# Motif Counting in Preferential Attachment Graphs

**Jan Dreier**
Department of Computer Science, RWTH Aachen University, Germany
https://tcs.rwth-aachen.de/~dreier
dreier@cs.rwth-aachen.de

**Peter Rossmanith**
Department of Computer Science, RWTH Aachen University, Germany
https://tcs.rwth-aachen.de
rossmani@cs.rwth-aachen.de

## Abstract

Network motifs are small patterns that occur in a network significantly more often than expected. They have gathered a lot of interest, as they may describe functional dependencies of complex networks and yield insights into their basic structure [22]. Therefore, a large amount of work went into the development of methods for network motif detection in complex networks [20, 28, 8, 31, 16, 1, 25]. The underlying problem of motif detection is to count how often a copy of a pattern graph $H$ occurs in a target graph $G$. This problem is #W[1]-hard when parameterized by the size of $H$ [14] and cannot be solved in time $f(|H|)n^{o(|H|)}$ under #ETH [7].

Preferential attachment graphs [3] are a very popular random graph model designed to mimic complex networks. They are constructed by a random process that iteratively adds vertices and attaches them preferentially to vertices that already have high degree. Preferential attachment has been empirically observed in real growing networks [24, 19].

We show that one can count subgraph copies of a graph $H$ in the preferential attachment graph $G_m^n$ (with $n$ vertices and $nm$ edges, where $m$ is usually a small constant) in expected time $f(|H|)m^{O(|H|^6)}\log(n)^{O(|H|^{12})}n$. This means the motif counting problem can be solved in expected quasilinear FPT time on preferential attachment graphs with respect to the parameters $|H|$ and $m$. In particular, for fixed $H$ and $m$ the expected run time is $O(n^{1+\varepsilon})$ for every $\varepsilon > 0$.

Our results are obtained using new concentration bounds for degrees in preferential attachment graphs. Assume the (total) degree of a set of vertices at a time $t$ of the random process is $d$. We show that if $d$ is sufficiently large then the degree of the same set at a later time $n$ is likely to be in the interval $(1 \pm \varepsilon)d\sqrt{n/t}$ (for $\varepsilon > 0$) for all $n \geq t$. More specifically, the probability that this interval is left is exponentially small in $d$.

**2012 ACM Subject Classification** Theory of computation → Parameterized complexity and exact algorithms

**Keywords and phrases** random graphs, motif counting, average case analysis, preferential attachment graphs

**Digital Object Identifier** 10.4230/LIPIcs.FSTTCS.2019.13

## 1 Introduction

Network motifs are small patterns that occur in a network significantly more often than expected. They are relevant for example in the analysis of biological networks such as transcription networks of bacteria [22]. Detecting network motifs is computationally very expensive and there exist numerous algorithms for this task [20, 28, 8, 31, 16, 1, 25]. The underlying problem of motif detection is to count how often a copy of a pattern graph $H$ occurs in a target graph $G$. This can be very hard, as counting perfect matchings is #P-hard [29]. One of the fastest algorithms by Curticapean, Dell, and Marx can count subgraph copies of a graph $H$ with $k$ edges in a graph $G$ of size $n$ in time $k^{O(k)}n^{0.174k+o(k)}$ [12]. When it comes to parameterized complexity, counting $k$-cliques is #W[1]-hard [14] and cannot be done in time $f(k)n^{o(k)}$ under #ETH [7].

39th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2019).
Editors: Arkadev Chattopadhyay and Paul Gastin; Article No. 13; pp. 13:1–13:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A general question is whether problems that are hard on general graphs can be solved efficiently in real-world networks. To this end, the average run time of algorithms on random graphs has been considered (see [15] for a survey from 1997). For example Janson, Łuczak and Norros show that in certain scale-free random graphs with exponent $\alpha > 2$ one can find a maximal clique in polynomial time [18].

Preferential attachment graphs [3] are random graphs designed to mimic complex networks. They are constructed by a random process that iteratively adds vertices and attaches them preferentially to vertices that already have high degree. Scale-free behaviour has been identified as a central property of many complex networks [6, 9] and the preferential attachment process is a widely recognized explanation [5] of this behaviour. Preferential attachment has been empirically observed in real growing networks [24, 19].

Recently, the behaviour of some algorithms on preferential attachment graphs has been analyzed. Let $G_m^n$ be the preferential attachment graph with $n$ vertices and $m$ edges per vertex by (see Section 2 or a rigorous definition). For example, Korula and Lattanzi present a reconciliation algorithm with proven 97% success in preferential attachment graphs [21] and Cooper and Frieze show that the cover time of a simple random walk on $G_m^n$ is with high probability asymptotic to $\frac{2m}{m-1} n \log(n)$ [10].

We show that the motif counting problem can be solved in expected quasilinear time on preferential attachment graphs by a simple algorithm for any motif of constant size. For simple graphs $G$ and $H$ let $\#\mathrm{Sub}(H, G)$ be the number of subgraphs of $G$ isomorphic to $H$. If the graph $G$ has loops or multi-edges (as preferential attachment graphs do) then the subgraphs is counted with respect to the simple graph corresponding to $G$. Our main result is the following.

▶ **Theorem 5.** *There exists a function $f$ such that for every graph $H$ and $n, m \in \mathbf{N}$ one can compute $\#\mathrm{Sub}(H, G_m^n)$ in expected time $f(|H|)m^{O(|H|^6)} \log(n)^{O(|H|^{12})} n$.*

This means one can compute $\#\mathrm{Sub}(H, G_m^n)$ in expected quasilinear FPT time on preferential attachment graphs with respect to the parameters $|H|$ and $m$. In particular, for fixed $H$ and $m$ the expected run time is $O(n^{1+\varepsilon})$ for every $\varepsilon > 0$. Our results can be easily extended to alternative definitions of $\#\mathrm{Sub}(H, G)$ for multigraphs.

Our result is obtained as follows: At first, we define a value $\gamma_l(G)$ for every graph $G$ and $l \in \mathbf{N}^+$ and present a simple algorithm to compute $\#\mathrm{Sub}(H, G)$ in time $f(|H|)\gamma_{|H|}(G)$ for some function $f$ (Lemma 3). We then bound $\gamma_l(G)$ by the number of subgraphs in $G$ of bounded size with at most two pendant vertices (Lemma 7). Using this insight, we can bound the expected value of $\gamma_l(G)$ in preferential attachment graphs by $\mathrm{E}[\gamma_l(G_m^n)] = m^{O(l^6)} \log(n)^{O(l^{12})} n$ (Theorem 4), which directly yields the efficient subgraph counting algorithm. This analysis is based on concentration bounds for vertex degrees, which are proven in Section 5.

### Concentration Bounds for Degrees in Preferential Attachment Graphs

A large part of the analysis of our motif counting algorithm is based on concentration bounds for degrees, which we believe to be of individual interest. Aspects of the degree distributions in preferential attachment graphs are well studied [5, 4, 2, 17, 23, 27, 26, 32]. For example, Bollobás et al. [5] show that the degree sequence follows a power law distribution and Peköz et al. [26, 27] bound the rate of convergence of the degree of individual vertices to a limit distribution. The resulting tail bounds for degrees of individual vertices, however, have only polynomial accuracy. We complement these results by providing exponentially strong

concentration bounds for vertices or sets of vertices with high degree. We believe these bounds to be useful for proving structural properties and analyzing algorithms on preferential attachment graphs beyond motif counting.

Let the vertices in a preferential attachment graph be $v_1, v_2, v_3, \dots$ in order of insertion. Let $t \in \mathbf{N}$ and $S \subseteq \{v_1, \dots, v_t\}$. We analyze the evolution of the degree of $S$ in the random process over time. For $n \geq t$ and $m \geq 1$ we define $d_m^n(S)$ to be the sum over all degrees of vertices in $S$ in $G_m^n$ (we define $d_m^n(v_i) := d_m^n(\{v_i\})$).

Assume the degree of $S$ at a time $t$ to be $d_m^t(S) = d$. It can be shown that the expected degree of $S$ at a later time $n \geq t$ of the same random process asymptotically approaches $\mathrm{E}[d_m^n(S) \mid d_m^t(S) = d] \sim \sqrt{\frac{n}{t}}d$ [30]. In general, the preferential attachment process is too unstable and chaotic to guarantee that the degree of $S$ closely centered around its expected value. We show, however, that if $d$ is sufficiently large then the degree of $S$ at time $n$ is likely to be in the interval $(1 \pm \varepsilon)\sqrt{\frac{n}{t}}d$ (for $\varepsilon > 0$) for all $n \geq t$. More specifically, the probability that this interval is left is exponentially small in $d$. This is formalized by the following theorem.

▶ **Theorem 19.** *For $t, m, d \in \mathbf{N}^+$, $0 < \varepsilon \leq 1/2$, $S \subseteq \{v_1, \dots, v_t\}$ with $\Pr[d_m^t(S) = d] \neq 0$ and $d \geq \log(\log(3tm))\varepsilon^{-200}$*

$$\Pr\left[(1-\varepsilon)\sqrt{\frac{n}{t}}d < d_m^n(S) < (1+\varepsilon)\sqrt{\frac{n}{t}}d \text{ for all } n \geq t \;\middle|\; d_m^t(S) = d\right] \geq 1 - e^{-\varepsilon^{200}d}.$$

Note that concentration is guaranteed for all $n \geq t$ simultaneously. This means especially that the degree of large sets of vertices is strongly concentrated at all times of the random process. The constants have been chosen to ease calculations and can be greatly improved.

## 2 Preliminaries

We will denote probabilities by $\Pr[*]$ and expectation by $\mathrm{E}[*]$. The logarithm is the natural logarithm. We use common graph theory notation [13]. The *order* of a graph is $|G| = |V(G)|$. The *size* of a graph is $\|G\| = |V(G) + E(G)|$. All graphs (except preferential attachment graphs) are simple graphs. The underlying simple graph of a multigraph is obtained by replacing multi-edges with a single edge and removing self-loops. In this work we focus on the *preferential attachment random graph model* [3]. The model generates random graphs by iteratively inserting new vertices and edges. It depends on a parameter $m$ that equals the number of edges attached to a newly created vertex. We follow the definition of Bollobás et al. [5]: For a fixed $m$, the random process is defined by starting with a single vertex and iteratively adding vertices, thereby constructing a sequence of graphs $G_m^1, G_m^2, \dots, G_m^t$, where $G_m^t$ has $t$ vertices and $mt$ edges. We define $d_m^t(v)$ to be the degree of vertex $v$ in the graph $G_m^t$. The random process for $m = 1$ works as follows. A random graph is started with one vertex $v_1$ that has exactly one self-loop. This graph is $G_1^1$. We then define the graph process inductively: Given $G_1^{t-1}$ with vertex set $\{v_1, \dots, v_{t-1}\}$, we create $G_1^t$ by adding a new vertex $v_t$ together with a single edge from $v_t$ to $v_i$, where $i$ is chosen at random from $\{1, \dots, t\}$ with

$$\Pr[i = s] = \begin{cases} d_1^{t-1}(v_s)/(2t-1) & 1 \leq s < t, \\ 1/(2t-1) & s = t. \end{cases}$$

This means we add an edge to a random vertex with a probability proportional to its degree at the time. For $m > 1$, the process can be defined by merging sets of $m$ consecutive vertices in $G_1^{mt}$ to single vertices in $G_m^t$ [5]. Let $v'_1, \dots, v'_{mt}$ be the vertices of $G_1^{mt}$. The graph $G_m^t$

with vertices $v_1, \dots, v_t$ is constructed by merging $v'_{(i-1)m+1}, \dots, v'_{im}$ into a single vertex $v_i$. The graph $G_m^t$ is a multigraph. The number of edges between vertices $v_i$ and $v_j$ in $G_m^t$ equals the number of edges between the corresponding sets of vertices in $G_1^{mt}$. Self-loops and multi-edges are allowed.

In this work we obtain concentration bounds for the total degree of a set of vertices $S \subseteq \{v_1, \dots, v_t\}$ during the random process. We define the degree of a set $S$ at time $n \geq t$ as $d_m^n(S) = \sum_{v \in S} d_m^n(v)$.

## 3 Subgraph Counting

We start by presenting a very simple algorithm that decides for a graph $G$ and a connected pattern graph $H$ if there exists a subgraph of $G$ isomorphic to $H$. Then Lemma 2 and 3 generalize this algorithm into a counting algorithm for arbitrary pattern graphs.

Assume there is a subgraph $H'$ of $G$ isomorphic to $H$ that we want to find and let $l = |H|$. Since $H'$ is a connected graph with at most $l$ vertices there exists a vertex $v \in V(G)$ such that $H'$ is contained in the $l$-neighborhood $G[N_l^G(v)]$ of $v$. We build a spanning tree $T$ of $G[N_l^G(v)]$. Since $T$ is a tree, it is fairly easy to find $H'$ if $H'$ is a subgraph of $T$. But what happens if $H'$ contains edges that are not contained in $T$? We call the edges of $G[N_l^G(v)]$ that are not in $T$ the *extra edges* of the $l$-neighborhood of $v$. Since $H$ has at most $\binom{l}{2}$ edges, there exists a subset $F$ of at most $\binom{l}{2}$ many extra edges such that $H$ is contained in $(V(T), E(T) \cup F)$. The graph $(V(T), E(T) \cup F)$ is a tree with at most $\binom{l}{2}$ extra edges and therefore has bounded treewidth. Using Courcelle's theorem [11] it is still easy to find $H'$ in $(V(T), E(T) \cup F)$. In summary, one can find the graph $H'$ by enumerating all $v \in V(G)$ and sets $F$ of at most $\binom{l}{2}$ extra edges in the $l$-neighborhood of $v$ in $G$, and then using Courcelle's theorem.

We define a value $\gamma_l(G)$ of a graph $G$, which can be obtained by multiplying the size of each $l$-neighborhood with the number of sets of extra edges of size at most $\binom{l}{2}$.

▶ **Definition 1.** *Let $G$ be a graph and $l \in \mathbf{N}^+$. We define*

$$\gamma_l(G) = \sum_{v \in V(G)} |N_l^G(v)| \sum_{k=0}^{\binom{l}{2}} \binom{\|G[N_l^G(v)]\| - |N_l^G(v)| - 1}{k}.$$

*For multigraphs $G$, $\gamma_l(G)$ is defined with respect to the simple underlying graph.*

We now show that $\gamma_l(G)$ captures the run time of the previously discussed algorithm (up to a factor independent of $G$). We start with counting connected patterns and generalize this afterwards to arbitrary patterns.

▶ **Lemma 2.** *There exists a function $f$ such that for every graph $G$ and connected graph $H$ one can compute $\#\mathrm{Sub}(H, G)$ in time $f(|H|)\gamma_{|H|}(G)$.*

**Proof.** Let $l = |H|$. We compute spanning trees $T_v$ of $G[N_l^G(v)]$ for $v \in V(G)$ in time $\sum_{v \in V(G)} O(\|G[N_l^G(v)]\|)$ by breadth-first searches. Let $\mathcal{F}_v := \{F \mid F \subseteq E(G[N_l^G(v)]) \setminus E(T_v), |F| \leq \binom{l}{2}\}$ be the set of all subsets of at most $\binom{l}{2}$ edges that are in $G[N_l^G(v)]$ but not in $T_v$. We construct the sets $\mathcal{F}_v$ for $v \in V(G)$ in time $\sum_{v \in V(G)} O(\|G[N_l^G(v)]\| + |\mathcal{F}_v| l^2)$.

Let $I$ be the set of all subgraphs of $G$ isomorphic to $H$. For $v \in V(G)$ and $F \in \mathcal{F}_v$ let $I_{v,F}$ be the set of subgraphs $H'$ of $(V(T_v), E(T_v) \cup F)$ such that $H'$ is isomorphic to $H$, $v \in V(H')$ and $F \subseteq E(H')$. We claim that $\#\mathrm{Sub}(H, G) = |I| = \sum_{v \in V} \sum_{F \in \mathcal{F}_v} \frac{|I_{v,F}(H)|}{|H|}$. Let $H'$ be a subgraph of $G$. If $H'$ is not isomorphic to $H$ then by definition $H' \notin I_{v,F}$ for all

$v \in V(G)$, $F \in \mathcal{F}_v$. Assume now that $H'$ is isomorphic to $H$. To prove the claim, need to make sure that $H'$ is counted exactly $|H|$ times. This is the case because $H' \in I_{v,F}$ if and only if $v \in V(H')$ and $F = E(H') \setminus E(T_v)$.

In order to compute $\#\mathrm{Sub}(H, G)$, it is now sufficient to iterate over all $v \in V$ and $F \in \mathcal{F}_v$ and compute $|I_{v,F}|$. The graph $(V(T_v), E(T_v) \cup F)$ is a tree with at most $\binom{l}{2}$ additional edges and therefore has treewidth at most $\binom{l}{2} + 1$. By Courcelle's theorem [11], there exists a function $f'$ such that one can compute $|I_{v,F}|$ in time $f'(l)|N_l^G(v)|$.

The run time of this procedure is dominated by the time taken to compute $T_v, \mathcal{F}_v$ for $v \in V(G)$ and $|I_{v,F}|$ for $v \in V, F \in \mathcal{F}_v$. Since $\|G[N_l^G(v)]\| \leq |N_l^G(v)| + |\mathcal{F}_v|$, this run time is bounded by

$$\sum_{v \in V(G)} O\Big(\|G[N_l^G(v)]\| + |\mathcal{F}_v| l^2 + |\mathcal{F}_v| f'(l) |N_l^G(v)|\Big) = O\big(f'(l) \sum_{v \in V} |\mathcal{F}_v| |N_l^G(v)|\big). \qquad \blacktriangleleft$$

▶ **Lemma 3.** *There exists a function $f$ such that for graphs $G$ and $H$ one can compute $\#\mathrm{Sub}(H, G)$ in time $f(|H|)\gamma_{|H|}(G)$.*

**Proof.** (Sketch) Let $\mathcal{H}$ be a representative set of all connected pairwise non-isomorphic graphs with at most $|H|$ vertices. We compute $\#\mathrm{Sub}(H', G)$ for every connected graph $H' \in \mathcal{H}$. Via inclusion-exclusion, we can compute $\#\mathrm{Sub}(H, G)$. We sketch how the procedure works if $H$ consists of two components. Via induction, it can be generalized to an arbitrary number of components. Let $C_1$ and $C_2$ be the components of $H$. The value $c = \#\mathrm{Sub}(C_1, G) \cdot \#\mathrm{Sub}(C_2, G)$ counts all ways in which the two components of $H$ can be embedded in $G$. However, $c$ might be larger than $\#\mathrm{Sub}(H, G)$ since it also counts all embeddings where the two components intersect in $G$ by sharing one or more vertices. Every intersection of the two components is connected, thus, we can count them and subtract them. ◀

We now have a subgraph counting algorithm with efficient run time if the function $\gamma_{|H|}(G)$ is small. If $G$ has bounded degree or is a tree, then $\gamma_{|H|}(G)$ is an fpt function for the parameter $|H|$. It remains to show that the function is also small for certain random graphs.

## 4 Bounding $\gamma_l$ in Preferential Attachment Graphs

The remainder of this paper is concerned with the analysis of the run time of the aforementioned algorithm on preferential attachment graphs. This is done by using our concentration bounds for degrees (Theorem 19) to prove the following theorem.

▶ **Theorem 4.** *Let $l, n, m \in \mathbf{N}^+$ with $n \geq 2$. Then $\mathrm{E}[\gamma_l(G_m^n)] = m^{O(l^6)} \log(n)^{O(l^{12})} n$.*

This is then sufficient to prove our main result.

▶ **Theorem 5.** *There exists a function $f$ such that for every graph $H$ and $n, m \in \mathbf{N}$ one can compute $\#\mathrm{Sub}(H, G_m^n)$ in expected time $f(|H|) m^{O(|H|^6)} \log(n)^{O(|H|^{12})} n$.*

**Proof.** Direct consequence of Lemma 3 and Theorem 4. ◀

We prove Theorem 4 via multiple steps. In Lemma 7, we bound for every graph $G$ and $l \in \mathbf{N}^+$, $\gamma_l(G) \leq 16 l^6 \, |B_{4l^3}^2(G)|$, where $B_l^b(G)$ is defined below.

▶ **Definition 6.** *For a graph $G$ and $l, b \in \mathbf{N}$ let $B_l^b(G)$ be the set of subgraphs in $G$ of size at most $l$ with no isolated vertices and at most $b$ pendant vertices. If $G$ is a multigraph then $B_l^b(G)$ is defined with respect to the simple underlying graph.*

Then we use the degree bounds from Theorem 19 to step by step (Lemma 8 – 11) bound the expected value of $|B_l^b(G)|$ in preferential attachment graphs.

▶ **Lemma 7.** *Let $G$ be a graph and $l \in \mathbf{N}^+$. Then $\gamma_l(G) \leq 16l^6 \, |B^2_{4l^3}(G)|$.*

**Proof.** For every $v \in V(G)$ let $T_v$ be a breadth-first spanning tree with root $v$ in $G[N_l^G(v)]$ and $\mathcal{F}_v := \{F \mid F \subseteq E(G[N_l^G(v)]) \setminus E(T_v), |F| \leq \binom{l}{2}\}$ be the set of all subsets of at most $\binom{l}{2}$ edges that are in $G[N_l^G(v)]$ but not in $T_v$. Clearly $\gamma_l(G) = \sum_{v \in V(G)} |N_l^G(v)||\mathcal{F}_v|$.

Let $v \in V(G)$, $w \in N_l^G(v)$, $F \in \mathcal{F}_v$. Let $U \subseteq V(G)$ be the set containing $v, w$ and all endpoints of the edges in $F$. We define a graph $H_{v,w,F}$ as follows: Start with the empty graph, add the vertices $U$, the edges $F$, and for every $u \in U$ the unique path in $T_v$ from $v$ to $u$. Since $T_v$ is a breadth-first spanning tree, every path in $T_v$ starting at $v$ contains at most $l+1$ vertices. Since also $|U| \leq 2\binom{l}{2} + 2$, we can bound $V(H_{v,w,F}) \leq (2\binom{l}{2} + 2)(l+1) \leq 4l^3$. Furthermore $H_{v,w,F}$ contains no vertices with degree zero and every vertex in $H_{v,w,F}$ except for $v$ and $w$ is guaranteed to have degree at least two. This implies $H_{v,w,F} \in B^2_{4l^3}(G)$.

Let further $v' \in V(G)$, $w' \in N_l^G(v)$. If there exists $F' \in \mathcal{F}(v')$ with $H_{v,w,F} = H_{v',w',F'}$ then $v \in V(H_{v,w,s})$ and $w \in V(H_{v,w,s})$. Also there exists at most one $F' \in \mathcal{F}_v$ such that $H_{v,w,F} = H_{v',w',F'}$. Thus, there are at most $16l^6$ choices for $v', w', F'$ such that $H_{v,w,F} = H_{v',w',F'}$. ◀

It is now sufficient to bound the expected value of $|B_l^b(G)|$ in preferential attachment graphs. At first, we use Theorem 19 to give an upper bound on the degrees of single vertices.

▶ **Lemma 8.** *There exists $h > 0$ such that for $a \in \mathbf{R}$, $n, t, d \in \mathbf{N}^+$ with $n \geq at$, and $a \geq h \log\log(3at)$ it holds that $\Pr\left[d_1^n(v_t) \geq a\sqrt{\frac{n}{t}}\right] \leq e^{-a/h}$.*

**Proof.** Let $S = \{v_t, \dots, v_{t+\lceil a/5 \rceil}\}$. Then $d_1^n(v_t) \leq d_1^n(S)$. We assume $h$ to be large enough that $a \geq 1000$. Therefore $t + \lceil a/5 \rceil \leq at \leq n$ and $a/5 \leq d_1^{t+\lceil a/5 \rceil}(S) \leq 2\lceil 1 + a/5 \rceil \leq a/2$. We use these inequalities to bound

$$\Pr\left[d_1^n(v_t) \geq a\sqrt{\frac{n}{t}}\right] \leq \sum_{d=\lceil a/5 \rceil}^{\lfloor a/2 \rfloor} \Pr\left[d_1^{t+\lceil a/5 \rceil}(S) = d\right] \Pr\left[d_1^n(S) \geq 2\sqrt{\frac{n}{t}}d \;\middle|\; d_1^{t+\lceil a/5 \rceil}(S) = d\right].$$

Let $\varepsilon = 1/2$. We choose $h$ large enough such that $a/5 \geq \log(\log(3(t + \lceil a/5 \rceil)))\varepsilon^{-200}$ and $\varepsilon^{200} a/5 \geq a/h$. Theorem 19 yields for $\lceil a/5 \rceil \leq d \leq \lfloor a/2 \rfloor$

$$\Pr\left[d_1^n(S) \geq (1+\varepsilon)\sqrt{\frac{n}{t + \lceil a/5 \rceil}}d \;\middle|\; d_1^{t+\lceil a/5 \rceil}(S) = d\right] \leq e^{-\varepsilon^{200} d} \leq e^{-a/h}.$$ ◀

While it is easy to use the expected degree of a vertex to show that the probability that a single edge $v_x v_y$ exists in $G_1^n$ is close to $1/\sqrt{xy}$, it is surprisingly involved to bound the probability that multiple edges occur. This is because the existence of some edges influences the degree. Lemma 8 helps us here. We first show the result for $m = 1$ (Lemma 9) and then lift it to arbitrary values of $m$ (Lemma 10).

▶ **Lemma 9.** *Let $n \geq 2$ and $E \subseteq \binom{\{v_1, \dots, v_n\}}{2}$. Then*

$$\Pr[E \subseteq E(G_1^n)] \leq \log(n)^{O(|E|)^2} \prod_{v_x v_y \in E} 1/\sqrt{xy}.$$

**Proof.** We can assume $E$ that $E = \{v_{x_1}v_{y_1}, \dots, v_{x_l}v_{y_l}\}$ with $x_i < y_i$ for $1 \leq i \leq l$ and $y_i < y_j$ if $i < j$. Also, we define for $k \leq l$, $E_k = \{v_{x_1}v_{y_1}, \dots, v_{x_k}v_{y_k}\}$ as the subset of the first $k$ edges. The chain rule gives us

$$\Pr[E \subseteq E(G_1^n)] = \prod_{k=1}^{l} \Pr[v_{x_k}v_{y_k} \in E(G_1^n) \mid E_{k-1} \subseteq E(G_1^n)].$$

We fix some $1 \leq k \leq l$ and set $x = x_k$, $y = y_k$. It is now sufficient to show that

$$\Pr[v_xv_y \in E(G_1^n) \mid E_{k-1} \subseteq E(G_1^n)] \leq \log(n)^{O(k)}/\sqrt{xy}.$$

If $d_1^{y-1}(v_x) = l$ for $l \in \mathbf{N}$ then the edge $v_xv_y$ is inserted with probability $l/(2y-1)$. Thus

$$\begin{aligned}\Pr[v_xv_y \in E(G_1^n) \mid E_{k-1} \subseteq E(G_1^n)] &= \sum_{l=1}^{\infty} l/(2y-1) \cdot \Pr[d_1^{y-1}(v_x) = l \mid E_{k-1} \subseteq E(G_1^n)] \\ &= 1/(2y-1) \cdot \mathrm{E}[d_1^{y-1}(v_x) \mid E_{k-1} \subseteq E(G_1^n)] \leq \mathrm{E}[d_1^{y}(v_x) \mid E_{k-1} \subseteq E(G_1^n)]/y. \quad (1)\end{aligned}$$

Let now $\lambda \in \mathbf{R}$, whose value we will specify later. Since $d_1^y(v_x) \leq 2y$, the law of total probability states

$$\begin{aligned}\mathrm{E}[d_1^y(v_x) \mid E_{k-1} \subseteq E(G_1^n)] &\leq \lambda + 2y \Pr[d_1^y(v_x) > \lambda \mid E_{k-1} \subseteq E(G_1^n)] \\ &\leq \lambda + 2y \Pr[d_1^y(v_x) > \lambda] / \Pr[E_{k-1} \subseteq E(G_1^n)]. \quad (2)\end{aligned}$$

We now need to find a lower bound for $\Pr[E_{k-1} \subseteq E(G_1^n)]$. For the first $y$ steps the summed degree of all vertices is at most $2y$. Also each vertex has degree at least one. This means that every individual edge has probability at least $1/2y$, independent of where previous edges are. This observation together with the chain rule yields

$$\Pr[E_{k-1} \subseteq E(G_1^n)] = \prod_{i=1}^{k-1} \Pr[v_{x_i}v_{y_i} \in E(G_1^n) \mid E_{i-1} \subseteq E(G_1^n)] \leq 1/(2y)^k. \quad (3)$$

Combining (1), (2), and (3) yields

$$\Pr[v_xv_y \in E(G_1^n) \mid E \subseteq E(G_1^n)] \leq \lambda/y + 2y \Pr[d_1^y(v_x) > \lambda](2y)^k/y. \quad (4)$$

Let $h$ be the constant from Lemma 8. We now set $\lambda = h \log(y)^{2k}\sqrt{y/x}$. Then (4) and Lemma 8 (with $a = h\log(y)^{2k}$ and $e^{-a/h} = y^{-2k}$) yield

$$\Pr[v_xv_y \in E(G_1^n) \mid E \subseteq E(G_1^n)] \leq h \log(y)^{2k}\sqrt{y/x}/y + 2y^{-2k}(2y)^k = \log(n)^{O(k)}/\sqrt{xy}. \quad \blacktriangleleft$$

▶ **Lemma 10.** *Let $n, m \in \mathbf{N}^+$, $n \geq 2$ and $E \subseteq \binom{\{v_1,\dots,v_n\}}{2}$. Then*

$$\Pr[E \subseteq E(G_m^n)] \leq \log(n)^{O(|E|)^2} m^{2|E|} \prod_{v_xv_y \in E} 1/\sqrt{xy}.$$

**Proof.** One can simulate $G_m^n$ via $G_1^{mn}$, by merging every $m$ consecutive vertices into a single one. For $v_xv_y \in E$ let $E_{xy} = \{v_{x'}v_{y'} \mid m(x-1)+1 \leq x' \leq mx, m(y-1)+1 \leq y' \leq my\}$. This means the edge $v_xv_y$ is present after the merge operation in $G_m^n$ if any edge from $E_{xy}$ is present in $G_1^{mn}$. The union bound and Lemma 9 yield

$$\begin{aligned}\Pr[E \subseteq E(G_m^n)] &\leq \log(n)^{O(|E|)^2} \prod_{v_xv_y \in E} \sum_{v_{x'}v_{y'} \in E_{xy}} 1/\sqrt{x'y'} \\ &\leq \log(n)^{O(|E|)^2} m^{2|E|} \prod_{v_xv_y \in E} 1/\sqrt{xy}. \quad \blacktriangleleft\end{aligned}$$

We can now bound $\mathrm{E}[|B_l^b(G_m^n)|]$ by iterating over all possible embeddings of graphs of size at most $l$ with no isolated vertices and $b$ pendant vertices into $G_m^n$. We use Lemma 10 to bound the probability that the edges required for this embedding are indeed present in $G_m^n$.

▶ **Lemma 11.** *Let $l, b, n, m \in \mathbf{N}^+$ with $n \geq 2$. Then $\mathrm{E}[|B_l^b(G_m^n)|] = n^{b/2}\log(n)^{O(l^4)}m^{O(l^2)}$.*

**Proof.** Let $H$ be a graph with at most $l$ vertices, at most $b$ pendant vertices and no isolated vertices. Let $p$ be the expected number of subgraphs of $G_m^n$ that are isomorphic to $H$. We want to give an upper bound for $p$. Let $V(H) = \{u_1, \dots, u_\gamma\}$ with $\gamma \leq l$ and let $\delta_1, \dots, \delta_\gamma$ be the degree sequence of $V(H)$. We compute the following bound for later

$$\sum_{x_i=1}^{n} \frac{1}{\sqrt{x_i^{\delta_i}}} \leq 1 + \int_1^n \frac{1}{\sqrt{x^{\delta_i}}}dx \leq 1 + \begin{cases} \log(n) & \text{if } \delta_i \geq 2, \\ 2\sqrt{n} & \text{if } \delta_i = 1. \end{cases} \tag{5}$$

For integers $1 \leq x_1, \dots, x_\gamma \leq n$, we consider an embedding of $H$ into $G_m^n$ that maps $u_i$ to $v_{x_i}$ (for $1 \leq i \leq \gamma$). According to Lemma 10, the probability that this embedding of $H$ is a subgraph of $G_m^n$ is at most $\log(n)^{O(l^4)}m^{O(l^2)}\prod_{i=1}^{\gamma}\frac{1}{\sqrt{x_i^{\delta_i}}}$. We sum over all possible embeddings and use (5) to bound $p$ by

$$\begin{aligned}\sum_{x_1=1}^{n}\cdots\sum_{x_\gamma=1}^{n}\log(n)^{O(l^4)}m^{O(l^2)}\prod_{i=1}^{\gamma}\frac{1}{\sqrt{x_i^{\delta_i}}} &= \log(n)^{O(l^4)}m^{O(l^2)}\sum_{x_1=1}^{n}\frac{1}{\sqrt{x_1^{\delta_1}}}\cdots\sum_{x_\gamma=1}^{n}\frac{1}{\sqrt{x_\gamma^{\delta_\gamma}}} \\ &\overset{(5)}{=} \log(n)^{O(l^4)}m^{O(l^2)}(1+\log(n))^{\gamma}(1+2\sqrt{n})^{b} = n^{b/2}\log(n)^{O(l^4)}m^{O(l^2)}.\end{aligned}$$

For an arbitrary but fixed graph $H$ with at most $l$ vertices, no isolated vertices and at most $b$ pendant vertices we have bound the expected number of occurrences $p$. There are no more than $2^{l^2}$ graphs with at most $l$ fixed vertices. Therefore, $\mathrm{E}[|B_l^b(G_m^n)|] \leq 2^{l^2}n^{b/2}\log(n)^{O(l^4)}m^{O(l^2)}$. ◀

At last, Theorem 4 is a direct consequence Lemma 7 and Lemma 11.

## 5 Degree Bounds

In this section we show that under certain conditions the degree of vertices is closely centered around their expected value. This is formalized in Theorem 19, which is proven at the end of this section. We separately show upper and lower bounds and then join these bounds together. These bounds are proven by first giving bounds that hold for a short interval of time (Section 5.1) and then extending these bounds for longer intervals of time (Section 5.2).

Let $n \geq t$ and $S \subseteq \{v_1, \dots, v_t\}$. Remember that $d_m^n(S)$ is the degree of a set $S$ in $G_m^n$. Due to the technical nature of this section, we sometimes consider the set $S \subseteq \{v_1, \dots, v_t\}$ to be fixed and write $D(n)$ as shorthand for $d_1^n(S)$ to avoid having large formulas as a superscript. We also define $D(n) := D(\lfloor n \rfloor)$ for $n \in \mathbf{R}$. For $n > t$ we can explicitly state the probability distribution of $D(n)$ under the condition $D(n-1)$ as

$$\Pr[D(n) = x \mid D(n-1)] = \begin{cases} D(n-1)/(2n-1) & x = D(n-1)+1 \\ 1 - D(n-1)/(2n-1) & x = D(n-1) \\ 0 & \text{otherwise.} \end{cases}$$

## 5.1 Short-Term Degree Bounds

Here we show that for small $\delta$ from time-step $t$ to $(1+\delta)t$ it is very likely that we increase the degree of the set $S$ by a factor of $1+\delta/2+O(\delta^2)$.

▶ **Lemma 12.** *Let $0<\delta<1$ and $t \geq \frac{2}{\delta^2}$. Then*

$$\Pr\Big[D\big((1+\delta)t\big) \leq \big(1+\frac{\delta}{2}-2\delta^2\big)D(t) \;\Big|\; D(t)\Big] \leq e^{-\frac{1}{16}\delta^3 D(t)}.$$

**Proof.** For every $t' \in \mathbf{R}$ $D(t') = D(\lfloor t' \rfloor)$. For every $t' \in \mathbf{N}$ either $D(t') = D(t'-1)$ or $D(t') = D(t'-1)+1$. Let $N$ be the number of integers between $t$ and $(1+\delta)t$. Let $\Delta_i$ with $1 \leq i \leq N$ be the Bernoulli variable indicating that $D(\lfloor t \rfloor + i) = D(\lfloor t \rfloor + i - 1) + 1$ and $\Delta = \Delta_1 + \cdots + \Delta_N$. Then $D(t) + \Delta = D((1+\delta)t)$. Furthermore

$$\Pr[\Delta_i = 1 \mid \Delta_1, \ldots, \Delta_{i-1}, D(t)] = \frac{D(\lfloor t \rfloor + i - 1)}{2(\lfloor t \rfloor + i) - 1} \geq \frac{D(t)}{2(1+\delta)t}.$$

Let $X = X_1 + \cdots + X_N$ be the sum of identically distributed Bernoulli variables with $\Pr[X_i = 1] = \frac{D(t)}{2(1+\delta)t}$. We consider two experiments: The first game is $N$ tosses of a fair coin. The second one is $N$ tosses of a biased coin, where the probability that the $i$th coin comes up head depends on the outcome of the previous coins but always is at least $1/2$. Obviously, the probability of at least $s$ heads in the second experiment is at least as high as the probability of at least $s$ heads in the first experiment. The same argument implies

$$\Pr[\Delta \leq s \mid D(t)] \leq \Pr[X \leq s \mid D(t)]. \tag{6}$$

With $t \geq \frac{2}{\delta^2}$ we get $N \geq \delta t - 1 \geq (\delta - \frac{1}{2}\delta^2)t$ and

$$E[X \mid D(t)] = N \Pr[X_i = 1 \mid D(t)] \geq \frac{(\delta - \delta^2/2)D(t)}{2(1+\delta)}. \tag{7}$$

In contrast to $\Delta$, we can directly apply Chernoff bounds to $X$:

$$\Pr\Big[X \leq (1-\delta)E\big[X \mid D(t)\big] \;\Big|\; D(t)\Big] \leq e^{-\frac{1}{2}\delta^2 E[X|D(t)]}. \tag{8}$$

Combining the above inequality with (7), (6) and (8) yields

$$\begin{aligned}\Pr\Big[\Delta \leq \frac{(1-\delta)(\delta-\delta^2/2)D(t)}{2(1+\delta)} \;\Big|\; D(t)\Big] &\overset{(6)(7)}{\leq} \Pr\Big[X \leq (1-\delta)E[X \mid D(t)] \;\Big|\; D(t)\Big] \\ &\overset{(8)}{\leq} e^{-\frac{1}{2}\delta^2 E[X|D(t)]} \overset{(7)}{\leq} e^{-\frac{\delta^3-\delta^4/2}{4(1+\delta)}D(t)} \leq e^{-\frac{1}{16}\delta^3 D(t)}. \end{aligned} \tag{9}$$

For $0 \leq \delta \leq 1$, $\frac{(1-\delta)(\delta-\delta^2/2)}{2(1+\delta)} \geq \frac{\delta}{2} - 2\delta^2$. Thus, by (9) and $D((1+\delta)t) = \Delta + D(t)$

$$\begin{aligned}\Pr\big[D((1+\delta)t) \leq (1+\delta/2-2\delta^2)D(t) \;\big|\; D(t)\big] &= \Pr\big[\Delta \leq (\delta/2-2\delta^2)D(t) \;\big|\; D(t)\big] \\ &\leq e^{-\frac{1}{16}\delta^3 D(t)}.\end{aligned}$$

◀

Unfortunately, an additional factor of $\log(2et)$ is introduced in the following upper bound. The proof is very similar to the previous one and is omitted for lack of space.

▶ **Lemma 13.** *Let $0 < \delta \leq \frac{1}{e^2}$ and $t \geq \frac{2}{\delta^2}$. Then*

$$\Pr\Big[D((1+\delta)t) \geq (1+\delta/2+2\delta^2)D(t) \;\Big|\; D(t)\Big] \leq \log(2et)e^{-\frac{1}{8}\delta^3 D(t)}.$$

## 5.2 Long-Term Degree Bounds

In the previous subsection we established bounds for a small interval from step $t$ to step $(1+\delta)t$ with an error of order $\delta^2$. In this subsection we combine these bounds into long-term bounds. We get these bounds by defining positions $t_0 = t$ and $t_{k+1} = (1+\delta_k)t_k$ with $k \in \mathbf{N}$ and using the union bound to guarantee that for each interval from time $t_k$ to $t_{k+1}$ the short-term bounds hold. The choice of $\delta_k$ is of high importance for the success of this strategy. It turns out that we need the product $\prod_{k=1}^{\infty}(1+\delta_k)$ to diverge, but the error $\prod_{k=1}^{\infty}(1+\delta_k^2)$ to converge. We settle for $\delta_k = \varepsilon/k^{2/3}$, which satisfies both conditions.

Lemma 14 and Lemma 15 bridge the gap between the bounds for small intervals and longer periods by stating that if the degree differs by a factor of $(1 \pm \varepsilon)$ from its expected value then there has been one interval where the allowed error $O(\delta^2)$ has been exceeded.

▶ **Lemma 14.** *Let $0 < \varepsilon \le 1/8$, $t > 0$, and $f\colon \mathbf{R} \to \mathbf{R}$ be an increasing function. For every $k \in \mathbf{N}$ let $\delta_k = \frac{\varepsilon}{k^{2/3}}$, $h_k = \prod_{i=1}^{k-1}(1+\delta_i)$, and $c_k = \prod_{i=1}^{k-1}(1+\frac{1}{2}\delta_i - 2\delta_i^2)$.*

*If there is an $n \in \mathbf{N}$, such that $t < n$ and $f(n) < (1-\varepsilon)\sqrt{\frac{n}{t}}f(t)$, then there is a $k \in \mathbf{N}$ such that $f((1+\delta_k)h_k t) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)f(h_k t)$ and $f(h_k t) \ge c_k f(t)$.*

**Proof.** Consider any $n \in \mathbf{N}$, $n \ge t$. Let $k(n) \in \mathbf{N}$ be the maximal value such that $h_{k(n)}t \le n$. Then $\frac{n}{1+\delta_{k(n)}} \le h_{k(n)}t$, because of the maximality of $k(n)$. Notice that

$$(1-\varepsilon)\sqrt{\frac{n}{t}} \le e^{-\frac{1}{2}\varepsilon}e^{-\frac{1}{2}\varepsilon}\sqrt{\frac{n}{t}} = e^{-\frac{1}{2}\varepsilon}\sqrt{\frac{n}{te^{\varepsilon}}} \le e^{-\frac{1}{2}\varepsilon}\sqrt{\frac{n}{t(1+\delta_{k(n)})}} \le e^{-\frac{1}{2}\varepsilon}\sqrt{h_{k(n)}}$$

and for all $k \in \mathbf{N}$

$$c_k \ge \prod_{i=1}^{k-1} e^{\frac{1}{2}\delta_i - 3\delta_i^2} \ge \Big(\prod_{i=1}^{k-1} e^{\delta_i}\Big)^{\frac{1}{2}} \prod_{i=1}^{\infty} e^{-\frac{3\varepsilon^2}{i^{4/3}}} \ge \Big(\prod_{i=1}^{k-1}(1+\delta_i)\Big)^{\frac{1}{2}} e^{-4\varepsilon^2} \ge e^{-\frac{1}{2}\varepsilon}\sqrt{h_k}.$$

Combining the upper two inequalities gives us $(1-\varepsilon)\sqrt{n/t} \le c_k$. We assumed $f(n) < (1-\varepsilon)\sqrt{\frac{n}{t}}f(t)$. Monotonicity of $f$ yields $f(h_{k(n)}t) \le f(n) < (1-\varepsilon)\sqrt{\frac{n}{t}}f(t) \le c_{k(n)}f(t)$.

Let $J = \{\, j \ge 0 \mid f(h_{j+1}t) < c_{j+1}f(t) \,\}$. The set $J$ is not empty because $k(n) - 1 \in J$ by the equation above. Furthermore, $0 \notin J$ because $h_1 = c_1 = 1$ and therefore $f(h_1 t) = f(t) = c_1 f(t)$. Let now $k$ be the minimal value in $J$. Then $k > 0$, $f(h_k t) \ge c_k f(t)$, and $f(h_{k+1}t) < c_{k+1}f(t)$. At last, we have

$$f((1+\delta_k)h_k t) = f(h_{k+1}t) < c_{k+1}f(t) = (1+\frac{1}{2}\delta_k - 2\delta_k^2)c_k f(t) \le (1+\frac{1}{2}\delta_k - 2\delta_k^2)f(h_k t). \quad \blacktriangleleft$$

The proof of Lemma 15 is similar to the one of Lemma 14 and is therefore omitted.

▶ **Lemma 15.** *Let $0 < \varepsilon \le 1/40$, $t > 0$, and $f\colon \mathbf{R} \to \mathbf{R}$ be an increasing function. For every $k \in \mathbf{N}$ let $\delta_k = \frac{\varepsilon}{k^{2/3}}$, $h_k = \prod_{i=1}^{k-1}(1+\delta_i)$, and $c_k = \prod_{i=1}^{k-1}(1+\frac{1}{2}\delta_i + 2\delta_i^2)$.*

*If there is an $n \in \mathbf{N}$, such that $t < n$ and $f(n) > (1+\varepsilon)\sqrt{\frac{n}{t}}f(t)$, then there is a $k \in \mathbf{N}$ such that $f((1+\delta_k)h_k t) > (1+\frac{1}{2}\delta_k + 2\delta_k^2)f(h_k t)$ and $f(h_k t) \le c_k f(t)$.*

▶ **Lemma 16.** *Let $0 < \varepsilon \le 1/40$, $t > \frac{1}{\varepsilon^6}$. For every $k \in \mathbf{N}$ let $\delta_k = \frac{\varepsilon}{k^{2/3}}$, $h_k = \prod_{i=1}^{k-1}(1+\delta_i)$, $c_k^+ = \prod_{i=1}^{k-1}(1+\frac{1}{2}\delta_i + 2\delta_i^2)$ and $c_k^- = \prod_{i=1}^{k-1}(1+\frac{1}{2}\delta_i - 2\delta_i^2)$. Then*

$$\Pr\Big[D((1+\delta_k)h_k t) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_k t), D(h_k t) \ge c_k^- D(t) \;\Big|\; D(t)\Big] \le e^{-\frac{1}{16}\delta_k^3 c_k^- D(t)},$$

$$\Pr\Big[D((1+\delta_k)h_k t) > (1+\frac{1}{2}\delta_k + 2\delta_k^2)D(h_k t), D(h_k t) \le c_k^+ D(t) \;\Big|\; D(t)\Big] \le \log(2et)e^{-\frac{1}{8}\delta_k^3 c_k^+ D(t)}.$$

**Proof.** At first we focus on the first bound. By the law of total probability

$$\Pr\Big[D((1+\delta_k)h_kt) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_kt), D(h_kt) \geq c_k^- D(t) \;\Big|\; D(t)\Big] \\ \leq \Pr\Big[D((1+\delta_k)h_kt) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_kt) \;\Big|\; D(h_kt) \geq c_k^- D(t)\Big].$$

The second line of this equation states the probability that the degree of a vertex is in the future below a certain threshold under the condition that it is currently above a certain threshold. We can bound this probability if we assume that it currently is not above, but exactly at the threshold:

$$\Pr\Big[D((1+\delta_k)h_kt) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_kt) \;\Big|\; D(h_kt) \geq c_k^- D(t)\Big] \\ \leq \Pr\Big[D((1+\delta_k)h_kt) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_kt) \;\Big|\; D(h_kt) = c_k^- D(t)\Big].$$

Similarly, the probability that the degree of a vertex is in the future above a certain threshold under the condition that it is currently below a certain threshold can be bounded by assuming that it is exactly at the threshold. Thus, it suffices to prove the following two bounds

$$\Pr\Big[D((1+\delta_k)h_kt) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_kt) \;\Big|\; D(h_kt) = c_k^- D(t)\Big] \leq e^{-\frac{1}{16}\delta_k^3 c_k^- D(t)},$$

$$\Pr\Big[D((1+\delta_k)h_kt) > (1+\frac{1}{2}\delta_k + 2\delta_k^2)D(h_kt) \;\Big|\; D(h_kt) = c_k^+ D(t)\Big] \leq \log(2et)e^{-\frac{1}{8}\delta_k^3 c_k^+ D(t)}.$$

Lemma 12 and 13 state that if $0 \leq \delta_k = e/k^{3/2} \leq 1/e^2$ and $h_kt \geq 2/\delta_k^2$ for every $k$ then these bounds are true. We observe that for $0 \leq \varepsilon \leq 1/8$ the first precondition is always satisfied. We will finish the proof by showing that $h_kt \geq 2/\delta_k^2$ for every $k$. Observe that for $0 \leq k \leq 1$ we have $h_kt \geq 2/\varepsilon^2 \geq 2/\delta_k^2$. We can therefore assume $k \geq 2$. First, we need a lower bound for $h_k$.

$$h_k = \prod_{i=1}^{k-1}(1+\frac{\varepsilon}{i^{2/3}}) \geq \prod_{i=1}^{k-1} e^{\frac{\varepsilon}{i^{2/3}}} \geq e^{3\varepsilon(k-1)^{1/3}} \geq e^{2\varepsilon k^{1/3}}$$

One can show that $e^x/x^4 \geq e^4/256$ for $x > 0$. We therefore get for $x = 2\varepsilon k^{1/3}$

$$h_k \geq e^{2\varepsilon k^{1/3}} = \frac{e^{2\varepsilon k^{1/3}}}{(2\varepsilon k^{1/3})^4}\frac{16\varepsilon^6}{\delta_k^2} \geq \frac{e^x}{x^4}\frac{16\varepsilon^6}{\delta_k^2} \geq \frac{e^4}{256}\frac{16\varepsilon^6}{\delta_k^2} \geq \frac{2\varepsilon^6}{\delta_k^2}.$$

Since $t \geq \frac{1}{\varepsilon^6}$ it follows that $h_kt \geq \frac{2}{\delta_k^2}$. ◀

▶ **Lemma 17.** *For* $0 < \varepsilon \leq 1/40$ *and* $\frac{1}{\varepsilon^6} < t \in \mathbf{N}$

$$\Pr\Big[(1-\varepsilon)\sqrt{\frac{n}{t}}D(t) < D(n) < (1+\varepsilon)\sqrt{\frac{n}{t}}D(t) \text{ for all } n \geq t \;\Big|\; D(t)\Big] \\ \geq 1 - \log(15t)\varepsilon^{-6}\exp\big(-\varepsilon^{15}10^{-24}D(t)\big).$$

**Proof.** Observe that

$$\Pr\Big[(1-\varepsilon)\sqrt{\frac{n}{t}}D(t) < D(n) < (1+\varepsilon)\sqrt{\frac{n}{t}}D(t) \text{ for all } n \geq t \;\Big|\; D(t)\Big] \geq 1 - (p^+ + p^-)$$

with

$$p^- := \Pr\Big[D(n) < (1-\varepsilon)\sqrt{\frac{n}{t}}D(t) \text{ for some } n \geq t \;\Big|\; D(t)\Big]$$
$$p^+ := \Pr\Big[D(n) > (1+\varepsilon)\sqrt{\frac{n}{t}}D(t) \text{ for some } n \geq t \;\Big|\; D(t)\Big].$$

We proceed by finding upper bounds for $p^+$ and $p^-$. For $k \in \mathbf{N}$ let $\delta_k = \frac{\varepsilon}{k^{2/3}}$, $h_k = \prod_{i=1}^{k-1}(1-\delta_i)$, $c_k^- = \prod_{i=1}^{k-1}(1-\frac{1}{2}\delta_i - 2\delta_i^2)$ and $c_k^+ = \prod_{i=1}^{k-1}(1-\frac{1}{2}\delta_i + 2\delta_i^2)$. Every function $f(t) : \mathbf{R} \to \mathbf{R}$ that is a realization of the random variables $D(t)$ is monotonically increasing. It follows using Lemma 14, Lemma 15, the union bound over all possible choices of $k$, and Lemma 16 that

$$p^- \leq \sum_{k=0}^{\infty} \Pr\Big[D((1+\delta_k)h_k t) < (1+\frac{1}{2}\delta_k - 2\delta_k^2)D(h_k t), D(h_k t) \geq c_k^- D(t) \;\Big|\; D(t)\Big]$$
$$\leq \sum_{k=0}^{\infty} e^{-\frac{1}{16}\delta_k^3 c_k^- D(t)},$$
$$p^+ \leq \sum_{k=0}^{\infty} \Pr\Big[D((1+\delta_k)h_k t) > (1+\frac{1}{2}\delta_k + 2\delta_k^2)D(h_k t), D(h_k t) \leq c_k^+ D(t) \;\Big|\; D(t)\Big]$$
$$\leq \sum_{k=0}^{\infty} \log(2et) e^{-\frac{1}{8}\delta_k^3 c_k^+ D(t)}.$$

It remains to show that $p^+ + p^- \leq \log(15t)\varepsilon^{-6}\exp\big(-\varepsilon^{15}10^{-24}D(t)\big)$. This last last step requires a longer calculation which we omit because of space limitations. ◀

The next lemma is a slight variant of Lemma 17. The proof is omitted for lack of space.

▶ **Lemma 18.** *For $t \in \mathbf{R}$, $t \geq 1$, $0 < \varepsilon \leq 1/2$, $d \in \mathbf{N}$ with $\Pr[D(t) = d] \neq 0$ and $d \geq \log(\log(3t))\varepsilon^{-200}$*

$$\Pr\Big[(1-\varepsilon)\sqrt{\frac{n}{t}}d < D(n) < (1+\varepsilon)\sqrt{\frac{n}{t}}d \text{ for all } n \geq t \;\Big|\; D(t) = d\Big] \geq 1 - e^{-\varepsilon^{200}d}.$$

At last, we generalize this result to different values of $m$.

▶ **Theorem 19.** *For $t, m, d \in \mathbf{N}^+$, $0 < \varepsilon \leq 1/2$, $S \subseteq \{v_1, \ldots, v_t\}$ with $\Pr[d_m^t(S) = d] \neq 0$ and $d \geq \log(\log(3tm))\varepsilon^{-200}$*

$$\Pr\Big[(1-\varepsilon)\sqrt{\frac{n}{t}}d < d_m^n(S) < (1+\varepsilon)\sqrt{\frac{n}{t}}d \text{ for all } n \geq t \;\Big|\; d_m^t(S) = d\Big] \geq 1 - e^{-\varepsilon^{200}d}.$$

**Proof.** As stated in the introduction, we can simulate $G_m^n$ via $G_1^{mn}$, by merging every $m$ consecutive vertices into a single one. Let $G_m^n$ be a graph with vertices $V = \{v_1, \ldots, v_n\}$. We can assume that this graph has been constructed from a graph $G_1^{mn}$ with vertex set $V' = \{v'_1, \ldots, v'_{mn}\}$ by merging $v'_{(i-1)m+1}, \ldots, v'_{im}$ into $v_i$ for $1 \leq i \leq n$. Let $S' \subseteq V'$ be the set of vertices in $G_1^{mn}$ that are merged into $S$. Since the graph allows multi-edges, $d_m^n(S)$ and $d_1^{mn}(S')$ have the same probability distribution. Lemma 17 states with $d_1^{mn}(S') = D(mn)$

$$\Pr\Big[(1-\varepsilon)\sqrt{\frac{n}{t}}d < d_1^{mn}(S') < (1+\varepsilon)\sqrt{\frac{n}{t}}d \text{ for all } nm \geq tm \;\Big|\; d_1^{tm}(S') = d\Big] \geq 1 - e^{-\varepsilon^{200}d}.$$ ◀

## References


1 Noga Alon, Phuong Dao, Iman Hajirasouliha, Fereydoun Hormozdiari, and S Cenk Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241–i249, 2008.

2 Agnes Backhausz et al. Limit distribution of degrees in random family trees. *Electronic Communications in Probability*, 16:29–37, 2011.

3 Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

4 Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.

5 Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The Degree Sequence of a Scale-free Random Graph Process. *Random Structures & Algorithms*, 18(3):279–290, May 2001.

6 Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1017, 2019.

7 Jianer Chen, Benny Chor, Mike Fellows, Xiuzhen Huang, David Juedes, Iyad A Kanj, and Ge Xia. Tight lower bounds for certain parameterized NP-hard problems. *Information and Computation*, 201(2):216–231, 2005.

8 Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 106–115. ACM, 2006.

9 Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

10 Colin Cooper and Alan Frieze. The cover time of the preferential attachment graph. *Journal of Combinatorial Theory, Series B*, 97(2):269–290, 2007.

11 Bruno Courcelle. The Monadic Second-Order Logic of Graphs I. Recognizable Sets of Finite Graphs. *Information and Computation*, 85(1):12–75, 1990.

12 Radu Curticapean, Holger Dell, and Dániel Marx. Homomorphisms Are a Good Basis for Counting Small Subgraphs. In *Proc. of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 210–223, New York, NY, USA, 2017. ACM. doi:10.1145/3055399.3055502.

13 R. Diestel. *Graph Theory*. Springer, Heidelberg, 2010.

14 Jörg Flum and Martin Grohe. The parameterized complexity of counting problems. *SIAM Journal on Computing*, 33(4):892–922, 2004.

15 Alan Frieze and Colin McDiarmid. Algorithmic theory of random graphs. *Random Structures & Algorithms*, 10(1-2):5–42, 1997.

16 Joshua A Grochow and Manolis Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *Annual International Conference on Research in Computational Molecular Biology*, pages 92–106. Springer, 2007.

17 Svante Janson. Limit theorems for triangular urn schemes. *Probability Theory and Related Fields*, 134(3):417–452, 2006.

18 Svante Janson, Tomasz Łuczak, and Ilkka Norros. Large cliques in a power-law random graph. *Journal of Applied Probability*, 47(4):1124–1135, 2010.

19 Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

20 Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

21 Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *Proc. of the VLDB Endowment*, 7(5):377–388, 2014.

22 Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

23 Tamás F Móri. The maximum degree of the Barabási–Albert random tree. *Combinatorics, Probability and Computing*, 14(3):339–348, 2005.

24 Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

25 Saeed Omidi, Falk Schreiber, and Ali Masoudi-Nejad. MODA: an efficient algorithm for network motif discovery in biological networks. *Genes & genetic systems*, 84(5):385–395, 2009.

26 Erol Peköz, Adrian Röllin, and Nathan Ross. Joint degree distributions of preferential attachment random graphs. *Advances in Applied Probability*, 49(2):368–387, 2017.

27 Erol A Peköz, Adrian Röllin, Nathan Ross, et al. Degree asymptotics with rates for preferential attachment random graphs. *The Annals of Applied Probability*, 23(3):1188–1218, 2013.

28 Falk Schreiber and Henning Schwöbbermeyer. Frequency concepts and pattern detection for the analysis of motifs in networks. In *Transactions on computational systems biology III*, pages 89–104. Springer, 2005.

29 Leslie G Valiant. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979.

30 Remco van der Hofstad. *Random graphs and complex networks*, volume 1. Cambridge University Press, 2016.

31 Sebastian Wernicke. Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(4):347–359, 2006.

32 Panpan Zhang, Chen Chen, and Hosam Mahmoud. Explicit characterization of moments of balanced triangular Pólya urns by an elementary approach. *Statistics & Probability Letters*, 96:149–153, 2015.