Counting Arbitrary Subgraphs in Data Streams

Daniel M. Kane[1], Kurt Mehlhorn[2], Thomas Sauerwald[2], and He Sun[2]

[1] Department of Mathematics, Stanford University, USA
[2] Max Planck Institute for Informatics, Germany

Abstract. We study the subgraph counting problem in data streams. We provide the first non-trivial estimator for approximately counting the number of occurrences of an *arbitrary* subgraph H of constant size in a (large) graph G. Our estimator works in the turnstile model, i.e., can handle both edge-insertions and edge-deletions, and is applicable in a distributed setting. Prior to this work, only for a few non-regular graphs estimators were known in case of edge-insertions, leaving the problem of counting general subgraphs in the turnstile model wide open (cf. [13, Problem 11]). We further demonstrate the applicability of our estimator by analyzing its concentration for several graphs H and the case where G is a power law graph.

1 Introduction

Counting (small) subgraphs in massive graphs is one of the fundamental tasks in algorithm design and has various applications, including analyzing the connectivity of networks, uncovering the structural information of large graphs, and indexing graph databases. The current best known algorithm for the simplest non-trivial version of the problem, counting the number of triangles, is based on matrix multiplication, and is infeasible even for a graph of medium size. Therefore it is natural to consider the problem in the data streaming setting, where the edges come sequentially and the algorithm is required to approximate the number of subgraphs without storing the whole graph.

Formally in this problem, we are given a set of items $s_1, s_2, \ldots$ in a data stream. These items arrive sequentially and represent edges of an underlying graph $G = (V, E)$. Two standard models [15] in this context are *the Cash Register Model* and *the Turnstile Model*. In the cash register model, each item s_i represents one edge and these arrived items form a graph G with edge set $E := \bigcup\{s_i\}$, where $E = \emptyset$ initially. The turnstile model generalizes the cash register model and is applicable to dynamic situations. Specifically, each item s_i in the turnstile model is of the form (e_i, sign_i), where e_i is an edge of G and $\mathrm{sign}_i \in \{+, -\}$ indicates that e_i is inserted to or deleted from G. That is, after reading the ith item, $E \leftarrow E \cup \{e_i\}$ if $\mathrm{sign}_i = +$, and $E \leftarrow E \setminus \{e_i\}$ otherwise.

In a more general distributed setting, there are k distributed sites, each receiving a stream S_i of elements over time, and every S_i is processed by a local host. When the number of subgraphs is asked for, these k hosts cooperate to give an approximation for the underlying graph formed by $\bigcup_{i=1}^{k} S_i$.

Our Results & Techniques. We present the first sketch for counting *arbitrary* subgraphs of constant size in data streams. While most of the previous algorithms are based on sampling techniques and cannot be extended to count more complex subgraphs, our algorithm can approximately count arbitrary (possibly directed) subgraphs. Moreover, our algorithm runs in the turnstile model and is applicable in the distributed setting.

More formally, for any fixed subgraph H with a constant number of edges, we present an algorithm that $(1 \pm \varepsilon)$-approximates $\#H$, the number of occurrences of H, in the underlying graph G. That is, for any constant $0 < \varepsilon < 1$, with probability at least $2/3$ the output Z of our algorithm satisfies $Z \in [(1-\varepsilon) \cdot \#H, (1+\varepsilon) \cdot \#H]$. For several families of graphs H, our algorithm achieves a $(1 \pm \varepsilon)$-approximation for the number of subgraphs H in G within sublinear space. Our result generalizes the previous works which can be only used for counting cycles in the turnstile model [11, 12], and answers the 11th open problem in the 2006 IITK Workshop on Algorithms for Data Steams [13].

Our sketch relies on a novel approach of designing random vectors that are based on different combinations of complex numbers. By using different roots of unity and random mappings from vertices in G to complex numbers, we obtain an unbiased estimator for $\#H$. This partially answers Problem 4 of the survey by Muthukrishnan [15], which asks for suitable applications of complex-valued hash functions in data streaming algorithms. Apart from counting subgraphs in streams, we believe that our new approach will find more applications.

We further consider the problem of counting stars in power law graphs, which include many practical networks from computer science, biology and sociology. We show that $O\left(\frac{1}{\varepsilon^2} \cdot \log n\right)$ bits suffice to get a $(1\pm\varepsilon)$-approximation for counting stars S_k, while the exact counting needs $n \cdot \log n$ bits of space. Our main results are summarized in Table 1 on the next page.

To demonstrate that for a large family of graphs G our algorithm achieves a $(1 \pm \varepsilon)$-approximation within sublinear space, we consider the random graph $G = G(n,p)$, where each edge is placed independently with a fixed probability $p \geqslant (1+\varepsilon) \cdot \ln(n)/n$. Random graphs are of interest for the performance of our algorithm, as the independent appearance of the edges in $G = G(n,p)$ reduces the number of particular patterns. In other words, if our algorithm has low space complexity for counting a subgraph H in $G(n,p)$, then the space complexity is even lower for counting a more frequently occurring graph in a real-world graph G which has the same density as $G(n,p)$.

Regarding the space complexity of our algorithm on random graphs, assume for instance that the subgraph H is a P_3 or S_3 (i.e., a path or a star with 3 edges). The expected number of occurrences of such a graph is of order $n^4p^3 \gg 1$. It can be shown by standard techniques (cf. [1, Section 4.4]) that the number of occurrences is also of this order with probability $1 - o(1)$ as $n \to \infty$. Assuming that this event occurs, Theorem 3.3 along with the facts that $m = \Theta(n^2p)$ and $\Delta(G) = \Theta(np)$ imply a $(1\pm\varepsilon)$-approximation algorithm for P_3 (or S_3) with space complexity $O(\frac{1}{\epsilon^2} \cdot n \cdot \log n)$. For stars S_k with any constant k, the result from Theorem 4.1 yields a $(1 \pm \varepsilon)$-approximation algorithm with space complexity

Conditions	Space Complexity	Reference
any graph G any graph H	$O\Big(\frac{1}{\varepsilon^2} \cdot \frac{m^k \cdot \Delta(G)^k}{(\#H)^2} \cdot \log n\Big)$	Theorem 3.3
any graph G H with $\delta(H) \geqslant 2$	$O\Big(\frac{1}{\varepsilon^2} \cdot \frac{m^k}{(\#H)^2} \cdot \log n\Big)$	Theorem 3.3
any graph G stars S_k	$O\left(\frac{n^{1-1/k}}{\varepsilon^2} \cdot \left(\frac{n^{3/2-1/(2k)} \cdot \Delta(G)^{2k}}{(\#S_k)^2} + 1\right) \cdot \log n\right)$	Theorem 4.1
Power law graph G stars S_k	$O\left(\frac{1}{\varepsilon^2} \cdot \log n\right)$	Theorem 4.2

Table 1. Space requirement for $(1 \pm \varepsilon)$-approximately counting an undirected and connected graph H with $k = O(1)$ edges. Here δ and Δ denote the minimum and maximum degree, respectively. Space complexity is measured in terms of bits.

$O\left(\frac{1}{\epsilon^2} \cdot \sqrt{n} \cdot \log n\right)$. Finally, for any cycle with $k = O(1)$ edges, Theorem 3.3 gives an algorithm with space complexity $O(\frac{1}{\epsilon^2} \cdot p^{-k} \cdot \log n)$, which is sublinear for sufficiently large values of p, e.g., $p = \Theta(1)$.

Related Work. Bar-Yossef, Kumar and Sivakumar were the first to consider the subgraph counting problem in data streams and presented an algorithm for counting triangles [3]. After that, the problem of counting triangles in data streams was studied extensively [4, 6, 11, 17]. The problem of counting other subgraphs was also addressed in the literature. Buriol et al. [7] considered the problem of estimating clustering indexes in data streams. Bordino et al. [5] extended the technique of counting triangles [6] to all subgraphs on three and four vertices. Manjunath et al. [12] considered counting cycles of constant size in data streams. Among these results, only two algorithms [11, 12] work in the turnstile model and these only hold for cycles.

Apart from designing algorithms in the streaming model, the subgraph counting problem has been studied extensively. Alon et al. [2] presented an algorithm for counting given-length cycles. Gonen et al. [10] showed how to count stars and other small subgraphs in sublinear time. In particular, several small subgraphs in a network, named network motifs, have been identified as the simple building blocks of complex biological networks and the distribution of their occurrences could reveal answers to many important biological questions [14, 19].

Notation. Let $G = (V, E)$ be an undirected graph without self-loops and multiple edges. The set of vertices and edges are represented by $V[G]$ and $E[G]$, respectively. We will assume that $V[G] = \{1, \ldots, n\}$ and n is known in advance. For any vertex $u \in V[G]$, the degree of u is denoted by $\deg(u)$. The maximum and minimum degree of G are denoted by $\Delta(G)$ and $\delta(G)$, respectively.

Given two directed graphs H_1 and H_2, we say that H_1 is *homomorphic* to H_2 if there is a mapping $i : V[H_1] \to V[H_2]$ such that $(u, v) \in E[H_1]$ implies $(i(u), i(v)) \in E[H_2]$. Graphs H_1 and H_2 are said to be *isomorphic* if there is a bijection $i : V[H_1] \to V[H_2]$ such that $(u, v) \in E[H_1]$ iff $(i(u), i(v)) \in E[H_2]$. For any graph H, let $\text{auto}(H)$ be the number of automorphisms of H.

For any graph H, we call a not necessarily induced subgraph H_1 of G an *occurrence* of H, if H_1 is isomorphic to H. Let $\#(H,G)$ be the number of occurrences of H in G. When the reference to G is clear from the context, we simply write $\#H$. A kth root of unity is any number of the form $\mathrm{e}^{2\pi\mathrm{i}\cdot j/k}$, $0 \leqslant j < k$.

Organization. The paper is structured as follows. Section 2 presents an unbiased estimator for counting arbitrary subgraphs, followed by the correctness and concentration analysis in Section 3. In Section 4 we consider two approaches to reduce the space complexity and give an improved algorithm for counting stars. Due to page limitations, some proofs and lemmas are omitted in this extended abstract and can be found in the appendix.

2 An Unbiased Estimator for Counting Subgraphs

We present a framework for counting general subgraphs. Suppose that H is a fixed graph with t vertices and k edges, and we want to count the number of occurrences of H in G. For the notation, we denote vertices of H by a, b and c, and vertices of G are denoted by u, v and w, respectively. Let the degree of vertex a in H be $\deg_H(a)$. We equip the edges of H with an arbitrary orientation, as this is necessary for the further analysis. Therefore, each edge in H together with its orientation can be expressed as $\overrightarrow{ab}$ for some $a, b \in V[H]$. For simplicity and with slight abuse of notation we will use H to denote such an oriented graph.

At a high level, our estimator maintains k complex variables $Z_{\overrightarrow{ab}}(G)$, $\overrightarrow{ab} \in E[H]$, and these variables are set to be zero initially. For every arriving edge $\{u,v\} \in E[G]$ we update each $Z_{\overrightarrow{ab}}(G)$ according to

$$Z_{\overrightarrow{ab}}(G) \leftarrow Z_{\overrightarrow{ab}}(G) + \mathcal{M}_{\overrightarrow{ab}}(u,v) + \mathcal{M}_{\overrightarrow{ab}}(v,u) \ ,$$

where $\mathcal{M}_{\overrightarrow{ab}} : V[G] \times V[G] \to \mathbb{C}$ can be computed in constant time. Hence

$$Z_{\overrightarrow{ab}}(G) = \sum_{\{u,v\} \in E[G]} \mathcal{M}_{\overrightarrow{ab}}(u,v) + \mathcal{M}_{\overrightarrow{ab}}(v,u) \ .$$

Intuitively $\mathcal{M}_{\overrightarrow{ab}}(u,v)$ expresses the event to give $\{u,v\}$ the orientation $\overrightarrow{uv}$ and map $\overrightarrow{uv}$ to $\overrightarrow{ab}$, and $\mathcal{M}_{\overrightarrow{ab}}(u,v) + \mathcal{M}_{\overrightarrow{ab}}(v,u)$ is used to express two different orientations of edge $\{u,v\}$. For every query of the number of subgraphs, the estimator simply outputs the real part of $\alpha \cdot \prod_{\overrightarrow{ab}} Z_{\overrightarrow{ab}}(G)$, where $\alpha \in \mathbb{R}^+$ is a scaling factor.

More formally, each $\mathcal{M}_{\overrightarrow{ab}}(u,v)$ is defined according to the degree of vertices a, b in graph H and consists of the product of three types of random variables $Q, X_c(w)$ and $Y(w)$, $c \in V[H]$ and $w \in V[G]$:

- Variable Q is a random τth root of unity, where $\tau := 2^t - 1$.
- For vertex $c \in V[H], w \in V[G]$, $X_c(w)$ is random $\deg_H(c)$th root of unity, and for each vertex $c \in V[H]$, $X_c : V[G] \to \mathbb{C}$ is chosen independently and uniformly at random from a family of $4k$-wise independent hash functions. Variables Q and $X_c, c \in V[H]$ are chosen independently.

– For every $w \in V[G]$, $Y(w)$ is a random element from $S := \{1, 2, 4, 8, \ldots, 2^{t-1}\}$ as part of a $4k$-wise independent hash function. Variables $Y(w), w \in V[G]$ and Q are chosen independently.

Given the notations above, we define each function $\mathcal{M}_{\overrightarrow{ab}}$ as

$$\mathcal{M}_{\overrightarrow{ab}}(u,v) := X_a(u)\ X_b(v)\ Q^{\frac{Y(u)}{\deg_H(a)}}\ Q^{\frac{Y(v)}{\deg_H(b)}}.$$

Estimator 1 gives the formal description of the update and query procedure.

Estimator 1 Counting $\#(H, G)$

Step 1 (Update): When an edge $e = \{u, v\} \in E[G]$ arrives, update each $Z_{\overrightarrow{ab}}$ w.r.t.

$$Z_{\overrightarrow{ab}}(G) \leftarrow Z_{\overrightarrow{ab}}(G) + \mathcal{M}_{\overrightarrow{ab}}(u,v) + \mathcal{M}_{\overrightarrow{ab}}(v,u). \tag{1}$$

Step 2 (Query): When $\#(H, G)$ is required, output the real part of

$$\frac{t^t}{t! \cdot \mathrm{auto}(H)} \cdot Z_H(G)\ , \tag{2}$$

where Z_H is defined by

$$Z_H(G) := \prod_{\overrightarrow{ab} \in E[H]} Z_{\overrightarrow{ab}}(G)\ . \tag{3}$$

Estimator 1 is applicable in a quite general setting: First, the estimator runs in the turnstile model. For simplicity the update procedure above is only described for the edge-insertion case. For every item of the stream that represents an edge-deletion, we replace "+" by "−" in (1). Second, our estimator also works in the distributed setting, where every local host maintains variables $Z_{\overrightarrow{ab}}$, $ab \in E[H]$, and does the update for every arriving item in the local stream. When the output is required, these variables located at different hosts are summed up and we return the estimated value according to (3). Third, the estimator above can be revised easily to count the number of directed subgraphs in a directed graphs. Since in this case we need to change the constant of (2) accordingly, in the rest of our paper we only focus on the case of counting undirected graphs.

3 Analysis of the Estimator

In this section we first prove that $Z_H(G)$ defined by (3) is an unbiased estimator for $\#(H, G)$. Then we analyze the variance of the estimator. The main result of our paper will be presented at the end of this section.

Let us first explain the intuition behind our estimator. By definition we have

$$Z_H(G) = \prod_{\overrightarrow{ab} \in E[H]} Z_{\overrightarrow{ab}}(G) = \prod_{\overrightarrow{ab} \in E[H]} \sum_{\{u,v\} \in E[G]} \left(\mathcal{M}_{\overrightarrow{ab}}(u,v) + \mathcal{M}_{\overrightarrow{ab}}(v,u) \right)\ . \tag{4}$$

Since H has k edges, $Z_H(G)$ is a product of k terms and each term is a sum over all edges of G each with two possible orientations. Hence, in the expansion of $Z_H(G)$ any k-tuple $(e_1,\ldots,e_k) \in E^k[G]$ contributes 2^k different terms to $Z_H(G)$ and each term corresponds to a certain orientation of $(e_1,\ldots,e_k)$. Let $\overrightarrow{T} = (\overrightarrow{e_1},\ldots,\overrightarrow{e_k})$ be an arbitrary orientation of $(e_1,\ldots,e_k)$ and $G_{\overrightarrow{T}}$ be the graph induced from $\overrightarrow{T}$.

At a high level, we use three types of variables to test if $G_{\overrightarrow{T}}$ is isomorphic to H. However, the roles of these variables are different. (i) Through function Y every vertex $u \in V_{\overrightarrow{T}}$ maps to one element in S randomly. If $|V_{\overrightarrow{T}}| = |S| = t$, then with constant probability, vertices in $V_{\overrightarrow{T}}$ map to different t numbers in S. Otherwise, $|V_{\overrightarrow{T}}| < t$ and vertices in $V_{\overrightarrow{T}}$ cannot map to different t elements. Since Q has the property that $\mathbb{E}[Q^i] \neq 0$ $(1 \leqslant i \leqslant \tau)$ iff $i = \tau$, where $\tau = \sum_{\ell \in S} \ell$, the combination of Q and Y guarantees that $G_{\overrightarrow{T}}$ contributes to $\mathbb{E}[Z_H(G)]$ only if graph H and $G_{\overrightarrow{T}}$ have the same number of vertices. (ii) For $c \in V[H], w \in V[G]$, random variable $X_c(w)$ guarantees that $G_{\overrightarrow{T}}$ contributes to $\mathbb{E}[Z_H(G)]$ only if there is a homomorphism from H to $G_{\overrightarrow{T}}$.

Now we analyze our sketch and prove that $Z_H(G)$ is an unbiased estimator for $\#(H,G)$.

Theorem 3.1. *Let H be a graph with t vertices and k edges. Assume that variables $X_c(w), Y(w), c \in V[H], w \in V[G]$ and Q are as defined above. Then*

$$\mathbb{E}[Z_H(G)] = \frac{t! \cdot \mathrm{auto}(H)}{t^t} \cdot \#(H,G) \ .$$

Proof. For a k-tuple $\overrightarrow{T} = (e_1,\ldots,e_k) \in E^k[G]$, let $G_{\overrightarrow{T}} = (V_{\overrightarrow{T}}, E_{\overrightarrow{T}})$ be the induced multi-graph from set $E_{\overrightarrow{T}}$, i.e., $G_{\overrightarrow{T}}$ has edge multi-set $E_{\overrightarrow{T}} = \{e_1,\ldots,e_k\}$. Let $\overrightarrow{T} = (\overrightarrow{e_1},\ldots,\overrightarrow{e_k})$ be an arbitrary orientation of $(e_1,\ldots,e_k)$, where $\overrightarrow{e_i} = \overrightarrow{u_i v_i}$. Consider the expansion of $Z_H(G)$ below:

$$Z_H(G) = \prod_{\overrightarrow{ab} \in E[H]} Z_{\overrightarrow{ab}}(G) = \prod_{\overrightarrow{ab} \in E[H]} \sum_{\{u,v\} \in E[G]} \left(\mathcal{M}_{\overrightarrow{ab}}(u,v) + \mathcal{M}_{\overrightarrow{ab}}(v,u) \right) \ .$$

The term corresponding to $(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})$ in the expansion of $Z_H(G)$ is

$$\prod_{i=1}^{k} \mathcal{M}_{\overrightarrow{a_i b_i}}(u_i,v_i) = \prod_{i=1}^{k} X_{a_i}(u_i)\ X_{b_i}(v_i)\ Q^{\frac{Y(u_i)}{\deg_H(a_i)}}\ Q^{\frac{Y(v_i)}{\deg_H(b_i)}} \ , \tag{5}$$

where $\overrightarrow{a_i b_i}$ is the ith edge of H (where we assume any order) and $\overrightarrow{u_i v_i}$ is the ith edge in $\overrightarrow{T}$. We show that the expectation of (5) is non-zero if and only if the graph induced by $\overrightarrow{T}$ is an occurrence of H in G. Moreover, if the expectation of (5) is non-zero, then its value is a constant.

For a vertex w of G and a vertex c of H, let

$$\theta_{\overrightarrow{T}}(c,w) = \left|\left\{ i \mid (u_i = w \text{ and } a_i = c) \text{ or } (v_i = w \text{ and } b_i = c) \right\}\right|$$

be the number of edges in $\overrightarrow{T}$ with head (or tail) w mapping to the edges in H with head (or tail) c. Since every vertex c of H is incident to $\deg_H(c)$ edges, for any $c \in V[H]$ it holds that $\sum_{w \in V_T} \theta_{\overrightarrow{T}}(c,w) = \deg_H(c)$. By the definition of $\theta_{\overrightarrow{T}}$, we rewrite (5) as

$$\left(\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right)\cdot\left(\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T}}} Q^{\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right).$$

Therefore

$$\begin{aligned}&Z_H(G)\\ &=\sum_{\substack{e_1,\ldots,e_k\\ e_i\in E[G]}}\sum_{\overrightarrow{T}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})}\left(\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right)\cdot\left(\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T}}} Q^{\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right),\end{aligned}$$

where the first summation is over all k-tuples of edges in $E[G]$ and the second summation is over all their possible orientations. By linearity of expectations of these random variables and the assumption that $X_c(w)$ $(c \in V[H], w \in V[G]), Y$ and Q have sufficient independence, we have

$$\begin{aligned}&\mathbb{E}[Z_H(G)]\\ &=\mathbb{E}\left[\sum_{\substack{e_1,\ldots,e_k\\ e_i\in E[G]}}\sum_{\overrightarrow{T}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})}\left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T}}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right)\cdot\left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T}}}} Q^{\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right)\right]\\ &=\sum_{\substack{e_1,\ldots,e_k\\ e_i\in E[G]}}\sum_{\overrightarrow{T}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})}\left(\prod_{c\in V[H]}\mathbb{E}\left[\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right]\right)\cdot\mathbb{E}\left[\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T}}}} Q^{\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right].\end{aligned}$$

Let

$$\alpha_{\overrightarrow{T}} := \underbrace{\left(\prod_{c\in V[H]}\mathbb{E}\left[\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right]\right)}_{A}\cdot\underbrace{\mathbb{E}\left[\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T}}} Q^{\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right]}_{B}. \quad (6)$$

We will next show that $\alpha_{\overrightarrow{T}}$ is either zero or a nonzero constant independent of $\overrightarrow{T}$. The latter is the case if and only if G_T, the undirected graph induced from edge set $\overrightarrow{T}$, is an occurrence of H in G.

We consider the product A at first. Assume $A \neq 0$. Using the same technique as [12], we construct a homomorphism from H to $G_{\overrightarrow{T}}$ under the condition $A \neq 0$. Remember that: (i) For any $c \in V[H]$ and $w \in V_{\overrightarrow{T}}$, $\theta_{\overrightarrow{T}}(c,w) \leqslant \deg_H(c)$, and (ii) $\mathbb{E}\left[X_c^i(w)\right] \neq 0$ if and only if $i = \deg_H(c)$ or $i = 0$. Therefore for any

fixed $\overrightarrow{T}$ and $c \in V[H]$, $\mathbb{E}\left[\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right] \neq 0$ if and only if $\theta_{\overrightarrow{T}}(c,w) \in \{0, \deg_H(c)\}$ for all w. Now assume that $\mathbb{E}\left[\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right] \neq 0$ for every $c \in V[H]$. Then $\theta_{\overrightarrow{T}}(c,w) \in \{0, \deg_H(c)\}$ for all $c \in V[H]$ and $w \in V[G]$. Since $\sum_w \theta_{\overrightarrow{T}}(c,w) = \deg_H(c)$ for any $c \in V[H]$, there must be a unique vertex $w \in V_{\overrightarrow{T}}$ such that $\theta_{\overrightarrow{T}}(c,w) = \deg_H(c)$. Define $\varphi : V[H] \to V_{\overrightarrow{T}}$ as $\varphi(c) = w$ for the vertex w satisfying $\theta_{\overrightarrow{T}}(c,w) = \deg_H(c)$. Then φ is a homomorphism, i.e. $ab \in E[H]$ implies $\varphi(a)\varphi(b) \in E[G_{\overrightarrow{T}}]$. Hence $A \neq 0$ implies H is homomorphic to $G_{\overrightarrow{T}}$, and

$$\prod_{c\in V[H]} \mathbb{E}\left[\prod_{w\in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right] = \prod_{c\in V[H]} \mathbb{E}\left[X_c^{\deg_H(c)}(\varphi(c))\right] = 1 \ . \tag{7}$$

Second we consider the product B. Our task is to show that, under the condition $A \neq 0$, $G_{\overrightarrow{T}}$ is an occurrence of H if and only if $B \neq 0$. Observe that

$$\mathbb{E}\left[\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T}}} Q^{\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right] = \mathbb{E}\left[Q^{\sum_{c\in V[H]}\sum_{w\in V_{\overrightarrow{T}}}\frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}}\right] \ .$$

Case 1: Assume that $G_{\overrightarrow{T}}$ is an occurrence of H in G. Then $|V[H]| = |V_{\overrightarrow{T}}|$ and function φ constructed above is a bijection, which implies that

$$\begin{aligned}&\sum_{c\in V[H]}\sum_{w\in V_{\overrightarrow{T}}} \frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)}\\ &= \sum_{c\in V[H]} \frac{\theta_{\overrightarrow{T}}(c,\varphi(c))Y(\varphi(c))}{\deg_H(c)} = \sum_{c\in V[H]} Y(\varphi(c)) = \sum_{w\in V_{\overrightarrow{T}}} Y(w) \ .\end{aligned}$$

Without loss of generality, let $V_{\overrightarrow{T}} = \{w_1, \ldots, w_t\}$. By considering all possible choices for $Y(w_1), \ldots, Y(w_t)$, denoted by $y(w_1), \ldots, y(w_t) \in S$, and independence between Q and $Y(w), w \in V[G]$,

$$\begin{aligned}B =& \sum_{j=0}^{\tau-1} \sum_{y(w_1),\ldots,y(w_t)\in S} \frac{1}{\tau}\left(\prod_{i=1}^{t} \mathbf{Pr}\left[\,Y(w_i) = y(w_i)\,\right]\right)\cdot \exp\left(\frac{2\pi \mathrm{i} j}{\tau}\sum_{\ell=1}^{t} y(w_\ell)\right)\\ =& \sum_{j=0}^{\tau-1} \sum_{\substack{y(w_1),\ldots,y(w_t)\in S\\ \vartheta:=y(w_1)+\cdots+y(w_t),\tau|\vartheta}} \frac{1}{\tau}\left(\frac{1}{t}\right)^t \exp\left(\frac{2\pi\mathrm{i}}{\tau}\cdot\vartheta\cdot j\right) + \\ & \sum_{j=0}^{\tau-1} \sum_{\substack{y(w_1),\ldots,y(w_t)\in S\\ \vartheta:=y(w_1)+\cdots+y(w_t),\tau\nmid\vartheta}} \frac{1}{\tau}\left(\frac{1}{t}\right)^t \exp\left(\frac{2\pi\mathrm{i}}{\tau}\cdot\vartheta\cdot j\right) \ .\end{aligned}$$

Applying Lemma A.2 with $R = \exp\left(\frac{2\pi \mathrm{i}}{\tau}\right)$, the second summation is zero. Hence by Lemma A.3 we have

$$B = \sum_{\substack{y(w_1),\ldots,y(w_t)\in S \\ \tau \mid y(w_1)+\cdots+y(w_t)}} \left(\frac{1}{t}\right)^t = \sum_{\substack{y(w_1),\ldots,y(w_t)\in S \\ y(w_1)+\cdots+y(w_t)=\tau}} \left(\frac{1}{t}\right)^t = \left(\frac{1}{t}\right)^t \cdot t! = \frac{t!}{t^t} \ .$$

Case 2: Assume that $G_{\overrightarrow{T}}$ is not an occurrence of H in G and let $V_{\overrightarrow{T}} = \{w_1, \ldots, w_{t'}\}$, $t' < t$. Then there is a vertex $w \in V_{\overrightarrow{T}}$ and different $b, c \in V[H]$, such that $\varphi(b) = \varphi(c) = w$. As before we have

$$\sum_{c\in V[H]} \sum_{w\in V_{\overrightarrow{T}}} \frac{\theta_{\overrightarrow{T}}(c,w)Y(w)}{\deg_H(c)} = \sum_{c\in V[H]} Y(\varphi(c)) \ .$$

By Lemma A.3, $\tau \nmid \sum_{c\in V[H]} Y(\varphi(c))$ regardless of the choices of $Y(w_1), \ldots, Y(w_{t'})$. Hence

$$\begin{aligned} B =& \sum_{j=0}^{\tau-1} \sum_{\substack{y(w_1),\ldots,y(w_{t'})\in S \\ \vartheta := \sum_{c\in V[H]} y(\varphi(c))}} \frac{1}{\tau} \left(\frac{1}{t}\right)^{t'} \exp\left(\frac{2\pi \mathrm{i} j}{\tau} \cdot \vartheta\right) \\ =& \sum_{j=0}^{\tau-1} \sum_{\substack{y(w_1),\ldots,y(w_{t'})\in S \\ \vartheta := \sum_{c\in V[H]} y(\varphi(c))}} \frac{1}{\tau} \left(\frac{1}{t}\right)^{t'} \exp\left(\frac{2\pi \mathrm{i}}{\tau} \cdot \vartheta \cdot j\right) = 0 \ , \end{aligned}$$

where the last equality follows from Lemma A.2 with $R = \exp\left(\frac{2\pi \mathrm{i}}{\tau}\right)$.

Let $\mathbf{1}_{G_{\overrightarrow{T}} \equiv H}$ be the indicator expression that is one if $G_{\overrightarrow{T}}$ and H are isomorphic and zero otherwise. By the definition of graph automorphism and (7)

$$\mathbb{E}[Z_H(G)] = \sum_{\substack{e_1,\ldots,e_k \\ e_i\in E[G]}} \sum_{\overrightarrow{T}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})} \frac{t!}{t^t} \cdot \left(\mathbf{1}_{G_{\overrightarrow{T}} \equiv H}\right) = \frac{t! \cdot \mathrm{auto}(H)}{t^t} \cdot \#(H, G) \ . \qquad \square$$

We can use a similar technique to analyze the variance of $Z_H(G)$ and then use Chebyshev's inequality to upper bound the number of trials required for the desired approximation. Since $Z_H(G)$ is complex-valued, we need to upper bound $Z_H(G) \cdot \overline{Z_H(G)}$.

Lemma 3.2. *Let G be any graph with m edges, H be any graph with k edges for a constant k. Random variables $X_c(w), c \in V[H], w \in V[G]$ and Q are defined as above. Then the following statements hold:*

1. *If $\delta(H) \geqslant 2$, then $\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right] = O\left(m^k\right)$.*
2. *Let H be a connected graph with $k \geqslant 2$ edges and $\mathcal{H}$ be the set of all subgraphs H' in G with the following properties: (i) H' has $2k$ edges, and*

(ii) every connected component of H' contains at least two edges. Then $\mathbb{E}\left[Z_H(G)\cdot\overline{Z_H(G)}\right]=O\left(|\mathcal{H}|\right)$. *In particular,*

$$\mathbb{E}\left[Z_H(G)\cdot\overline{Z_H(G)}\right]=O\left(m^k\cdot\Delta(G)^k\right)\ .$$

By using Chebyshev's inequality, we can get a $(1\pm\varepsilon)$-approximation by running our estimator in parallel and returning the average of the output of these individual estimators. This leads to our main result for counting $\#(H,G)$.

Theorem 3.3. *Let G be any graph with m edges and H be any graph with $k=O(1)$ edges. For any constant $0<\varepsilon<1$, there is an algorithm to $(1\pm\varepsilon)$-approximate $\#(H,G)$ using (i) $O\big(\frac{1}{\varepsilon^2}\cdot\frac{m^k}{(\#H)^2}\cdot\log n\big)$ bits if $\delta(H)\geqslant 2$, and (ii) using $O\big(\frac{1}{\varepsilon^2}\cdot\frac{m^k\cdot(\Delta(G))^k}{(\#H)^2}\cdot\log n\big)$ bits for any H.*

Discussion. Statement (i) of Theorem 3.3 extends the main result of [12, Theorem 1] which requires H to be a cycle. Note that a naïve sampling-based approach would choose a random k-tuple of edges and require $m^k/(\#H)$ space. Theorem 3.3 improves upon this approach, in particular if the graph G is sparse and the number of occurrences of H is a growing function in n.

4 Extensions

We have developed a general framework for counting arbitrary subgraphs of constant size. Our algorithm is general in the sense that, for any fixed subgraph H, one only needs to (i) give vertices of H an arbitrary labeling, and edges of H an arbitrary orientation; (ii) choose hash functions according to H, and (iii) run Estimator 1 multiple times in parallel. A reasonable number of executions guarantees a $(1\pm\varepsilon)$-approximation.

For several typical applications we can further improve the space complexity by grouping the sketches or using certain properties of the underlying graph G. For the ease of the discussion we only focus on counting stars throughout the section.

Grouping Sketches. The space complexity in Theorem 3.3 relies on the number of edges that the sketch reads. To reduce the variance, a natural way is to use multiple copies of the sketches, and every sketch is only responsible for the updates of the edges from a certain subgraph.

To formulate this intuition, we partition $V=\{1,\ldots,n\}$ into $g:=n^{1-1/(2k)}$ subsets $\mathcal{V}_1,\ldots,\mathcal{V}_g$, and $\mathcal{V}_i:=\left\{j\colon(i-1)\cdot n^{1/(2k)}+1\leqslant j\leqslant i\cdot n^{1/(2k)}\right\}$. Without loss of generality we assume that $n^{1/(2k)}\in\mathbb{N}$. Associated with every $\mathcal{V}_i$, we maintain a sketch $\mathcal{C}_i$, whose description is essentially the same as Estimator 1 and can be found in the appendix. For every arriving edge $e=\{u,v\}$ in the stream, we update sketch $\mathcal{C}_i$ if $u\in\mathcal{V}_i$ or $v\in\mathcal{V}_i$. Since (i) the central vertex of every occurrence of S_k is in exactly one subset $\mathcal{V}_i$, and (ii) every edge adjacent to

one vertex in $\mathcal{V}_i$ is taken into account by sketch $\mathcal{C}_i$, we know that every occurrence of S_k in G is only counted by one sketch $\mathcal{C}_i$.

More formally, let $\widetilde{\#}(S_k, G|_{\mathcal{V}_i})$ be the number of S_k whose central vertex is in $\mathcal{V}_i$. Then it holds that $\#(S_k, G) = \sum_{i=1}^{g} \widetilde{\#}(S_k, G|_{\mathcal{V}_i})$. This indicates that if every $\mathcal{C}_i$ is unbiased for $\widetilde{\#}(S_k, G|_{\mathcal{V}_i})$, then we can use the sum of returned values of different $\mathcal{C}_i$ to approximate $\#(S_k, G)$. See Section B.1 for a detailed analysis.

Theorem 4.1. *Let G be a graph with n vertices. For any constants $0 < \varepsilon < 1$ and k, there is an algorithm to $(1 \pm \varepsilon)$-approximate $\#(S_k, G)$ with space complexity*

$$O\left(\frac{n^{1-1/(2k)}}{\varepsilon^2} \cdot \left(\frac{n^{3/2-1/(2k)} \cdot \Delta(G)^{2k}}{(\#S_k)^2} + 1\right) \cdot \log n\right).$$

To illustrate Theorem 4.1, let us consider graphs G with $\Delta(G)/\delta(G) = o(n^{1/(4k)})$ and $\delta(G) \geqslant k$. Since $\#(S_k, G) = \Omega\left(n \cdot \delta(G)^k\right)$, Theorem 4.1 implies that $o\left(\frac{1}{\varepsilon^2} \cdot n \cdot \log n\right)$ bits suffice to give a $(1 \pm \varepsilon)$-approximation.

Counting on Power Law Graphs. Besides organizing the sketches into groups, we will see now that the space complexity can be also reduced if we assume the underlying graph G to have certain properties. One of the most important properties shared by many biological, social or technological networks is the so-called *Power Law* degree distribution, i.e., the number of vertices with degree d, denoted by $f(d) := |\{v \in V \colon \deg(v) = d\}|$, satisfies $f(d) \sim d^{-\beta}$, where $\beta > 0$ is the power law exponent. For many technological networks, experimental studies indicate that β is between 2 and 3 (see [16]).

Formally, we use the following model from [18] based on the cumulative degree distribution. For given constants $\sigma \geqslant 1$ and $d_{\min} \in \mathbb{N}$, we say that G has an approximate power law degree distribution with exponent $\beta \in (2, 3)$, if

$$\sum_{d=k}^{n-1} f(d) \in \left[\lfloor \sigma^{-1} \cdot n \cdot k^{-\beta+1} \rfloor, \sigma \cdot n \cdot k^{-\beta+1}\right] \qquad \forall\ k \geqslant d_{\min}\ .$$

Hence, the degree distribution below the threshold $d_{\min}$ is arbitrary. Note that for this model, the maximum degree $\Delta(G)$ is of order $n^{1/(\beta-1)}$ (see also [8, 16] for some justification why this should be the right order of the maximum degree).

We now state our result for counting stars on power law graphs.

Theorem 4.2. *Assume that G has an approximate power law degree distribution with exponent $\beta \in (2, 3)$. Then, for any constants $0 < \varepsilon < 1$ and k, we can $(1 \pm \varepsilon)$-approximate $\#(S_k, G)$ using $O\left(\frac{1}{\varepsilon^2} \cdot \log n\right)$ bits.*

Bibliography

[1] N. Alon and J. Spencer. *The Probabilistic Method.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, 3rd edition, 2008.

[2] N. Alon, R. Yuster, and U. Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.
[3] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Reductions in streaming algorithms, with an application to counting triangles in graphs. In *Proc. 13th Symp. on Discrete Algorithms (SODA)*, pages 623–632, 2002.
[4] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *Proc. 14th Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, pages 16–24, 2008.
[5] I. Bordino, D. Donato, A. Gionis, and S. Leonardi. Mining large networks with subgraph counting. In *Proc. 8th Intl. Conf. on Data Mining (ICDM)*, pages 737–742, 2008.
[6] L. S. Buriol, G. Frahling, S. Leonardi, A. Marchetti-Spaccamela, and C. Sohler. Counting triangles in data streams. In *Proc. 25th Symp. Principles of Database Systems (PODS)*, pages 253–262, 2006.
[7] L. S. Buriol, G. Frahling, S. Leonardi, and C. Sohler. Estimating clustering indexes in data streams. In *Proc. 15th European Symposium on Algorithms (ESA)*, pages 618–632, 2007.
[8] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85:4626–4628, 2000.
[9] S. Ganguly. Estimating frequency moments of data streams using random linear combinations. In *Proc. 8th Intl. Workshop on Randomization and Comput. (RANDOM)*, pages 369–380, 2004.
[10] M. Gonen, D. Ron, and Y. Shavitt. Counting stars and other small subgraphs in sublinear-time. *SIAM J. Disc. Math.*, 25(3):1365–1411, 2011.
[11] H. Jowhari and M. Ghodsi. New streaming algorithms for counting triangles in graphs. In *Proc. 11th Intl. Conf. Computing and Combinatorics (COCOON)*, pages 710–716, 2005.
[12] M. Manjunath, K. Mehlhorn, K. Panagiotou, and H. Sun. Approximate counting of cycles in streams. In *Proc. 19th European Symposium on Algorithms (ESA)*, pages 677–688, 2011.
[13] A. McGregor. Open Problems in Data Streams and Related Topics, IITK Workshop on Algorithms For Data Sreams 2006. `http://www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf`.
[14] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298 (5594):824–827, 2002.
[15] S. Muthukrishnan. Data Streams: Algorithms and Applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.
[16] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
[17] R. Pagh and C. E. Tsourakakis. Colorful triangle counting and a mapreduce implementation. *Inf. Process. Lett.*, 112(7):277–281, 2012.
[18] R. van der Hofstad. Random Graphs and Complex Networks. `http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf`, 2012. Book in preparation.
[19] E. Wong, B. Baur, S. Quader, and C. Huang. Biological network motif detection: principles and practice. *Briefings in Bioinformatics*, pages 1–14, June 2011.

A Omitted Details from Section 3

In this section, we provide the omitted details from Section 3. In Section A.1, we list three auxiliary lemmas that are used to show that $Z_H(G)$ is an unbiased estimator. Section A.2 gives several upper bounds on the number of subgraphs, which are used to bound the variance in Section A.3.

A.1 Auxiliary Algebraic Tools

The following three lemmas are used to prove Theorem 3.1.

Lemma A.1 ([9]). *Let $X_c, c \in V[H]$ be a randomly chosen $\deg_H(c)$th root of unity. Then for any $1 < i \leqslant \deg_H(c)$, it holds that*

$$\mathbb{E}[X_c^i] = \begin{cases} 1, & i = \deg_H(c) \ , \\ 0, & 1 \leqslant i < \deg_H(c) \ . \end{cases}$$

In particular, $\mathbb{E}[X_c] = 1$ if $\deg_H(c) = 1$.

Lemma A.2. *Let R be a primitive τth root of unity and $k \in \mathbb{N}$. Then*

$$\sum_{\ell=0}^{\tau-1} (R^k)^\ell = \begin{cases} \tau, & \tau \mid k \ , \\ 0, & \tau \nmid k \ . \end{cases}$$

Proof. (1) If $\tau \mid k$, then $R^k = 1$ and $\sum_{\ell=0}^{\tau-1} (R^k)^\ell = \sum_{\ell=0}^{\tau-1} 1 = \tau$. (2) If $\tau \nmid k$, then $R^k \neq 1$ and

$$\sum_{\ell=0}^{\tau-1} (R^k)^\ell = \frac{1 - (R^k)^\tau}{1 - R^k} = \frac{1 - R^{k \cdot \tau}}{1 - R^k} = 0. \qquad \square$$

Lemma A.3. *Let $x_i \in \mathbb{Z}_{\geqslant 0}$ and $\sum_{i=0}^{t-1} x_i = t$. Then $2^t - 1 \mid \sum_{i=0}^{t-1} 2^i \cdot x_i$ if and only if $x_0 = \cdots = x_{t-1} = 1$.*

Proof. Let $x_0, \ldots, x_{t-1}$ be sequence so that (i) Not all $x_i = 0$, (ii) $2^t - 1 \mid \sum_{i=0}^{t-1} 2^i x_i$, and (iii) $\sum_{i=0}^{t-1} x_i$ is minimal among all sequences $\{x_i\}$ satisfying the conditions (i) and (ii).

It is easy to see that $x_i < 2$ for each i. This is because otherwise decreasing x_i by 2 and increasing x_{i+1} by 1 (or decreasing x_{t-1} by 2 and increasing x_0 by 1) preserves $\sum_{i=0}^{t-1} 2^i x_i \mod 2^t - 1$ and decreases $\sum_{i=0}^{t-1} x_i$. On the other hand, if $x_i \leqslant 1$ then $\sum_{i=0}^{t-1} 2^i x_i \leqslant 2^t - 1$ with equality iff $x_i = 1$ for all i. Thus $x_0 = \ldots = x_{t-1} = 1$ is the unique solution to $2^t - 1 \mid \sum_{i=0}^{t-1} 2^i x_i$ for $x_i \geqslant 0$ and $0 < \sum_{i=0}^{t-1} x_i \leqslant t$. $\square$

A.2 Bounding the Number of Subgraphs

In the following, we list four lemmas that all give upper bounds on the number of subgraphs with certain properties in G. These bounds are used later to analyze the variance of our unbiased estimator.

Lemma A.4 ([12]). *Let G be a graph with m edges and $\mathcal{H}$ be a set of subgraphs of G such that every $H \in \mathcal{H}$ has the following properties: (1) H has k edges, where k is a constant. (2) Each connected component of H is an Eulerian circuit. Then $|\mathcal{H}| = O(m^{k/2})$.*

Lemma A.5. *Let G be a graph with m edges and H be any fixed graph with k edges and $\delta(H) \geqslant 2$, where k is a constant. Then $\#(H,G) = O\left(m^{k/2}\right)$.*

Proof. Case 1: Every vertex in H has even degree. Assume that H has ℓ connected components $H_1, \ldots, H_\ell$ where each H_i has k_i edges. Note that $\#(H,G) \leqslant \prod_{i=1}^{\ell} \#(H_i,G)$. Since every H_i is Eulerian, by Lemma A.4 we have $\#(H_i,G) = O\left(m^{k_i/2}\right)$, and $\#(H,G) = \prod_{i=1}^{\ell} O\left(m^{k_i/2}\right) = O\left(m^{k/2}\right)$.

Case 2: There is a vertex in H with odd degree. We prove the statement by induction on $|E[H]|$. By the handshake lemma, at least two vertices must be of odd degree. Let $P = (v_0, \ldots, v_\ell)$ be a path of minimal length between vertices of odd degree in H. Thus $\deg(v_0), \deg(v_\ell)$ are odd, and $\deg(v_j)$ is even for all $1 \leqslant j \leqslant \ell - 1$. Let H' be the graph obtained from H by removing the edges in P and the vertices v_j of degree exactly 2 (which would upon removal of edges in P have degree 0). Then $\delta(H') \geqslant 2$. Thus by our inductive hypothesis, $\#(H',G) = O(m^{(k-\ell)/2})$. We split into two cases based upon whether ℓ is even or odd.

- If ℓ is even, then a homomorphism $H \to G$ is determined by a homomorphism $H' \to G$ along with the images of the edges $\{v_1,v_2\}, \{v_3,v_4\}, \ldots, \{v_{\ell-1},v_\ell\}$. Thus the number of such homomorphisms is at most $O(m^{(k-\ell)/2}) \cdot m^{\ell/2} = O(m^{k/2})$.
- If ℓ is odd, then a homomorphism $H \to G$ is determined by a homomorphism $H' \to G$ along with the images of the edges $\{v_1,v_2\}, \{v_3,v_4\}, \ldots, \{v_{\ell-2},v_{\ell-1}\}$. Hence

$$\#(H,G) \leqslant \#(H',G) \cdot m^{(\ell-1)/2} = O(m^{(k-\ell)/2}) \cdot m^{(\ell-1)/2} = O(m^{k/2}). \qquad \square$$

We now prove a general upper bound on the number of connected subgraphs with k edges in terms of the k-th moment of the degree distribution.

Lemma A.6. *Let G be any graph and H be an arbitrary connected subgraph of G with k edges (possibly with multiple edges), where k is a constant. Then $\#(H,G) = O\left(\sum_{u \in V} (\deg(u))^k\right)$.*

Proof. For any $u \in V[G]$, define $\mathcal{H}_u$ to be the set consisting of all subgraphs H' with k edges in G so that $u = \operatorname{argmax}_{w \in V[H']} \{\deg(w)\}$ (If there is more than one such vertex u, pick an arbitrary one). Since we can specify any $H \in \mathcal{H}_u$ by

(i) choosing all edges incident to u, (ii) choosing all edges incident to u, or the other endpoint of the edges chosen in (i), and so on, we have

$$|\mathcal{H}_u| \leqslant \prod_{i=1}^{k} (i \cdot \deg(u)) = O\left((\deg(u))^k\right).$$

Let H_1 be an occurrence of H in G and let $u = \operatorname{argmax}_{w \in V[H_1]} \{\deg(w)\}$. Note that $H_1 \in \mathcal{H}_u$. Therefore

$$\#(H, G) \leqslant \sum_{u \in V[G]} |\mathcal{H}_u| = O\left(\sum_{u \in V} (\deg(u))^k\right) \qquad \square$$

By partitioning the subgraph H into connected components, we easily obtain the following result.

Lemma A.7. *Let G be any graph with m edges and H be an arbitrary subgraph of G with k edges (possibly with multiple edges), where k is a constant and every connected component of H contains at least two edges. Then $\#(H, G) = O(m^{k/2}\Delta^{k/2})$.*

Proof. Assume w.l.o.g. that k is even. Partitioning all subgraphs with k edges into at most $k/2$ connected components with $\ell_1, \ldots, \ell_{k/2}$ edges each and applying Lemma A.6 to every component separately, we conclude that

$$\begin{aligned}
\#(H, G) &= \sum_{\substack{\ell_1,\ldots,\ell_{k/2} \in \{0,2,3,\ldots,k\} \\ \sum_{j=1}^{k/2} \ell_j = k}} \prod_{j\colon \ell_j \geqslant 2} O\left(\sum_{u \in V} \deg(u)^{\ell_j}\right) \\
&= \sum_{\substack{\ell_1,\ldots,\ell_{k/2} \in \{0,2,3,\ldots,k\} \\ \sum_{j=1}^{k} \ell_j = k}} \prod_{j\colon \ell_j \geqslant 2} O\left(\Delta^{\ell_j - 1} \cdot \sum_{u \in V} \deg(u)\right) \\
&= \sum_{\substack{\ell_1,\ldots,\ell_{k/2} \in \{0,2,3,\ldots,k\} \\ \sum_{j=1}^{k/2} \ell_j = k}} O\left(\Delta^{k - |\{j\colon \ell_j \geqslant 2\}|} \cdot m^{|\{j\colon \ell_j \geqslant 2\}|}\right) \\
&= \sum_{\substack{\ell_1,\ldots,\ell_{k/2} \in \{0,2,3,\ldots,k\} \\ \sum_{j=1}^{k/2} \ell_j = k}} O\left(\Delta^{k} \cdot (m/\Delta)^{|\{j\colon \ell_j \geqslant 2\}|}\right) \\
&= O\left(\Delta^{k/2} \cdot m^{k/2}\right). \qquad \square
\end{aligned}$$

A.3 Bounding the Variance

After the preparations in Section A.1 and Section A.2, we are now ready to prove Lemma 3.2 which provides two upper bounds on the variance of our estimator.

Lemma 3.2 (from page 9)**.** *Let G be any graph with m edges, H be any graph with k edges for a constant k. Random variables $X_c(w), c \in V[H], w \in V[G]$ and Q are defined as above. Then the following statements hold:*

1. *If* $\delta(H) \geqslant 2$, *then* $\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right] = O\left(m^k\right)$.
2. *Let* H *be a connected graph with* $k \geqslant 2$ *edges and* $\mathcal{H}$ *be the set of all subgraphs* H' *in* G *with the following properties: (i)* H' *has* $2k$ *edges, and (ii) every connected component of* H' *contains at least two edges. Then* $\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right] = O\left(|\mathcal{H}|\right)$. *In particular,*

$$\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right] = O\left(m^k \cdot \Delta(G)^k\right) \quad .$$

Proof. Note that there is a trivial bound $\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right] = O(m^{2k})$, since in the expansion of $Z_H(G) \cdot \overline{Z_H(G)}$, each term corresponds to a tuple of $2k$ edges with an arbitrary orientation and each such term is bounded by a constant. To obtain the two statements of the lemma, we have to bound the number of occurring subgraphs more carefully. We start to prove the first statement now.

As in our analysis for $\mathbb{E}[Z_H(G)]$, we analyze the subgraphs that contribute to $\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right]$. However, instead of considering subgraphs of G with k edges, we need to analyze the subgraphs in G with $2k$ edges. By the definition of $Z_H(G)$, we express $Z_H(G) \cdot \overline{Z_H(G)}$ as follows:

$$\begin{aligned}
&Z_H(G) \cdot \overline{Z_H(G)} \\
&= \left[\sum_{e_1,\ldots,e_k} \sum_{\overrightarrow{T_1}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})} \left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T_1}}}} X_c^{\theta_{\overrightarrow{T_1}}(c,w)}(w)\right) \cdot \left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T_1}}}} Q^{\frac{\theta_{\overrightarrow{T_1}}(c,w)\cdot Y(w)}{\deg_H(c)}}\right)\right] \cdot \\
&\qquad \left[\sum_{e'_1,\ldots,e'_k} \sum_{\overrightarrow{T_2}=(\overrightarrow{e'_1},\ldots,\overrightarrow{e'_k})} \left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T_2}}}} X_c^{-\theta_{\overrightarrow{T_2}}(c,w)}(w)\right) \cdot \left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T_2}}}} Q^{-\frac{\theta_{\overrightarrow{T_2}}(c,w)Y(w)}{\deg_H(c)}}\right)\right] \\
&= \sum_{e_1,\ldots,e_k} \sum_{\overrightarrow{T_1}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})} \sum_{e'_1,\ldots,e'_k} \sum_{\overrightarrow{T_2}=(\overrightarrow{e'_1},\ldots,\overrightarrow{e'_k})} \left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T_1}}\cup V_{\overrightarrow{T_2}}}} X_c^{\theta_{\overrightarrow{T_1}}(c,w)-\theta_{\overrightarrow{T_2}}(c,w)}(w)\right) \cdot \\
&\qquad \left(Q^{\sum_{c\in V_H}\left(\sum_{w\in V_{\overrightarrow{T_1}}}\theta_{\overrightarrow{T_1}}(c,w)Y(w)-\sum_{w\in V_{\overrightarrow{T_2}}}\theta_{\overrightarrow{T_2}}(c,w)Y(w)\right)\Big/\deg_H(c)}\right) \quad .
\end{aligned}$$

By the linearity of expectations and the condition that random variables $X_c(w)$ ($c \in V[H], w \in V[G]$) are $4k$-wise independent, and $X_c, c \in V[H], Q$ are chosen inde-

pendently, we have

$$\mathbb{E}\left[Z_H(G)\cdot\overline{Z_H(G)}\right]$$

$$=\sum_{e_1,\ldots,e_k}\sum_{\overrightarrow{T_1}=(\overrightarrow{e_1},\ldots,\overrightarrow{e_k})}\sum_{e'_1,\ldots,e'_k}\sum_{\overrightarrow{T_2}=(\overrightarrow{e'_1},\ldots,\overrightarrow{e'_k})}\left(\prod_{\substack{c\in V[H]\\ w\in V_{\overrightarrow{T_1}}\cup V_{\overrightarrow{T_2}}}}\mathbb{E}\left[X_c^{\theta_{\overrightarrow{T_1}}(c,w)-\theta_{\overrightarrow{T_2}}(c,w)}(w)\right]\right)\cdot$$

$$\mathbb{E}\left[Q^{\sum_{c\in V_H}\left(\sum_{w\in V_{\overrightarrow{T_1}}}\theta_{\overrightarrow{T_1}}(c,w)Y(w)-\sum_{w\in V_{\overrightarrow{T_2}}}\theta_{\overrightarrow{T_2}}(c,w)Y(w)\right)\Big/\deg_H(c)}\right].$$

Let us define

$$\alpha_{\overrightarrow{T_1},\overrightarrow{T_2}}:=\prod_{c\in V[H]}\prod_{w\in V_{\overrightarrow{T_1}}\cup V_{\overrightarrow{T_2}}}\mathbb{E}\left[X_c^{\theta_{\overrightarrow{T_1}}(c,w)-\theta_{\overrightarrow{T_2}}(c,w)}(w)\right].$$

Since

$$\mathbb{E}\left[Q^{\sum_{c\in V_H}\left(\sum_{w\in V_{\overrightarrow{T_1}}}\theta_{\overrightarrow{T_1}}(c,w)Y(w)-\sum_{w\in V_{\overrightarrow{T_2}}}\theta_{\overrightarrow{T_2}}(c,w)Y(w)\right)\Big/\deg_H(c)}\right]=O(1),$$

we only need to consider the number of graphs induced by $\overrightarrow{T_1},\overrightarrow{T_2}$ with the property $\alpha_{\overrightarrow{T_1},\overrightarrow{T_2}}\neq 0$. Remember that: (i) For any $c\in V_H$ and $w\in V_{\overrightarrow{V_1}\cup\overrightarrow{V_2}}$, $\mathbb{E}[X_c^i(w)]\neq 0$ if and only if i is divisible by $\deg_H(c)$, and (ii) for any $c\in V_H$ and $w\in V_{\overrightarrow{V_1}\cup\overrightarrow{V_2}}$, it holds that $0\leqslant\theta_{\overrightarrow{T_1}}(c,w)\leqslant\deg_H(c)$ and $0\leqslant\theta_{\overrightarrow{T_2}}(c,w)\leqslant\deg_H(c)$. Therefore $\alpha_{\overrightarrow{T_1},\overrightarrow{T_2}}\neq 0$ if and only if for every $c\in V[H]$ and every $w\in V_{\overrightarrow{V_1}\cup\overrightarrow{V_2}}$ one of the following three conditions holds: (i) $\theta_{\overrightarrow{T_1}}(c,w)=\deg_H(c)$ and $\theta_{\overrightarrow{T_2}}(c,w)=0$; (ii) $\theta_{\overrightarrow{T_1}}(c,w)=0$ and $\theta_{\overrightarrow{T_2}}(c,w)=\deg_H(c)$; (iii) $\theta_{\overrightarrow{T_1}}(c,w)=\theta_{\overrightarrow{T_2}}(c,w)$. We partition $V_{\overrightarrow{T_1}\cup\overrightarrow{T_2}}$ into three disjoint subsets A,B and C defined by

$$A:=V_{\overrightarrow{T_1}}\setminus V_{\overrightarrow{T_2}},\qquad B:=V_{\overrightarrow{T_2}}\setminus V_{\overrightarrow{T_1}},\qquad C:=V_{\overrightarrow{T_1}}\bigcap V_{\overrightarrow{T_2}}\ .$$

Sets A,B,C are defined corresponding to the above conditions (i), (ii) and (iii). By the assumption $\delta(H)\geqslant 2$, the degree of every vertex in sets A and B is at least two. By condition (iii), every vertex in C has nonzero and even degree, and hence the degree is also at least two. Hence the edges in $\overrightarrow{T_1}\bigcup\overrightarrow{T_2}$ form a homomorphic copy of some graph H' with $2k$ edges and $\delta(H')\geqslant 2$. Clearly there are at most $O(1)$ isomorphism classes of such graph H', and by Lemma A.5 each isomorphism class embeds in at most $O(m^k)$ ways. Therefore the number of such subgraphs in G is bounded by $O(m^k)$ and $\mathbb{E}\left[Z_H(G)\cdot\overline{Z_H(G)}\right]=O\left(m^k\right)$. This finishes the proof of the first statement.

We now show the second statement of the lemma. Fix $\overrightarrow{T_1}$ and $\overrightarrow{T_2}$ with $\alpha_{\overrightarrow{T_1},\overrightarrow{T_2}}\neq 0$. We first show that there is no connected component consisting of a single edge in $G_{\overrightarrow{T_1}\cup\overrightarrow{T_2}}$.

Assume for contradiction that there is one connected component in $G_{\overrightarrow{T_1} \cup \overrightarrow{T_2}}$ that consists of a single edge $e := \{u, v\}$. Since every vertex in set C has even degree, it holds $\{u, v\} \subseteq A \cup B$. By the proof of the first statement, there is no edge in $G_{\overrightarrow{T_1} \cup \overrightarrow{T_2}}$ connecting A and B directly. Therefore vertices u, v are in the same set. For simplicity assume $\{u, v\} \subseteq A$. Then there is a mapping between edges in $E[H]$ and edges in $G_{\overrightarrow{T_1}}$. Because H is connected, one of u, v has degree at least 2. Therefore there is another edge incident to edge e, which contradicts the assumption. Since $\alpha_{\overrightarrow{T_1}, \overrightarrow{T_2}} = O(1)$, $\mathbb{E}\left[Z_H(G) \cdot \overline{Z_H(G)}\right] = O(|\mathcal{H}|)$ by the definition of set $\mathcal{H}$.

Using this and Lemma A.7, we obtain the second statement. □

Theorem 3.3 (from page 10)**.** *Let G be any graph with m edges and H be any graph with $k = O(1)$ edges. For any constant $0 < \varepsilon < 1$, there is an algorithm to $(1 \pm \varepsilon)$-approximate $\#(H, G)$ using (i) $O\big(\frac{1}{\varepsilon^2} \cdot \frac{m^k}{(\#H)^2} \cdot \log n\big)$ bits if $\delta(H) \geqslant 2$, and (ii) using $O\big(\frac{1}{\varepsilon^2} \cdot \frac{m^k \cdot (\Delta(G))^k}{(\#H)^2} \cdot \log n\big)$ bits for any H.*

Proof. We run s parallel and independent copies of our estimator, and take the average value $Z^* = \frac{1}{s} \sum_{i=1}^{s} Z_i$, where each Z_i is the output of the i-th instance of the estimator. Therefore $\mathbb{E}[Z^*] = \mathbb{E}[Z_H(G)]$ and a straightforward calculation shows

$$\mathbb{E}[Z^* \overline{Z}^*] - |\mathbb{E}[Z^*]|^2 = \frac{1}{s} \left(\mathbb{E}[Z_H(G) \cdot \overline{Z_H(G)}] - |\mathbb{E}[Z_H(G)]|^2\right) .$$

By Chebyshev's inequality for complex-valued random variables (see, e.g., [12, Lemma 3]), we have

$$\mathbf{Pr}\left[\,|Z^* - \mathbb{E}[Z^*]| \geqslant \varepsilon \cdot |\mathbb{E}[Z^*]|\,\right] \leqslant \frac{\mathbb{E}[Z_H(G) \cdot \overline{Z_H(G)}] - \mathbb{E}[Z_H(G)] \cdot \overline{\mathbb{E}[Z_H(G)]}}{s \cdot \varepsilon^2 \cdot |\mathbb{E}[Z_H(G)]|^2} .$$

For Statement (i), by Lemma 3.2, Statement 1, we have

$$\mathbb{E}[Z_H(G) \cdot \overline{Z_H(G)}] - \mathbb{E}[Z_H(G)] \cdot \overline{\mathbb{E}[Z_H(G)]} \leqslant \mathbb{E}[Z_H(G) \cdot \overline{Z_H(G)}] = O(m^k) .$$

By choosing $s = O\left(\frac{1}{\varepsilon^2} \cdot \frac{m^k}{(\#H)^2}\right)$, we get $\mathbf{Pr}\left[\,|Z^* - \mathbb{E}\left[Z^*\right]| \geqslant \varepsilon \cdot |\mathbb{E}[Z^*]|\,\right] \leqslant 1/3$. Hence the overall space complexity is $O\left(\frac{1}{\varepsilon^2} \cdot \frac{m^k}{(\#H)^k} \cdot \log n\right)$. By using the same analysis and Lemma 3.2, Statement 2, we obtain (ii). □

B Omitted Details from Section 4

B.1 Grouping Sketches

This part gives the omitted details of counting stars by grouping sketches.

Lemma B.1 (Decomposition Lemma). *Let $\mathcal{V}_1, \ldots, \mathcal{V}_g$ be a partition of V. For any fixed $\mathcal{V}_i \subseteq V$, define $G|_{\mathcal{V}_i} = (V_i', E_i')$ as the subgraph of G such that: (i) $V_i' \subseteq \mathcal{V}_i \cup \partial\mathcal{V}_i$, where ∂S is the set of neighbors of vertices in S; (ii) For every $\{u, v\} \in E_i'$ iff $u \in \mathcal{V}_i$ or $v \in \mathcal{V}_i$. Then for any star S_k it holds that*

$$\#(S_k, G) = \sum_{i=1}^{g} \widetilde{\#}(S_k, G|_{\mathcal{V}_i}),$$

where $\widetilde{\#}(S_k, G|_{\mathcal{V}_i})$ is the number of stars in $G|_{\mathcal{V}_i}$ whose central point is in $\mathcal{V}_i$.

Proof. Assume that H is an occurrence of S_k in G and the central vertex of star H is u. Since $\mathcal{V}_1, \ldots, \mathcal{V}_g$ is a partition of V, vertex u is in exactly one set $\mathcal{V}_i$. By definition, H appears in $G|_{\mathcal{V}_i}$. Hence graph H is an occurrence of S_k in G iff H is an occurrence of S_k in exactly one $G|_{\mathcal{V}_i}$. By the definition of $\widetilde{\#}(S_k, G|_{\mathcal{V}_i})$, the statement follows. □

With the help of Lemma B.1, we revise our estimator as follows. In the initialization step, we have $g \times k$ random variables, where $g := n^{1-1/(2k)}$. Every k variables form a group $\mathcal{S}^{(i)}$ which is used for counting $\widetilde{\#}(H, G|_{\mathcal{V}_i})$. That is, we have g identical copies of the estimator and the i-th copy only uses random variables in group $\mathcal{S}^{(i)}$. Then for every coming edge $(u, v) \in E[G]$, we update the random variables of group $\mathcal{S}^{(i)}$ if $u \in \mathcal{V}_i$ or $v \in \mathcal{V}_i$. In the end, instead of computing $Z_{S_k}(G)$, we calculate $Z_{S_k}(G|_{\mathcal{V}_1}), \ldots, Z_{S_k}(G|_{\mathcal{V}_g})$ and output

$$Z_{S_k}(G) := \sum_{i=1}^{g} Z_{S_k}\left(G|_{\mathcal{V}_i}\right) \tag{8}$$

as the approximation of $\#(H, G)$. Estimator 2 presents a revised estimator for counting $\#(S_k, G|_{\mathcal{V}_i})$, where vertex a is the central vertex of the star.

Estimator 2 Counting $\#(S_k, G|_{\mathcal{V}_i})$, update procedure

---

Step 1 (Update): When an edge $e = \{u, v\} \in E[G]$ arrives, update each variable $Z_{\overrightarrow{ab}}$:
(a) If $u \in \mathcal{V}_i$ and $v \in \mathcal{V}_i$, then

$$Z_{\overrightarrow{ab}}(G) \leftarrow Z_{\overrightarrow{ab}}(G) + \mathcal{M}_{\overrightarrow{ab}}(u, v) + \mathcal{M}_{\overrightarrow{ab}}(v, u).$$

(b) If $u \in \mathcal{V}_i$ and $v \in \partial\mathcal{V}_i$, then

$$Z_{\overrightarrow{ab}}(G) \leftarrow Z_{\overrightarrow{ab}}(G) + \mathcal{M}_{\overrightarrow{ab}}(u, v).$$

(c) If $u \in \partial\mathcal{V}_i$ and $v \in \mathcal{V}_i$, then

$$Z_{\overrightarrow{ab}}(G) \leftarrow Z_{\overrightarrow{ab}}(G) + \mathcal{M}_{\overrightarrow{ab}}(v, u).$$

Lemma B.2. *Let H be an S_k for any constant k. Then $\mathbb{E}[Z_H(G)] = \#(H,G)$ and*

$$\mathbb{E}[Z_H(G)\cdot\overline{Z_H(G)}] = O\left(n^{2-1/(2k)}\cdot\Delta^{2k}\right) + \left(\#(H,G)\right)^2.$$

Proof. Since for any star S_k in G with central point w, S_k is only counted by exactly one $G|_{\mathcal{V}_i}$ with the property that $w\in\mathcal{V}_i$, it holds that

$$\#(H,G) = \sum_{i=1}^{g}\widetilde{\#}(H,G|_{\mathcal{V}_i}). \tag{9}$$

On the other hand, by linearity of expectations we know that

$$\mathbb{E}\left[Z_H(G)\right] = \sum_{i=1}^{g}\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\right]. \tag{10}$$

Combing (9), (10) and the fact that $\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\right] = \widetilde{\#}(H,G|_{\mathcal{V}_i})$ yields $\mathbb{E}\left[Z_H(G)\right] = \#(H,G)$. We bound $\mathbb{E}[Z_H(G)\cdot\overline{Z_H(G)}]$ as follows.

$$\begin{aligned}
&\mathbb{E}[Z_H(G)\cdot\overline{Z_H(G)}]\\
&= \mathbb{E}\left[\left(\sum_{i=1}^{g}Z_H\left(G|_{\mathcal{V}_i}\right)\right)\cdot\left(\sum_{i=1}^{g}\overline{Z_H\left(G|_{\mathcal{V}_i}\right)}\right)\right]\\
&= \mathbb{E}\left[\sum_{i=1}^{g}Z_H\left(G|_{\mathcal{V}_i}\right)\cdot\overline{Z_H\left(G|_{\mathcal{V}_i}\right)} + \sum_{1\leqslant i\neq j\leqslant g}Z_H\left(G|_{\mathcal{V}_i}\right)\cdot\overline{Z_H\left(G|_{\mathcal{V}_j}\right)}\right]\\
&= \sum_{i=1}^{g}\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\cdot\overline{Z_H\left(G|_{\mathcal{V}_i}\right)}\right] + \sum_{1\leqslant i\neq j\leqslant g}\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\cdot\overline{Z_H\left(G|_{\mathcal{V}_j}\right)}\right].
\end{aligned}$$

Note that every $G|_{\mathcal{V}_i}$ has at most $n^{1/(2k)}\cdot\Delta$ edges. Since $H = S_k$, we know by Lemma 3.2, Statement 2, that every graph contributing to $\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\cdot\overline{Z_H\left(G|_{\mathcal{V}_i}\right)}\right]$ only consist of connected components having at least two edges. Hence,

$$\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\cdot\overline{Z_H\left(G|_{\mathcal{V}_i}\right)}\right] = O\left(\left(n^{1/(2k)}\Delta\right)^k\cdot\Delta^k\right) = O\left(n^{1/2}\cdot\Delta^{2k}\right).$$

Therefore

$$\begin{aligned}
&\mathbb{E}[Z_H(G)\cdot\overline{Z_H(G)}]\\
&= O\left(n^{1-1/(2k)}\cdot n^{1/2}\cdot\Delta^{2k}\right) + \sum_{1\leqslant i\neq j\leqslant g}\mathbb{E}\left[Z_H\left(G|_{\mathcal{V}_i}\right)\right]\cdot\mathbb{E}\left[\overline{Z_H\left(G|_{\mathcal{V}_j}\right)}\right]\\
&= O\left(n^{1-1/(2k)}\cdot n^{1/2}\cdot\Delta^{2k}\right) + \left(\sum_{i=1}^{g}\#\left(H,G|_{\mathcal{V}_i}\right)\right)^2\\
&= O\left(n^{3/2-1/(2k)}\cdot\Delta^{2k}\right) + \left(\#(H,G)\right)^2,
\end{aligned}$$

which completes the proof. □

Theorem 4.1 (from page 11)**.** *Let G be a graph with n vertices. For any constants $0 < \varepsilon < 1$ and k, there is an algorithm to $(1 \pm \varepsilon)$-approximate $\#(S_k, G)$ with space complexity*

$$O\left(\frac{n^{1-1/(2k)}}{\varepsilon^2} \cdot \left(\frac{n^{3/2-1/(2k)} \cdot \Delta(G)^{2k}}{(\#S_k)^2} + 1\right) \cdot \log n\right).$$

Proof. Since the space requirement of a single execution of all g estimators is $O(g \cdot \log n) = O(n^{1-1/(2k)} \cdot \log n)$ bits, following the proof of Theorem 3.3 we know that in order to achieve $(1 \pm \varepsilon)$-approximation, we run $O\left(\frac{1}{\varepsilon^2} \cdot \left(\frac{n^{3/2-1/(2k)} \cdot \Delta^{2k}}{(\#S_k)^2} + 1\right)\right)$ independent executions of each estimator and the total space requirement is

$$O\left(\frac{n^{1-1/(2k)}}{\varepsilon^2} \cdot \left(\frac{n^{3/2-1/(2k)} \cdot \Delta^{2k}}{(\#S_k)^2} + 1\right) \cdot \log n\right).$$

□

B.2 Counting Stars on Power Law Graphs

We apply Lemma A.6 to graphs with approximate power law degree distribution (See Section 4 for the precise definition). As it turns out, the number of connected subgraphs with k edges is $\Theta(\Delta^k)$, which is matched for stars with k edges.

Lemma B.3. *Let G be a graph that has an approximate power law degree distribution with exponent $\beta \in (2, 3)$. Let H be any connected graph with $k \geqslant 2$ edges (some of which may be multi-edges). Then $\#(H, G) = O(n^{k/(\beta-1)})$. Moreover, if $H = S_k$, then $\#(H, G) = \Theta\left(n^{k/(\beta-1)}\right)$.*

Proof. We use Lemma A.6 and the approximate Power law degree distribution to obtain that

$$\begin{aligned}
\#(H, G) &= O\left(\sum_{u \in V} \left(\deg(u)^k\right)\right) \\
&= O\left(\sum_{d=1}^{d_{\min}} n \cdot (d_{\min})^k + \sum_{d=d_{\min}+1}^{\Delta} f(d) \cdot d^k\right) \\
&= O\left(n + \sum_{r=1}^{\lceil \log_2(\Delta/d_{\min}) \rceil} \sum_{d=d_{\min} \cdot 2^{r-1}+1}^{d_{\min} \cdot 2^r} f(d) \cdot d^k\right) \\
&= O\left(n + \sum_{r=1}^{\lceil \log_2(\Delta/d_{\min}) \rceil} (d_{\min} \cdot 2^r)^k \cdot \sum_{d=d_{\min} \cdot 2^{r-1}+1}^{\Delta} f(d)\right) \\
&= O\left(n + \sum_{r=1}^{\lceil \log_2(\Delta/d_{\min}) \rceil} (d_{\min} \cdot 2^r)^k \cdot n \cdot (d_{\min} \cdot 2^{r-1})^{-\beta+1}\right) \\
&= O\left(n + n \cdot \Delta^{k-(\beta-1)}\right) = O\left(n^{k/(\beta-1)}\right),
\end{aligned}$$

where in the last line we used our assumptions $k \geqslant 2, \beta \in (2,3)$ and the fact that $\Delta = \Theta(n^{1/(\beta-1)})$. Similarly, we can prove if $H = S_k$, then

$$\#(H,G) = \Omega\left(\Delta^k\right) = \Omega\left(n^{k/(\beta-1)}\right). \qquad \square$$

We point out that Lemma B.3 implies that we can get a constant factor approximation for $\#(S_k,G)$ by simply estimating the maximum degree of G. However, even if we knew the maximum degree of G *exactly*, then this would only yield an approximation of $\#(S_k,G)$ up to some large constant factor.

Based on Lemma B.3 and the results from Section 3, we can now easily prove that the unbiased estimator from Section 2 requires only logarithmic space for approximately counting the number of occurrences of S_k in power law graphs.

Theorem 4.2 (from page 11)**.** *Assume that G has an approximate power law degree distribution with exponent $\beta \in (2,3)$. Then, for any constants $0 < \varepsilon < 1$ and k, we can $(1 \pm \varepsilon)$-approximate $\#(S_k,G)$ using $O\left(\frac{1}{\varepsilon^2} \cdot \log n\right)$ bits.*

Proof. Let us consider $\mathbb{E}\left[\, Z_H(G) \cdot \overline{Z_H(G)} \,\right]$. If $H = S_k$, then by Lemma 3.2, this term is upper bounded by the number of subgraphs of G with $2k$ edges so that every connected component consists of at least two edges (possibly, multi-edges). By partitioning all graphs with that property into the at most k connected components and applying Lemma B.3 to each connected component with $\ell_i \geqslant 2$ edges separately, we conclude that

$$\mathbb{E}\left[\, Z_H(G) \cdot \overline{Z_H(G)} \,\right] = \sum_{\substack{\ell_1,\ell_2,\ldots,\ell_k \in \{0,2,3,\ldots,2k\} \\ \sum_{i=1}^{k} \ell_i = 2k}} \prod_{i\colon \ell_i \geqslant 2} O\left(n^{\ell_i/(\beta-1)}\right) = O\left(n^{2k/(\beta-1)}\right).$$

Since $\#(S_k,G) = \Omega\left(n^{k/(\beta-1)}\right)$, we obtain the result as in Theorem 3.3 by taking the average of $O\left(\frac{1}{\varepsilon^2}\right)$ independent executions and applying Chebyshev's inequality. $\square$