Approximate Counting of Cycles in Streams*

Madhusudan Manjunath[1], Kurt Mehlhorn[1],
Konstantinos Panagiotou[1], and He Sun[1,2]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
[2] Fudan University, Shanghai, China

Abstract. We consider the subgraph counting problem in data streams and develop the first non-trivial algorithm for approximately counting cycles of an arbitrary but fixed size. Previous non-trivial algorithms could only approximate the number of occurrences of subgraphs of size up to six. Our algorithm is based on the idea of computing instances of *complex-valued* random variables over the given stream and improves drastically upon the naïve sampling algorithm. In contrast to most existing approaches, our algorithm works in a distributed setting and for the turnstile model, i. e., the input stream is a sequence of edge insertions and deletions.

1 Introduction

Counting the number of occurrences of a graph H in a graph G has wide applications in uncovering important structural characteristics of the underlying network G, revealing information of the most frequent patterns, and so on. We are interested in the situation where G is very large. It is then natural to assume that G is given as a data stream, i.e., the edges of the graph G arrive consecutively and the algorithm uses only limited space to return an approximate value. Exact counting is not an option for massive input graphs. Already counting triangles exactly requires to store the entire graph.

Formally speaking, let $\mathcal{S} = s_1, s_2, \cdots, s_N$ be a stream that represents a graph $G = (V, E)$, where N is the length of the stream and each item s_i is associated with an edge in G. Typical models [12] in this topic include *the Cash Register Model* and *the Turnstile Model*. In the cash register model, each item s_i expresses one edge in G, and in the turnstile model each item s_i is represented by (e_i, sign_i) where e_i is an edge of G and $\text{sign}_i \in \{+, -\}$ indicates that e_i is inserted to or deleted from G. As a generalization of the cash register model, the turnstile model supports the dynamic insertions and deletions of the edges.

In a distributed setting the stream $\mathcal{S}$ is partitioned into sub-streams $S_1, \ldots, S_t$ and each S_i is fed to a different processor. At the end of the computation, the processors collectively estimate the number of occurrences of H with a small amount of communication.

* The third author was supported by the Alexander von Humboldt-Foundation.

C. Demetrescu and M.M. Halldórsson (Eds.): ESA 2011, LNCS 6942, pp. 677–688, 2011.

Our Results. We present a general framework for counting cycles of arbitrary size in a massive graph. Our algorithm runs in the turnstile model and the distributed setting, and for any constants $0 < \varepsilon, \delta < 1$, our algorithm achieves an (ε, δ)-approximation, i. e. , the returning value Z of the algorithm and the exact value $Z^* = \#C_k$, the number of occurrences of C_k, satisfy $\Pr[|Z - Z^*| > \varepsilon \cdot Z^*] < \delta$. We also provide an unbiased estimator for general d-regular graphs. This considerably extends the class of graphs that can be counted in the data streaming model and answers partially an open problem proposed by many references, see for example the extensive survey by Muthukrishnan [12] and the 11th open question in the 2006 IITK Workshop on Algorithms for Data Streams [11].

Because the problem of counting the number of cycles of length k, parameterized by k, is $\#\mathbf{W}[1]$-complete [7], our result demonstrates that efficient approximations for $\#\mathbf{W}[1]$-complete problems are possible under certain conditions, even if only a restricted amount of space can be used.

Besides that, we initiate the study of complex-valued hash functions in counting subgraphs. Complex-valued estimators have been successfully applied in other contexts such as approximating the permanent, see [6,10]. In the data streaming setting, Ganguly [8] used a complex-valued sketch to estimate frequency moments. Our main result is as follows:

Theorem 1. *Let G be a graph with n vertices and m edges. For any k, there is an algorithm using S bits of space to (ε, δ)-approximate the number of occurrences of C_k in G provided that $S = \Omega\big(\frac{1}{\varepsilon^2} \cdot \frac{m^k}{(\#C_k)^2} \cdot \log n \cdot \log \frac{1}{\delta}\big)$. The algorithm works in the turnstile model.*

Discussion: A naïve approach for counting the number of occurrences of a k-cycle would either sample independently k vertices (if possible) or k edges from the stream. Since the probability of k vertices (or k edges) forming a cycle is $\#C_k/n^k$ (or $\#C_k/m^k$), this approach needs space $\Omega\left(\frac{n^k \log n}{\#C_k}\right)$ and $\Omega\left(\frac{m^k \log n}{\#C_k}\right)$, respectively. Thus, our algorithm improves upon these two approaches, especially for sparse graphs with many k-cycles, and has the additional benefit that it is applicable in the turnstile model and the distributed setting. Moreover, note that our bound is essentially tight, as there are graphs where the space complexity of the algorithm is $O(\log n)$; consider for example the "extremal graph" with a clique on $\Theta(\sqrt{m})$ vertices, where all other vertices are isolated. Moreover, as a corollary of Theorem 1, when the number of occurrences of C_k is $\Omega\left(m^{k/2-\alpha}\right)$ for $0 \leq \alpha < 1/2$, our algorithm with sub-linear space $O\left(\frac{1}{\varepsilon^2} \cdot m^{2\alpha} \cdot \log \frac{1}{\delta}\right)$ suffices to give a good approximation.

Related Work: Counting subgraphs in a data stream was first considered in a seminal paper by Bar-Yossef, Kumar, and Sivakumar [1]. There, the triangle counting problem was reduced to the problem of computing frequency moments. After that, several algorithms for counting triangles have been proposed [2,4,9].

Jowhari and Ghodsi presented three algorithms in [9], one of which is applicable in the turnstile model. Moreover, the problem of counting subgraphs different from triangles has also been investigated in the literature. Bordino, Donato, Gionis, and Leonardi [3] extended the technique of counting triangles [4] to

all subgraphs on three and four vertices. Buriol, Fahling, Leonardi and Sohler [5] presented a streaming algorithm for counting $K_{3,3}$, the complete bipartite graph with three vertices in each part. However, except the one presented in [9], most algorithms are based on sampling techniques and do not apply to the turnstile model.

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 $V[G] = \{1, \cdots, n\}$ and n is known in advance.

Given two directed graphs H_1 and H_2, we say that H_1 and H_2 are *homomorphic* if there is a mapping $i : V[H_1] \to V[H_2]$ such that $(u, v) \in E[H_1]$ if and only if $(i(u), i(v)) \in E[H_2]$. Furthermore, H_1 and H_2 are said to be *isomorphic* if the mapping i is a bijection.

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. We use $\#(H, G)$ to denote the number of occurrences of H in G. When G is the input graph, for simplicity we use $\#H$ to express $\#(H, G)$. Moreover, let C_ℓ be a cycle on ℓ edges.

Organization. Section 2 reviews Jowhari and Ghodsi's algorithm for counting triangles in streams. We generalize Jowhai and Ghodsi's approach in Sect. 3 and get an unbiased estimator for general d-regular graphs. Section 4 discusses the space complexity for counting cycles with arbitrary size. We end this paper with some open problems in Sect. 5.

2 A Review of Jowhari and Ghodsi's Algorithm

We give a brief account of Jowhari and Ghodsi's algorithm [9] in order to prepare the reader for our extension of their approach. Jowhari and Ghodsi estimate the number of triangles in a graph G. Let X be a $\{-1, +1\}$-valued random variable with expectation zero. They associate with every vertex w of G an instance $X(w)$ of X; the $X(w)$'s are 6-wise independent. They compute $Z = \sum_{\{u,v\} \in E[G]} X(u)X(v)$ and output $Z^3/6$ as the estimator for $\#C_3$.

Lemma 1 ([9]). $\mathbb{E}[Z^3] = 6 \cdot \#C_3$.

Proof. For any triple $T \in E^3[G]$ of edges and any vertex w of G, let $\deg_T(w)$ be the number of edges in T incident to w, then $\deg_T(w)$ is an integer no larger than 3. Also

$$\begin{aligned}\mathbb{E}[Z^3] &= \mathbb{E}\left[\left(\sum_{\{u,v\} \in E[G]} X(u)X(v)\right)^3\right] \\ &= \mathbb{E}\left[\sum_{T=(\{u_1,v_1\},\{u_2,v_2\},\{u_3,v_3\}) \in E^3} X(u_1)X(v_1)X(u_2)X(v_2)X(u_3)X(v_3)\right]\end{aligned}$$

Let V_T be the set of vertices that are incident to the edges in T. Then

$$\mathbb{E}[Z^3] = \mathbb{E}\left[\sum_{T\in E^3}\prod_{w\in V_T} X(w)^{\deg_T(w)}\right]$$

By the 6-wise independence of the $X(w), w \in V$, we have

$$\mathbb{E}[Z^3] = \sum_{T\in E^3}\prod_{w\in V_T}\mathbb{E}\left[X(w)^{\deg_T(w)}\right] = \sum_{T\in E^3}\prod_{w\in V_T}\mathbb{E}\left[X^{\deg_T(w)}\right]$$

Since $\mathbb{E}\left[X^{\deg_T(w)}\right] = 1$ if $\deg_T(w)$ is even and $\mathbb{E}\left[X^{\deg_T(w)}\right] = 0$ if $\deg_T(w)$ is odd, we know that $\prod_{w\in V_T}\mathbb{E}\left[X^{\deg_T(w)}\right] = 1$ if and only if the edges in T form a triangle. Since each triangle is counted six times, we have $\mathbb{E}[Z^3] = 6 \cdot \#C_3$. □

The crucial ingredients of the proof are (1) 6-wise independence guarantees that the expectation-operator can be pulled inside, and (2) random variable X is defined such that only vertices with even degree in T have nonzero expectation.

3 Algorithm Framework

We now generalize the algorithm in Section 2 and present an algorithm framework for counting general d-regular graphs. Suppose that H is a d-regular graph with k edges and we want to count the number of occurrences of H in G. The vertices of H are expressed by a, b and c, etc., and the vertices of G are expressed by u, v and w, etc., respectively. We will 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 express such an oriented graph.

For each oriented edge $\overrightarrow{ab}$ in H our algorithm maintains a complex-valued variable $Z_{\overrightarrow{ab}}(G)$, which is initialized to zero. The variables are defined in terms of random variables $Y(w)$ and $X_c(w)$, where c is a node of H and w is a node of G. The random variables $Y(w)$ are instances of a random variable Y and the random variables $X_c(w)$ are instances of a random variable X. The range of both random variables is a finite subset of complex numbers. We will realize the random variables by hash functions from $V[G]$ to $\mathbb{C}$; this explains why we indicate the dependence on w by functional brackets. We assume that the variables $X_c(w)$ and $Y(w)$ have sufficient independence as detailed below.

Our algorithm performs two basic steps: First, when an edge $e = \{u, v\} \in E[G]$ arrives, we update each variable $Z_{\overrightarrow{ab}}$ according to

$$Z_{\overrightarrow{ab}}(G) \leftarrow Z_{\overrightarrow{ab}}(G) + \big(X_a(u)\cdot X_b(v) + X_b(u)\cdot X_a(v)\big)\cdot Y(u)\cdot Y(v). \quad (1)$$

Second, when the number of occurrences of a graph H is required, the algorithm returns the real part of $Z/(\alpha \cdot \mathrm{aut}(H))$, where Z is defined via

$$Z := Z_H(G) = \prod_{\overrightarrow{ab}\in E[H]} Z_{\overrightarrow{ab}}(G), \quad (2)$$

α and $\mathrm{aut}(H)$ are constant numbers for any given H and will be determined later.

Remark 1. For simplicity, the algorithm above is only for the edge-insertion case. An edge deletion amounts to replacing '+' by '−' in (1).

Remark 2. The first step may be carried out in a distributed fashion, i. e., we have several processors each processing a subset of edges. In the second step the counts of the different processors are combined.

Theorem 2. *Let H be a d-regular graph with k edges. Let us assume that the random variables defined above satisfy the following two properties:*

1. *The random variables $X_c(w)$ and $Y(w)$, where $c \in V[H]$ and $w \in V[G]$, are instances of random variables X and Y, respectively. The random variables are $4k$-wise independent.*
2. *Let Z be any one of $X_c, c \in V[H]$ or Y. Then for any $1 \le i \le 2k$, $\mathbb{E}\left[Z^i\right] \neq 0$ if and only if $i = d$.*

Then $\mathbb{E}[Z_H(G)] = \alpha \cdot \mathrm{aut}(H) \cdot \#(H, G)$, where $\alpha = \left(\mathbb{E}\left[X^d\right]\mathbb{E}\left[Y^d\right]\right)^{2k/d} \in \mathbb{C}$ and $\mathrm{aut}(H)$ is the number of permutations and orientations of the edges in H such that the resulting graph is isomorphic to H.

The theorem above shows that $Z_H(G)$ is an unbiased estimator for any d-regular graph H, assuming that there exist random variables $X_c(w)$ and $Y(w)$ with certain properties. We will prove Theorem 2 at first, and then construct such random variables.

Proof (of Theorem 2). We first introduce some notations. For a k-tuple $T = (e_1, \ldots, e_k) \in E^k[G]$, let $G_T = (V_T, E_T)$ be the induced multi-graph, i.e., G_T has edge multi-set $E_T = \{e_1, \ldots, e_k\}$. By definition, we have

$$\begin{aligned} Z_H(G) &= \prod_{\overrightarrow{ab} \in E[H]} Z_{\overrightarrow{ab}}(G) \\ &= \prod_{\overrightarrow{ab} \in E[H]} \left(\sum_{\{u,v\} \in E[G]} (X_a(u) \cdot X_b(v) + X_a(v) \cdot X_b(u)) \cdot Y(u) \cdot Y(v) \right). \end{aligned}$$

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. Thus, in the expansion of $Z_H(G)$, any k-tuple $(e_1, \cdots, e_k) \in E^k[G]$ contributes 2^k different terms to $Z_H(G)$ and each term corresponds to a certain orientation of $(e_1, \cdots, e_k)$. Let $\overrightarrow{T} = (\overrightarrow{e_1}, \cdots, \overrightarrow{e_k})$ be an arbitrary orientation of $(e_1, \cdots, e_k)$, where $\overrightarrow{e_i} = \overrightarrow{u_i v_i}$. So the term in $Z_H(G)$ corresponding to $(\overrightarrow{e_1}, \cdots, \overrightarrow{e_k})$ is

$$\prod_{i=1}^{k} X_{a_i}(u_i) \cdot X_{b_i}(v_i) \cdot Y(u_i) \cdot Y(v_i) \ , \tag{3}$$

where (a_i, b_i) is the i-th edge of H and $\overrightarrow{u_i v_i}$ is the i-th edge in $\overrightarrow{T}$. We show that (3) is non-zero if and only if the graph induced by $\overrightarrow{T}$ is isomorphic to H (i. e. it also preserves the orientations of the edges).

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

$$\theta_{\overrightarrow{T}}(c, w) = \left|\left\{i \mid (u_i = w \text{ and } a_i = c) \text{ or } (v_i = w \text{ and } b_i = c)\right\}\right| \ . \tag{4}$$

Thus for any $c \in V[H]$, $\sum_{w \in V_T} \theta_{\overrightarrow{T}}(c, w) = d$ since every vertex c of H appears in exactly d edges (a_i, b_i); recall that H is d-regular. Using the definition of $\theta_{\overrightarrow{T}}$, we may rewrite (3) as

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

where $\deg_{\overrightarrow{T}}(w)$ is the number of edges in $\overrightarrow{T}$ incident to w. Therefore

$$Z_H(G)$$
$$= \sum_{\substack{e_1, \cdots, e_k \\ e_i \in E[G]}} \sum_{\overrightarrow{T} = (\overrightarrow{e_1}, \cdots, \overrightarrow{e_k})} \left(\prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right) \cdot \left(\prod_{w \in V_{\overrightarrow{T}}} Y^{\deg_{\overrightarrow{T}}(w)}(w)\right),$$

where the first summation is over all the k-tuples of edges in $E[G]$ and the second summation is over all their possible orientations. Since each term of Z_H is the product of $4k$ random variables, which by assumption are $4k$-wise independent, we infer by linearity of expectation that

$$\mathbb{E}[Z_H(G)]$$
$$= \mathbb{E}\left[\sum_{\substack{e_1, \cdots, e_k \\ e_i \in E[G]}} \sum_{\overrightarrow{T} = (\overrightarrow{e_1}, \cdots, \overrightarrow{e_k})} \left(\prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T}}} X_c^{\theta_{\overrightarrow{T}}(c,w)}(w)\right) \cdot \left(\prod_{w \in V_{\overrightarrow{T}}} Y^{\deg_{\overrightarrow{T}}(w)}(w)\right)\right]$$
$$= \sum_{\substack{e_1, \cdots, e_k \\ e_i \in E[G]}} \sum_{\overrightarrow{T} = (\overrightarrow{e_1}, \cdots, \overrightarrow{e_k})} \prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}\left[X^{\theta_{\overrightarrow{T}}(c,w)}\right] \cdot \prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}\left[Y^{\deg_{\overrightarrow{T}}(w)}\right] \ .$$

Let

$$\alpha(\overrightarrow{T}) := \prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}\left[X^{\theta_{\overrightarrow{T}}(c,w)}\right] \cdot \prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}\left[Y^{\deg_{\overrightarrow{T}}(w)}\right] \ .$$

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 is an occurrence of H in G.

We have $\mathbb{E}\left[X^i\right] \neq 0$ if and only if $i = d$ or $i = 0$. Therefore for any $\overrightarrow{T}$ and $c \in V[H]$, $\prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}[X^{\theta_{\overrightarrow{T}}(c,w)}] \neq 0$ if and only if $\theta_{\overrightarrow{T}}(c, w) \in \{0, d\}$ for all w. Since $\sum_w \theta_{\overrightarrow{T}}(c, w) = \deg_H(c) = d$, there must be a unique vertex $w \in V_{\overrightarrow{T}}$ such that

$\theta_{\overrightarrow{T}}(c,w) = d$. Define $\varphi : V[H] \to V_{\overrightarrow{T}}$ as $\varphi(c) = w$. Then φ is a homomorphism and

$$\prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}\left[X^{\theta_{\overrightarrow{T}}(c,w)}\right] = \prod_{c \in V[H]} \mathbb{E}\left[X^d\right] = \mathbb{E}\left[X^d\right]^{|V[H]|}.$$

Since $\mathbb{E}[Y^i] \neq 0$ if and only if $i = d$ or $i = 0$, so for any $\overrightarrow{T}$, $\prod_{w \in V_{\overrightarrow{T}}} \mathbb{E}[Y^{\deg_{\overrightarrow{T}}(w)}] \neq 0$ if and only if every vertex $w \in V_{\overrightarrow{T}}$ has degree d in the graph with edge set T. Thus $|V_{\overrightarrow{T}}| = 2k/d = |V[H]|$, which implies that φ is an isomorphism mapping.

We have now shown that $\alpha(\overrightarrow{T})$ is either zero or the nonzero constant

$$\alpha = \left(\mathbb{E}\left[X^d\right]\mathbb{E}\left[Y^d\right]\right)^{2k/d}.$$

The latter is the case if and only if $G_{\overrightarrow{T}}$ is an occurrence of H in G. Let $(G_{\overrightarrow{T}} \equiv H)$ be the indicator expression that is one if $G_{\overrightarrow{T}}$ and H are isomorphic and zero otherwise. Then

$$\mathbb{E}[Z_H(G)] = \sum_{\substack{e_1,\cdots,e_k \\ e_i \in E[G]}} \sum_{\overrightarrow{T}=(\overrightarrow{e_1},\cdots,\overrightarrow{e_k})} \alpha(\overrightarrow{T}) \cdot \left(G_{\overrightarrow{T}} \equiv H\right) = \alpha \cdot \mathrm{aut}(H) \cdot \#(H,G) \ .$$

□

For the case of cycles, we have $\mathrm{aut}(H) = 2k$. We turn to construct hash functions needed in Theorem 2. The basic idea is to choose a $8k$-wise independent hash function $h : D \to \mathbb{C}$ and map the values in D to complex numbers with certain properties. We first show a simple lemma about roots of polynomials of a simple form.

Lemma 2. *For positive interger r, let $P_r(z) = 2 + z^r$ and $z_j = 2^{1/j} \cdot \mathrm{e}^{\frac{\pi i}{j}}$. The complex number z_j is a root of the polynomial $P_r(z)$ if and only if $j = r$.*

Proof. We first verify that z_r is a root of the polynomial $P_r(z)$: since $z_r^r = 2 \cdot \mathrm{e}^{\pi \cdot i} = -2$, we have $z_r^r + 2 = 0$. To show the converse, we consider z_j^r for $r \neq j$ and verify that $|z_j^r| = \left|2^{r/j}\mathrm{e}^{\frac{\pi \cdot j \cdot r}{j}}\right| = 2^{r/j}$. Since $2^{r/j} \neq 2$ if $j \neq r$, the claim follows. □

Let z_j as in Lemma 2 and define random variable H_j as

$$H_j = \begin{cases} 1, & \text{with probability } 2/3, \\ z_j, & \text{with probability } 1/3. \end{cases} \tag{5}$$

Then $\mathbb{E}[H_j^\ell] = \left(2 + z_j^\ell\right)/3 = P_\ell(z_j)/3$ which is nonzero if $j \neq \ell$.

Theorem 3. *For positive integers d and k, let*

$$H = \prod_{1 \leq j \leq 2k, j \neq d} H_j$$

where the H_j are independent. For all integers ℓ between 1 and $2k$, $\mathbb{E}[H^\ell] \neq 0$ if and only if $d = \ell$.

Proof. By independence, $\mathbb{E}[H^\ell] = \prod_{1 \leq j \leq 2k, j \neq d} \mathbb{E}[H_j^\ell]$. This product is nonzero if ℓ is different from all j that are distinct from d, i. e., $\ell = d$. □

4 Proof of the Main Theorem

Now we bound the space of the algorithm for the case of cycles of arbitrary length. The basic idea is to use the second moment method on the complex-valued random variable Z. We first note a couple of lemmas that turn out to be useful: the first lemma is a generalization of Chebyshev's inequality for a complex-valued random variable and the second lemma is an upper bound on the number of closed walks of a given length in terms of the number of edges of the graph. Recall that the conjugate of a complex number $z = a + ib$ is denoted by $\overline{z} := a - ib$.

Lemma 3. *Let X be a complex-valued random variable with finite support and let $t > 0$. We have that*

$$\Pr[|X - \mathbb{E}[X]| \geq t \cdot |\mathbb{E}[X]|] \leq \frac{\mathbb{E}[X\overline{X}] - \mathbb{E}[X]\,\overline{\mathbb{E}[X]}}{t^2 |\mathbb{E}[X]|^2} \ .$$

Proof. Since $|X - \mathbb{E}[X]|^2 = (X - \mathbb{E}[X])\overline{(X - \mathbb{E}[X])}$ is a positive-valued random variable, we apply Markov's inequality to obtain

$$\begin{aligned}\Pr\left[|X - \mathbb{E}[X]| \geq t \cdot |\mathbb{E}[X]|\right] &= \Pr\left[|X - \mathbb{E}[X]|^2 \geq t^2 \cdot |\mathbb{E}[X]|^2\right] \\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])\overline{(X - \mathbb{E}[X])}]}{t^2 |\mathbb{E}[X]|^2} \ .\end{aligned}$$

Expanding $\mathbb{E}[(X - \mathbb{E}[X])\overline{(X - \mathbb{E}[X])}]$ we obtain that

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X])\overline{(X - \mathbb{E}[X])}] &= \mathbb{E}[X\overline{X}] - \mathbb{E}[X\overline{\mathbb{E}[X]}] - \mathbb{E}[\overline{X}\mathbb{E}[X]] + \overline{\mathbb{E}[X]}\mathbb{E}[X] \\ &= \mathbb{E}\left[X\overline{X}\right] - \mathbb{E}[X]\overline{\mathbb{E}[X]} \ .\end{aligned}$$

The last equality uses the linearity of expectation and that $\mathbb{E}[\overline{X}] = \overline{\mathbb{E}[X]}$. □

We now show an upper bound on the number of closed walks of a given length in a graph. This upper bound will control the space requirement of the algorithm.

Lemma 4. *Let G be an undirected graph with n vertices and m edges. Then the number of closed walks W_k with length k in G is at most $\frac{2^{k/2-1}}{k} \cdot m^{k/2}$.*

Proof. Let A be the adjacency matrix of G with eigenvalues $\lambda_1, \cdots, \lambda_n$. Since G is undirected, A is real symmetric and each eigenvalue λ_i is a real number. Then $W_k = \frac{1}{2k} \cdot \sum_{i=1}^{n} (A^k)_{ii}$ where for a matrix M, M_{ij} is the ij-th entry of the matrix. Because $\sum_{i=1}^{n} (A^k)_{ii} = \mathrm{tr}(A^k) = \sum_{i=1}^{n} \lambda_i^k \leq \sum_{i=1}^{n} |\lambda_i|^k$ and $\left(\sum_{i=1}^{n} |\lambda_i|^k\right)^{1/k} \leq \left(\sum_{i=1}^{n} |\lambda_i|^2\right)^{1/2} = (2m)^{1/2}$ for any $k \geq 2$, we have $W_k \leq \frac{1}{2k} \cdot \left(\sum_{i=1}^{n} |\lambda_i|^2\right)^{k/2} = \frac{2^{k/2-1}}{k} \cdot m^{k/2}$. □

Corollary 1. *Let G be a graph on m edges and $\mathcal{H}$ be a set of subgraphs of G such that every $H \in \mathcal{H}$ has properties: (1) H has k edges, where k is a constant. (2) Each connected-component of H is an Eulerian circuit. Then $|\mathcal{H}| = O(m^{k/2})$.*

Proof. Fix an integer $r \in \{1, \ldots, k\}$ and consider graphs in $\mathcal{H}$ that have r connected components. By Lemma 4, the number of such graphs is at most

$$\sum_{\substack{k_1,\cdots,k_r \\ k_1+\cdots+k_r=k}} \prod_{i=1}^{r} W_{k_i} \leq \sum_{\substack{k_1,\cdots,k_r \\ k_1+\cdots+k_r=k}} \prod_{i=1}^{r} \frac{2^{k_i/2-1} \cdot m^{k_i/2}}{k_i} \leq f(k) \cdot (2m)^{k/2},$$

where $f(k)$ is a function of k. Because there are at most k choices of r, we have $|\mathcal{H}| = O(m^{k/2})$. □

Observe that the expansion of $\mathbb{E}[Z_H(G)\overline{Z_H(G)}]$ consists of m^{2k} terms and the modulus of each term is upper bounded by a constant. So a naïve upper bound for $\mathbb{E}[Z_H(G)\overline{Z_H(G)}]$ is $O(m^{2k})$. Now we only focus on the case of cycles and use the "cancellation" properties of the random variables to get a better bound for $\mathbb{E}[Z_H(G)\overline{Z_H(G)}]$.

Theorem 4. *Let H be a cycle C_k with an arbitrary orientation and suppose that the following properties are satisfied:*

1. *The random variables $X_c(w)$ and $Y(w)$, where $c \in V[H]$ and $w \in V[G]$ are $8k$-wise independent.*
2. *Let Z be any one of $X_c, c \in V[H]$ or Y. Then for any $1 \leq i \leq 2k$, $\mathbb{E}\left[Z^i\right] \neq 0$ if and only if $i = 2$.*

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

Proof. By the definition of $Z_H(G)$ we express $Z_H(G)\overline{Z_H(G)}$ as

$$\sum_{\substack{\overrightarrow{T_1}=(\overrightarrow{e_1},\cdots,\overrightarrow{e_k}) \\ \overrightarrow{T_2}=(\overrightarrow{e'_1},\cdots,\overrightarrow{e'_k}) \\ e_i,e'_i \in E[G]}} \left(\prod_{\substack{c \in V[H] \\ w \in V_{\overrightarrow{T_1}}}} X_c(w)^{\theta_{\overrightarrow{T_1}}(c,w)} \right) \cdot \left(\prod_{w \in V_{\overrightarrow{T_1}}} Y(w)^{\deg_{\overrightarrow{T_1}}(w)} \right) \cdot \left(\prod_{\substack{c \in V[H] \\ w \in V_{\overrightarrow{T_2}}}} \overline{X_c(w)}^{\theta_{\overrightarrow{T_2}}(c,w)} \right) \cdot \left(\prod_{w \in V_{\overrightarrow{T_2}}} \overline{Y(w)}^{\deg_{\overrightarrow{T_2}}(w)} \right),$$

where the function $\theta_{\overrightarrow{T}}(\cdot,\cdot)$ is defined in (4). Using the linearity of expectations and the $8k$-wise independence of the random variables $X_c(w)$ and $Y(w)$, we obtain

$$\mathbb{E}\left[Z_H(G)\overline{Z_H(G)}\right] = \sum_{\substack{\overrightarrow{T_1}=(\overrightarrow{e_1},\cdots,\overrightarrow{e_k}) \\ \overrightarrow{T_2}=(\overrightarrow{e'_1},\cdots,\overrightarrow{e'_k}) \\ e_i,e'_i \in E[G]}} Q_{\overrightarrow{T_1},\overrightarrow{T_2}},$$

where

$$Q_{\overrightarrow{T_1},\overrightarrow{T_2}} = \left(\prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}} \mathbb{E}\left[X_c(w)^{\theta_{\overrightarrow{T_1}}(c,w)} \overline{X_c(w)}^{\theta_{\overrightarrow{T_2}}(c,w)} \right] \right) \cdot \left(\prod_{w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}} \mathbb{E}\left[Y(w)^{\deg_{\overrightarrow{T_1}}(w)} \overline{Y(w)}^{\deg_{\overrightarrow{T_2}}(w)} \right] \right) .$$

For any $c \in V[H]$ and $w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}$, we write

$$R_{\overrightarrow{T_1},\overrightarrow{T_2}}(c,w) = \mathbb{E}\left[X_c(w)^{\theta_{\overrightarrow{T_1}}(c,w)} \overline{X_c(w)}^{\theta_{\overrightarrow{T_2}}(c,w)} \right].$$

Let $R_{\overrightarrow{T_1},\overrightarrow{T_2}} = \prod_{c \in V[H]} \prod_{w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}} R_{\overrightarrow{T_1},\overrightarrow{T_2}}(c,w)$. Then

$$Q_{\overrightarrow{T_1},\overrightarrow{T_2}} = R_{\overrightarrow{T_1},\overrightarrow{T_2}} \cdot \prod_{w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}} \mathbb{E}\left[Y(w)^{\deg_{\overrightarrow{T_1}}(w)} \overline{Y(w)}^{\deg_{\overrightarrow{T_2}}(w)} \right].$$

We claim that if the term $Q_{\overrightarrow{T_1},\overrightarrow{T_2}} \neq 0$, then every vertex in $V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}$ has even degree in the undirected sense. First, we show that using this claim we can finish the proof of the theorem. Note that $\mathbb{E}[Z_H(G)\overline{Z_H(G)}] = \sum_{G_{\overrightarrow{T_1},\overrightarrow{T_2}} \in \mathcal{E}_{2k}} Q_{\overrightarrow{T_1},\overrightarrow{T_2}}$ where $\mathcal{E}_{2k}$ is the set of directed subgraphs of G on $2k$ edges with every vertex having even degree in the undirected sense. Observing that the undirected graph defined by $G_{\overrightarrow{T_1},\overrightarrow{T_2}}$ is a Eulerian circuit, by Corollary 1 we get $\mathbb{E}[Z_H(G)\overline{Z_H(G)}] \leq \sum_{G_{\overrightarrow{T_1},\overrightarrow{T_2}} \in \mathcal{E}_{2k}} |Q_{\overrightarrow{T_1},\overrightarrow{T_2}}| \leq c \cdot m^k$. Note that an upper bound for the constant c is $\max_{G_{\overrightarrow{T_1},\overrightarrow{T_2}} \in \mathcal{E}_{2k}} |Q_{\overrightarrow{T_1},\overrightarrow{T_2}}|$.

Let us now prove that $Q_{\overrightarrow{T_1},\overrightarrow{T_2}} \neq 0$ implies that every vertex in $V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}$ has even degree in the undirected sense. We first make the following observations: For any vertex c of C_k and w in $V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}$ we have: $\mathbb{E}\left[X_c^i(w)\right] \neq 0$ if and only if $i = 2$. After expanding $Z_H(G)$ and $\overline{Z_H(G)}$, $X_c(\cdot), c \in V[H]$ appears twice in each term, so we have $\sum_{w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}} \theta_{\overrightarrow{T_1}}(c,w) + \theta_{\overrightarrow{T_2}}(c,w) = 4$. Consider a subgraph $G_{\overrightarrow{T_1},\overrightarrow{T_2}}$ on $2k$ edges such that $R_{\overrightarrow{T_1},\overrightarrow{T_2}} \neq 0$. Assume for the sake of contradiction that $G_{\overrightarrow{T_1},\overrightarrow{T_2}}$ has a vertex w of odd degree. This implies that there is a vertex $c \in C_k$ such that $\theta_{\overrightarrow{T_1}}(c,w) + \theta_{\overrightarrow{T_2}}(c,w)$ is either one or three. However $\theta_{\overrightarrow{T_1}}(c,w) + \theta_{\overrightarrow{T_2}}(c,w)$ cannot be one since in this case both $R_{\overrightarrow{T_1},\overrightarrow{T_2}}$ and $Q_{\overrightarrow{T_1},\overrightarrow{T_2}}$ must vanish. Now consider the case where $\theta_{\overrightarrow{T_1}}(c,w) + \theta_{\overrightarrow{T_2}}(c,w) = 3$. This means that $R_{\overrightarrow{T_1},\overrightarrow{T_2}}(c,w)$ is either $\mathbb{E}[X_c^2(w)\overline{X_c(w)}]$ or the symmetric variant $\mathbb{E}\big[X_c(w)\overline{X_c(w)}^2\big]$. Assume that $R_{\overrightarrow{T_1},\overrightarrow{T_2}}(c,w) = \mathbb{E}[X_c^2(w)\overline{X_c(w)}]$. Since $\sum_{w \in V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}} \theta_{\overrightarrow{T_1}}(c,w) + \theta_{\overrightarrow{T_2}}(c,w) = 4$, there must be a vertex $w' \neq w$ in $V_{\overrightarrow{T_1}} \cup V_{\overrightarrow{T_2}}$ such that $R_{\overrightarrow{T_1},\overrightarrow{T_2}}(c,w') = \mathbb{E}[\overline{X_c(w')}]$. This implies that $R_{\overrightarrow{T_1},\overrightarrow{T_2}}$ vanishes and hence $Q_{\overrightarrow{T_1},\overrightarrow{T_2}}$ must also vanish, which leads to a contradiction. □

Now we prove Theorem 1.

Proof (of Theorem 1). First, observe that

$$\frac{\mathbb{E}[Z_H(G)\overline{Z_H(G)}] - \mathbb{E}^2[Z_H(G)]}{|\mathbb{E}[Z_H(G)]|^2} \leq \frac{\mathbb{E}[Z_H(G)\overline{Z_H(G)}]}{|\mathbb{E}[Z_H(G)]|^2}.$$

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

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

By Chebyshev's inequality (Lemma 3), we have

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

Observe that

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

By choosing $s = O\left(\frac{1}{\varepsilon^2} \cdot \frac{m^k}{(\#C_k)^2}\right)$, we get $\Pr\left[|Z^* - \mathbb{E}[Z^*]| \geq \varepsilon \cdot |\mathbb{E}[Z^*]|\right] \leq 1/3$.

The probability of success can be amplified to $1 - \delta$ by running in parallel $O\left(\log \frac{1}{\delta}\right)$ copies of the algorithm and outputting the median of those values.

Since storing each random variable requires $O(\log n)$ space and the number of random variables used in each trial is $O(1)$, so the overall space complexity is as claimed. □

5 Conclusions

In this paper we presented an unbiased estimator for counting the number of occurrences of any d-regular graph H in a graph G. For the special case $d = 2$, we proved that the variance of the computed random variables is not too big, thus obtaining an efficient algorithm for computing approximate estimates for the quantities in question. Our work raises a number of challenging open questions.

1. Is it possible to generalize the proposed approach to count other subgraphs, such as for example general cliques? Our results provide an unbiased estimator. However, is there any way of keeping the variance of the underlying random variables small?
2. We used complex-valued hash functions to achieve the desired result. However, there might be other possibilities. Can we use hash functions that take values from other structures, such as Clifford algebras, to obtain better upper bounds for the space complexity of the algorithm?
3. Our algorithm improves significantly upon the naïve sampling algorithms. Unfortunately, it is not clear at all what the optimal memory consumption of an algorithm is. So another fundamental research direction is to obtain lower bounds for counting subgraphs in the turnstile model.

Acknowledgement. The authors would like to thank Divya Gupta for helping them with the implementation of an earlier version of the algorithm. Madhusudan Madhusudan thanks Girish Varma for stimulating discussions.

References

1. Bar-Yossef, Z., Kumar, R., Sivakumar, D.: Reductions in streaming algorithms, with an application to counting triangles in graphs. In: Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 623–632 (2002)
2. Becchetti, L., Boldi, P., Castillo, C., Gionis, A.: Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 16–24 (2008)
3. Bordino, I., Donato, D., Gionis, A., Leonardi, S.: Mining large networks with subgraph counting. In: Proceedings of the 8th IEEE International Conference on Data Mining, pp. 737–742 (2008)
4. Buriol, L.S., Frahling, G., Leonardi, S., Marchetti-Spaccamela, A., Sohler, C.: Counting triangles in data streams. In: Proceedings of the 25th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 253–262 (2006)
5. Buriol, L.S., Frahling, G., Leonardi, S., Sohler, C.: Estimating clustering indexes in data streams. In: Arge, L., Hoffmann, M., Welzl, E. (eds.) ESA 2007. LNCS, vol. 4698, pp. 618–632. Springer, Heidelberg (2007)
6. Chien, S., Rasmussen, L.E., Sinclair, A.: Clifford algebras and approximating the permanent. Journal of Computer and System Sciences 67(2), 263–290 (2003)
7. Flum, J., Grohe, M.: The parameterized complexity of counting problems. SIAM Journal on Computing 33(4), 892–922 (2004)
8. Ganguly, S.: Estimating frequency moments of data streams using random linear combinations. In: Jansen, K., Khanna, S., Rolim, J.D.P., Ron, D. (eds.) RANDOM 2004 and APPROX 2004. LNCS, vol. 3122, pp. 369–380. Springer, Heidelberg (2004)
9. Jowhari, H., Ghodsi, M.: New streaming algorithms for counting triangles in graphs. In: Wang, L. (ed.) COCOON 2005. LNCS, vol. 3595, pp. 710–716. Springer, Heidelberg (2005)
10. Karmarkar, N., Karp, R., Lipton, R., Lovasz, L., Luby, M.: A Monte-Carlo algorithm for estimating the permanent. SICOMP: SIAM Journal on Computing 22, 284–293 (1993)
11. McGregor, A.: Open Problems in Data Streams and Related Topics. In: IITK Workshop on Algoriths For Data Sreams (2006), `http://www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf`
12. Muthukrishnan, S.: Data Streams: Algorithms and Applications. Foundations and Trends in Theoretical Computer Science 1(2) (2005)