On Small-World generating Models *

Michael Kaufmann, Katharina A. Lehmann and Hendrik Post

Wilhelm-Schickard-Institut für Informatik, Universität Tübingen, Sand 14, 72076 Tübingen
email: mk / lehmannk / post@informatik.uni-tuebingen.de

[...] the best material model for a cat is another, or preferably the same cat. - N. Wiener and A. Rosenblueth [24]

Abstract. What exactly is a *small-world*? Watts and Strogatz [23] define every network with a high clustering coefficient and a low diameter to be a small-world. We will show here that for this classic definition there are some counter-intuitive examples where either false-negative or false-positive classifications of network models occur.
To bring forth a new definition for small-world generating network models, we will first introduce a slightly varied small-world network model. This model is based on a regular grid graph and an added $G(n,p)$[13] random graph. We will then give an upper bound for the diameter of the generated networks dependent on p and n. This upper bound is generalized to combinations of a so-called 'locally clustered' graph family with a $G(n,p)$ graph. On the basis of this general method we propose a new definition for small-world generating models.

1 Introduction

The 'small-world effect' has long been a part of folklore. It describes the fact that most of us are tightly knit into small social clusters while on the other hand we need just a short chain of acquaintances to connect us to any other human on the world. Milgram estimated the number of persons in such a chain to be around six [19] which is why this observation is also known under the title 'six degrees of separation'. The first formal approach to explain this astonishing result was made by Watts and Strogatz in a seminal paper [23] in which they gave a rough definition of small-world networks and presented a model for their generation. They defined a small-world to be every network with a high clustering coefficient and a low diameter.

Following their publication, several real-world networks such as the WWW or file-sharing communities were analyzed and shown to be small-worlds (e.g. [1, 2, 11, 14]). A second research area deals with network-based processes on small-world networks, like the behavior of neural networks on small-worlds [17] or disease spreading in small-worlds [21]. Other directions of research tried to find

* Partially supported by DFG-Grant Ka812/11-1

more rigorous analytical results on the properties of either the classic small-world model or on variants of the small-world model that were easier to analyze or captured new aspects of small-worlds [20, 12, 16, 15, 3, 10, 4, 7].

In the classical model of Watts and Strogatz, n vertices are placed equidistantly on a ring and every vertex is connected with its k next neighbors. Every edge has a probability of p' to be rewired, i.e., one of the endpoints is fixed and the edge is rewired to a new, randomly drawn target vertex [23]. For $p' = 1$ a certain kind of random graph emerges. A variant proposed by Watts together with Newman [20] is based on the same basic ring graph but instead of rewiring, a special kind of random graph is added to this basic graph. It is not quite clear how this random graph is generated but the number of its edges is restricted to pkn which makes it impossible that a proper $G(n, p)$ random graph instance is added for $k < n - 1$. Watts and Newman state that there is a single threshold value for p' such that the scaling of the diameter changes from a linear behaviour to a logarithmic scaling.

We see three main objections regarding the classic small-world model [23] and the variant in [20]:

Random Graph Component In the classic small-world model, the rewired random edges are not building an instance of the commonly known and analyzed random graph families $G(n, m)$ or $G(n, p)$, not even for $p = 1$ [8, 6]. In a $G(n, m)$ instance, a fixed number of m random edges is drawn between n vertices, in a $G(n, p)$ instance every possible edge between the vertices exists with probability p. The diameter of these graph classes is $O(\log n)$. In the classic small-world model [23], not all possible instances of the $G(n, m)$ or $G(n, p)$ sets can be realized by the described process. This is mainly due to the fixation of one of the endpoints in the rewiring process. Nonetheless, it is claimed in [23] that the emerging random graph for $p = 1$ has a diameter that scales with $O(\log n)$. In the variant of [20], the type of the added random graph is unclear. Thus, the scaling of the diameter of this random graph component is possibly not given by $O(\log n)$ but the calculated threshold value depends heavily on this assumption.

Clustering Coefficient and Diameter as Indicators for Small-Worldness The model Watts and Strogatz proposed for generating small-world networks is based on the rewiring process in which edges are generated randomly. The main problem with the classic definition of small-world networks is that it cannot determine whether some random process was involved in the generation of a given network. We will illustrate this point:

It is very easy to construct a network with a high clustering coefficient and a small diameter by first building a balanced tree, e.g., a quaternary tree, and then adding edges between those leaves of the tree that have only distance 2 to each other (Fig. 1). The balanced tree will then provide a diameter of $O(\log n)$ for the whole network and more than half of the vertices are leaves with a clustering coefficient of 1, which generates an average clustering coefficient of > 0.5 for the whole graph. This network model is based on a hierarchical tree that

connects locally dense clusters with each other. For example, the power grid of the U.S. may rather be based on a hierarchical backbone that spans the network rather than that it is the result of a process where random edges are added to a locally clustered network [23]. This example shows that the clustering coefficient identifies networks as small-worlds that are generated by another network model, and thus it identifies **false-positives**.

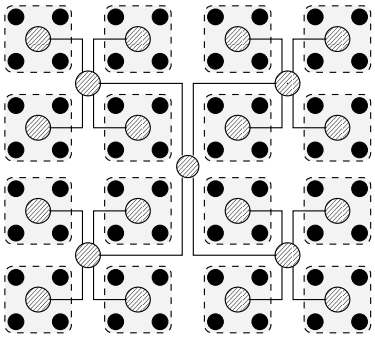

Fig. 1. Hybrid graphs of small cliques and a balanced tree as a counterexample for the classic definition of small-worlds: Each block of four vertices constitutes a clique in which every vertex has a clustering coefficient of 1. These blocks are combined to a connected graph by a quaternary tree where every vertex has exactly four children and most of the vertices in the tree have a clustering coefficient of 0. The clique component consists of 64 vertices, the added tree component consists of 21 vertices and thus the combined graph shows an average clustering coefficient of at least 0.75. The diameter of this graph is determined by the diameter of the tree component which scales with $O(\log n)$.

Also, networks with a small clustering coefficient can be considered as small-world networks. We want to illustrate this point with some citations: Regular grid graphs with a degree of $2d$ show a clustering coefficient of 0. Nonetheless, also those graphs constitute a small-world if they are combined with random graphs: Kleinberg bases his searchable small-worlds on regular grid graphs with a clustering coefficient of 0 [15, 16]. Networks representing mostly hetero-sexual relationships contain very little triangles. Rather, the graph forms nearly a bi-partite graph between males and females [5]. Nonetheless these networks are normally classified as small-worlds because of their combination of mainly local relationships with some additional long range contacts [18, 22]. Chung et al. generalize the idea behind the clustering coefficient to include more models into the general framework of small-world generating models [10, 4]: Their small-world model combines a so-called (k, l)-*local* graph with a power-law random graph. A graph is a (k, l)-local graph if for each edge $e = (v, w)$, v and w are connected by at least k edge-disjoint paths of at most length l. Summarizing, the intu-

ition is that not all small-worlds can be detected with the help of the clustering coefficient: it gives **false-negatives**.

Missing Generalizability The small-world model of Watts and Strogatz is attractive because of its simplicity. On the other hand it lacks some extendability to create more practical small-world generating models [10, 4].

Our goal is to build a generalized framework for small-world generating models that removes the objections given above. To deal with the first point we will simplify the small-world model of Newman and Watts by adding a $G(n, p)$[13] instance to a regular, d-dimensional grid. We will give an upper bound on the diameter of this basic hybrid graph. The notion of the *clustering coefficient* will be semantically replaced by the notion of *locally clustered* graph families. Then, the upper bound on the diameter of our basic hybrid graphs can be generalized to any combination of a *locally clustered* graph family with certain random graph families. Thus, we offer possible building blocks for constructing small-world generating models that are based on two components: a local and a random component. This allows for more complex small-world generating models that may be the basis for the simulation of real-world complex systems. We will further use the building-block framework to give a new definition of small-world generating models that incorporates all classic small-world models.

The paper is organized as follows: In Sec. 2 we give some basic definitions needed in the concourse of the article. Sec. 3 is structured into three subsections: Subsec. 3.1 introduces the main model, Subsec. 3.2 gives the upper bound for this model and in Subsec. 3.3 we generalize the analysis to any combination of locally clustered graph families with $G(n, p)$. In Sec. 4 the new definition for small-world generating models is introduced and discussed. Sec. 5 concludes with a summary and discussion of the results.

2 Definitions

A graph family $G(n)$ in this article denotes any set of graphs generated by the same algorithm and parameterized by the number of vertices in it. For non-random graph families and a fixed set of parameters only one specific graph is generated. For graph families generated partly by probabilistic processes, $G(n)$ is defined as the set of all possible realizations. Statements about $G(n)$ are then interpreted as statements about expected characteristics of this set. We will use the notation $G(n)$ interchangeably for the set or a specific realization of this set.

A regular d-dimensional, equilateral grid (hypercubical lattice) $G_d(n)$ is defined as a set of vertices, placed on integer positions in d dimensions. $a \in \mathbb{N}$ denotes the number of vertices placed in each of the d dimensions. The number of vertices in this grid is then given by $n = a^d$, where every possible position - identified by a d-dimensional vector $(1 \leq b_1 \leq a, 1 \leq b_2 \leq a, \ldots, 1 \leq b_d \leq a)$ - is occupied with one vertex. The degree $deg(v)$ of a vertex is defined as the number of incident edges and equals the number of direct neighbors of v. Every

vertex v is connected by an edge to those vertices that differ in their position by exactly one in exactly one dimension from the position of v, i.e., every vertex has at most degree $2d$. For these grids, the graph theoretic distance $d(v, w)$ of any two vertices v, w, i.e., the minimal number of traversed edges to walk from v to w, coincides with the Manhattan distance $d_M(v, w)$ of these vertices which is defined by:

$$d_M(v, w) = \sum_{1 \leq i \leq d} |b_i(v) - b_i(w)| \tag{1}$$

The diameter $D(G)$ of any graph G is defined as the maximal distance of any two vertices within the graph. The diameter $D(G_d(n))$ is given by the maximal Manhattan distance of any two vertices in $G_d(n)$ and can be calculated by:

$$D(G_d(n)) = \sum_{1 \leq i \leq d} a - 1 = d(a - 1) \tag{2}$$

A graph is *connected* if there is a way from every vertex v to any other vertex w.

The clustering coefficient $C(v)$ of a vertex v is defined as the number of edges $e(v)$ between direct neighbors of v and the maximal possible number of edges between direct neighbors [23]:

$$C(v) = \frac{e(v)}{deg(v)(deg(v) - 1)} \tag{3}$$

The clustering coefficient $C(G)$ of a graph G is defined as the average clustering coefficient of G's vertices.

A $G(n, p)$ random graph is defined as an instance of all possible graphs with n vertices where every of the $\binom{n}{2}$ edges exists with probability p [13]. A $G(n, m)$ random graph is defined as an instance of all possible graphs with n vertices and exactly m edges, drawn uniformly at random from all possible edges.

We will use the following theorem on the diameter of random graphs $G(n, p)$ [8]:

Theorem 1. *If* $pn/\log n \to \infty$ *and* $\log n/\log(np) \to \infty$ *then* $D(G(n, p))$ *is asymptotically equal to* $\log n/\log(np)$ *with high probability.*

Note that this theorem implicitly includes that the random graph is connected with high probability. To simplify the following proofs we will use a stricter version of the theorem and require additionally that $p \geq (\log n)^{1+\epsilon}/n$.

3 A Framework for Small-World generated Models

3.1 A first Starting Point

As argued in the introduction, the Watts-Strogatz- and the Newman-Watts-model suffer from some problems. We replace their models by a simplified version composed of a random graph $G(n, p)$ and a regular d-dimensional grid in the

following way: The basic regular graph is the d-dimensional grid of n vertices, where each vertex is connected to its $2d$ next neighbors, combined with a $G(n,p)$ random graph on the same n vertices. We will denote by $G_d(n,p)$ a graph from our model, which is given by the combination of a $G_d(n)$ regular grid and a random graph $G(n,p)$.

The remaining part of this section gives answers to the following question: How does the diameter of regular networks combined with a small set of random edges scale?

Since the basic network is a d-dimensional grid, the diameter of it without any added random edges will scale with $a-1$ for a fixed dimension d: $D(G_d(n)) = d \cdot (a-1)$. If the added random graph has a probability of $(\log p)^{1+\epsilon}/n$ then the combined graph will have a diameter that is dominated by the diameter of the random graph and thus is asymptotical to at most $\log n / \log(np)$ (Theorem 1).

What happens in the regime where p lies below $(\log p)^{1+\epsilon}/n$? When will the diameter of the combined graphs scale at most (poly-) logarithmically?

In Theorem 5 we will give a detailed upper bound for the diameter of the combined graph for a given number of random edges.

3.2 The Diameter of $G_d(n,p)$-Graphs

For the above given model of a graph $G_d(n,p)$ the following lemma holds:

Lemma 2. *For $p = \frac{1}{cn}$, $c \in \mathbb{R}^+$ the diameter of $G_d(n,p)$ is asymptotically bounded by at most*

$$d \cdot \left(\left\lceil \sqrt[d]{c \cdot (\log n)^{1+\epsilon}} \right\rceil - 1 \right) \cdot \left(\frac{\log n}{(1+\epsilon)\log\log n - \log 2} + 1 \right) \tag{4}$$

Proof. The proof proceeds in four steps:

1. To prove the lemma we partition $G_d(n,p)$ into n_S connected d-dimensional equilateral subgraphs $S_i, 1 \leq i \leq n_S$ with a side length l such that each subgraph contains at least $s = l^d \geq c(\log n)^{1+\epsilon}$ vertices (Figure 2).
2. For any a, we can only build $\lfloor a/l \rfloor$ full subgraphs per dimension. n^* denotes the number of all vertices contained in a full subgraph. We will show that the $n - n^*$vertices that are not contained in any full subgraph build a vanishing fraction of all vertices for $n \to \infty$. We will thus base our proof on a reduced regular d-dimensional grid of size n^* that contains only the full subgraphs.
3. We construct a supergraph $G_S(n_S) = (S, E')$ where each vertex $v_i \in S$ uniquely represents the subgraph S_i for $1 \leq i \leq n_S$. $e = (v_i, v_j) \in E'$ iff there is at least one random edge from any vertex in S_i to any vertex in S_j. We will prove that Theorem 1 is applicable on $G_S(n_S)$.
4. Then we will expand $G_S(n_S)$ to gain a bound on the diameter of the original but reduced graph $G_d(n^*,p)$. The diameter of $G_d(n^*,p)$ is bounded by the product of the diameter of the subgraphs $D(S_i)$ and the diameter $D(G_S)$. We will show that there are numerous partitions of $G_d(n,p)$ into n_S subgraphs. Especially, for any pair of vertices v, w there is at least one partition of

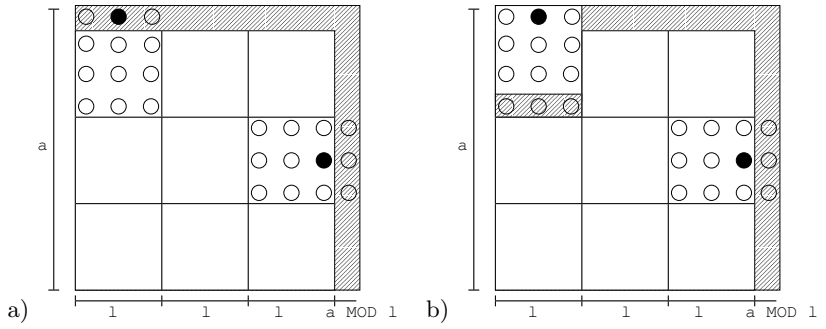

Fig. 2. Valid partitions for a 2-dimensional grid with side length a. Full equilateral subgraphs with side length l may be placed arbitrarily as long as their number is maximal. Therefore numerous partitions exist and for each pair of vertices numerous partitions can be found where both are contained in full subgraphs.

$G_d(n)$ such that both, v and w, are contained in full subgraphs. Since every supergraph based on a possible partition obeys Theorem 1, we will therefore have shown that the whole graph $G_d(n,p)$ obeys Lemma 2 and the case is proven.

We will start by partitioning a $G_d(n,p)$ graph. Let $S_i, 1 \leq i \leq n_S$ denote an equilateral subgraph that has a side length of $l = \left\lceil \sqrt[d]{c \cdot (\log n)^{1+\epsilon}} \right\rceil$ in each dimension. The number s of vertices contained in one (full) subgraph is bounded by:

$$c \cdot (\log n)^{1+\epsilon} \leq s = \left\lceil \sqrt[d]{c \cdot (\log n)^{1+\epsilon}} \right\rceil^d < 2^d \cdot c \cdot (\log n)^{1+\epsilon} \tag{5}$$

We will now partition $G_d(n,p)$ into the subgraphs as shown in Figure 2. Obviously, incomplete subgraphs exist if a/l is not integer. The leftover vertices can be placed arbitrarily between full subgraphs as indicated in Figure 2 b). For simplicity we will consider instead of $G_d(n)$ a smaller hypercube $G_d(n^*)$ containing all full subgraphs. Note that now a^* with $\sqrt[d]{n}^* = a^* \leq a$ is the maximal integer smaller than a that is a multiple of l. Let $q = a^*/l$ denote the number of subgraphs in one dimension.

The relative fraction of vertices not contained in full subgraphs is approaching 0 for $n \to \infty$:

$$\frac{n - n^*}{n} \leq \frac{(l \cdot (q+1))^d - (l \cdot q)^d}{(l \cdot q)^d} \tag{6}$$

$$= \left(\frac{q+1}{q}\right)^d - 1 \tag{7}$$

Since $q \to \infty$ for $n \to \infty$, the relative fraction of ignored vertices is asymptotically 0. Note that $n_S = \frac{n^*}{s} \geq \frac{n}{2^d \cdot c \cdot (\log n)^{1+\epsilon}} \to \infty$. Thus for $n \to \infty$ we may

safely use

$$n \geq n^* > n/2 \tag{8}$$

In $G_d(n^*, p)$ there are s^2 possible random edges between any vertex from subgraph S_i and any vertex from subgraph S_j. Each of these edges exists independently with probability p. It follows that for G_S the probability p_S is exactly $\frac{s^2}{cn}$.

We will now prove that Theorem 1 can be applied to $G_S(n)$. A basic observation is that for $n \to \infty$, also $n_S \to \infty$. Additionally, we must show that $\frac{p_S n_S}{\log n_S} \to \infty$ and $\frac{\log n_S}{\log(n_S p_S)} \to \infty$ for $n_S \to \infty$.

Regarding, that for all $n_S > 1$, $n^* > n/2$ (eq. 8) the following two equations hold:

$$\begin{aligned} \frac{p_S \cdot n_S}{\log n_S} &= \frac{s^2}{cn} \cdot \frac{n^*}{s} \cdot \frac{1}{\log \frac{n^*}{s}} && (9)\\ &\geq \frac{s}{2c(\log n^* - \log s)} && (10)\\ &\geq \frac{(\log n)^{1+\epsilon}}{2\log n - 2\log s} && (11) \end{aligned}$$

such that $\frac{p_S \cdot n_S}{\log n_S} \to \infty$ for $n \to \infty$ and

$$\begin{aligned} \frac{\log n_S}{\log(p_S \cdot n_S)} &= \frac{\log \frac{n^*}{s}}{\log \frac{s \cdot n^*}{cn}} && (12)\\ &\geq \frac{\log n/2 - \log(2^d \cdot c(\log n)^{1+\epsilon})}{\log(2^d (\log n)^{1+\epsilon})} && (13) \end{aligned}$$

such that also $\frac{\log n_S}{\log(p_S \cdot n_S)} \to \infty$. By theorem 1 we know that thus G_S has a diameter asymptotical to $\frac{\log n_S}{\log(p_S \cdot n_S)}$. Regarding that $n^*/n > 1/2$ this is bounded by

$$\begin{aligned} D(G_S) &= \frac{\log n_S}{\log(p_S \cdot n_S)} && (14)\\ &\leq \frac{\log n}{\log \frac{s}{2c}} && (15)\\ &\leq \frac{\log n}{(1+\epsilon)\log\log n - \log 2} && (16)\\ &\leq \frac{\log n}{\log\log n} && (17) \end{aligned}$$

Where the last inequality is valid for all n with $\epsilon \log\log n > \log 2$.

We will now expand $G_S(n)$ in order to get an upper bound for the diameter of $G_d(n, p)$.

Let v and w be two vertices in the original graph $G_d(n, p)$. First, we will reduce $G_d(n, p)$ to $G_d(n^*, p)$ in such a way that v and w are contained in $G_d(n^*, p)$.

Then we know that there is a path from subgraph S_i containing v to subgraph S_j containing w with a length of no more than $D(G_S)$. This path is denoted by $(e_1, e_2, ..., e_k)$, the sequence of edges to traverse to walk from S_i to S_j.

To use this path in the original graph $G_d(n,p)$, we will first have to walk from vertex v to that vertex v' from S_i that is attached to e_1. This will at most take $D(S_i) = d \cdot (\left\lceil \sqrt[d]{c \cdot (\log n)^{1+\epsilon}} \right\rceil - 1)$ steps. For every entered subgraph S_x on the way to subgraph S_j, an additional distance of $D(S_x)$ has at most to be added to get from the random edge entering the subgraph to the edge leaving this subgraph. Thus, the distance of v, w in the original graph $G_d(n,p)$ is asympotically given by at most

$$D(S_i) \cdot (D(G_S)+1) \leq d \cdot \left(\left\lceil \sqrt[d]{c \cdot (\log n)^{1+\epsilon}} \right\rceil - 1 \right) \cdot \left(\frac{\log n}{(1+\epsilon)\log\log n - \log 2} + 1 \right) \tag{18}$$

With this, Lemma 2 is proven. □

In the following we want to discuss what happens if the degree of the underlying grid graph is enlarged.

As stated in Lemma 2, the diameter of a $G_d(n,p)$ graph is asymptotically at most $D(S_i) \cdot (D(G_S) + 1)$. Let $G_d(n,k,p)$ denote an extended regular grid, in which every vertex is connected to its k next neighbors, combined with an additional $G(n,p)$ graph. The diameter $D(S_i)$ depends on the degree of the vertices in the underlying grid graph. Thus, if we want to reduce the diameter of the $G_d(n,k,p)$ graph we just have to add some more edges to the grid. For example, $D(S_i)$ is reduced to 1 if for $p = \frac{1}{c \cdot n}$ we add edges from every vertes to its $c \cdot (\log n)^{1+\epsilon}$ next neighbors. The combined graph $G_d(n, (\log n)^{1+\epsilon}, p)$ has now at most the diameter of G_S.

3.3 Generalizing the Small-World Model

In this section we will generalize Lemma 2 in two ways:

1. The probability p of the added random graph $G(n,p)$ can be as small as $\frac{1}{f(n) \cdot n}$ as long as $\frac{1}{(\log n)^{1+(\epsilon/2)}} \leq f(n) \leq \frac{1}{n^{1-\delta}}$ for some constants $\delta, \epsilon > 0$ and $n \to \infty$.
2. The basic regular d-dimensional grid can be replaced by certain graph families. This was already indicated at the end of section 3.2.

These two extensions lead finally to our generalized theorem on the diameter of small-worlds generated by our model.

Generalizing the random graph component At first, we explain in which range p can be chosen, such that the proof-technique can still be applied. In section 3.2, we kept $p = 1/cn$. For smaller $p = \frac{1}{f(n) \cdot n}$, the size s of the subgraphs has to be chosen larger such that Theorem 1 can be applied. Let again n_S denote the size of the supergraph.

For simplicity we assume that $n_S = n/s \in \mathbb{N}$ and $p \cdot s \cdot n = (\log n)^{1+\epsilon} \in \mathbb{N}$. The general case follows the argumentation above.

The number of nodes in each subgraph will be chosen such that $s = \frac{(\log n)^{1+\epsilon}}{p \cdot n} = f(n) \cdot (\log n)^{1+\epsilon}$. Again, Lemma 2 requires the validity of

$$\frac{p_S n_S}{\log n_S} \to \infty \tag{19}$$

and

$$\frac{\log n_S}{\log(n_S p_S)} \to \infty \tag{20}$$

As before $p_S = s^2 \cdot p$. We first analyze the condition given in equation (19):

$$\begin{aligned} \frac{p_S n_S}{\log n_S} &= s^2 \cdot p \cdot \frac{n}{s} \cdot \frac{1}{\log n_S} && (21)\\ &> s \cdot p \cdot \frac{n}{\log n} && (22)\\ &= (\log n)^{\epsilon} && (23) \end{aligned}$$

which tends to infinity for increasing n. The second condition (20) simplifies to

$$\begin{aligned} \frac{\log n_S}{\log(n_S p_S)} &= \frac{\log\left(\frac{n}{s}\right)}{\log\left(\frac{n}{s} \cdot s^2 \cdot p\right)} && (24)\\ &= \frac{\log\left(\frac{\frac{n}{f(n)}}{(\log n)^{1+\epsilon}}\right)}{\log(\log n)^{1+\epsilon}} && (25)\\ &= \frac{\log\left(\frac{n}{f(n)}\right)}{\log(\log n)^{1+\epsilon}} - 1 && (26) \end{aligned}$$

which tends to infinity for $f(n) \leq \frac{1}{n^{1-\delta}}, \delta > 0$. Therefore both conditions are met and Theorem 1 can be applied to G_S. If $f(n)$ is too low than there will be many random edges per vertex such that the subgraph size $s < 1$. To avoid this, we restrict $f(n) \geq \left(\frac{1}{(\log n)^{1+(\epsilon/2)}}\right)$ in order to guarantee a reasonable size for the subgraphs.

We summarize this result in

Lemma 3. *For any function $\frac{1}{(\log n)^{1+(\epsilon/2)}} \leq f(n) \leq \frac{1}{n^{1-\delta}}, \epsilon, \delta > 0$ and $p = \frac{1}{f(n) \cdot n}$, we can partition the grid graph within a $G_d(n,p)$ graph into $n_S = \frac{n}{s}$ subgraphs S_i of size $s = f(n) \cdot (\log n)^{1+\epsilon}$ such that $G_d(n,p)$ shows a diameter of asymptotically at most (Eq. 26)*

$$\underbrace{d \cdot \left(\left\lceil \sqrt[d]{c \cdot (\log n)^{1+\epsilon}} \right\rceil - 1\right)}_{D(S_i)} \cdot \underbrace{\left(\frac{\log(n/f(n))}{\log(\log n)^{1+\epsilon}}\right)}_{D(G_S)} \tag{27}$$

Possible replacements of the regular grid graph In the general proof we have used the following two properties of regular grid graphs: First, regular grid graphs are partitionable for every n into $\Theta(n/s(n))$ subgraphs of size $s(n)$ for any function $s(n) \leq n$ such that each of these subgraphs is a connected graph. The second property used is that for any pair of vertices v, w there must be at least one partition such that v and w are contained in any of the subgraphs.

To abstract from this special graph family to all graph families with these to properties we introduce the following definition:

Definition 4. Let $G_L(n)$ be a graph family with the following two properties:

1. $G_L(n)$ is partitionable for every n into $\Theta(n/s(n))$ subgraphs of size $s(n)$ for any function $s(n) \leq n$ such that each of these subgraphs is a connected graph
2. For any pair of vertices v, w and every n there must be at least one partition as described such that v and w are contained in proper subgraphs

$G_L(n)$ is called a *locally clustered* graph family.

Furthermore, a graph family can be *restricted locally clustered* with respect to some function $s(n) \leq n$ if for every n and every pair v, w $G_L(n)$ is partitionable into $\Theta(n/s(n))$ connected subgraphs of size $s(n)$ such that v and w are contained in proper subgraphs.

The notion of (restricted) local clusters in a graph thus can simply be interpreted as that every vertex in $G_L(n)$ can directly or indirectly reach at least $s(n)$ other vertices.

Classical small-world models are either based on the 1-dimensional ring lattice [23, 12, 7] or on d-dimensional regular grids [20, 16] and thus are based on *locally clustered* graph families. k-next neighborhood graphs in which n vertices are distributed uniformly in a unit-square and where every vertex is connected to its k geometrically next neighbors are also a *locally clustered* graph family. The proof for this statement is kind of lengthy, so the interested reader will find it in the appendix in Sec. 6.

Note that every graph family $G_L(n)$ is *restricted locally clustered* for at least $s(n) = 1$. Let $s_{max}(n)$ be that function $s'(n)$ that has the fastest growth of all functions $s(n)$ for which $G_L(n)$ is restricted locally clustered. If now $s_{max}(n) = k,\ k \in \mathbb{N}$ for $G_L(n)$ and $G_L(n)$ replaces the regular grid then it is clear that the size of the subgraphs is also at most k to obey $\Theta(n/s)$. This implies that p of the added random graph must be at least $O\left(\frac{(\log n)^{1+\epsilon}}{n}\right)$ in order to achieve a supergraph that obeys Theorem 1. It follows that the diameter is reduced to the diameter of a random graph because we added a random graph with the wanted diameter. This is certainly not a very interesting combination of graph classes. We will discuss this point further in Sec. 4.

We conclude this section with a theorem on the diameter of generalized small-world models combining a locally clustered graph family with a thin random graph:

Theorem 5. *Let $G_L(n,p)$ denote the combination of instances of a locally clustered graph family $G_L(n)$ and a $G(n,p)$ graph where $p = \frac{1}{f(n)\cdot n}$, $\frac{1}{(\log n)^{1+(\epsilon/2)}} \leq f(n) \leq \frac{1}{n^{1-\delta}}, \epsilon > 0, \delta > 0$. $D(s(n,p))$ denotes the maximal diameter of any subgraph of $G_L(n,p)$ with size $s(n,p) = \frac{(\log n)^{1+\epsilon}}{p\cdot n}$, $\epsilon > 0$, the diameter of $G_L(n,p)$ is asymptotically at most:*

$$D\left(s\left(n, \frac{1}{f(n)\cdot n}\right)\right) \cdot \underbrace{\left(\frac{\log(n/f(n)}{\log(\log n)^{1+\epsilon}}\right)}_{D(G_S)} \tag{28}$$

As we will discuss in the next section, the surprise in the whole small-world discussion lies in the fact that both graph components alone will have a much higher diameter than $O(\log n)$. Nonetheless, in our small-world model that can be based on any *locally clustered* graph family, also the use of cliques (complete graphs) is possible. There, the diameter is still bound from above by Theorem 5 but this bound is not tight: the diameter is 1. This brings us to the questions: What are small-worlds? We will discuss this question in the next section.

4 A new definition for small-world network models

We have shown in the introduction that the classic definition for small-worlds is somewhat erroneous and misleading. Our impression is that this stems from the following: The small-world generating process proposed by Watts and Strogatz required a random process for creating short-cuts [23]. But the classification of networks as small-worlds was not based on the recognition of this process but on a combination of structural measurements, i.e., the high clustering coefficient and the small diameter. This combination of network characteristics is not able to differentiate between those networks that include a local network and a random network and networks generated by other processes. In summary, the small-world network generation was process-oriented whereas the classification of small-world networks was result-oriented and there is no direct one-to-one matching between both sets of networks.

The drastic interest in the small-world phenomenon based on the classic small-world model seems to be based on two effects:

1. There is no centralized organization of the small-world network: every vertex builds random edges independently from others
2. The hybrid graph of a local and a random network component has a significantly lower diameter than the minimum of the diameter of boths components

To cover the first point there needs to be a *process-oriented classification of small-world networks* because the result may not always tell which generation process was behind it. The second effect excludes all those network models as small-world models in which the small diameter is an inbuilt feature of one component: If we have n unconnected vertices and add a dense random graph to it,

there is no surprise that the hybrid graph's diameter scales with $O(\log n)$. Also, the addition of some random edges to a clique will show the same diameter as the clique alone.
These two aspects lead us to the following new definition for small-world generating models:

Definition 6. A small-world model is defined as any combination $G_{LR}(n)$ of a *restricted locally clustered* graph family $G_L(n)$ and a random graph family $G_R(n)$ where the diameter $D(G_{LR}(n))$ is at most scaling poly-logarithmically and where the following relations hold for $n \to \infty$:

$$\frac{D(G_L(n))}{D(G_{LR}(n))} \to \infty \quad \text{and} \quad \frac{D(G_R(n))}{D(G_{LR}(n))} \to \infty \tag{29}$$

A small-world network is then a network that is generated by a small-world generating model - in real-world systems one may rather speak of small-world generating processes. Note that for *restricted locally clustered* graph families not any combination with a $G(n,p)$ graph may be appropriate. If it is a *locally clustered* graph family then it can be combined with any $G(n,p)$ under the conditions given in Theorem 5.

This definition removes the above given problems and captures - in our opinion - both effects that are the basis for the small-world phenomenon. It includes all classical small-world models because they are based on *(restricted) locally clustered* graphs [23, 15, 16, 12, 10, 4]. Note that these models may not be classified as small-world generating models for all combinations of their parameters. For the classic Watts-Strogatz-model or the Newman-Watts-model a very high k, e.g., $k = n-1$ would result in a clique as the local component. The above given definition restricts the models to those cases where the small-world effect is the result of the combination of both components and not of one of the components alone. The restriction on those cases reflects a similar decision made in the classic definition where the number of local edges per vertex k was required to be much smaller than n and also $p \ll 1$. We want to discuss two small-world generating models a bit more detailed with respect to the above given definition: In the case of the Kleinberg-small-world model the random graph component has not yet been analyzed with respect to its diameter to our knowledge. Kleinberg uses a random graph in which for each vertex v, q edges are added. Every vertex has a position in a 2-dimensional grid and edge $e = (v, w)$ is drawn with a probability proportional to $(d(v, w))^{-2}$ [16]. Kleinberg shows that a small-world network emerges for $q = 1$. Although there is no analysis available on the diameter of this random graph family, it seems quite likely that the described component with only one random edge per vertex is not even connected and will thus show a diameter of ∞.

The other small-world model to be analyzed was given by Chung et al. [10, 4]. In their basic model, the random graph component is a power-law random graph. Following [9], this component alone has almost surely a diameter of $O(\log n)$. On the other hand, Chung et al. have shown that for (k, l)-local graphs with certain properties, the resulting hybrid graph has a diameter of, e.g., $O(\log \log n)$. This is

clearly a new, emerging characteristic of the hybrid graph that is not dominated by one of the components alone. Thus, in the tradition of Watts and Strogatz we regard only the second, specialized cases as small-world generating models. We just want to mention here that it might be possible to make the power-law random graph sparser in the general Chung-small-world model but this analysis has not yet been conducted.

We summarize that these classic small-world models are represented by the above given definition.

5 Discussion

In this paper we have proposed a general framework for the design of small-world generating models: We have shown that they can be combined of (restricted) locally clustered and random graph models. This provides high flexibility in tuning a model to simulate a given real-world complex system. We have given a generalized theorem that describes an upper bound for the diameter of these hybrid graphs in dependence of the structure of both, the local and random component. Based on this framework we have proposed a very broad definition of small-world generating models, incorporating all classic small-world models.

Watts and Strogatz have provided us with the first formal model for generating small-world networks. The beauty of their model lies in its simplicity and clarity. Although we have questioned the usefulness of measuring the clustering coefficient as an indicator for small-worldness of real-world networks we have not given an alternative with which real-world networks can be identified as small-worlds. Admittedly, this is much more complicated in our framework than in the simple model of Watts and Strogatz. The only way to decide whether a real-world network is a small-world network in the above given definition is to analyze its generating process. If this is based on a local and a random component than the network should be regarded as a small-world if the diameter is short and each component alone has a high diameter.

Chung and Lu proposed to partition a given real-world network in a local and a global, random component [4]. They provide an approximative partitioning algorithm which works fine but is depending on two parameters that may not always be known in advance.

The partitioning into two components is also our suggestion: If there is any additional information about the network this could be used to partition it into its local and random component. For example, if the building of edges in a real-network is associated with a cost this could be regarded as a measure of distance. Then, both components can be separated and analyzed. Here, we can rehabilitate the clustering coefficient: If the global - presumably random - component shows a high clustering coefficient then it can be safely concluded that with high probability this component is not the result of a classic random process.

Of course, our definition of small-world generating components is somewhat influenced by our personal impression of what small-worlds really are. Since we are aware of that problem we want to conclude our discussion with the

introductory quotation of Wiener and Rosenblueth in "The Role of Models in Science":

> *[...] the best material model for a cat is another, or preferably the same cat.* - N. Wiener and A. Rosenblueth [24]

References

1. L. Adamic. The small world web. In *Proceedings of ECDL'99 - Lecture Notes on Computer Science*, pages 443–452, 1999.
2. L. Adamic and E. Adar. How to search a social network, 2004.
3. L.A.N. Amaral, A. Scala, M. Barthelemy, and H.E. Stanley. Classes of small-world networks. *PNAS*, 97(21):11149–11152, 2000.
4. R. Andersen, F. Chung, and L. Lu. Analyzing the small world phenomenon using a hybrid model with local network flow. In *Algorithms and Models for the Web-Graph; Third International Workshop, WAW 2004*, LNCS 3243, pages 19–30, 2004.
5. P.S. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110:44–91, 2004.
6. B. Bollobás. *Random Graphs.* Cambridge Studies in Advanced Mathematics 73. Cambridge University Press, London, 2nd edition, 2001.
7. B. Bollobás and F. Chung. The diameter of a cycle plus a random matching. *SIAM Journal on Discrete Mathematics*, 1:328–333, 1998.
8. B. Bollobás and O. M. Riordan. Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks (S. Bornholdt, H.G. Schuster (eds.))*, pages 1–34, 2003.
9. F. Chung and L. Lu. The average distances in random graph with given expected degrees. *PNAS*, 99(25):15879–15882, 2002.
10. F. Chung and L. Lu. The small world phenomenon in hybrid power law graphs. In *Complex Networks (E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (eds.))*, pages 91–106, 2004.
11. P.S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003.
12. S.N. Dorogovtsev and J.F.F. Mendes. Exactly solvable analogy of small-world networks. *Europhys. Lett.*, 50, 2000.
13. E. N. Gilbert. Random graphs. *Anual Math. Statist.*, 30:1141–1144, 1959.
14. A. Iamnitchi, M. Ripeanu, and I. Foster. Small-world file-sharing communities. In *Proceedings of the IEEE INFOCOM 2004*, 2004.
15. J. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.
16. J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.
17. L.F. Lago-Fernandez, R. Huerta, F. Corbacho, and J.A. Siguenza. Fast response and temporal coherent oscillations in small-world networks. *Phys. Rev. Letters*, 84:2758–2761, 2000.
18. F. Liljeros. Sexual networks in contemporary western societies. *Physica A*, 338:238–245, 2004.
19. S. Milgram. The small world problem. *Psychology Today*, 1:61–67, 1967.
20. M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Phys. Lett. A*, 263:341–346, 1999.

21. M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60:7332–7342, 1999.
22. M.C. Gőpfert and D. Robert. The web of human sexual contacts. *Nature*, 411:907–908, 2001.
23. D. J. Watts and S. Strogatz. Collective dynamics of small world networks. *Nature*, 393:440–442, 1998.
24. N. Wiener and A. Rosenblueth. The role of models in science. *Philosophy of Science*, 12(4):316–321, 1945.

6 Appendix: *k*-next neighborhood graphs are *locally clustered*

The k-next neighbor graph family or in short, knn-graphs, belongs to the class of *locally clustered* network families as we will show in this section. The proof is applied only to 2-dimensional knn-graphs defined as:

A k-next neighborhood graph $G(n,k)$ is any possible instance of n vertices, uniformly distributed in a two-dimensional unit-square, where every vertex is connected with its k geometrically next vertices.

Note that the relation is not symmetric and therefore the knn-graph is a directed graph and thus we differentiate between the outgoing and ingoing edges of a vertex.

In order to prove the given property we will proceed in the following steps:

1. Bound the distance to next neighbors from above and below.
2. Prove that a knn-graph is highly likely connected.
3. Show that a generic partition procedure yields the required n/s subgraphs.

6.1 A bound for the maximum distance of nearest neighbors

Let the knn-disk of any vertex v be defined as the minimal disk which contains all of its k next neighbors. Note that the disc-radius is equal to the maximum distance of any connected nearest neighbor to v.

The probability for any vertex v to be placed in some area A within the unit square is exactly A. Thus, the placement of vertices into a given area is a Bernoulli trial with $p = A$ and $q = 1 - A$. Therefore the Chernoff bound may be applied to yield an upper bound for the diameter. Moreover, the expected radius is given by $\bar{r} = \sqrt{\frac{k}{\pi \cdot n}}$. The result for the upper bound is given in

Lemma 7. *Let $\hat{r} = \sqrt{\hat{c}} \cdot \bar{r}$ denote a knn-disk radius with $\hat{c} = 3 + \sqrt{8}$. Further let $k \geq \log n$.*
With high probability ($Pr[...] \geq 1 - 1/n$), no disk with radius $\hat{r}$ around any vertex v exists that does not contain at least k vertices.

Proof. Let D_v denote a knn-disk around v with an expected number of vertices lying in that disk equal to $\bar{k} = c \cdot k$. X_k denotes the number of vertices lying inside of D_v. Now we apply a relaxed version of the Chernoff inequality for independent Bernoulli trials. With $\mu = c \cdot k$ and $\delta = 1 - \frac{1}{c}$

$$Pr[X_k < (1-\delta)\mu] < e^{-\frac{1}{2}\mu\delta^2} = e^{-\frac{ck}{2}(1-\frac{1}{c})^2} < n^{-\frac{c}{2}(1-\frac{1}{c})^2} \tag{30}$$

For $c = \hat{c} = 3 + \sqrt{8}$ we yield

$$Pr[X_k < (1-\delta)\mu] < \frac{1}{n^2} \tag{31}$$

Hence, the probability that there is a knn-disk with radius larger than $\hat{r} = \sqrt{3+\sqrt{8}} \cdot \bar{r}$ in a knn-graph is $< 1/n$.

The interpretation of this result is that it is almost impossible for $n \to \infty$ that any knn-disk exists with a radius larger than $\hat{r}$. Therefore in our following theorems, we consider the radius of knn-discs to be bound by $\hat{r} = \sqrt{\hat{c}} \cdot \bar{r}$.

Note that theses equations are only valid for disks that do not intersect with the unit squares border. If a vertex v_c were to be positioned in a corner of the unit disc, a factor of 2 would apply to the radius.

6.2 Connectedness of knn-graphs

We will now show that a knn-graph is **whp** connected.

The proof for the following lemma will be ommited. Here we will just sketch it shortly: As can be seen in Fig. 6.2, in every unconnected knn-graph the pair of closest vertices lying in different components have an angle of at least 120° in which none of their k next neighbors is placed. A simple stochastic argument shows that the probability for this is given by $(2/3)^k$. Equating this with the probability bound of $1/n$ and solving the equation to k yields the needed k such that with high probability not even one vertex with the above mentioned property exists. This leads to the following lemma:

Lemma 8. *A knn-graph is connected with high probability for* $k \geq \frac{\log n}{\log(3/2)}$

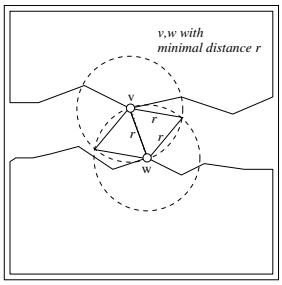

Fig. 3. v and w are two vertices from different connected components of a *knn*-graph having minimal euclidian distance to each other. Two circles are drawn around v and w, respectively, with a radius that equals the euclidian distance between v and w. The figure shows that none of the k-next neighbors of neither v nor w can exist in the intersection of these circles without contradicting the condition that v and w are the pair of vertices from different connected components with minimal euclidian distance.

Note that the probability for an unconnected *knn*-graph is smaller than $1/n$ since the existence of a vertex with the above given property is only necessary for an unconnected graph but certainly not sufficient.

Having argued that knn-graphs are connected, we will now show that connected commensurate partitions can be found.

6.3 Constructing the partition

The following procedure constructs partitions as required by the definition for *locally clustered* graph families. The size of the subgraphs is depending on the added random graph family. Nevertheless, if the random graph is given, one can easily calculate the fixed subgraph of size s. The definition requires that for each pair of vertices a partition into $\Theta(n/s)$ subgraphs must exist, so that both vertices are included in full subgraphs.

For each pair v, w of vertices we construct slightly different partitions. For each of them, we start with a geometric partition, based on squares containing at least $4/\pi \cdot s$ vertices. The exact positions for the squares are chosen such that both vertices are contained in full subgraphs. Beside this requirement the positions of the squares are arbitrary as long as the number of squares placed completely inside the unit square is maximal. Note, that a constant relative fraction of vertices may exist, that is not contained in any subgraph. Each of the squares covers an area A_s so that with high probability at least $4/\pi \cdot s$ vertices are geometrically contained in each of them. The area is given by $A_s > 4/\pi \cdot \pi \cdot \hat{r}^2$, where $\hat{r}$ denotes the maximal expected knn-disc radius (Lemma 7).
The maximal (centered) circle (Fig. 6.3) within each A_s contains only vertices from the same connected component. Otherwise at least one vertex would have an arc of more than 120° without any knn-edge which is highly unlikely as was already shown in Lemma 6.2. The area of this circle covers $\pi/4$ of A_s. We expect therefore that at least $\pi/4 \cdot 4/\pi \cdot s = s$ vertices from the same connected component for each A_s. As we explained before, for each constructed partition a constant fraction of vertices can be disregarded.

The overall result of this section is summarized by

Lemma 9. *The family of k-nearest neighbor graphs $G_k(n, k)$ on a point set in a 2-dimensional Euclidean space is locally clustered.*

Figure 6.3 shows an example for a partition for $s = 2$.

Note that the expected diameter $D(S_i)$ of the subgraphs is expectedly scaling with $O(\sqrt[d]{s})$ as it is the case with grid graphs.

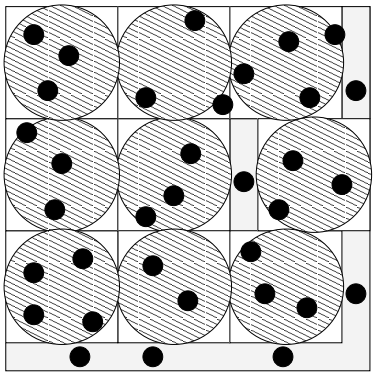

Fig. 4. This figure shows the result of the partitioning procedure for $s = 2$ as described in 6.3. Each square contains more than $4/\pi \cdot s \approx 2.5$ vertices. Each circle within any quadratic region contains at least $s = 2$ vertices that must form a connected subgraph. Note that the distribution of points is only schematic.