HAL
open science

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

Jens Gustedt

To cite this version:

Jens Gustedt. Generalized Attachment Models for the Genesis of Graphs with High Clustering Coefficient. Santo Fortunato and Giuseppe Mangioni and Ronaldo Menezes and Vincenzo Nicosia. Complex Networks - Results of the 2009 International Workshop on Complex Networks (CompleNet 2009), 207, Springer Berlin / Heidelberg, pp.99-113, 2009, Studies in Computational Intelligence, 978-3-642-01205-1. 10.1007/978-3-642-01206-8_9 . inria-00312059v1

HAL Id: inria-00312059

https://inria.hal.science/inria-00312059v1

Submitted on 25 Aug 2008 (v1), last revised 3 Jun 2009 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

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

Jens Gustedt

N° 6622

2008-08-19

Thème NUM

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6622--FR+ENG

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

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

Jens Gustedt

Thème NUM — Systèmes numériques
Équipe-Projet AlGorille

Rapport de recherche n° 6622 — 2008-08-19 — 16 pages

Abstract: Commonly used techniques for the random generation of graphs 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 well-known procedures such as those of Erdős & Rény and Barabási & Albert. 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

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

Modèles généralisés d'attachment préférentiel pour la genèse de graphes avec coefficient de clustering élevé

Résumé : Les techniques courrament utilisés pour la génération aléatoire de graphes ont deux inconvéniants ; notement leur manque de biais concernant l'évolution du graphe, et, leur incapacité de produire des familles de graphes dont le coefficient de clustering ne disparait pas ou même pourrait être préscrit à une valeur choisie. Dans ce travail nous proposons un modèle pour la genèse de graphes qui s'attaque ces deux questions. Interprété en tant que procédure de génération aléatoire il généralise deux procédure bien connues, d'un coté celle d'Erdős & Rény et celle de Barabási & Albert de l'autre ; juste vue en tant que schéma de composition de graphes il généralise les schéma d'élimination parfaite des graphes chordaux. Itérativement, le modéle ajoute des *contextes* qui introduisent des dépendences explicites à l'évolution du graphe. Ainsi il reflettent un biais historique pendant l'évolution qui va plus loin que les contraintes simples de l'attachement préférentiel. En fixant certains quantités statistiques durant la genèse nous sommes capable de générer des familles de graphes dont le coefficient de clustering peut être effectivement borné par le bas, substentiellement au dessus de zéro.

Mots-clés : modèles d'attachement préferentiels, génération aléatoire de graphes, k-trees

Rev. 2477 as of 2008-08-19 21:17:04

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.

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.

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* $cc(G)$ 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* $lcc(G, v)$ at v is the quotient between the amount of edges inside the neighborhood of v and the maximum possible amount of such edges. In closed form these definitions read

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

$$cc(G) = \frac{1}{|V(G)|} \sum_{v \in V(G)} lcc(G, v) \tag{2}$$

Observe that *lcc* is always well defined, even if $deg_G(v)$ is 0 or 1 since then v is simplicial.

With this definition for the border cases of degree 1 we have in particular that *cc* will be high for trees: for any tree G without vertices of degree 2, $cc(G) > 0.5$. More generally, 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 1.1 *For some $k > 0$ let G be a k-tree with a k-tree-decomposition that has no vertices of degree 2. Then $cc(G) \geq \frac{1}{2} - \frac{k-3}{2|V(G)|}$.*

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 straight forward estimations.

The bound given in Observation 1.1 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 1.1 puts a restriction on the permissible k-tree decomposition that G should have. It is easy to see that for the path with n vertices $cc(P_n) = 2/n$. 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 lcc. Since the amount 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 3 then shows the relationship to previously known graph classes and proves some basic properties of the construction.

In Section 5 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 1.1, the proof 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 1.2 *Let G be a k-tree for some $k > 1$. Then $cc(G) \geq \dfrac{k-1}{4k^2}$.*

Here, the restriction on k is necessary because of examples like P_n as mentioned above. If a graph contains an abundant amount of induced P_2 its clustering coefficient can be brought arbitrarily low.

2 The Model and its Basic Propeties

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 has been verified for a large number of application graphs, see Guillaume and Latapy (2004). 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' 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. What their modeling failed to explain 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.

2.1 Objects and their Contexts

So 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 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

$$E = E_{Ob,Co} := \{\{ob_1, ob_2\} \mid \exists co \in Co, ob_1 \neq ob_2 \in co\}\,.$$

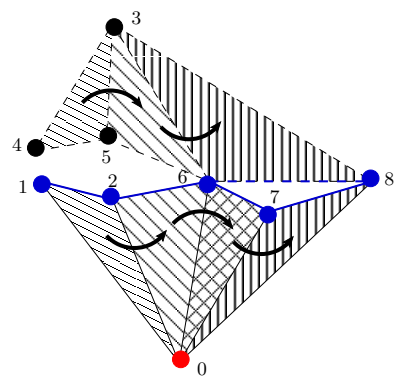
First, observe that from that definition (Ob, E) has no loops.

Second, observe that 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 a the relationship among the papers, linking two papers if they share a common author.

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

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

(a) A sequence of contexts



(b) The graphical representation

Figure 1: A graph that is generated by a sequence of contexts

parameter τ can be thought of as discretized time, $\aleph > 0$ defines a number of pre-existent contexts, and we have that

$$\begin{array}{ccccccc} Ob_{-\aleph+1} & \cdots & \subseteq & Ob_0 & \subseteq & Ob_1 & \subseteq & \cdots \\ Co_{-\aleph+1} & \cdots & \subseteq & Co_0 & \subseteq & Co_1 & \subseteq & \cdots \end{array}$$

In terms of the link graph this defines a growing sequence of graphs

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

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 amount of vertices (resp. edges) at time τ are denoted with n_τ and m_τ respectively, *i.e*

$$\begin{aligned} \mathrm{n}_\tau &= |Ob_\tau| \\ \mathrm{m}_\tau &= |E_\tau| \,. \end{aligned}$$

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

$$Ob_\tau = \bigcup_{t \leq \tau} co_t \tag{3}$$

$$Co_\tau = \bigcup_{t \leq \tau} \{co_t\}. \tag{4}$$

The potentially infinite base set of objects and contexts are denoted as

$$Ob_\infty = \bigcup_{-\aleph < t} co_t \tag{5}$$

$$Co_\infty = \bigcup_{-\aleph < t} \{co_t\}. \tag{6}$$

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

$$Ob_\kappa \setminus Ob_{\tau-1} \not\subset Ob_\tau \setminus Ob_{\tau-1}, \tag{7}$$

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. By (7) 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 of co_τ and $\mathcal{L}_\tau$ 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, $create_\tau = Ob_\tau \setminus Ob_{start_\tau - 1}$.

2.2 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 initial existence of some of them, such that those that are then created may refer to. 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 amount of vertices and edges that we assume present before the genesis starts, and which we assume to be finite numbers.

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

$$stab_\tau = co_{\rho(\tau)} \cap co_\tau \tag{8}$$

$$old_\tau = co_{\rho(\tau)} \setminus co_\tau \tag{9}$$

$$new_\tau = co_\tau \setminus co_{\rho(\tau)} \tag{10}$$

in general we have that

$$stab_\tau \neq \emptyset \tag{11}$$

$$new_\tau \neq \emptyset \tag{12}$$

For constructing co_τ a pre-existing $co_{\rho(\tau)}$, 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 $co_{\rho(\tau)}$ 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, $|old_\tau|$

$\mathcal{N}_\tau$, the number of replacing objects, $|new_\tau|$

$\mathcal{S}_\tau$, the number of *sporadic* objects that had been present before this stable subsequence but are reintroduced into the new context, *i.e* $|new_\tau \setminus create_\tau|$.

$\mathcal{L}_\tau$, the length of the maximal stable sequence containing co_τ as defined above

3 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.

$(2, 1, 1, 0, t)$ 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 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.

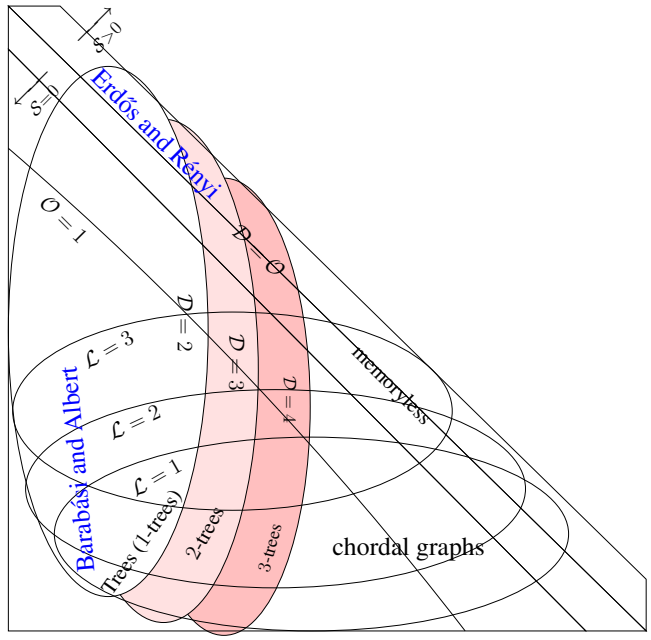


Figure 2: Special Cases of Graph Classes

$(2, 2, 2, 1, *)$ 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 chosing 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.

4 Estimating the amount of edges

An important feature of most real world graphs are 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 4.1 *Let $\mathcal{S} = 0$ and $\tau = \mathit{start}_\tau$, i.e τ is the starting point of a maximal stable sequence. Then the amount of edges $|E_\tau \setminus E_{\tau-1}|$ added by context co_τ is $|\mathit{stab}_\tau| \cdot |\mathit{create}_\tau| + \binom{|\mathit{create}_\tau|}{2}$.*

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

Now let $\mathit{ob} \in \mathit{create}_\tau$ be any new object. By definition all edges induced for it in co_τ must be new. So we get $|\mathit{stab}_\tau|$ new edges for it that link to older elements.

Among the objects in create_τ 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 $\mathit{co}_{\rho(\tau)}$ to co_τ replaces *exactly one* object $\mathit{ob} \in \mathit{old}_\tau$ by the unique element $\mathit{ob}' \in \mathit{create}_\tau$.

For the simplicity of the arguments we will first assume that the choice of ρ is such that for all $\tau < \tau'$ such that co_τ and $\mathit{co}_{\tau'}$ are in the same stable sequence ($\mathit{create}_\tau = \mathit{create}_{\tau'}$) their predecessors by ρ are mutually disjoint

$$\mathit{co}_{\rho(\tau)} \cap \mathit{co}_{\rho(\tau')} = \emptyset. \tag{13}$$

We call such a sequence *distinctive*.

Lemma 4.2 *For a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$, the amount of edges $|E_\tau \setminus E_{\tau-1}|$ added by context co_τ is $\mathcal{D}_\tau - 1$.*

Proof: If $\tau = \mathit{start}_\tau$, *i.e* is first in its stable subsequence, this is Lemma 4.1, since $\mathcal{O} = |\mathit{create}_\tau| = 1$, $|\mathit{stab}_\tau| = |\mathit{co}_\tau| - 1$ and thus $|\mathit{stab}_\tau| \cdot |\mathit{create}_\tau| = |\mathit{co}_\tau| - 1$.

If $\tau \neq \mathit{start}_\tau$, we have that $\mathit{co}_h \cap \mathit{co}_\tau = \mathit{create}_\tau$ for all $\mathit{start}_\tau \leq h < \tau$. Thus no edge that is induced by co_τ may have been created previously. □

Corollary 4.3 *For a distinctive sequence with $\mathcal{O} = 1$ and $\mathcal{S} = 0$, we have that*

$$|E_\tau| = m_0 - \tau + \sum_{\theta \leq \tau} \mathcal{D}_\theta \tag{14}$$

$$= m_0 + \tau(\widetilde{\mathcal{D}}_\tau - 1). \tag{15}$$

Lemma 4.4 *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^+. Then the average degree of G_τ is*

$$\delta_\tau = \frac{1}{2} \cdot \frac{m_0 + \tau\left(\widetilde{\mathcal{D}}_\tau - 1\right)}{n_0 + \frac{\tau}{\widetilde{\mathcal{L}}_\tau}} \tag{16}$$

$$\approx \frac{1}{2} \cdot \widetilde{\mathcal{L}}_\tau \left(\widetilde{\mathcal{D}}_\tau - 1\right). \tag{17}$$

More precisely

$$\left|\delta_\tau - \frac{1}{2}\cdot\widetilde{\mathcal{L}}_\tau\left(\widetilde{\mathcal{D}}_\tau - 1\right)\right| \quad = \quad O\left(\frac{1}{\tau}\right) \tag{18}$$

Proof: Equation (16) follows directly from Corollary 4.3. For (18) we have with some constant C that only depends on $\lambda^+, \delta^+, \mathrm{m}_0$ and n_0

$$\left|\delta_\tau - \frac{1}{2}\cdot\widetilde{\mathcal{L}}_\tau\left(\widetilde{\mathcal{D}}_\tau - 1\right)\right| \quad = \quad \frac{1}{2}\cdot\left|\frac{\mathrm{m}_0 + \tau\left(\widetilde{\mathcal{D}}_\tau - 1\right)}{\mathrm{n}_0 + \frac{\tau}{\widetilde{\mathcal{L}}_\tau}} - \widetilde{\mathcal{L}}_\tau\left(\widetilde{\mathcal{D}}_\tau - 1\right)\right| \tag{19}$$

$$= \quad \frac{1}{2}\cdot\left|\frac{\widetilde{\mathcal{L}}_\tau\mathrm{m}_0 - \mathrm{n}_0\widetilde{\mathcal{L}}_\tau^{\,2}\left(\widetilde{\mathcal{D}}_\tau - 1\right)}{\widetilde{\mathcal{L}}_\tau\mathrm{n}_0 + \tau}\right| \tag{20}$$

$$\leq \quad \frac{C}{\tau} \tag{21}$$

(17) is just a reformulation of (18). □

For the case that not only the parameters themselves are bounded an even stronger statement. 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. 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. But for our graph genesis we obtain:

Lemma 4.5 *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*

$$a_\tau \leq \max\left\{a_{\tau-1},\, \mathcal{L}_\tau \cdot (\mathcal{D}_\tau - 1)\right\}. \tag{22}$$

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.

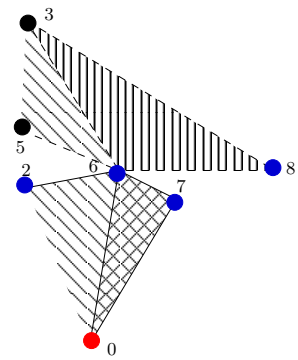
5 Restrictions to the Neighborhood of a Vertex

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.

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

τ	-2	-1	0	1	2	3
ob_0	$\boxed{0}$			0	0	
ob_2	$\boxed{2}$			2		
ob_3		$\boxed{3}$	3			3
ob_5		$\boxed{5}$	5			
ob_6			$\boxed{6}$	$\boxed{6}$	6	6
ob_7					$\boxed{7}$	
ob_8						$\boxed{8}$

(a) The restricted sequence of contexts



(b) The induced graph

Figure 3: The restriction $\left(co_{\dots}^{[ob_6]}\right)$ to ob_6.

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 5.1 *For a distinctive sequence $(co_{\dots})$ with $\mathcal{O} = 1$ and $\mathcal{S} = 0$ and an object $ob \in Ob_\infty$ let $\left(co_{\dots}^{[ob]}\right)$ the subsequence of contexts that contains 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 5.2 *For a distinctive sequence $(co_{\dots})$ with $\mathcal{O} = 1$ and $\mathcal{S} = 0$ and an object $ob \in Ob_\infty$ let $\left(co_{\dots}^{[ob]}\right)$ 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*

$$\alpha + \beta + \tau' \left(\widetilde{\mathcal{D}^{[ob]}}_{\tau'} - 2 \right) \tag{23}$$

	$\deg_4$	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

with $\alpha = |E_{-1} \cap N(ob)|$ the number of edges that were initially present in the neighborhood of ob and where

$$\beta = \sum_{\tau_0 \leq \tau \leq \tau_0 + \mathcal{L}_{\tau_0}} \binom{\mathcal{D}_\tau - 1}{2} \tag{24}$$

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 the terms in (23) are always well defined since $\mathcal{D} \geq 2$ and therefore $\widetilde{\mathcal{D}} \geq 2$, too. For (24) and contexts τ with $\mathcal{D}_\tau = 2$ the contribution to the sum is 0. So $\mathcal{D}_\tau = 2$ clearly is a borderline where we have a important property change. In consequence the following theorem needs a restriction to $\mathcal{D}_\tau > 2$.

Theorem 1 *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 $2 < \delta^- \leq \mathcal{D}_\tau$ and $1 \leq \lambda^- \leq \mathcal{L}_\tau$ for all τ and some integers δ^-, λ^-. Then for*

$$\begin{aligned} R^- &= \lambda^-(\delta^- - 2) \\ \text{and } \; R^+ &= \lambda^+(\delta^+ - 1) \end{aligned}$$

the clustering coefficient is bounded from below, namely for τ sufficiently large we have

$$cc(G_\tau) \geq \frac{1}{4} \cdot \frac{R^-}{R^{+2}}.$$

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 1 evaluates to 0.0225, which is far from the real value of 0.7.

Proof: From Lemma 4.4 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

$$\begin{aligned}
\frac{1}{n_\tau}\sum_{ob\in Ob_\tau}\frac{\mathrm{NE}_\tau(ob)}{\binom{\deg_\tau(ob)}{2}} &\geq \frac{1}{n_\tau}\sum_{\substack{ob\in Ob_\tau\\ \deg_\tau(ob)\leq 2R^+}}\frac{\mathrm{NE}_\tau(ob)}{\binom{2R^+}{2}}\\
&\geq \frac{1}{n_\tau}\sum_{\substack{ob\in Ob_\tau\\ \deg_\tau(ob)\leq 2R^+}}\frac{\lambda^-(\delta^- - 2)}{\binom{2R^+}{2}} \qquad (25)\\
&\geq \frac{1}{n_\tau}\cdot\frac{n_\tau}{2}\cdot\frac{\lambda^-(\delta^- - 2)}{\binom{2R^+}{2}}\\
&\geq \frac{\lambda^-(\delta^- - 2)}{4{R^+}^2}
\end{aligned}$$

And the claim follows. □

Notice that (25) only uses part of (23).

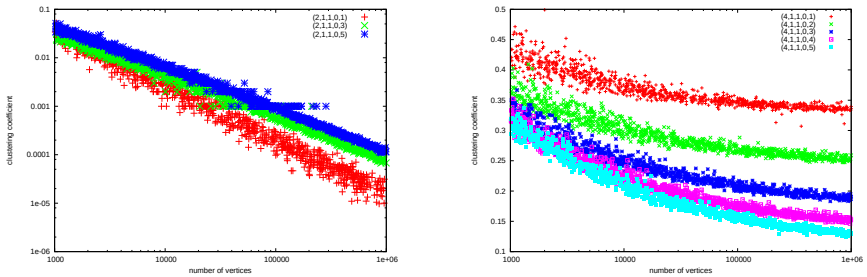
6 Some Experimental Results

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 Schimit (2008). Here we just like to emphasize on the threshold of $\mathcal{D} = 2$ that is apparent in Theorem 1. 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.

7 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



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

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

Figure 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 $cc(G)$. The graphs are generated such that $\log_{10} |V(G)|$ is uniformly distributed in $[3, 6)$.

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.

Acknowledgement

The author likes to thank Matthieu Latapy for interesting discussions on the subject and Pedro Schimit for a first implementation of the genesis as it is presented here.

References

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

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

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

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

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

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

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

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

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

Schimit, P., Aug. 2008. Numerical results for the generalized attachment models for the genesis of graphs. Internship report, INRIA.

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