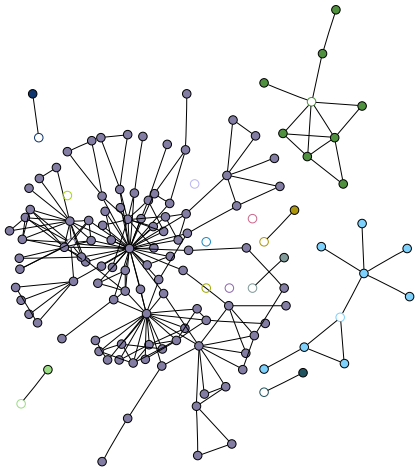


(d) The lengths $\mathcal{L}$ of the stable sequences are chosen with a shifted Poisson distribution with minimum of 1 and a mean value of 1.5.

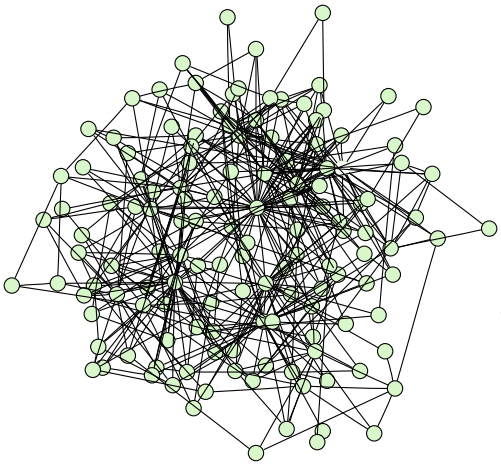
Figure 1: Graphs with 128 vertices, generated with fixed values for the context size (=2), the number of new vertices per context (=1), and context intersection (=1). There is no aging process, all existing contexts have the same probability to be chosen as a paragon.



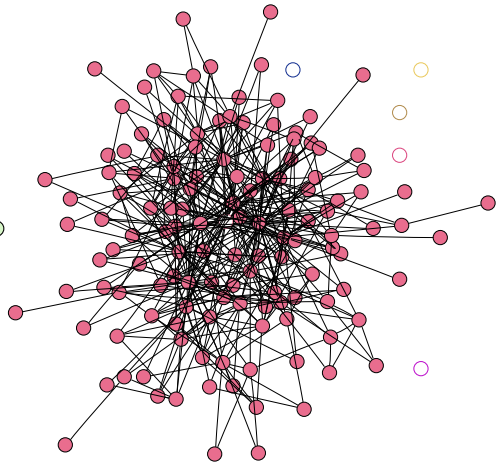
(a) The lengths $\mathcal{L}$ of the stable sequences are fixed to 1 thus the resulting graph is chordal.



(b) An instance with a slight perturbation of the rule for (a) such that sequences with length $\mathcal{L}$ longer than 1 may appear with low probability



(c) The lengths $\mathcal{L}$ of stable sequences are 2.



(d) The lengths $\mathcal{L}$ of the stable sequences are chosen with a shifted Poisson distribution with minimum of 1 and a mean value of 2.

Figure 2: Graphs with 128 vertices, generated with fixed number of new vertices per context (= 1) and context intersection that is distributed with a Poisson distribution of mean 2 (top two) or 2.5 (bottom two). There is no aging process, all existing contexts have the same probability to be chosen as a paragon.

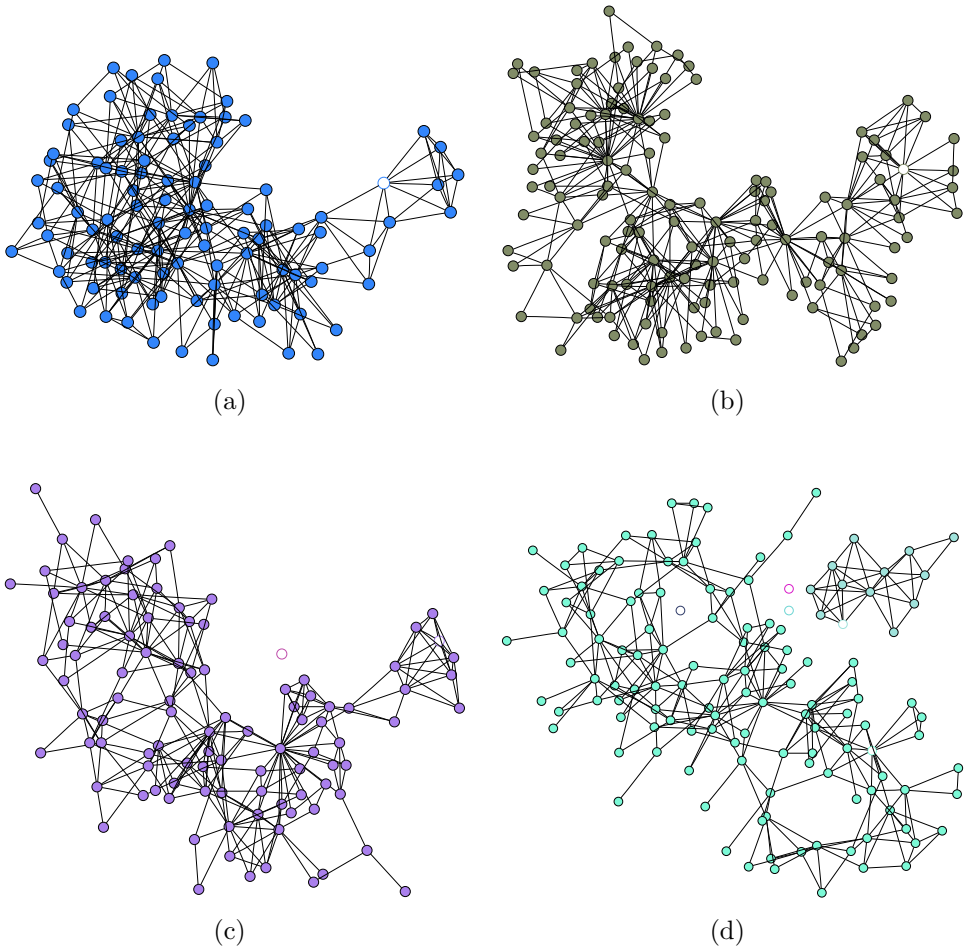


Figure 3: Graphs with 128 vertices by applying an aging process that chooses existing contexts with a Poisson distribution with mean value at 10% of the actual time length of the process. Graphs are generated with 1 new vertex per context and with a mean value of the intersection size between old and new context of 2. For the top two this intersection size is constant and for the lower two a Poisson distribution. The length $\mathcal{L}$ of the stable sequence is chosen with a mean value of 1.5. For the graphs on the left this is achieved by a Bernouilli distribution (values 1 and 2 equally likely) and for the right by a shifted Poisson distribution with minimum value 1.

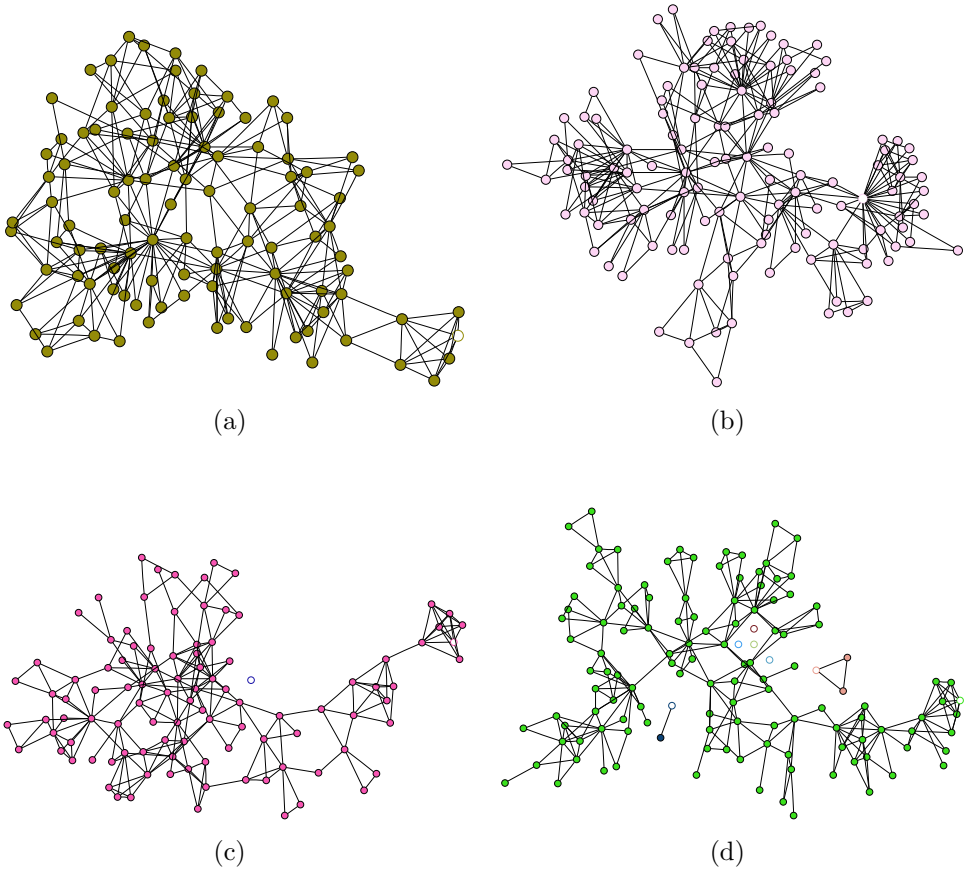
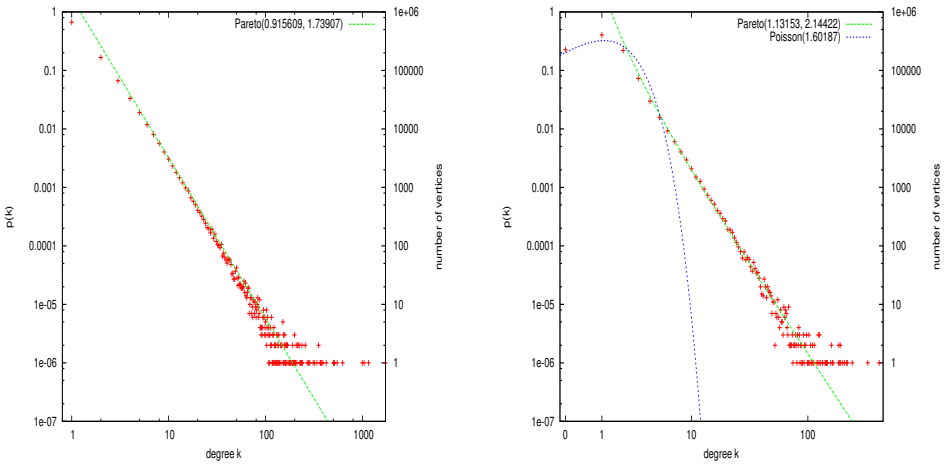


Figure 4: Same as Figure 3 only by replacing the distribution of the aging process with a Binomial distribution with same mean value.



(a) The size of the intersection between old an new contexts is fixed to 1.

(b) The size of the intersection between old an new contexts is distributed with Poisson with mean value 1.

Figure 5: The degree sequences of 2 graphs with 10^6 vertices generated with 1 new vertex per context and a stable sequence length $\mathcal{L}$ of 1. There is no aging process, all existing contexts have the same probability to be chosen as a paragon.

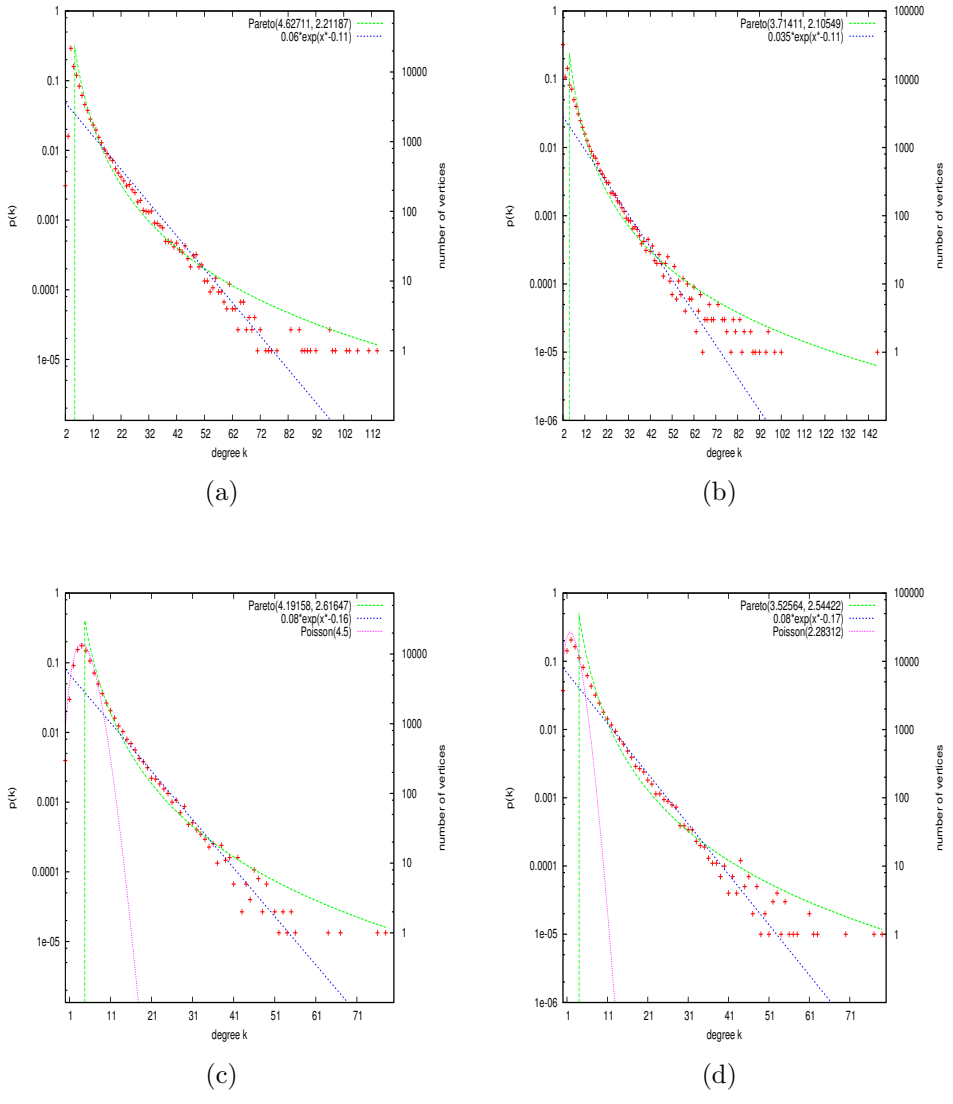


Figure 6: Degree distributions of 4 graphs with 10^6 vertices by applying the same rules as in Figure 4.

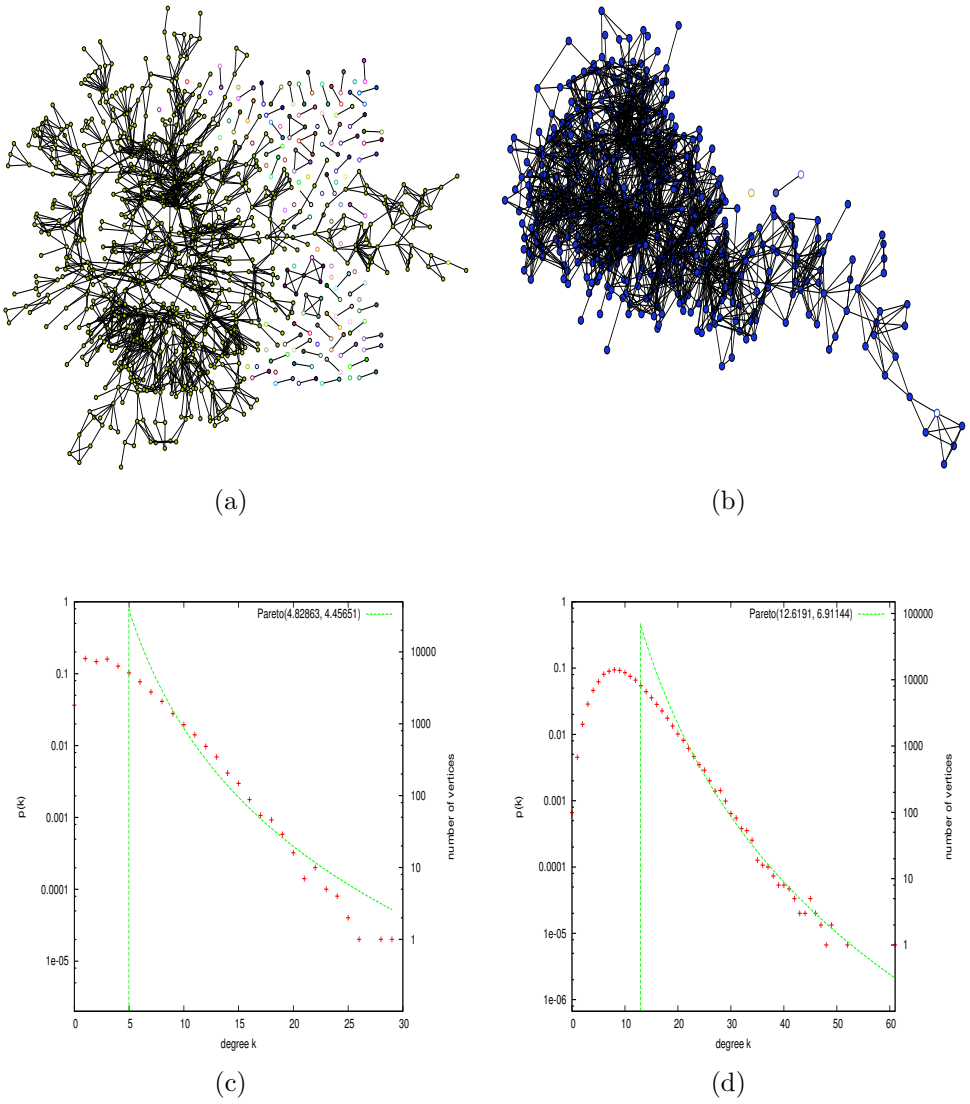


Figure 7: Graphs sampled according to known statistics of co-author graphs as given by Newman [2001b]. The graphs of the left side emulate statistical properties of the *Los Alamos e-Print Archive*, those on the right the *Medline* data base. Graphs on the top are sampled with 500 vertices, those on the bottom in the same order as the real world graphs.

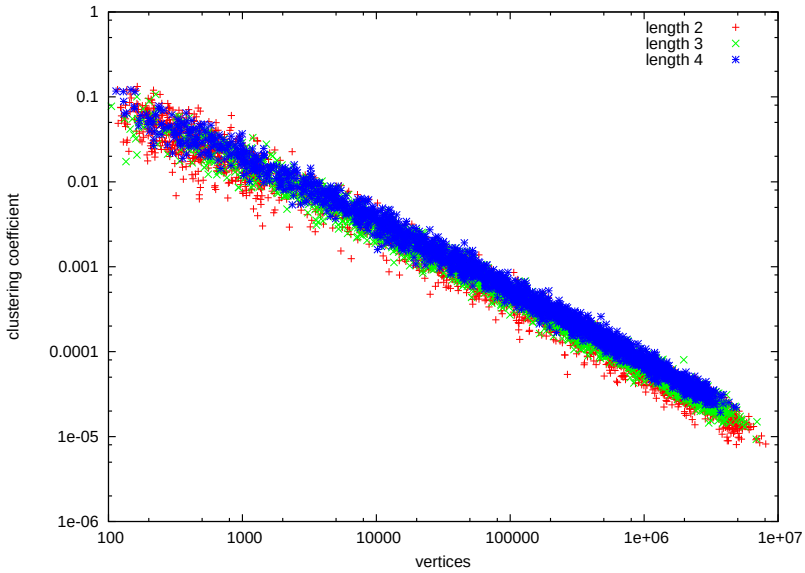


Figure 8: Clustering coefficient of families of graphs with fixed context size of 2 and varying stable sequence length $\mathcal{L}$

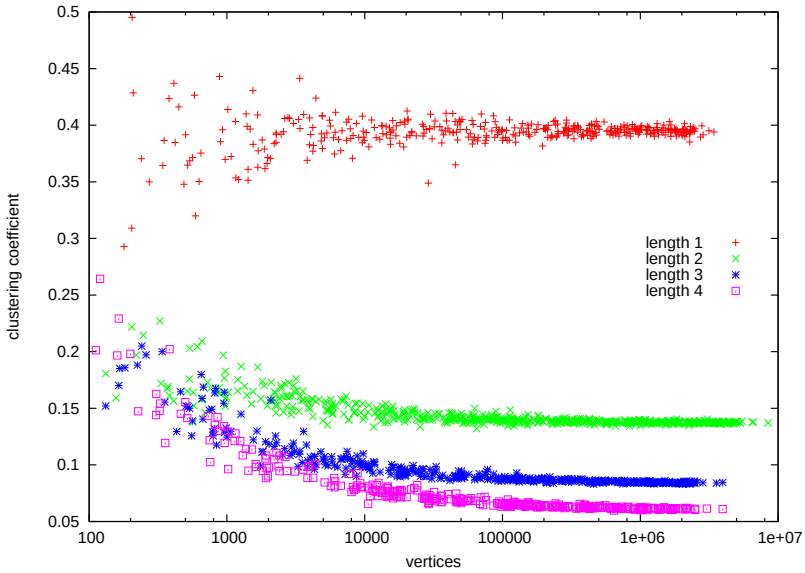


Figure 9: Clustering coefficient of families of graphs with fixed context size 3, fixed intersection 2, and varying stable sequence length $\mathcal{L}$

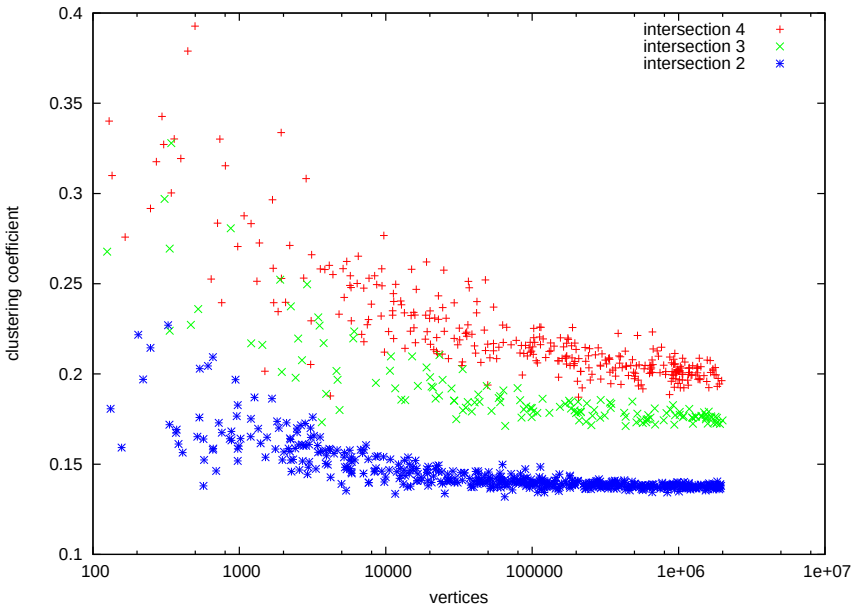


Figure 10: Clustering coefficient of families of graphs, contexts with fixed stable sequence length $\mathcal{L} = 2$ and varying context size and intersection

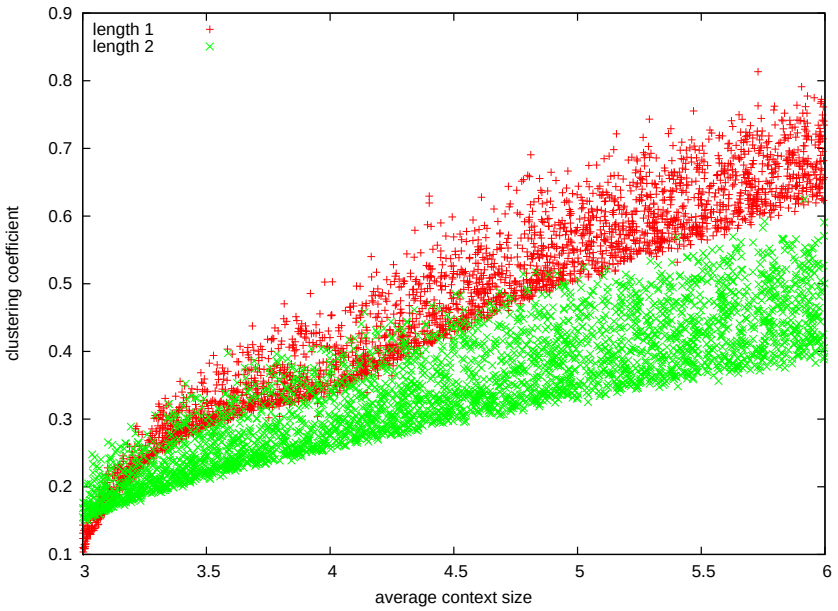


Figure 11: Clustering coefficient in function of the average size of the contexts for random vertex numbers.

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