Approximately Counting Cliques

Lars Eilstrup Rasmussen*
University of California at Berkeley

August 1, 1996

Abstract

We present a very simple, randomized approximation algorithm for determining the number of cliques in a random graph.

*Supported in part by NSF Grant CCR-9505448 and a UC Berkeley Faculty Research Grant

1 Introduction

1.1 Cliques

Let $G = ([n], E)$ be an undirected graph on vertex set $[n] = \{1, 2, \ldots, n\}$. By a *clique* of G we shall mean any complete subgraph of G. In particular, the empty subgraph, as well as the n singleton subgraphs, are cliques of G, and it is therefore trivial to decide whether a given graph contains a clique. However, the study of cliques, most notably of maximal and maximum cliques, has attracted considerable attention in the past decades. For example, it is known that, for almost all sufficiently large n, the size, $\mathrm{cl}(G)$, of a maximum clique of a *random* graph, G, exactly equals its expectation, $\mathrm{cl}(n)$, with probability tending to 1 (see e.g. [2], Chapter XI). Nevertheless, when G is a *general* graph, $\mathrm{cl}(G)$ cannot be approximated even to within a factor $n^{1-\epsilon}$ for any $\epsilon > 0$, unless NP=coR [4]. Likewise, no efficient procedure for *finding* a clique of size significantly greater than $\frac{1}{2}\mathrm{cl}(G)$ is known even for random graphs (indeed, it has been conjectured that no such procedure exists [5]).

In this paper, we shall consider the problem of counting the total number, $\#\mathrm{cl}(G)$, of cliques of G. We note that a simple construction ([13], Theorem 1.17) coupled with [4] demonstrates that even *approximating* $\#\mathrm{cl}(G)$ can not be done for a general graph to within a factor $2^{n^{1-\epsilon}}$ for any $\epsilon > 0$ unless NP=coR. Weakening the requirements further, we ask whether there exists an efficient (randomized) approximation algorithm for $\#\mathrm{cl}(G)$ that works for a random graph with high probability (i.e., with probability tending to 1). In this paper, we answer the question in the affirmative.

1.2 Randomized approximation schemes

We shall formalize the notion of "efficient approximation algorithm" as follows: A *randomized approximation scheme* [7] for $\#\mathrm{cl}(G)$ is a probabilistic algorithm which, when given an input graph, G, and a real number, $0 < \epsilon < 1$, outputs a number X_G (a random variable) such that

$$\Pr\{(1-\epsilon)\#\mathrm{cl}(G) \leq X_G \leq (1+\epsilon)\#\mathrm{cl}(G)\} \geq \tfrac{3}{4}.$$

The success probability may be boosted to $1-\delta$ for any $0 < \delta < 1$ by running the algorithm $\mathrm{O}(\log \delta^{-1})$ times and taking the median of the results [6]. Such a scheme is said to be *fully polynomial* if its execution time is bounded by some polynomial in the size of G and ϵ^{-1}. We shall henceforth contract the phrase *fully polynomial randomized approximation scheme* to *fpras*. In this paper, we will describe an algorithm which, when run on a uniformly chosen graph on n vertices, satisfies the requirements of a fpras with probability (over the choice of graph) tending to 1 with n.

Often, as will be the case in this paper, a fpras is constructed from an *unbiased estimator*: we shall construct a polynomial time, probabilistic algorithm whose output on a graph G is a random variable X_G such that $\mathrm{E}[X_G] = \#\mathrm{cl}(G)$. We then simply run a number of independent copies of the algorithm on the same input, and output the average of the results. A straightforward application of Chebyshev's inequality shows that $\mathrm{O}\left(\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2}\epsilon^{-2}\right)$ copies of the estimator suffice to constitute a randomized approximation scheme for $\#\mathrm{cl}(G)$. From this it is clear that if the *critical ratio*, $\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2}$, is bounded above by a polynomial in the size of G, then we have constructed an fpras for $\#\mathrm{cl}(G)$. We will show that, for some very small constant γ ($\leq 10^{-8}$), the critical ratio of our estimator for a random graph of size n is bounded by n^γ with probability (over the choice of graph) tending to 1 with n. Since a single run of the unbiased estimator can be completed in time $\mathrm{O}(n^2)$, we get, for a random graph, a fpras with running time $\mathrm{O}(n^{2+\gamma}\epsilon^{-2})$.

Our algorithm hinges on a certain self-reducibility property of cliques which allows us to naturally identify a graph with a rooted, binary tree, such that each sub-tree corresponds to the cliques of a certain subgraph, and such that the leaves are in one-to-one correspondence with the cliques of the original graph. We then approximate the number of leaves in this tree by traversing a single path from the root to a leaf, making random choices at each node. When a leaf is reached, we compute the probability that the algorithm reached that particular leaf, and output the inverse of that probability. It is not hard to see that the expected value of this experiment is exactly the number of leaves in the tree (since each leaf contributes exactly 1 to the expectation). Similarly, the critical ratio of the experiment (which is a measure of the variance relative to the square of the mean) will be sufficiently small if the distribution induced by the experiment on the leaves of the tree is not too far from uniform.

The algorithm follows a method used previously by Hammersley [3], Knuth [8, 9] and Rasmussen [12] in other settings. However, as we shall see, the application of the method to the problem of counting cliques requires, in contrast to the above applications, that the random choices made at each node in the tree are biased according to pre-estimates of the sizes of the left and right subtrees. This biasing necessitates a considerably more involved analysis of the critical ratio of the resulting estimator.

We shall specify our algorithm in Section 2, present the analysis of its runtime in Section 3, and make a few concluding remarks in Section 4.

2 The algorithm

In this section, we will specify the algorithm and state the main result regarding its runtime. In preparation, we introduce some simple notation:

Definition 1 *Let $G = ([n], E)$ be an undirected graph and denote by*

$$\#\mathrm{cl}_i(G) = |\{C \subset [n] \ : \ |C| = i \ \wedge \ \forall v \neq w \in C \ : \ \{v, w\} \in E\}|$$

the number of cliques of size i in G, by

$$\#\mathrm{cl}(G) = \sum_i \#\mathrm{cl}_i(G)$$

the total number of cliques in G, by

$$\Gamma(v) = \{w \in [n] - v \ : \ \{v, w\} \in E\}$$

the set of neighbors of vertex v, by

$$d(v) = |\Gamma(v)|$$

the degree of v, and, for $S \subset [n]$, denote by

$$G_S = (S, \{\{v, w\} \in E \ : \ v, w \in S\})$$

the subgraph of G induced by S. Finally, let

$$\mathcal{G}(n) = \{G = ([n], E) \;:\; E \subset [n]^2\}$$

denote the set of all n-vertex graphs, and say that a property, P, holds for random *graphs if* $\Pr\{P(G_n)\} \to 1$ *as $n \to \infty$ when G_n is chosen u.a.r. from $\mathcal{G}(n)$.* □

We begin by noticing that the set of cliques of G is naturally partitioned into those cliques containing a given vertex, v, and those not containing v:

$$\#\mathrm{cl}(G) = \#\mathrm{cl}(G_{\Gamma(v)}) + \#\mathrm{cl}(G_{[n]/v}). \tag{1}$$

Expanding out this recurrence immediately gives an exponential time, deterministic algorithm for computing $\#\mathrm{cl}(G)$ exactly (which would exhaustively enumerate each leaf in the binary tree defined by the recurrence). The idea for our approximation algorithm is to randomly choose only *one* of the right-hand terms in (1) according to some bias, and recursively *estimate* its size. From that we then compute an estimate of $\#\mathrm{cl}(G)$ by assuming that the bias accurately reflects the relative sizes of the two terms. As we shall see, the expected value of this experiment is exactly the desired number, $\#\mathrm{cl}(G)$, *regardless* of how the bias is constructed, whereas the variance of the estimator depends on how accurate is our pre-estimate of the ratio of the two terms.

More precisely, our algorithm assigns to all graphs G a r.v. X_G as follows:

```
if n = 0 then X_G = 1
else flip a p-q-coin
     if heads then X_G = p^{-1} X_{G_{Γ(v)}}
     else          X_G = q^{-1} X_{G_{[n]/v}}
```

Figure 1: The Algorithm.

Here, $p = p(G)$ is a function of the input graph, G, a p-q-coin is one which lands heads up with probability p and tails up with probability $q = 1 - p$, and v is an arbitrary vertex (chosen independently of the edges of G). We shall specify an appropriate value for p presently.

Theorem 2 *Let $G \in \mathcal{G}(n)$, and p any function from graphs to $[0,1]$ such that*

$$p(G) = 0 \;\Rightarrow\; \#\mathrm{cl}(G_{\Gamma(v)}) = 0$$

and

$$p(G) = 1 \;\Rightarrow\; \#\mathrm{cl}(G_{[n]/v}) = 0.$$

Then

$$\mathrm{E}[X_G] = \#\mathrm{cl}(G).$$

Proof: A straightforward induction on n. □

To control the variance of the estimator, X_G, we simply let p and q reflect the *expected* fractions of cliques in the two terms in (1) for a random graph, given the degree, $d = d(v)$, of v. To specify this precisely, we introduce some further notation:

Definition 3 *Choose G_n u.a.r. from $\mathcal{G}(n)$. Define*

$$E_n \stackrel{\mathrm{def}}{=} \mathrm{E}[\#\mathrm{cl}(G_n)] = \sum_{i=0}^{n} \binom{n}{i} 2^{-\binom{i}{2}} \stackrel{\mathrm{def}}{=} \sum_{i=0}^{n} E_{n,i}.$$

□

The specification of the algorithm in Figure 1 is now completed by letting

$$p = p(G) = \frac{E_d}{E_d + E_{n-1}}$$

and

$$q = 1 - p = \frac{E_{n-1}}{E_d + E_{n-1}}.$$

We are now ready to state the main theorem:

Theorem 4 *For a random graph, G, the critical ratio of the estimator, X_G, in Figure 1 is bounded by*

$$\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2} \leq \mathrm{O}(n^\gamma),$$

where $\gamma \leq 10^{-8}$.

Proof: See Section 3. $\Box$

Corollary 5 *For a random input graph in $\mathcal{G}(n)$ and tolerance $\epsilon > 0$, independent repetitions of the algorithm in Figure 1 yield a fpras with runtime $\mathrm{O}(n^{2+\gamma}\epsilon^{-2})$, where $\gamma \leq 10^{-8}$.*

Proof: Let $\epsilon > 0$, and choose an input graph, G, uniformly at random from $\mathcal{G}(n)$. Run $t = 4\epsilon^{-2}\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2}$ independent copies, $X_G^1, X_G^2, \ldots, X_G^t$, of the algorithm and output the observed average, $S_G = \frac{1}{t}\sum_{i=1}^t X_G^i$. Then, by Chebyshev's inequality,

$$\Pr\left\{|S_G - \#\mathrm{cl}(G)| \leq \epsilon\#\mathrm{cl}(G)\right\} \geq 1 - \frac{\mathrm{E}[X_G^2] - \mathrm{E}[X_G]^2}{t\epsilon^2\mathrm{E}[X_G]^2} \geq \tfrac{3}{4}.$$

The result now follows from Theorem 4 since it is clear from Figure 1 that each run of the algorithm can be completed in time $\mathrm{O}(n^2)$. $\Box$

3 The analysis

In this section, we present the analysis of the runtime of the algorithm in Figure 1, specifically by proving Theorem 4. Again, we start by introducing some simple notation:

Definition 6 *Let $f = f(n)$ and $g = g(n)$ be functions of n. Write*

$$f \sim g \qquad \textit{when} \qquad \lim_{n\to\infty}\frac{f(n)}{g(n)} = 1$$

and
$$f \prec g \qquad \textit{when} \qquad \limsup_{n\to\infty} \frac{f(n)}{g(n)} < 1.$$
$\square$

Notation 7 *Throughout this paper,* log *denotes the base 2 logarithm,* ln *the natural logarithm.* $\square$

The study of the sum E_n will play a key role in our analysis. In particular, we shall extend the domain of $E_{n,i}$, viewed as a function on i, to the real numbers using Euler's Gamma function, and demonstrate that E_n converges to an almost constant multiple of the value of $E_{n,i}$ at a particular point, $i = e_n$, which is close to the function's maximum (Lemma 11).

The small error, γ, in Theorem 4 originates from this "almost constant" limit, which in turn originates from the even smaller error term in the following lemma:

Lemma 8 *Let* $x \in [\frac{1}{4}, 1]$*. Then*
$$\frac{\partial}{\partial x} \sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} = \frac{\log x + \frac{1}{2}}{x} \sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} + \gamma$$
where $\gamma = \gamma(x) \in [-10^{-9}, 10^{-9}]$*.*

Proof: The identity follows from a special case of the transformation formula (see, e.g., [11], Chapter 10) for ϑ_3:
$$\vartheta_3(v,\tau) \stackrel{\text{def}}{=} \sum_{k=-\infty}^{\infty} e^{\pi i k^2 \tau} e^{2\pi i k v} = \sqrt{\tfrac{i}{\tau}}\; e^{-\frac{\pi i v^2}{\tau}} \vartheta_3(\tfrac{v}{\tau}, -\tfrac{1}{\tau}).$$

Let $\tau = \frac{i \ln 2}{2\pi}$, $v = \frac{\ln(x\sqrt{2})}{2\pi i}$. Then
$$\sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} \quad = \quad \sum_{k=-\infty}^{\infty} \left(\tfrac{1}{\sqrt{2}}\right)^{k^2} (x\sqrt{2})^k$$

$$
\begin{aligned}
&= \vartheta_3(v,\tau) \\
&= \sqrt{\tfrac{i}{\tau}}\; e^{-\frac{\pi i v^2}{\tau}}\, \vartheta_3(\tfrac{v}{\tau}, -\tfrac{1}{\tau}) \\
&= 2^{\frac{5}{8}}\sqrt{\tfrac{\pi}{\ln 2}}\; x^{\frac{\log x+1}{2}}\, S(x),
\end{aligned}
$$

where $S(x) = \sum\limits_{k=-\infty}^{\infty} e^{-\frac{2\pi^2 k^2}{\ln 2} - 2\pi i k \frac{\ln(x\sqrt{2})}{\ln 2}}$. Thus,

$$\frac{\partial}{\partial x}\sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} = \frac{\log x + \frac{1}{2}}{x}\sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} + \gamma,$$

where

$$\gamma = 2^{\frac{5}{8}}\sqrt{\tfrac{\pi}{\ln 2}}\; x^{\frac{\log x+1}{2}} \sum_{k=-\infty}^{\infty} -\frac{e^{-\frac{2\pi^2 k^2}{\ln 2} - 2\pi i k \frac{\ln(x\sqrt{2})}{\ln 2}}\, 2\pi i k}{x\ln 2} \in [-10^{-9}, 10^{-9}].$$

□

For simplicity, we shall carry out the remainder of the analysis leaving out the tedious details of tracking this error. That is, we shall pretend that

$$\frac{\partial}{\partial x}\sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} = \frac{\log x + \frac{1}{2}}{x}\sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} \tag{2}$$

and, based on this approximate equality, prove in place of Theorem 4 as stated above, that, for a random graph G and *any* $\xi > 0$:

$$\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2} \leq \mathrm{O}(n^{\xi}).$$

The reader is encouraged to check that the error, $\gamma \in [-10^{-9}, 10^{-9}]$, of Lemma 8 does indeed turn into the error, $\gamma \leq 10^{-8}$, of Theorem 4.

We now show how Euler's Gamma function can be used to obtain a sharp estimate of the sum E_n:

Definition 9 (Euler's Gamma function) *For $x \in \mathbf{R}$, let*

$$x! \stackrel{\text{def}}{=} \Gamma(x+1) \stackrel{\text{def}}{=} \int_0^\infty t^x e^{-t} dt.$$

Furthermore, for $y \in \mathbf{R}$, let

$$(x)_y = \frac{x!}{(x-y)!} \qquad \textit{and} \qquad \binom{x}{y} = \frac{(x)_y}{y!}.$$

□

The Gamma function possesses simple poles at $x = 0, -1, -2\ldots$ (see e.g. [1], page 255-266 for properties of the Gamma function). However, those poles will not interfere with our use of the function in this paper. A simple application of integration by parts shows, for $x \in \mathbf{R}$, that $x! = x(x-1)!$. Thus, since $0! = \Gamma(1) = 1$, the Gamma function does indeed extend the natural factorial function

$$n! = n(n-1)(n-2)\ldots 1.$$

Also note that, for $x \in \mathbf{R}$ and $k \in \mathbf{N}$,

$$(x)_k = x(x-1)(x-2)\ldots(x-k+1).$$

In Lemma 14, we will make use of the fact that

$$\Psi(x) \stackrel{\text{def}}{=} \frac{\partial}{\partial x} \ln \Gamma(x)$$

is monotonically increasing on the positive real axis.

As already advertised, Definition 9 allows us to extend the domain of $E_{n,i}$ (Definition 3) to non-integer values of i. We now prove the claim indicated before Lemma 8 that E_n converges to an almost constant multiple of the value of $E_{n,i}$ at a particular point. We will also do this for the quantity H_n, which is the difference between consecutive E_n's:

Definition 10

$$H_n \stackrel{\text{def}}{=} E_{n+1} - E_n = \sum_{i=0}^{n} \binom{n}{i} 2^{-\binom{i+1}{2}} \stackrel{\text{def}}{=} \sum_{i=0}^{n} H_{n,i}.$$

□

Lemma 11 *Let $e = e_n$ be defined by $E_{n,e} = E_{n,e+1}$, and $h = h_n$ by $H_{n,h} = H_{n,h+1}$. Then, for some constant C,*

$$E_n \sim CE_{n,e} \qquad \text{and} \qquad H_n \sim CH_{n,h},$$

Furthermore,

$$e \sim \log n - \log\log n,$$

and $e > h > e-1$ for all sufficiently large n.

Remark: As explained in the remarks following Lemma 8, the statement of this lemma is based on the approximate equality (2). In fact, the quantity C is not constant, but varies with n by a small factor (at most 1 ± 10^{-9}). In particular, the second $\sim$ in (3) below is off by this factor, as the reader may readily check.

Proof: Note that $\frac{E_{n,i}}{E_{n,i+1}}$ is monotonically decreasing in i, so e is well defined. We first solve for e:

$$\frac{E_{n,e+1}}{E_{n,e}} = \frac{n-e}{e+1} 2^{-e} = 1$$

$$\Leftrightarrow \quad \log n - \log\log n < e < \log n - \log\log n + \mathrm{O}\left(\frac{\log\log n}{\log n}\right)$$

Similarly, $e > h > e-1$. Note in particular that all of e, h, $n-e$ and $n-h$ tend to infinity with n.

We now proceed in two steps to prove that

$$CE_{n,e} \sim \sum_{k=-\lfloor e \rfloor}^{n-\lfloor e \rfloor} E_{n,e+k} \sim E_n. \tag{3}$$

Consider first the values of $E_{n,i}$ at (constant) integer distance from e: for any fixed $k \in \mathbf{Z}$ such that $0 \leq e+k \leq n$, we have

$$\frac{E_{n,e+k}}{E_{n,e}} = 2^{-\binom{k}{2}-ek} \times \frac{(n-e)_k}{(e+k)_k} = 2^{-\binom{k}{2}} \times \frac{(n-e)_k}{(n-e)^k} \times \frac{(e+1)^k}{(e+k)_k} \sim 2^{-\binom{k}{2}}$$

Since, for any $\epsilon > 0$, there is a constant l such that, for all sufficiently large n,

$$\sum_{k=-\lfloor e \rfloor}^{n-\lfloor e \rfloor} E_{n,e+k} \leq (1+\epsilon) \sum_{k=-l}^{l} E_{n,e+k} \qquad \text{and} \qquad \sum_{k=-\infty}^{\infty} 2^{-\binom{k}{2}} \leq (1+\epsilon) \sum_{k=-l}^{l} 2^{-\binom{k}{2}},$$

the first $\sim$ in (3) follows with

$$C = \sum_{k=-\infty}^{\infty} 2^{-\binom{k}{2}} \qquad \left[\approx 2^{\frac{5}{8}}\sqrt{\tfrac{\pi}{\ln 2}} \quad \text{by the proof of Lemma 8} \right].$$

To obtain the second $\sim$ in (3), we prove that, for all $a \in [0,1]$,

$$\lim_{n\to\infty} \frac{\frac{\partial}{\partial a}\sum_{i=0}^{n} E_{n,i+a}}{E_n} = \lim_{n\to\infty} \frac{\sum_{i=0}^{n} \frac{\partial}{\partial i} E_{n,i+a}}{E_n} = 0.$$

Let $e' = \lceil e \rceil + a$. Since $E_{n,e'} \leq E_n$, it suffices to show that

$$\lim_{n\to\infty} \sum_{i=0}^{n} \frac{\partial}{\partial i} \frac{E_{n,i+a}}{E_{n,e'}} = \lim_{n\to\infty} \sum_{k=-\lfloor e' \rfloor}^{n-\lfloor e' \rfloor} \frac{\partial}{\partial k} \frac{E_{n,e'+k}}{E_{n,e'}} = 0 \tag{4}$$

(where the first $=$ comes from substituting $k = i - \lfloor e' \rfloor$). Let

$$x = x_{e'} = \frac{E_{n,e'+1}}{E_{n,e'}} = \frac{n-e'}{e'+1} 2^{-e'}.$$

Then, as above,

$$\frac{E_{n,e'+k}}{E_{n,e'}} = 2^{-\binom{k}{2} - e'k} \times \frac{(n-e')_k}{(e'+k)_k} = x^k 2^{-\binom{k}{2}} \times \frac{(n-e')_k}{(n-e')^k} \times \frac{(e'+1)^k}{(e'+k)_k} \sim x^k 2^{-\binom{k}{2}},$$

and, again since all but a negligible part of the sum in (4) is contained within a constant distance from $k = 0$, it remains to show that

$$\sum_{k=-\infty}^{\infty} \frac{\partial}{\partial k} x^k 2^{-\binom{k}{2}} = \ln 2 \sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} (\log x + \tfrac{1}{2} - k) = 0,$$

or equivalently, that

$$\frac{\partial}{\partial x} \sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}} = \frac{\log x + \frac{1}{2}}{x} \sum_{k=-\infty}^{\infty} x^k 2^{-\binom{k}{2}}.$$

Since $x_{e'}$ decreases with e',

$$1 = x_e \geq x_{e'} > x_{e+2} = \tfrac{1}{4} \times \frac{n-e-2}{n-e} \times \frac{e+1}{e+3} \to \tfrac{1}{4},$$

and the result follows by Lemma 8 (actually from the approximate equality (2)). The proof for H_n is similar. $\square$

Before moving on, we shall need the following elementary technical results:

Lemma 12 (Chernoff) *Let D be a binomial random variable with parameters n and $\frac{1}{2}$, and let $T = \frac{1}{2}n + \sqrt{2bn \ln n}$. Then* $\Pr\{D > T\} < n^{-b}$.

Proof: Employ a standard Chernoff bound (see e.g. [10] p. 86) □

Lemma 13 *Let $m, k, \frac{m}{k}, \frac{n}{k} \to \infty$ as $n \to \infty$. Then*

$$\binom{n}{k}\binom{m}{k}^{-1} \sim \left(\frac{n}{m}\right)^k$$

Proof:

$$\binom{n}{k}\binom{m}{k}^{-1} = \frac{n!}{(n-k)!} \times \frac{(m-k)!}{m!} \sim n^k m^{-k},$$

where the last $\sim$ follows from e.g. [1], 6.1.46. □

We are now ready to prove the bound on the critical ratio of the estimator when run on a random graph. We shall take a slight detour. In particular, we shall first (Lemma 14) bound a suitable ratio of expectations, and then prove the main Theorem 4 using a crude bound on the variance of $\#\mathrm{cl}(G)$.

Lemma 14 *Choose $G = G_n$ u.a.r. from $\mathcal{G}(n)$, and let X_G be the corresponding estimator. Then, for any $\xi > 0$,*

$$\frac{\mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G^2]]}{\mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G]]^2} = \mathrm{O}(n^\xi),$$

where $\mathrm{E}_{\mathcal{G}}$ denotes the expectation over graphs, and E_σ the expectation over coin-tosses performed by the estimator.

Remark: The proof of this lemma relies on Lemma 11, which in turns relies on the approximate equality (2). Carrying forward the small error in Lemma 11 leads to a corresponding small error in Lemma 14, which means that the upper bound holds only for $\xi > 10^{-8}$ rather than for $\xi > 0$.

Proof: By Theorem 2,

$$\mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G]] = \mathrm{E}_{\mathcal{G}}[\#\mathrm{cl}(G)] = E_n.$$

Let F_n denote $\mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G^2]]$. For convenience of notation, choose for the remainder of the proof $G = G_{n+1}$ u.a.r. from $\mathcal{G}(n+1)$, and let D be the degree of v. By conditioning on D, we obtain the recurrence

$$F_0 = 1$$

$$F_{n+1} = \sum_{d=0}^{n} \Pr\{D = d\} \left(p^{-1} F_d + q^{-1} F_n\right),$$

and the identity

$$H_n = \mathrm{E}[\#\mathrm{cl}(G_{\Gamma(v)})] = \sum_{d=0}^{n} \Pr\{D = d\} E_d$$

(recall from Definition 10 that $H_n \stackrel{\mathrm{def}}{=} E_{n+1} - E_n$). We can now prove by induction on n that $F_n = \mathrm{O}(n^\xi E_n^2)$: this will complete the proof of the lemma. In particular, we will prove that $F_n \le n^\xi E_n^2 \Rightarrow F_{n+1} \le (n+1)^\xi E_{n+1}^2$ for n sufficiently large. We have

$$\begin{aligned} F_{n+1} &\le \sum_{d=0}^{n} \Pr\{D = d\} \left(p^{-1} d^\xi E_d^2 + q^{-1} n^\xi E_n^2\right) \\ &= \sum_{d=0}^{n} \Pr\{D = d\} \left(d^\xi E_d^2 + (d^\xi + n^\xi) E_d E_n + n^\xi E_n^2\right) \\ &\le \sum_{d=0}^{n} \Pr\{D = d\} d^\xi E_d^2 + 2n^\xi H_n E_n + n^\xi E_n^2. \end{aligned}$$

Since $E_{n+1}^2 = H_n^2 + 2E_n H_n + E_n^2$, it suffices to prove that

$$\sum_{d=0}^{n} \Pr\{D = d\} d^\xi E_d^2 \le n^\xi H_n^2.$$

We will split the sum around $T = T_n = \frac{1}{2}n + \sqrt{4n \ln n}$ by letting $\alpha = 2^{-\frac{\xi}{2}}$, and proving that

$$\sum_{d \ge T} \Pr\{D = d\} d^\xi E_d^2 \le (1-\alpha) n^\xi H_n^2, \tag{5}$$

and that

$$\sum_{d<T} \Pr\{D=d\} d^\xi E_d^2 \leq \alpha n^\xi H_n^2. \tag{6}$$

But first, observe that

$$\frac{\partial}{\partial i} \ln E_{n,i} = -\Psi(i+1) + \Psi(n-i+1) - \frac{2i-1}{2} \ln 2$$

is monotonically decreasing (see the remarks following Definition 9). Therefore,

$$e_n \leq i \leq e_n + 1 \;\Rightarrow\; E_{n,e_n} \leq E_{n,i} \tag{7}$$

Furthermore,

$$\frac{H_{n,i+1}}{H_{n,i}} \leq \frac{E_{T,i+1}}{E_{T,i}}$$

whenever $i \leq 4\sqrt{n \ln n}$. Since both sides of the inequality are decreasing in i (as in the proof of Lemma 11), this implies that $h_n \leq e_T$. Since e_n is increasing in n, and since $e_n - 1 \leq h_n$ (by Lemma 11), we have

$$e_T \leq e_n \leq h_n + 1 \leq e_T + 1 \leq e_n + 1,$$

which, by (7), implies that

$$E_{n,e_n} \leq E_{n,h_n+1} = 2^{h_n+1} H_{n,h_n+1} = 2^{h_n+1} H_{n,h_n},$$

and

$$E_{T,e_T} \leq E_{T,h_n+1}.$$

Now, consider the sum's upper part, i.e., the inequality in (5). Since $d^\xi E_d^2$ increases with d,

$$\begin{aligned}
& \sum_{d>T}^{n} \Pr\{D=d\} d^\xi E_d^2 \leq (1-\alpha) n^\xi H_n^2 \\
\Leftarrow \quad & E_n^2 \sum_{d>T}^{n} \Pr\{D=d\} \leq (1-\alpha) H_n^2 \\
\Leftarrow \quad & E_{n,e_n}^2 \sum_{d>T}^{n} \Pr\{D=d\} \prec (1-\alpha) H_{n,h_n}^2 \quad \text{(by Lemma 11)} \\
\Leftarrow \quad & \sum_{d>T}^{n} \Pr\{D=d\} \prec (1-\alpha) \left(\frac{\log n}{2n}\right)^2 \leq (1-\alpha) 2^{-2(h_n+1)},
\end{aligned}$$

which holds by Lemma 12 and the choice of T.

Since $\sum_{d \leq T} \Pr\{D = d\} < 1$, it suffices for the lower part of the sum (inequality (6)) to prove that

$$\begin{aligned} & T^{\frac{\xi}{2}} E_{T,e_T} \prec \sqrt{\alpha} n^{\frac{\xi}{2}} H_{n,h_n} \quad \text{(by Lemma 11)} \\ \Leftarrow\quad & T^{\frac{\xi}{2}} E_{T,h_n+1} \prec n^{\frac{\xi}{2}} 2^{-(h_n+1+\frac{\xi}{4})} E_{n,h_n+1} \\ \Leftarrow\quad & \left(\frac{T}{n}\right)^{h_n+1+\frac{\xi}{2}} \prec 2^{-(h_n+1+\frac{\xi}{4})} \quad \text{(by Lemma 13)}, \end{aligned}$$

which, since $\log\left(\frac{T}{n}\right) \leq \frac{2\sqrt{4n\ln n}}{n} - 1$, follows from

$$(h_n + 1 + \tfrac{\xi}{2}) \frac{2\sqrt{4n\ln n}}{n} \prec \tfrac{\xi}{4},$$

and the proof is complete. $\square$

Finally, we assemble the pieces to prove our main technical theorem, which we restate:

Theorem 4 *For a random graph, G, the critical ratio of the estimator, X_G, in Figure 1 is bounded by*

$$\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2} \leq \mathrm{O}(n^{\gamma}),$$

where $\gamma \leq 10^{-8}$.

Proof: In fact, following our previous strategy, we shall assume the approximate equality (2) and prove the stronger bound

$$\frac{\mathrm{E}[X_G^2]}{\mathrm{E}[X_G]^2} \leq \mathrm{O}(n^{\xi})$$

for any $\xi > 0$. Since the errors introduced by this assumption are small (see remarks following Lemmas 11 and 14), we get the result claimed in the theorem.

We shall prove separately that, for a random graph, the denominator of the critical ratio will not be too much smaller than E_n^2, and that the numerator will not be too much larger than its expectation. Combined with Lemma 14, this will complete the proof of the theorem. Thus, choose $G = G_n$ u.a.r. from $\mathcal{G}(n)$.

To bound the denominator, we follow the method of [2], Chapter XI, by considering pairs of cliques of size $e = e_n$. Let $Y_i = \#\mathrm{cl}_i(G)$. We have

$$\mathrm{E}[Y_e^2] = \sum_{l=0}^{e} \binom{n}{e}\binom{e}{l}\binom{n-e}{e-l} 2^{-2\binom{e}{2}+\binom{l}{2}},$$

and thus

$$\frac{\mathrm{E}[Y_e^2]}{\mathrm{E}[Y_e]^2} = \binom{n}{e}^{-1} \sum_{l=0}^{e} \binom{e}{l}\binom{n-e}{e-l} 2^{\binom{l}{2}} \stackrel{\text{def}}{=} \binom{n}{e}^{-1} \sum_{l=0}^{e} f_l.$$

For $l = 0, 1, \ldots, e-1$, let

$$\Delta_l \stackrel{\text{def}}{=} \frac{f_l}{f_{l+1}} = \frac{l+1}{e-l} \times \frac{n-2e+l+1}{e-l} \times 2^{-l}.$$

In particular, for large n,

$$\Delta_0 \geq \frac{n - 2\log n}{\log^2 n} \geq \log^2 n, \tag{8}$$

and, very crudely,

$$\Delta_{e-3} \geq 1.$$

Furthermore,

$$\Delta_l = \Delta_{l-1} \times \frac{l+1}{l} \times \left(\frac{e-l+1}{e-l}\right)^2 \times \frac{n-2e+l+1}{n-2e+l} \times 2^{-1},$$

which implies, for large n,

$$\Delta_{e-3} \leq \Delta_l \qquad \text{for } l = 1 \ldots e-1. \tag{9}$$

Combining (8) and (9) gives

$$f_l \leq \frac{f_0}{\log^2 n} \qquad \text{for } l = 1 \ldots e,$$

and hence
$$\frac{\mathrm{E}[Y_e^2]}{\mathrm{E}[Y_e]^2} \leq \binom{n}{e}^{-1} f_0 \left(1 + \mathrm{o}(1)\right) \leq 1 + \mathrm{o}(1).$$
Now, by Chebyshev's inequality,
$$\Pr\left\{n^{\frac{\xi}{4}} C^{-1} Y_e < \mathrm{E}[Y_e]\right\} \to 0,$$
and thus, by Lemma 11,
$$\Pr\left\{n^{\frac{\xi}{4}} \#\mathrm{cl}(G_n) < E_n\right\} \to 0. \tag{10}$$
To bound the numerator, simply employ Markov's inequality:
$$\Pr\left\{\mathrm{E}_\sigma[X_G^2] > n^{\frac{\xi}{4}} \mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G^2]]\right\} \to 0. \tag{11}$$
Finally, (10) and (11) together imply that
$$\Pr\left\{\frac{\mathrm{E}_\sigma[X_G^2]}{\mathrm{E}_\sigma[X_G]^2} \geq n^{\frac{3\xi}{4}} \frac{\mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G^2]]}{\mathrm{E}_{\mathcal{G}}[\mathrm{E}_\sigma[X_G]]^2}\right\} \to 0,$$
and the result follows from Lemma 14. □

4 Too simple

In this short final section, we offer some circumstantial justification for our approach by demonstrating that certain "obvious," naïve methods fail.

Note that it proved essential to the success of our estimator that $\#\mathrm{cl}(G_n)$ is fairly concentrated around its mean, E_n (see the proof of Theorem 4). At first sight, it might seem that this concentration means that the following trivial algorithm satisfies our requirements: on input a graph of size n, simply output the expectation E_n. That approach would, for any *fixed* $\epsilon > 0$, produce an ϵ-approximation of $\#\mathrm{cl}(G)$ for random graphs (in time not depending on ϵ). We required, however, the quantifiers in a different order: for random graphs, the approximation scheme must be able to produce, for any $\epsilon > 0$, an ϵ-approximation in time polynomial in $n\epsilon^{-1}$.

A slightly more sophisticated approach would be to run the following algorithm, for a suitable function $\epsilon_0 = \epsilon_0(n)$:

```
if ε > ε_0 then
    compute E_n
else
    brute force count the cliques of G_n
```

Figure 2: Threshold Algorithm.

Clearly, this algorithm will work for a graph G_n (i.e., produce an ϵ-approximation in time polynomial in $n\epsilon^{-1}$ for any $\epsilon > 0$) only if

$$(1 - \epsilon_0)\#\mathrm{cl}(G) \leq E_n \leq (1 + \epsilon_0)\#\mathrm{cl}(G), \tag{12}$$

and

$$\#\mathrm{cl}(G) \leq \mathrm{poly}(n\epsilon_0^{-1}). \tag{13}$$

To argue that this approach also fails for random graphs, consider the fact that a random graph will have more than $\frac{1}{2}n + \Omega(1)$ edges with at least constant probability. In that case, the graph would, roughly, be a uniform sample from $\mathcal{G}(n, p)$, the probability space of graphs on n vertices with individual edges present independently with probability $p = \frac{1}{2} + \Omega(n^{-1})$. Thus, the expected number of cliques in the graph would be:

$$E_{n,p} = \sum_{i=0}^{n} \binom{n}{i} p^{\binom{i}{2}},$$

which, ignoring the first two terms, is certainly larger than $(1 + \Omega(n^{-1}))E_n$. Therefore, and since $E_n \gg n$, a random graph fails to satisfy conditions (12) and (13) with at least constant probability.

5 Acknowledgments

I am thoroughly grateful to Nati Linial without whose input, in particular on the central Lemma 11, this paper would not be. Likewise, Gert Almkvist and Doron Zeilberger deserve many thanks for showing me how to prove Lemma 8, and Mike Saks for input on Section 4. I also wish to take this opportunity to thank Alistair Sinclair and Mike Luby for valuable input and invaluable encouragement throughout.

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Wiley, New York, 1972.

[2] B. Bollobás. *Random Graphs.* Academic Press, London, 1985.

[3] J. M. Hammersley. Existence theorems and Monte Carlo methods for the monomer-dimer problem. *Research Papers in Statistics*, pages 125–146, 1966.

[4] J. Håstad. Testing of the long code and hardness for clique. In *Proceedings of the 28th annual symposium on the theory of computing*, pages 11–19, 1996.

[5] M. Jerrum. Large cliques elude the Metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[6] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, pages 169–188, 1986.

[7] R. Karp and M. Luby. Monte-Carlo algorithms for enumeration and reliability problems. In *Proceedings of the 24th IEEE Symposium on Foundations of Computer Science*, pages 56–64, 1983.

[8] D. Knuth. Estimating the efficiency of bactrack programs. *Mathematics of Computations*, 29(129):121–136, Jan 1975.

[9] D. Knuth. Mathematics and computer science: Coping with finiteness. *Science*, 194(4271):1235–1242, December 1976.

[10] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, 1995.

[11] H. Rademacher. *Topics in analytic number theory*. Springer-Verlag, 1973.

[12] L. E. Rasmussen. Approximating the permanent: a simple approach. *Random Structures & Algorithms*, 5(2):349–361, April 1994.

[13] A. Sinclair. *Algorithms for random generation and counting: a Markov chain approach*. Birkhäuser, Boston, 1993.