Generalized Attachment Models for the Genesis of Graphs with High Clustering Coefficient*

Jens Gustedt

Abstract Commonly used techniques for the random generation of graphs such as those of Erdős & Rényi and Barabási & Albert have two disadvantages, namely their lack of bias with respect to history of the evolution of the graph, and their incapability to produce families of graphs with non-vanishing prescribed clustering coefficient. In this work we propose a model for the genesis of graphs that tackles these two issues. When translated into random generation procedures it generalizes the above mentioned procedures. When just seen as composition schemes for graphs they generalize the perfect elimination schemes of chordal graphs. The model iteratively adds so-called *contexts* that introduce an explicit dependency to the previous evolution of the graph. Thereby they reflect a historical bias during this evolution that goes beyond the simple degree constraint of preference edge attachment. Fixing certain simple statical quantities during the genesis leads to families of random graphs with a clustering coefficient that can be bounded away from zero.

Key words: attachment models, random graph generation, k-trees

1 Introduction and Overview

Modeling the genesis of graphs has become an important task in many application domains, most prominent are probably graphs of computer networks and social networks. Such models are an important prerequisite for the random generation of realistic large networks that are needed for simulations in the framework of the application domains and also for testing graph algorithms and their implementations on examples of a realistic size.

Application graphs that follow a vein of 'families of hidden cliques defining a graph' have been identified by Guillaume and Latapy [2004]: protein 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. The goal of this paper here is to give a random model that covers the inter-relationship between

INRIA Nancy – Grand Est, Villers-lés-Nancy, France. e-mail: jens.gustedt@loria.fr

* This paper appeared in Ronaldo Menezes et al. *Complex Networks*, Studies in Computational Intelligence Volume 207/2009, Springer Verlag pp. 99-113. The original publication is available at www.springerlink.com

those cliques, *e.g* the non-trivial overlap of those cliques, and the temporal evolution of these networks respectively their genesis.

During the genesis of such graphs, the dependence of the process from previous choices is an important detail that we try to handle. We propose a relatively simple model, in which each newly introduced clique depends on a previously known one. 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.

Classical random graph models, such as promoted by Erdős and Rényi [1960], usually do not fulfill the necessities from the application domains since the expected structure of the generated graphs is too far from what is observed in practical settings. Because of that, starting form the work of Barabási and Albert [1999] in the last decade a lot of attempts to provide more realistic models have been undertaken, see *e.g* Latapy [2007] for an overview, or Dorogovtsev and Mendes [2003] for a textbook.

These models try to capture different statistical properties of the generated graphs, such as the degree distribution or the expected distance between arbitrary pairs of vertices. Here, we will concentrate on an important property that is desired for random graphs, namely their density: it was observed that real world graphs are generally sparse ('*have much less edges than they could*') but are usually locally quite dense ('*the probability that two neighbors of a vertex are also connected is high*').

The *clustering coefficient* $\overline{cc}(G)$ of a graph G is meant to capture this feature. It measures how close or how far the vertices are to being simplicial: a vertex v is called *simplicial* if the neighborhood of v is a clique. If v is not simplicial, the *neighborhood density* $cc(G, v)$ at v is the quotient between the number of edges inside the neighborhood of v and the maximum number of such edges. Formally $\overline{cc}(G)$ is given as the average over all vertices $v \in V$ of

$$cc(G, v) = \begin{cases} 1 & \text{if } v \text{ is simplicial} \\ \frac{|E_v G|}{\binom{\deg_G(v)}{2}} & \text{otherwise} \end{cases} \tag{1}$$

Observe that for the (practically unimportant) border case of vertices of degree 0 or 1 this definition is different from what can be usually found in the literature, see *e.g* Dorogovtsev and Mendes [2003], but it will prove convenient in the sequel. Using only the quotient would not be well defined since the denominator then would be 0 for these special cases. With our particular choice we have that $\overline{cc}$ will be high for trees: for any tree G without vertices of degree 2, $\overline{cc}(G) > 0.5$. This high value fits well to what we obtain when we investigate other graphs that show a 'tree-like' structure.

A k-tree, see Arnborg [1985], Robertson and Seymour [1986], is a graph that can be obtained from a clique of size $k + 1$ by iteratively joining new vertices to cliques of size k. The so-called k-tree-decomposition of such a graph is easily defined from this iterative definition.

Observation 1 *For some $k > 0$ let G be a k-tree with a k-tree-decomposition that has no vertices of degree 2. Then $\overline{cc}(G) \geq \frac{1}{2} - \frac{k-2}{2|V(G)|}$.*

A proof for this statement can be obtained by counting the leaves of the decomposition: they correspond to simplicial vertices in the graph. Because of the restriction for the tree-decomposition at least half of the nodes of that decomposition are leaves and thus almost one half of the vertices of the graph are simplicial. The observation then follows by straightforward estimations.

The bound given in Observation 1 is not sharp. It accounts only for those vertices that are in fact simplicial and the impact of the other vertices is neglected. But on the other hand it shows us some reason why some graph G might have a high clustering coefficient namely that it might have '*a lot*' (here one half) of '*boundary*' vertices that are simplicial.

Also, Observation 1 puts a restriction on the permissible k-tree decomposition that G should have. It is easy to see that for the path with n edges $\overline{cc}(P_n) = 2/n$. So the bound or a similar one can't hold without that restriction and for $k = 1$. But other vertices, if not simplicial, still might contribute with a high value of cc. Since the number of potential edges in the neighborhood grows quadratically with the size of that neighborhood low degree vertices will in fact easier fulfill such a condition. One aim of this paper is to make such an observation into a precise counting argument which will prove that in some classes of sparse graphs there will always be enough low degree vertices that contribute with high values to the clustering coefficient.

To fulfill our goal of randomly generating graphs that have a high clustering coefficient we will generalize the constructions that lead to k-trees and chordal graphs. Our construction uses so-called contexts that are analogous to the cliques of a tree-decomposition. Section 2 introduces our model in detail. Section 2 then shows the relationship to previously known graph classes and proves some basic properties of the construction.

In Section 3 we will prove a more general bound for the clustering coefficient that is obtained by this construction. By this we can guarantee that all random graphs that are generated by it will have a non-vanishing clustering coefficient. In contrast to Observation 1, this will also take non-simplicial vertices into account.

To formulate the full statement of the bound as it is given there is not possible without the notations that are introduced in Section 2. But translated to the special case of k-trees it reads as follows.

Lemma 1. *Let G be a k-tree for some $k > 1$. Then $\overline{cc}(G) \geq \dfrac{k-1}{4k^2}$.*

Observe, that compared to Observation 1 here we got rid of the restriction on the shape of the tree-decomposition. The lower bound on k is necessary because of examples like P_n as mentioned above. If a graph contains an abundant number of induced P_2 its clustering coefficient can be brought arbitrarily low.

2 The Model and its Basic Properties

We will attempt to model the genesis of large interconnection networks. As mentioned above, our model will be an extension of the construction that leads to k-trees and more generally to chordal graphs. The idea of this paper is that we will distinguish the *observable*, generally a graph of relations, from an implicit family of *concepts* or *contexts* that define it, but which are in general not or only partially observable. These contexts correspond to the cliques of the tree-decomposition of chordal graphs. The genesis of these structures will be described as a *process*, *i.e* as an evolution of a combinatorial object *in time*.

This idea that the edges of the graph under investigation come from a more or less hidden structure of cliques was already implicit in Ravasz and Barabási [2003] and has been verified for a large number of application graphs, see Guillaume and Latapy [2004]. Implicitly it also is present in [Dorogovtsev and Mendes, 2003, Sec. 5.13 ff.] where techniques for the growth of graphs are introduced and for which a bound of $\overline{cc}$ is given. In fact, in our terminology this technique boils down to the generation of k-trees and what will be described in the sequel can be viewed as a generalization of that approach. As a basic example throughout this paper we will use an interconnection network of which we, as scientists, are all concerned: the graph of co-authorship. In that graph the vertex set is formed by the 'objects' *Ob* under investigation which in the example in fact are 'subjects', namely the authors that have been contributing to a specific scientific domain. We add an edge $e = \{ob_1, ob_2\} \in E$ if ob_1 and ob_2 have co-authored a scientific paper of the domain.

As we already see in this basic example, the implicit structure that we investigate is richer than the just the graph (Ob, E). In particular we have an important family of implicit objects or co-objects *Co* which are the scientific papers. Each such paper $co \in Co$ describes the context of a collaboration between a set of colleagues and the relational structure E is derived from them.

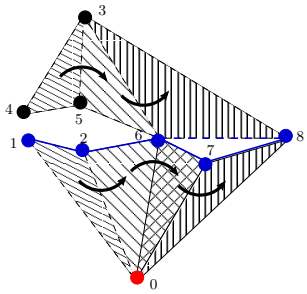
Other structures that follow the same vein of 'families of hidden cliques defining a graph' can be found in application graphs:

- For metabolic networks bio-chemical reactions define cliques.
- For the Internet on IP level a direct IP connection between two hosts is usually provided through a connection to the same switch or router, or by listening to a shared physical medium (radio or ethernet). Thus the Internet graph can be seen as generated from the local 'link layer' networks that each form local cliques.
- For social networks, connection between individuals can often be attributed to the common membership of in a social structure such as the family, the work place, school classes, church etc.
- For semantic networks two concepts (scientific papers, web pages, ...) are linked if they are co-referenced in some text (other paper, other web page, ...). Here the referring context defines a clique of all its referred objects.

What the modeling in previous papers failed to explain (up to our best knowledge) was the inter-relationship between those cliques, *e.g* the non-trivial overlap of those cliques, and the temporal evolution of these networks respectively their genesis.

τ	-2	-1	0	1	2	3	4
ob_0	0			0	0		0
ob_1	1						
ob_2	2			2			
ob_3		3	3			3	
ob_4		4					
ob_5		5	5				
ob_6			6	6	6	6	
ob_7					7		7
ob_8						8	8

(a) A sequence of contexts



(b) The graphical representation

Fig. 1 A graph that is generated by a sequence of contexts. On the left, columns corresponds to contexts and are given in generation order. Lines denote the contexts in which an object appears. The boxed entries denote the initial subsequence where a new element is '*created*'. On the right, the contexts correspond to triangles. The edges of the graph are the edges of these triangles.

Objects and their Contexts. More formally, we will investigate pairs (Ob, Co) where Ob is a set (usually finite) and $Co \subseteq 2^{Ob}$ is some family of subsets over Ob. We will refer to Ob as the set of *objects*, *e.g* the members of a scientific community, and to Co as *contexts* in which these objects appear together, *e.g* the scientific papers that they co-author. We will say that $ob_1, ob_2 \in Ob$ are *linked* if there is some $co \in Co$ such that $\{ob_1, ob_2\} \subseteq co$. The set of edges, relations or links E is then defined by $E_{Ob,Co} = \{\{ob_1, ob_2\} \mid \exists co \in Co, ob_1 \text{ and } ob_2 \in co \text{ with } ob_1 \neq ob_2\}$.

First, observe that from that definition (Ob, E) has no loops. Second, in this formal definition Ob and Co play 'opposite' sites in a bipartite relation that is defined by the containment relation $\in$. In view of the combinatorial structure the emphasis of Ob being the first set of the pair and Co the second is arbitrary ("just" given by the application). For the example, we could equally well be interested in the relationship among the papers, linking two papers if they share a common author. A context for papers then corresponds to the oeuvre of a scientist.

These pairs (Ob, Co) are considered as parts of a process (the 'genesis') of a growing structure. Namely we look at sequences $((Ob_\tau, Co_\tau))_{\tau=(-\aleph+1)\ldots,0,1,\ldots}$ where the parameter τ can be thought of as discretized time, $\aleph > 0$ defines a number of pre-existent contexts, and we have that

$$Ob_{-\aleph+1} \cdots \subseteq Ob_0 \subseteq Ob_1 \subseteq \cdots \quad \text{and} \quad Co_{-\aleph+1} \cdots \subseteq Co_0 \subseteq Co_1 \subseteq \cdots$$

In terms of the link graph this defines a growing sequence of graphs $G_i = (Ob_i, E_i)$, with $E_i = E_{Ob_i, Co_i}$:

$$(\emptyset, \emptyset) = (Ob_{-\aleph}, E_{-\aleph}) \subset (Ob_{-\aleph+1}, E_{-\aleph+1}) \cdots \subseteq (Ob_0, E_0) \subseteq (Ob_1, E_1) \subseteq \cdots$$

Figure 1 shows an example of a sequence of contexts and the resulting graphs.

What is usually observed in applications is only part of the genesis, *e.g* some or just one of the graphs. The number of vertices (resp. edges) at time τ are denoted with n_τ and m_τ respectively, *i.e*

$$\mathrm{n}_\tau = |Ob_\tau| \qquad \text{and} \qquad \mathrm{m}_\tau = |E_\tau| .$$

To describe such a genesis we will assume that one step from (Ob_τ, Co_τ) to $(Ob_{\tau+1}, Co_{\tau+1})$ is given by exactly one new context. That is, there is an enumeration of the contexts $\ldots, co_0, co_1, co_2, \ldots$ such that

$$Ob_\tau = \bigcup_{t\le\tau} co_t \qquad \text{and} \qquad Co_\tau = \bigcup_{t\le\tau} \{co_t\}.$$

The potentially infinite base sets for $\tau \to \infty$ are denoted Ob_∞ and Co_∞. Generally, we will also suppose that the sequence has no *redundancy*, *i.e* that for all τ there are $ob, ob' \in Ob_\tau$ such that $\{ob, ob'\} \notin E_{\tau-1}$. For all τ we will denote this set of non-redundant edges $\bar{E}_\tau = E_\tau \setminus E_{\tau-1}$ for which we thus have $\bar{E}_\tau \neq \emptyset$. Another property that we assume for the sequence is that it respects inclusion in the following sense. For $\tau < \kappa$, the new elements that appear in Ob_τ fulfill

$$Ob_\kappa \setminus Ob_{\tau-1} \not\subset Ob_\tau \setminus Ob_{\tau-1}, \tag{2}$$

i.e no context appearing later than co_τ in the sequence will add less elements to $Ob_{\tau-1}$ than co_τ.

Even with this property the exact ordering of the contexts will be arbitrary. In fact, if $Ob_\tau = Ob_{\tau+1}$ the contexts co_τ and $co_{\tau+1}$ can be considered *interchangeable*. A subsequence $co_\tau, \ldots, co_{\tau+\ell}$ in $(co_i)_{i=0,\ldots}$ is *stable* if all adjacent elements are interchangeable, or, in other words $Ob_\tau = \cdots = Ob_{\tau+\ell}$. It is *maximally stable* if it is stable and may not be extended to the left or right without loosing that property. We then also have that $Ob_\tau \cap Ob_{\tau-1} = \cdots = Ob_{\tau+\ell} \cap Ob_{\tau-1}$.

With that definition we may subdivide our sequence uniquely into maximally stable subsequences. For each τ, $start_\tau$ denotes the start index of the maximal stable subsequence to which co_τ belongs. $\mathcal{L}_\tau$ denotes the number of contexts in that subsequence. Both values are independent of the particular ordering of the subsequence. Also we associate to each such maximal stable subsequence the set of newly introduced objects, $create_\tau = Ob_\tau \setminus Ob_{start_\tau - 1}$.

The starting point of the genesis. In a genesis as we attempt to describe here, new objects and contexts will emerge from ones that previously exist. Clearly this is only possible if we assume the existence of some of them initially. In a sequence of contexts we will thus assume that there is a finite number $\aleph$ of predefined contexts $co_{-(\aleph-1)}, \ldots, co_0$. The parameters n_0 and m_0 are thus the number of vertices and edges that we assume present before the genesis starts, and which we assume to be finite numbers.

Besides some more or less obvious requirements (*e.g* that we only may connect to a clique of a required size if there is one), the statistical properties of the graphs that

will result below will not much depend on the initial choice of contexts. They will be dominated by the many other choices during the genesis. Such as the cristallization germ of a snowflake is very important for it to form initially but by itself it will not have much influence on the final shape.

The bias introduced by imitation. In our genesis, the dependence of the process from previous choices is an important detail that we have to handle. We propose a relatively simple model, in which each new $co \in Co$ depends on one previously known other element. In our example of the graph of authorship, a new work 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.

So in general we suppose that for each $\tau > 0$ there is $\rho(\tau) < \tau$, such that for

$$\mathit{stab}_\tau = \mathit{co}_{\rho(\tau)} \cap \mathit{co}_\tau \qquad \mathit{old}_\tau = \mathit{co}_{\rho(\tau)} \setminus \mathit{co}_\tau \qquad \mathit{new}_\tau = \mathit{co}_\tau \setminus \mathit{co}_{\rho(\tau)} \tag{3}$$

we have that $\mathit{stab}_\tau \neq \emptyset$ and $\mathit{new}_\tau \neq \emptyset$. For constructing co_τ a pre-existing $\mathit{co}_{\rho(\tau)}$, the *paragon*, is chosen, copied into a new set co_τ in which old_τ is replaced by new_τ.

Now, the type of transformations that are permitted when going from $\mathit{co}_{\rho(\tau)}$ to co_τ will be much dependent on the particular domain; different sets of rules will lead to specific families of graphs. We will investigate simple deterministic and statistical properties of such rules. Therefore, we will introduce some parameters on the sizes of these sets that could describe the evolution in different application domains, either by following some deterministic rule, or just by some statistical correlation. These parameters may then be used to describe an observed sequence or to randomly sample a 'typical' member of a specific family.

- $\mathcal{D}_\tau$ the *size* of co_τ
- $\mathcal{O}_\tau$ the number of replaced objects, $|\mathit{old}_\tau|$
- $\mathcal{N}_\tau$ the number of replacing objects, $|\mathit{new}_\tau|$
- $\mathcal{S}_\tau$ the number of *sporadic* objects that had been present before this stable subsequence but are reintroduced into the new context, *i.e* $|\mathit{new}_\tau \setminus \mathit{create}_\tau|$.
- $\mathcal{L}_\tau$ the length of the maximal stable sequence containing co_τ as defined above

We expect that applications usually could provide sensible values (resp. distributions) for these parameters. In the case of the co-authorship graph, e.g, it should be possible to describe the distribution of the number of authors of a paper ($\mathcal{D}$), the typical number of papers to which a young PhD student contributes at the beginning of his career ($\mathcal{L}$), the scientific heritage from the initial environment (relation ρ).

Special cases for fixed parameters. In the following we will look at cases such that the parameters from the previous section are equal to some constant for all $\tau > 0$. A sequence which fulfills these constraints for fixed values of $\mathcal{D}, \mathcal{O}, \mathcal{N}, \mathcal{S}$, and $\mathcal{L}$ will be called a $(\mathcal{D}, \mathcal{O}, \mathcal{N}, \mathcal{S}, \mathcal{L})$-*sequence*. Whenever we fix only some of these parameters we will replace those that are not fixed by the symbol '*'.

Fixing the values for some of these parameters lead to the genesis of families of graphs that are already well studied. Here we always assume that the initial start of

the genesis is always given by just one initial context co_0. A schema for the interrelationship of such graph families is given in Figure 2. In particular, known classes correspond to the following values:

$(\mathcal{D}, \mathcal{O}, \mathcal{N}, \mathcal{S}, \mathcal{L})$

$(2, 1, 1, 0, t)$ The contexts themselves are just edges, one of the objects of the paragon is replaced by the new object and every such new object appears as new in a stable sequence of length exactly t for some constant $t > 0$. This leads to the genesis by preferential edge attachment as described by Barabási and Albert [1999]. Suppose that for each τ the choice of ρ_τ is uniform among all possible values. Then, each vertex v is chosen with a probability that is proportional to its degree at instance τ.

$(2, 2, 2, 1, *)$ This is the limit case for which a complete renewal for each new edge that is added is performed. It leads to random graphs similar to the model of Erdős and Rényi [1960]; at any point choosing any of the pairs (v, w) is equally likely.

$(*, *, *, 0, 1)$ Each new context introduces new vertices and doesn't push old vertices into new contexts. This leads to the genesis of chordal graphs. It is easy to see that our genesis is a reverted elimination ordering, that ρ defines a tree and that that tree is a tree decomposition of the graph. In the restricted case that $\mathcal{D}_\tau \equiv k + 1$ is also fixed and $\mathcal{O}_\tau \equiv 1$, *i.e* $(k + 1, 1, 1, 0, 1)$-sequences, the graphs are the k-trees. For $\mathcal{D}_\tau = 2$, *i.e* $(2, 1, 1, 0, 1)$, we then just have trees.

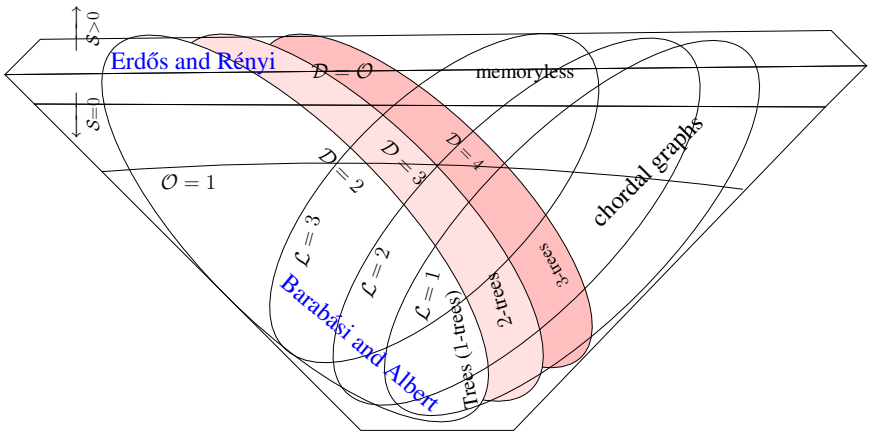


Fig. 2 Special cases of graph classes in function of the parameters $\mathcal{D}$, $\mathcal{O}$, $\mathcal{S}$ and $\mathcal{L}$. The topline '*memoryless*' refers to the fact that in the case that $\mathcal{D} = \mathcal{O}$ (*e.g* Erdős and Rényi graphs) each choice is independent from previous choices. The blob with $\mathcal{L} = 1$ are the chordal graphs, and intersecting this blob with those for $\mathcal{D} = 2, 3, \ldots$ gives rise to k-trees.

3 Parameter Estimations

Estimating the number of edges. An important feature of most real world graphs is their sparseness, *i.e* the fact that they usually have an average degree that is bounded by some small constant. In this section we will show how such a claim holds for our proposed genesis. In the next section, such bounds then will be a major ingredient to prove a bound on the clustering coefficient.

An important case for bounding the clustering coefficient from below will be sequences that don't have sporadic occurences of objects, *i.e* with $\mathcal{S} = 0$. As seen above these occur as generalizations of well studied classes of graphs, so studying the resulting families might be of interest of its own.

Lemma 2. *Let $\mathcal{S} = 0$ and $\tau = start_\tau$, i.e τ is the starting point of a maximal stable sequence. Then the number of edges $|E_\tau \setminus E_{\tau-1}|$ added by context co_τ is $|stab_\tau| \cdot |create_\tau| + \binom{|create_\tau|}{2}$.*

Proof. Let $stab_\tau = co_\tau \cap co_{\rho(\tau)} = \{ob_1, \ldots, ob_d\}$ be the objects that had been copied into co_τ. Since all of them are already present in $co_{\rho(\tau)}$, no new edge between them is added at time τ.

Now let $ob \in create_\tau$ be any new object. By definition all edges induced for it in co_τ must be new. So we get $|stab_\tau|$ new edges for it that link to older elements.

Among the objects in $create_\tau$ we create all pairs of possible edges. □

Let us now restrict even further by fixing $\mathcal{O} = 1$ and $\mathcal{S} = 0$, *i.e* classes where going from $co_{\rho(\tau)}$ to co_τ replaces *exactly one* object $ob \in old_\tau$ by the unique element $ob' \in create_\tau$.

For the simplicity of the arguments we will first assume that the following property holds for ρ: if $\tau < \tau'$ are such that co_τ and $co_{\tau'}$ are members of the same stable sequence ($create_\tau = create_{\tau'}$) then their predecessors by ρ are mutually disjoint.

$$co_{\rho(\tau)} \cap co_{\rho(\tau')} = \emptyset \tag{4}$$

We call such a sequence *distinctive*.

Lemma 3. *For a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$, the number of edges $|E_\tau \setminus E_{\tau-1}|$ added by context co_τ is $\mathcal{D}_\tau - 1$.*

Proof. If $\tau = start_\tau$, *i.e* is first in its stable subsequence, this is Lemma 2, since $\mathcal{O} = |create_\tau| = 1$, $|stab_\tau| = |co_\tau| - 1$ and thus $|stab_\tau| \cdot |create_\tau| = |co_\tau| - 1$.

If $\tau \neq start_\tau$, we have that $co_h \cap co_\tau = create_\tau$ for all $start_\tau \leq h < \tau$. Thus no edge that is induced by co_τ may have been created previously. □

Corollary 1. *For a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$, we have that*

$$|E_\tau| = m_0 - \tau + \textstyle\sum_{\theta \leq \tau} \mathcal{D}_\theta = m_0 + \tau(\widetilde{\mathcal{D}}_\tau - 1). \tag{5}$$

Lemma 4. *Suppose that for a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$ there are constants δ^+ and λ^+ with $\widetilde{\mathcal{D}}_\tau \leq \delta^+$ and $\widetilde{\mathcal{L}}_\tau \leq \lambda^+$. Then the average degree of G_τ is $\delta_\tau \approx \frac{1}{2} \cdot \widetilde{\mathcal{L}}_\tau \left(\widetilde{\mathcal{D}}_\tau - 1\right)$. More precisely $\left|\delta_\tau - \frac{1}{2} \cdot \widetilde{\mathcal{L}}_\tau \left(\widetilde{\mathcal{D}}_\tau - 1\right)\right| = O\left(\frac{1}{\tau}\right)$.*

Proof. With some constant C that only depends on $\lambda^+, \delta^+, \mathrm{m}_0$ and n_0 we have that

$$\left|\delta_\tau - \frac{1}{2} \cdot \widetilde{\mathcal{L}}_\tau \left(\widetilde{\mathcal{D}}_\tau - 1\right)\right| = \frac{1}{2} \cdot \left| \frac{\mathrm{m}_0 + \tau \left(\widetilde{\mathcal{D}}_\tau - 1\right)}{\mathrm{n}_0 + \frac{\tau}{\widetilde{\mathcal{L}}_\tau}} - \widetilde{\mathcal{L}}_\tau \left(\widetilde{\mathcal{D}}_\tau - 1\right) \right| \tag{6}$$

$$= \frac{1}{2} \cdot \left| \frac{\widetilde{\mathcal{L}}_\tau \mathrm{m}_0 - \mathrm{n}_0 \widetilde{\mathcal{L}}_\tau^{\,2} \left(\widetilde{\mathcal{D}}_\tau - 1\right)}{\widetilde{\mathcal{L}}_\tau \mathrm{n}_0 + \tau} \right| \leq \frac{C}{\tau} \tag{7}$$

□

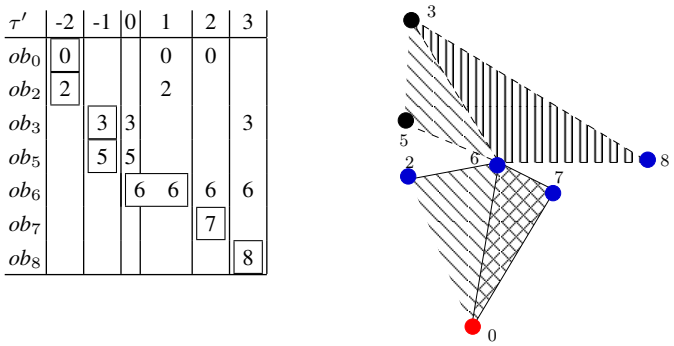
A graph G has *arboricity* a if its edges can be subdivided into a family $F_1, \ldots, F_a$ of forests over the same vertex set. Many families of graphs that have been traditionally studied have bounded arboricity, namely all classes that are closed under the so-called graph-minor operation, see Mader [1967]. From the definition follows that then necessarily $|E(G)| \leq a \cdot |V(G)|$, so G must be sparse to have low arboricity. The converse is generally not true: there are sparse graphs that have high arboricity, so having low arboricity is a stronger requirement than being sparse. For our graph genesis we obtain the following lemma. It shows that not only the density of the graphs is bounded in terms of parameters of the sequence, but also their arboricity.

Lemma 5. *Suppose that we have a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$. Denote the arboricity of the initial graph G_0 by a_0. Then for $\tau > 0$ the arboricity of G_τ is $a_\tau \leq \max\{a_{\tau-1}, \mathcal{L}_\tau \cdot (\mathcal{D}_\tau - 1)\}$.*

For a proof just observe that at each τ the newly created edges that link the new vertex are no more than $\mathcal{L}_\tau \cdot (\mathcal{D}_\tau - 1)$ and can each be assigned to a different forest $F_1, \ldots, F_{a_\tau}$ in the subdivision.

Restrictions to the Neighborhood of a Vertex. The aim of this section is to show that for a large variety of choices of the parameters the graph that is induced by the neighborhood of a vertex has interesting properties. In particular it contains a large subgraph that again can be obtained from the same generating process (but with other parameters). This 'recursive' local structure can then be used to bound the clustering coefficient from below. Notable exceptions from this structural property are the "*classical*" models of random graphs, in our terminology when $\mathcal{D} = 2$. On the other hand whenever $\widetilde{\mathcal{D}}$ is bounded away from 2 the cluster coefficient can be shown to be bounded in consequence.

Let $(co_\ldots)$ be a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$. Let $ob \in Ob_\infty$ be an object and let $\left(co_\ldots^{[ob]}\right)$ be the subsequence of contexts that contain ob. Figure 3 shows the case for 6 from Figure 1.



(a) The restricted sequence of contexts

(b) The induced graph

Fig. 3 The restriction $\left(co_{\dots}^{[ob_6]}\right)$ to ob_6. Only columns touching ob_6 appear and only lines corresponding to ob_6 and its neighbors.

Observe that this definition only talks about the subsequence that is induced by the membership of *ob*. Other edges between the neighbors of *ob* might by induced by contexts that do not include *ob* and will thus occur sporadically in the genesis. Figure 1 illustrates such cases: when adding the two contexts for object 8 a new edge $\{7, 8\}$ appears in the neighborhood of 6. Another example is edge $\{6, 8\}$ that appears in the neighborhood of 0. The existence of these edges is not deducible from the contexts in which 6 (resp. 0) is involved directly. We will call such edges *locally sporadic* for the corresponding vertex. Sporadic edges may not only occur after an object has been newly introduced. In Figure 1 the edges $\{0, 6\}$ and $\{6, 7\}$ are sporadic for 8.

Lemma 6. *Let $(co_{\dots})$ be a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$. Let $ob \in Ob_\infty$ be an object and let $\left(co_{\dots}^{[ob]}\right)$ be the subsequence of contexts that contain ob. Then $\left(co_{\dots}^{[ob]}\right)$ is a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$, too.*

Because of this possible occurrence of locally sporadic edges the following lemma only gives a lower bound on the number of edges in the neighborhood of an object.

Lemma 7. *Let $(co_{\dots})$ be a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$. Let $ob \in Ob_\infty$ be an object, let $\left(co_{\dots}^{[ob]}\right)$ be the subsequence of contexts that contain ob, and let $\mathcal{L}^{[ob]}$ and $\mathcal{D}^{[ob]}$ the respective parameters of that subsequence. Then the number of edges in the neighborhood of ob at time τ' is at least*

$$\alpha + \beta + \tau' \left(\widetilde{\mathcal{D}^{[ob]}}_{\tau'} - 2 \right) \text{ with } \alpha = \left| E_{-1} \cap N(ob) \right|, \beta = \sum_{\tau=\tau_0}^{\tau_0+\mathcal{L}_{\tau_0}} \binom{\mathcal{D}_\tau - 1}{2}. \tag{8}$$

α is the number of edges that were initially present in the neighborhood of *ob* and β is the number of edges in the neighborhood that are added before *ob* is created at time τ_0 and that would be induced by the generating contexts of *ob*.

Observe that here the time variable τ' (and thus the averaging) only accounts for those events in which *ob* is involved in the original sequence. In particular, this means that convergence to the approximate value will be much 'slower' compared to the original sequence.

Also notice that the terms in (8) are always well defined since $\mathcal{D} \geq 2$ and therefore $\widetilde{\mathcal{D}} \geq 2$, too. For contexts τ with $\mathcal{D}_\tau = 2$ the contribution to the sum in β is 0. So $\mathcal{D}_\tau = 2$ clearly is a borderline case where we have a important property change. In consequence the following theorem needs a restriction to $\mathcal{D}_\tau > 2$.

Theorem 2. *Suppose that for a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$, we have that $\widetilde{\mathcal{D}}_\tau \leq \delta^+$ and $\widetilde{\mathcal{L}}_\tau \leq \lambda^+$ for some constants δ^+ and λ^+. Suppose in addition that there are some integers δ^-, λ^- with $2 < \delta^- \leq \mathcal{D}_\tau$ and $1 \leq \lambda^- \leq \mathcal{L}_\tau$ for all τ. Then for $R^- = \lambda^-(\delta^- - 2)$ and $R^+ = \lambda^+(\delta^+ - 1)$ the clustering coefficient is bounded from below, namely for τ sufficiently large we have $\overline{cc}(G_\tau) \geq \frac{1}{4} \cdot \frac{R^-}{{R^+}^2}$.*

Proof. From Lemma 4 we know that if τ is large enough (*i.e* $\tau \gg C$) at least half of the vertices have a degree of at most $2R^+$. Therefore for the clustering coefficient we get

$$\begin{aligned}\frac{1}{n_\tau} \sum_{ob \in Ob_\tau} \frac{\mathrm{NE}_\tau(ob)}{\binom{\deg_\tau(ob)}{2}} &\geq \frac{1}{n_\tau} \sum_{\substack{ob \in Ob_\tau \\ \deg_\tau(ob) \leq 2R^+}} \frac{\mathrm{NE}_\tau(ob)}{\binom{2R^+}{2}} \\ &\geq \frac{1}{n_\tau} \sum_{\substack{ob \in Ob_\tau \\ \deg_\tau(ob) \leq 2R^+}} \frac{\lambda^-(\delta^- - 2)}{\binom{2R^+}{2}} \\ &\geq \frac{\lambda^-(\delta^- - 2)}{4{R^+}^2}\end{aligned} \tag{9}$$

And the claim follows. □

Notice that (9) only uses part of (8).

Consider the graph from Figure 1, again. Table 1 shows the statistics for the individual vertices. It turns out that the average clustering coefficient of this graph is quite high, namely 0.7. For the example graph we have $\delta^- = \delta^+ = 3$, $\lambda^- = 1$ and $\lambda^+ = 1.7$. The bound of Theorem 2 evaluates to 0.0225, which is far from the real value of 0.7.

	$\deg_4$	NE_4	$\binom{\deg_4}{2}$	cc_4
0	5	5	10	0.5
1	2	1	1	1.0
2	3	2	3	0.7
3	4	3	6	0.5
4	2	1	1	1.0
5	3	2	3	0.7
6	6	6	15	0.4
7	3	3	3	1.0
8	4	4	6	0.7
av	3.6	3		0.7

Table 1 degree statistics of the graph in Figure 1

4 Some Experimental Results

A first implementation of the random generation processes that are described here have been undertaken. They show that our approach is feasible for a large variety of parameters. It has been used to generate graphs in the range from several thousand to several million vertices. The experiments were only limited by two factors, namely the memory requirements to store the resulting graphs and the computing time that is needed for the approximation of the clustering coefficient, see Schank and Wagner [2005].

A full description of those results will be reported in a separate paper, a first preliminary report is available, see Gustedt and Schimit [2008]. Here we just like to emphasize on the threshold of $\mathcal{D} = 2$ that is apparent in Theorem 2. Figure 4 illustrates the fundamental difference in the properties of the resulting graphs for $\mathcal{D} = 2$ in Figure 4(a) (logarithmic scale in y) and $\mathcal{D} = 4$ for 4(b) (linear scale in y). For both we plot the approximation of the clustering coefficient for a set of randomly generated graphs. Whereas in the first case we have an exponential decrease of the clustering coefficient as the size of the graphs grows, the second clearly shows that we attain a non-zero limit.

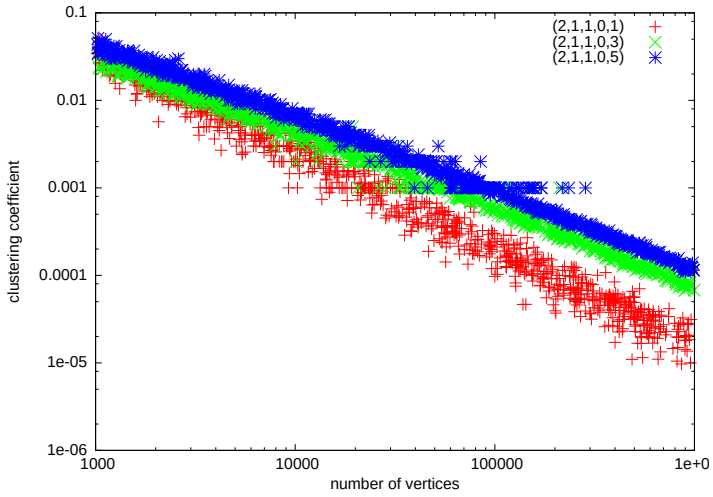
5 Conclusion and Outlook

In this paper we have presented a modeling framework for the genesis of graphs. It can used for the random generation of sparse graphs for which we are able to guarantee a high local density namely a high clustering coefficient. Our approach is practical and provides quick access to large families of sample graphs that can be used for simulation and for testing.

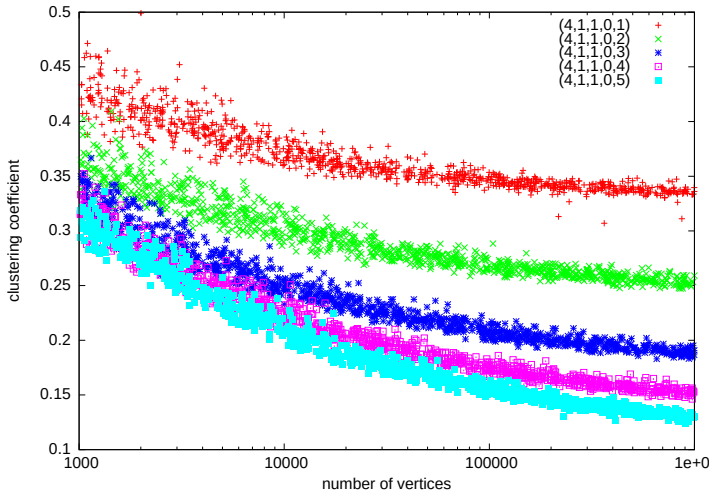
Still, a more profound investigation of the properties of the generated graphs has to be undertaken. In particular, it will be interesting to emulate the degree and clique constraints that have been observed in different application domains. In addition, the paragon relation $\rho(\tau)$ among the contexts as it is introduced here needs more study. We have to check to what extent it models the historical bias during a graph genesis realistically.

Another direction of future research has become apparent during the first experimental studies of our approach: compared to some application graphs such as the Internet graph and social networks the graphs that we have been able to generate were quite '*small*', they had only a million vertices. To generate larger sample graphs that could be the test input for large scale tests of distributed algorithms we will need to parallelize our approach.

Acknowledgements The author likes to thank Matthieu Latapy for interesting discussions and pointers on the subject and Pedro Schimit for a first implementation. Comments and suggestions by the anonymous reviewers have been much helpful to improve the paper for the final version.



(a) $\mathcal{D} = 2$ with $\mathcal{L} = 1, 3$ and 5



(b) $\mathcal{D} = 4$ with $\mathcal{L} = 1$ to 5

Fig. 4 Experimental results of the clustering coefficient in terms of the number of vertices (10^3 to 10^6). Each data point corresponds to a randomly generated graph G and plots the number of vertices $|V(G)|$ against the clustering coefficient $\overline{cc}(G)$. The graphs are generated such that $\log_{10} |V(G)|$ is uniformly distributed in $[3, 6)$.

References

S. Arnborg. Efficient algorithms for combinatorial problems on graphs with bounded decomposability – A survey. *BIT*, 25:2–23, 1985.

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

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, Feb. 2006. URL http://cogimage.dsi.cnrs.fr/publications/2006/BLMCH06.

U. Brandes, M. Gaertler, and D. Wagner. Engineering graph clustering: Models and experimental evaluation. *ACM Journal of Experimental Algorithms*, 12, 2007.

L. da Fontoura Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56:167, 2007. URL doi:10.1080/00018730601170527.

S. Dorogovtsev and J. 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.

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

J. Gustedt and P. 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.

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

W. Mader. Homomorphieeigenschaften und mittlere Kantendichte von Graphen. *Math. Ann.*, 174: 265–268, 1967.

Z. Nikoloski, N. Deo, and L. Kucera. Degree-correlation of scale-free graphs. In S. Felsner, editor, *2005 European Conference on Combinatorics, Graph Theory and Applications (EuroComb '05)*, volume AE of *DMTCS Proceedings*, pages 239–244. Discrete Mathematics and Theoretical Computer Science, 2005. URL http://www.dmtcs.org/dmtcs-ojs/index.php/proceedings/article/view/dmAE0148.

E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67 (2):026112, Feb 2003. doi: 10.1103/PhysRevE.67.026112.

N. Robertson and P. Seymour. Graph minors II, algorithmic aspects of tree-width. *J. Algorithms*, 7:309–322, 1986.

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

P. R. Villas Boas, F. A. Rodrigues, G. Travieso, and L. da Fontoura Costa. Chain motifs: The tails and handles of complex networks. *Physical Review E*, 77(2):026106, 2008. doi: 10.1103/PhysRevE.77.026106. URL http://link.aps.org/abstract/PRE/v77/e026106.