Yet Another Analysis on "Finding Hidden Cliques in Linear Time" for $k = c\sqrt{np(1-p)}$

UPDATED: This is a note on the paper "Finding Hidden Cliques in Linear Time" by *Uriel Feige* and *Dorit Ron*, based on which my Chinese paper "Algorithm for Relatively Small Planted Clique with Small Edge Probability" was written. This note contains only the main body of my paper, excluding introduction, conclusion, and most of the preliminaries.

1. Preliminaries

Theorem 1.1 (Chernoff Bounds). *Let* $\mathbf{X} \sim B(n, p)$. *Then*

(i) $\mathbf{Pr}[\mathbf{X} - np \geq t] \leq \exp\left(-\frac{t^2}{4np}\right)$, *if* $0 < t < 2np$;

(ii) $\mathbf{Pr}[\mathbf{X} - np \geq t] \leq \exp\left(-\frac{t}{2}\right)$, *if* $t \geq 2np$.

Theorem 1.2 (Bernstein's Inequality). *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be independent random variables with* $|\mathbf{X}_i - \mathbf{E}[\mathbf{X}_i]| \leq b$ *for some* $b \in \mathbb{R}_+$ *for all* $1 \leq i \leq n$. *Let* $\mathbf{X} := \sum_{i=1}^{n} \mathbf{X}_i$, *then for any* $t > 0$,

$$\mathbf{Pr}[\mathbf{X} > \mathbf{E}[\mathbf{X}] + t] \leq \exp\left(-\frac{t^2}{2\mathbf{Var}[\mathbf{X}] + \frac{2}{3}bt}\right).$$

Theorem 1.3 (Kolmogorov's Inequality). *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be independent random variables with zero means. Define* $\mathbf{S}_k := \sum_{i=1}^{k} \mathbf{X}_i$ *for all* $1 \leq k \leq n$, *then for any* $t > 0$,

$$\mathbf{Pr}\left[\max_{1 \leq k \leq n} |\mathbf{S}_k| \geq t\right] \leq \frac{\mathbf{Var}[S_n]}{t^2}.$$

2. The Algorithm

The following algorithm is proposed in the paper "Finding Hidden Cliques in Linear Time" by *Uriel Feige* and *Dorit Ron*:

$\mathbb{A}$ //Input: A graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}) \in \mathcal{G}(n, \frac{1}{2}, k)$.
//Output: The set of vertices in the planted clique $\mathbf{K}$.

(Removal Phase)
Set $\mathbf{G}_0 \leftarrow \mathbf{G}$, $r \leftarrow 0$
For each vertex $v \in \mathbf{G}_r$, calculate its degree
while true **do**
    if all the vertices in $\mathbf{G}_r$ have the same degree $n - r$
        then go to inclusion phase
    Choose the vertex u with lowest degree and then lowest index to break ties
    Set $r \leftarrow r + 1$, let $\mathbf{G}_r$ be the graph $\mathbf{G}_{r-1}$ on all its vertices but u
    Update the degrees of vertices in $\mathbf{G}_r$
(Inclusion Phase)
(Arbitrarily) rename the vertices in $\mathbf{G} \setminus \mathbf{G}_r$ as $v_0, v_1, \cdots, v_{r-1}$
Set $\mathbf{K'}_0$ as the set of vertices in $\mathbf{G}_r$
for $i = 0$ **to** $r - 1$ **do**
    if $\{v_i, u\} \in \mathbf{E}$ for every vertex $u \in \mathbf{K'}_i$
        then Set $\mathbf{K'}_{i+1} \leftarrow \mathbf{K'}_i \cup \{v_i\}$
        else Set $\mathbf{K'}_{i+1} \leftarrow \mathbf{K'}_i$
Output $\mathbf{K'}_{r-1}$.

3. The Theorem

In this note, we analyze the above algorithm for $G(n, p(n), k(n))$ in more general setting:

Theorem 3.1. *For a sufficiently large constant $c \in \mathbb{N}$, with probability at least $\frac{2}{3}$ the original algorithm will find the hidden clique of size $k(n) \geq c\sqrt{n \cdot p(n) \cdot (1 - p(n))}$ in $G(n, p(n), k(n))$, if $p(n) = n^{-\alpha(n)}$ satisfies the following conditions:*

- $\alpha(n) = \frac{1}{t(n)}$ *for some function $t : \mathbb{N} \mapsto \mathbb{N}$ such that $t(n) = o(\log n)$;*
- $\lim_{n\to\infty} \alpha(n) = 0$.

Throughout this essay we always assume that $k(n)$ and $\alpha(n)$ satisfy the above requirements.

4. Proofs

Definition 4.1. For our convenience, define:

- $\mathbf{G}'$ as the input random graph without the planted clique, and $\mathbf{K}$ the set of vertices in the planted clique;
- the *initial degree* of a vertex v as its degree in $\mathbf{G}'$;
- $n_i = n - i$ as the number of vertices in the graph after i-th step of removal phase $\mathbf{G}_i$;
- $\mathbf{K}_i$ as the set of vertices in $\mathbf{G}_i \cap \mathbf{K}$, $k_i = |\mathbf{K}_i|$;
- $\mathbf{F}_i$ as the set of vertices in $\mathbf{G}_i \setminus \mathbf{K}$, $f_i = |\mathbf{F}_i|$.

Lemma 4.2. $p(n) = o(1)$ *thus for n sufficiently large we have $\frac{1}{2} \leq 1 - p(n) \leq 1$.*

Proof: For the sake of contradiction, assume $p(n) \geq \varepsilon$ for some $\varepsilon \in (0,1)$ infinitely often. On one hand for such n we have $n^{-\alpha(n)} \geq \varepsilon$ thus $\alpha(n) \leq \frac{-\log \varepsilon}{\log n}$, hence $t(n) \geq -\frac{1}{\log \varepsilon} \cdot \log n$. On the other hand we know $t(n) = o(\log n)$, a contradiction.

Theorem 4.3. *For a sufficiently large constant $c_0 \in \mathbb{N}$, almost surely $\big($with probability $1 - O(1/n)\big)$ for every vertex v its degree into $\mathbf{K}$ in $\mathbf{G}'$ is at most $k(n) \cdot p(n) + c_0\sqrt{k(n) \cdot \ln n}$.*

Proof: For a fixed vertex v, let $\mathbf{X}$ denote its degree into $\mathbf{K}$ in $\mathbf{G}'$, then $\mathbf{X} \sim B(k(n), p(n))$. For n sufficiently large, we have $\alpha(n) \leq \frac{1}{6}$ so $\frac{3}{2}\alpha(n) \leq \frac{1}{4}$, thus

$$\frac{c_0\sqrt{k(n) \cdot \ln n}}{k(n) \cdot p(n)} = \frac{c_0}{\sqrt{k(n)}} \cdot n^{\alpha(n)}\sqrt{\ln n} \leq \frac{c_0}{\sqrt{c}} \cdot n^{\frac{3}{2}\alpha(n)-\frac{1}{2}} \cdot \sqrt{\ln n} \leq \frac{c_0}{\sqrt{c}} \cdot n^{-\frac{1}{4}}\sqrt{\ln n} = o(1).$$

Hence by Theorem 1.1 (i) and union bound, the probability that in $\mathbf{G}'$ there is a vertex with degree into $\mathbf{K}$ exceeding $k(n) \cdot p(n) + c_0\sqrt{k(n) \cdot \ln n}$ is at most

$$\begin{aligned} n \cdot \mathbf{Pr}\left[\mathbf{X} - k(n) \cdot p(n) \geq c_0\sqrt{k(n) \cdot \ln n}\right] &\leq n \cdot \exp\left(-\frac{c_0^2}{4} \cdot \frac{k(n) \cdot \ln n}{k(n) \cdot p(n)}\right) = \exp\left(-\left(\frac{c_0^2}{4} \cdot \frac{1}{p(n)} - 1\right) \cdot \ln n\right) \\ &< \exp\left(-\left(\frac{c_0^2}{4} - 1\right) \cdot \ln n\right) = O\left(\frac{1}{n}\right), \end{aligned}$$

for $c_0 > \sqrt{8}$. ◻

Theorem 4.4. *For a sufficiently large constant $c_1 \in \mathbb{N}$ depending on c, almost surely $\Big($with probability $1 - O\left(e^{-n^{0.8}}\right)\Big)$ for all $f_i \geq \frac{1}{c} \cdot k(n)$, there is a vertex $v \in \mathbf{F}_i$ with degree in $\mathbf{G}_i$ at most $n_i \cdot p(n) + c_1\sqrt{n \cdot p(n) \cdot (1 - p(n))}$.*

Proof: For any vertex v in $\mathbf{F}_i$, its degree in $\mathbf{G}_i$ is at most its degree in $\mathbf{G}'$, thus it is sufficient to prove this theorem using their degrees in $\mathbf{G}'$.

For fixed n_i, f_i, $\mathbf{F}_i$, and $\mathbf{K}_i$, let $\mathbf{X}$ denote the sum of degrees of vertices from $\mathbf{F}_i$ in $\mathbf{G}'$. Note that edges with exactly one end in $\mathbf{F}_i$ are counted once and edges with both ends in $\mathbf{F}_i$ are counted twice in this sum. Therefore $\mathbf{X} \sim B(k_i \cdot f_i, p(n)) + 2 \cdot B(\frac{f_i \cdot (f_i - 1)}{2}, p(n))$, and

$$\begin{aligned} \mathbf{E}[\mathbf{X}] &= k_i \cdot f_i \cdot p(n) + 2 \cdot \frac{f_i \cdot (f_i - 1)}{2} \cdot p(n) \\ &= (n_i - 1) \cdot f_i \cdot p(n), \\ \mathbf{Var}[\mathbf{X}] &= k_i \cdot f_i \cdot p(n) \cdot (1 - p(n)) + 4 \cdot \frac{f_i \cdot (f_i - 1)}{2} \cdot p(n) \cdot (1 - p(n)) \\ &\leq 2 \cdot (n_i - 1) \cdot f_i \cdot p(n) \cdot (1 - p(n)) \\ &\leq 2 \cdot n_i \cdot f_i \cdot p(n) \cdot (1 - p(n)). \end{aligned}$$

Let $t := c_1 \cdot f_i\sqrt{n \cdot p(n) \cdot (1 - p(n))}$ for some constant $c_1 \in \mathbb{N}$ to be determined. If $\mathbf{X} - \mathbf{E}[\mathbf{X}] \leq t$, i.e. the sum of degrees of vertices from $\mathbf{F}_i$ is at most $(n_i - 1) \cdot f_i \cdot p(n) + c_1 \cdot f_i\sqrt{n \cdot p(n) \cdot (1 - p(n)}$, then there must exist a vertex in $\mathbf{F}_i$ with degree in $\mathbf{G}'$ at most

$$\frac{1}{f_i}\left((n_i - 1) \cdot f_i \cdot p(n) + c_1 \cdot f_i\sqrt{n \cdot p(n) \cdot (1 - p(n)}\right) \leq n_i \cdot p(n) + c_1\sqrt{n \cdot p(n) \cdot (1 - p(n))}.$$

Now we bound the probability that $\mathbf{X} - \mathbf{E}[\mathbf{X}] > t$. By Theorem 1.2 we have

$$\begin{aligned} \mathbf{Pr}[\mathbf{X} > \mathbf{E}[\mathbf{X}] + t] &\leq \exp\left(-\frac{t^2}{2\mathbf{Var}[\mathbf{X}] + \frac{2}{3} \cdot 2t}\right) \\ &\leq \exp\left(-\frac{t^2}{4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1 - p(n)) + \frac{4}{3}t}\right). \end{aligned}$$

Note that

$$\frac{\frac{4}{3}t}{4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n))} = \frac{c_1 \cdot f_i \sqrt{n \cdot p(n) \cdot (1-p(n))}}{3 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n))} = \frac{c_1}{3} \cdot \frac{1}{n_i} \cdot \sqrt{\frac{n}{p(n) \cdot (1-p(n))}},$$

thus we can distinguish two cases assuming $f_i \geq \frac{1}{c} \cdot k(n)$.

Define $n' := \frac{c_1}{3}\sqrt{\frac{n}{p(n)\cdot(1-p(n))}}$.

(1) if $n_i > n'$, we have $\frac{4}{3}t < 4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n))$. Therefore

$$\begin{aligned}
\mathbf{Pr}\left[\mathbf{X} > \mathbf{E}\left[\mathbf{X}\right] + t\right] &\leq \exp\left(-\frac{t^2}{4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n)) + \frac{4}{3}t}\right) \\
&< \exp\left(-\frac{t^2}{2 \cdot 4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n))}\right) \\
&= \exp\left(-\frac{c_1^2}{8} \cdot \frac{f_i}{n_i} \cdot n\right) \\
&\leq \exp\left(-\frac{c_1^2}{8} \cdot \frac{1}{1+c} \cdot n\right), \qquad (1)
\end{aligned}$$

where in the last inequality we use

$$\begin{aligned}
\frac{f_i}{n_i} &= \frac{f_i}{f_i + k_i} = \frac{1}{1 + k_i/f_i} \\
&\geq \frac{1}{1 + k(n)/\frac{1}{c} \cdot k(n)} \\
&\quad \left(\text{by } k_i \leq k(n) \text{ and the assumption } f_i \geq \frac{1}{c} \cdot k(n)\right) \\
&= \frac{1}{1+c}
\end{aligned}$$

(2) if $n_i \leq n'$, then we have $\frac{4}{3}t \geq 4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n))$. Therefore

$$\begin{aligned}
\mathbf{Pr}\left[\mathbf{X} > \mathbf{E}\left[\mathbf{X}\right] + t\right] &\leq \exp\left(-\frac{t^2}{4 \cdot n_i \cdot f_i \cdot p(n) \cdot (1-p(n)) + \frac{4}{3}t}\right) \\
&\leq \exp\left(-\frac{t^2}{2 \cdot \frac{4}{3}t}\right) \\
&= \exp\left(-\frac{3}{8}c_1 \cdot f_i \cdot \sqrt{n \cdot p(n) \cdot (1-p(n))}\right) \\
&\leq \exp\left(-\frac{3}{8}c_1 \cdot n \cdot p(n) \cdot (1-p(n))\right) \qquad (2)
\end{aligned}$$

since $f_i \geq \frac{1}{c} \cdot k(n) \geq \sqrt{n \cdot p(n) \cdot (1-p(n))}$.

A vertex is either in $\mathbf{F}_i$, or in $\mathbf{K}_i$, or even not in $\mathbf{G}_i$. If the membership of all the vertices is fixed, then we get the corresponding n_i and f_i. Hence, the number of possible choices for n_i, f_i, $\mathbf{F}_i$ and $\mathbf{K}_i$ is at most 3^n. By union bound and inequality for case (1), the probability that there is some n_i, f_i, $\mathbf{F}_i$, and $\mathbf{K}_i$ with $f_i \geq \frac{1}{c} \cdot k(n)$ and $n_i > n'$ such that the corresponding $\mathbf{X}$ exceeds $\mathbf{E}\left[\mathbf{X}\right] + t$ is at most

$$3^n \cdot \exp\left(-\frac{c_1^2}{8} \cdot \frac{1}{1+c} \cdot n\right) = \exp\left(-\left(\frac{c_1^2}{8} \cdot \frac{1}{1+c} - \ln 3\right) \cdot n\right) = O(e^{-n}),$$

for c_1 sufficiently large so that

$$\frac{c_1^2}{8} \cdot \frac{1}{1+c} - \ln 3 \geq 1. \qquad (3)$$

More specifically, if we restrict $n_i \le n'$, then the number of possible choices is at most

$$\binom{n}{n'} 3^{n'} \le n^{n'} 3^{n'} = \exp\left((\ln n + \ln 3) \cdot n'\right).$$

By union bound and inequality for case (2), the probability that there is some n_i, f_i, $\mathbf{F}_i$, and $\mathbf{K}_i$ with $f_i \ge \frac{1}{c} \cdot k(n)$ and $n_i \le n'$ such that the corresponding $\mathbf{X}$ exceeds $\mathbf{E}[\mathbf{X}] + t$ is at most

$$\begin{aligned}
&\exp\left((\ln n + \ln 3) \cdot n'\right) \cdot \exp\left(-\frac{3}{8} c_1 \cdot n \cdot p(n) \cdot (1 - p(n))\right) \\
=&\exp\left(-\frac{3}{8} c_1 \cdot n \cdot p(n) \cdot (1 - p(n)) + (\ln n + \ln 3) \cdot \frac{c_1}{3} \sqrt{\frac{n}{p(n) \cdot (1 - p(n))}}\right) \\
\le&\exp\left(-\frac{3}{16} c_1 \cdot n^{1-\alpha(n)} + (\ln n + \ln 3) \cdot \frac{2}{3} c_1 \cdot n^{\frac{1}{2} + \frac{1}{2}\alpha(n)}\right) && \big(\text{by Lemma 4.2}\big) \\
\le&\exp\left(-\frac{3}{16} c_1 \cdot n^{0.8} + (\ln n + \ln 3) \cdot \frac{2}{3} c_1 \cdot n^{0.6}\right) && \big(\alpha(n) \le 0.2 \text{ for } n \text{ sufficiently large}\big) \\
\le&\exp\left(-\frac{c_1}{16} \cdot n^{0.8}\right) && \big((\ln n + \ln 3) \cdot n^{0.6} = o\left(n^{0.8}\right)\big) \\
=&O\left(e^{-n^{0.8}}\right)
\end{aligned}$$

for $c_1 \ge 16$.

Therefore for $f_i \ge \frac{1}{c} \cdot k(n)$, the probability that there is some n_i, f_i, $\mathbf{F}_i$, and $\mathbf{K}_i$ with $\mathbf{X}$ exceeds $\mathbf{E}[\mathbf{X}] + t$ is at most $O(e^{-n}) + O\left(e^{-n^{0.8}}\right) = O\left(e^{-n^{0.8}}\right)$.

Combining $c_1 \ge 16$ with (3), we can take

$$c_1 := \max\left\{16, \left\lceil 2\sqrt{2(1 + \ln 3)(1 + c)} \right\rceil\right\}. \tag{4}$$

Note that c_1 depends on c, which may depends on c_1 in the following proofs. We have to wait till the proof of the main theorem to see that we can set c right to satisfy all the requirements for c and c_1. □

Definition 4.5. For our convenience, define the following random variables:

– define binary random variable $\mathbf{x}_i$ indicating whether the vertex removed in step i belongs to $\mathbf{F}_i$.

– define random variable $\mathbf{r}_i$ to be the first step r such that $\sum_{j=1}^{r} \mathbf{x}_j = i$.

– for a vertex $v \in \mathbf{K}$, define:

  – binary random variable $\mathbf{y}_i(v)$ indicating whether at step $\mathbf{r}_i$, the vertex removed is a neighbour of v;

  – random variable $\mathbf{r}$(v) to be the step where v is removed, or the last step if it is never removed.

  – *surplus* of v as $\mathbf{s}_i(v) = \left(\sum_{j=1}^{i} \mathbf{y}_j(v)\right) - i \cdot p(n)$ for $i \le \sum_{j=1}^{\mathbf{r}(v)} \mathbf{x}_j$.

Theorem 4.6. *For a sufficiently large constant $c_2 \in \mathbb{N}$, for an arbitrary vertex $v \in \mathbf{K}$, we have*

$$\mathbf{Pr}\left[\max_{i \le \sum_{j=1}^{\mathbf{r}(v)} \mathbf{x}_j} \mathbf{s}_i(v) \ge c_2 \sqrt{n \cdot p(n) \cdot (1 - p(n))}\right] \le \frac{1}{100}.$$

Proof: Consider the generation of $\mathbf{G}$ as the following random process: first generate randomly the clique $\mathbf{K}$ and all its edges to cover the vertex v, then add the remaining edges independently at probability $p(n)$ except for those connecting v and u for $u \in \mathbf{F}$. Those edges will be sampled dynamically on demand in the new stochastic algorithm, which is obtained by changing the line in the original algorithm that chooses the lowest degree vertex into the following iteration:

```
Iteration:
while true do
    Choose the vertex u with lowest degree and then lowest index to break ties
    if edge {u, v} is already sampled
        then break
    flip the biased coin to sample the edge {u, v}
    if {u, v} ∉ E
        then break
```

By Principle of Deferred Decision, it is easy to see that the above stochastic algorithm will have the same sequence of removed vertices as the original one upto step $\mathbf{r}(\mathrm{v})$.

In this stochastic algorithm, upto step $r \leq \mathbf{r}(v)$ we may have sample many edges $\{u, v\}$ for many $u \in \mathbf{F}$. The crucial points here are:

– each edge $\{u, v\}$ is independently sampled according to $B(p(n))$;

– whenever the vertex $u \in \mathbf{F}$ is removed, edge $\{u, v\}$ must already have been sampled.

Therefore $\mathbf{s}_i(v)$ is equivalent to $\sum_{j=1}^{i} \mathbf{p}_j(v)$ where $\mathbf{p}_j$'s are i.i.d. according to the following distribution:

$$\mathbf{p}_j = \begin{cases} 1 - p(n) & \text{with probability } p(n), \\ -p(n) & \text{with probability } 1 - p(n). \end{cases}$$

Hence

$$\begin{aligned} \mathbf{E}\left[\mathbf{p}_j\right] &= \big(1 - p(n)\big) \cdot p(n) - p(n) \cdot \big(1 - p(n)\big) = 0, \\ \mathbf{Var}\left[\mathbf{p}_j\right] &= \mathbf{E}\left[\mathbf{p}_j^2\right] - \mathbf{E}\left[\mathbf{p}_j\right]^2 = p(n) \cdot \big(1 - p(n)\big)^2 + \big(1 - p(n)\big) \cdot p(n)^2 = p(n)\big(1 - p(n)\big). \end{aligned}$$

Then by Theorem 1.3,

$$\begin{aligned} \mathbf{Pr}\left[\max_{i \leq \sum_{j=1}^{\mathbf{r}(v)} \mathbf{x}_j} \mathbf{s}_i(v) \geq c_2 \sqrt{n \cdot p(n) \cdot (1 - p(n))}\right] &\leq \mathbf{Pr}\left[\max_{1 \leq i \leq n} |\mathbf{s}_i(v)| \geq c_2 \sqrt{n \cdot p(n) \cdot (1 - p(n))}\right] \\ &\leq \frac{n \cdot p(n) \cdot (1 - p(n))}{\left(c_2 \sqrt{n \cdot p(n) \cdot (1 - p(n))}\right)^2} \\ &= \frac{1}{c_2^2} \leq \frac{1}{100} \end{aligned}$$

for $c_2 \geq 10$. □

Theorem 4.7. *For a sufficiently large constant $c_3 \in \mathbb{N}$, for an arbitrary vertex v,*

$$\mathbf{Pr}\left[\textit{initial degree of } v \leq n \cdot p(n) - c_3 \sqrt{n \cdot p(n) \cdot (1 - p(n))}\right] \leq \frac{1}{100}.$$

Proof: Let $\mathbf{X}$ denote the initial degree of v, then $\mathbf{X} \sim B\big(n, p(n)\big)$. We have $\mathbf{E}\left[\mathbf{X}\right] = n \cdot p(n)$ and $\mathbf{Var}\left[\mathbf{X}\right] = n \cdot p(n) \cdot (1 - p(n))$. Define $t := c_3 \sqrt{n \cdot p(n) \cdot (1 - p(n))}$ for some constant $c_3 \in \mathbb{N}$ to be determined. Hence by Chebyshev's inequality we have

$$\mathbf{Pr}\left[\mathbf{X} \leq \mathbf{E}\left[\mathbf{X}\right] - t\right] \leq \mathbf{Pr}\left[|\mathbf{X} - \mathbf{E}\left[\mathbf{X}\right]| \geq t\right] \leq \frac{\mathbf{Var}\left[\mathbf{X}\right]}{t^2} = \frac{1}{c_3^2} \leq \frac{1}{100}$$

for $c_3 \geq 10$. □

Theorem 4.8. *Assume the almost-surely-happening events in Theorem 4.3 and 4.4 happen, then:*

(1) if we have $f_r = 0$ and $k_r \geq \frac{4}{5} \cdot k(n)$, where step r is the last step of removal phase, then the inclusion phase will succeed;

(2) for c sufficiently large, if there is a step $\hat{r}$ such that $f_{\hat{r}} = \frac{1}{c} \cdot k(n)$ and $k_{\hat{r}} \geq \frac{4}{5} \cdot k(n)$, then at the end of removal phase we will have $f_r = 0$ and $k_r = k_{\hat{r}}$;

(3) for c sufficiently large, choose step $\hat{r}$ to be the first step with $f_{\hat{r}} = \frac{1}{c} \cdot k(n)$, then with high probability $\left(\text{at least } 1 - \frac{1}{5}\right)$ we have $k_{\hat{r}} \geq \frac{4}{5} \cdot k(n)$.

Proof:

(1) For every vertex in $\mathbf{K}_r$, its degree is $k_r \geq \frac{4}{5} \cdot k(n)$. For every vertex from $\mathbf{F}$, its degree in the graph in the inclusion phase is at most its degree in $\mathbf{G}'$, which is at most $k(n) \cdot p(n) + c_0\sqrt{k(n) \cdot \ln n}$ by Theorem 4.3. For n sufficiently large we have $\alpha(n) \leq \frac{1}{2}$, thus by Lemma 4.2 we have

$$\frac{\sqrt{k(n) \cdot \ln n}}{k(n)} = \sqrt{\frac{\ln n}{k(n)}} \leq \frac{\sqrt{\ln n}}{\sqrt[4]{\frac{1}{2} \cdot n^{1-\alpha(n)}}} \leq \frac{\sqrt{\ln n}}{\sqrt[4]{\frac{1}{2} \cdot n^{\frac{1}{8}}}} = o(1), \tag{5}$$

hence $k(n) = \omega\left(\sqrt{k(n) \cdot \ln n}\right)$. Again by Lemma 4.2 we have

$$\frac{4}{5} \cdot k(n) - p(n) \cdot k(n) \geq \left(\frac{4}{5} - \frac{1}{2}\right) \cdot k(n) = \omega\left(c_0\sqrt{k(n) \cdot \ln n}\right),$$

hence $k_r > k(n) \cdot p(n) + c_0\sqrt{k(n) \cdot \ln n}$. Thus in the inclusion phase, every candidate vertex in $\mathbf{F}$ connects to strictly less than k_r vertices, so they are not common neighbours of vertices in $\mathbf{G}_r$, which is exactly $\mathbf{K}_r$ by our assumption, hence they won't be added into the final solution. Obviously all the vertices in $\mathbf{K}$ will be recovered, hence the inclusion phase will succeed.

(2) By induction at step $i \geq \hat{r}$, with induction hypothesis that we have $k_i = k_{\hat{r}} \geq \frac{4}{5} \cdot k(n)$, which is immediately true for the basis case $i = \hat{r}$.

Assume that induction hypothesis is true for step i, then vertices in $\mathbf{K}_i$ have degrees at least $\frac{4}{5} \cdot k(n)$. By Theorem 4.3, Lemma 4.2 and our assumption about $f_{\hat{r}}$, vertex in $\mathbf{F}_i$ have degrees at most

$$f_i + k(n) \cdot p(n) + c_0\sqrt{k(n) \cdot \ln n} \leq f_{\hat{r}} + k(n) \cdot p(n) + o(k) \leq \left(\frac{1}{c} + \frac{1}{2} + \frac{1}{5}\right) \cdot k(n) < \frac{4}{5} \cdot k(n),$$

for sufficiently large c such that $\frac{1}{c} < \frac{1}{10}$, i.e. $c > 10$. Thus at step $i + 1$ one vertex in $\mathbf{F}_i$ will be removed, while k_i remains unchanged, i.e. $k_{i+1} = k_i = k_{\hat{r}} \geq \frac{4}{5} \cdot k(n)$.

(3) – Let $\mathbf{L}$ be the set of vertices v in $\mathbf{K}$ such that the initial degree of v is not larger than $n \cdot p(n) - c_3\sqrt{n \cdot p(n) \cdot (1 - p(n))}$. For each $v \in \mathbf{K}$ define

$$\mathbf{L}(v) := \mathbb{I}\left[\text{initial degree of } v \leq n \cdot p(n) - c_3\sqrt{n \cdot p(n) \cdot (1 - p(n))}\right],$$

then $\mathbf{E}\left[\mathbf{L}(v)\right] = \mathbf{Pr}\left[\text{initial degree of } v \leq n \cdot p(n) - c_3\sqrt{n \cdot p(n) \cdot (1 - p(n))}\right]$ and $\mathbf{E}\left[|\mathbf{L}|\right] =$

$\sum_{v\in\mathbf{K}} \mathbf{E}[\mathbf{L}(v)]$ by linearity of expectations. By Theorem 4.7 and Markov inequality, we have

$$\begin{aligned}
\mathbf{Pr}\left[|\mathbf{L}| \geq \frac{1}{10}\cdot k(n)\right] &\leq \frac{\mathbf{E}[|\mathbf{L}|]}{\frac{1}{10}\cdot k(n)} \\
&= \frac{\sum_{v\in\mathbf{K}} \mathbf{E}[\mathbf{L}(v)]}{\frac{1}{10}\cdot k(n)} \\
&= \frac{k(n)\cdot \mathbf{Pr}\left[\text{initial degree of } v \leq n\cdot p(n) - c_3\sqrt{n\cdot p(n)\cdot(1-p(n))}\right]}{\frac{1}{10}\cdot k(n)} \\
&\leq \frac{\frac{1}{100}\cdot k(n)}{\frac{1}{10}\cdot k(n)} \\
&= \frac{1}{10},
\end{aligned}$$

thus with high probability we have $|\mathbf{L}| \leq \frac{1}{10}\cdot k(n)$.

– Let $\mathbf{S}$ be the set of vertices v in $\mathbf{K}$ that ever suffer a surplus larger than $c_2\sqrt{n\cdot p(n)\cdot(1-p(n))}$ during the first $\hat{r}$ steps. Then by Theorem 4.6 and Markov inequality, similarly we have $\mathbf{Pr}\left[|\mathbf{S}| \geq \frac{1}{10}\cdot k(n)\right] \leq \frac{1}{10}$, thus with high probability we have $|\mathbf{S}| \leq \frac{1}{10}\cdot k(n)$.

Assume $|\mathbf{L}| \leq \frac{1}{10}\cdot k(n)$ and $|\mathbf{S}| \leq \frac{1}{10}\cdot k(n)$, then $|\mathbf{L}\cup\mathbf{S}| \leq \frac{1}{5}\cdot k(n)$ and $|\mathbf{K}\setminus(\mathbf{L}\cup\mathbf{S})| \geq \frac{4}{5}\cdot k(n)$. We can prove vertices in $\mathbf{K}\setminus(\mathbf{L}\cup\mathbf{S})$ won't be removed in the first $\hat{r}$ steps:

– for each vertex in $\mathbf{K}\setminus(\mathbf{L}\cup\mathbf{S})$, by Theorem 4.3, Lemma 4.2 and equation (5), its degree into $\mathbf{K}$ in $\mathbf{G}'$ is at most

$$k(n)\cdot p(n) + c_0\sqrt{k(n)\cdot\ln n} < \frac{1}{2}\cdot k(n)$$

hence its degree in $\mathbf{G}$ is at least

$$\begin{aligned}
&n\cdot p(n) - c_3\sqrt{n\cdot p(n)\cdot(1-p(n))} - \left(k(n)\cdot p(n) + c_0\sqrt{k(n)\cdot\ln n}\right) + k(n) \\
&\geq n\cdot p(n) - c_3\sqrt{n\cdot p(n)\cdot(1-p(n))} + \frac{1}{2}\cdot k(n) \\
&= n\cdot p(n) + \left(\frac{c}{2} - c_3\right)\sqrt{n\cdot p(n)\cdot(1-p(n))}.
\end{aligned}$$

In the calculation of this degree, we ignore all the edges bewteen $\mathbf{K}\setminus(\mathbf{L}\cup\mathbf{S})$ and $\mathbf{L}\cup\mathbf{S}$, hence it becomes at least

$$n\cdot p(n) + \left(\frac{c}{2} - c_3 - \frac{c}{5}\right)\sqrt{n\cdot p(n)\cdot(1-p(n))}. \tag{6}$$

– by induction at step $0 \leq i \leq \hat{r}$, with induction hypothesis that such vertex has degree at least

$$n_i\cdot p(n) + \left(\frac{c}{2} - c_3 - c_2 - \frac{c}{5}\right)\sqrt{n\cdot p(n)\cdot(1-p(n))},$$

and none of these vertices is deleted in the first i steps, which is clearly satisfied for the basis case $i = 0$ by the equation (6).

Assume that the induction hypothesis for step i is true. Then in step $i+1$, if we can choose sufficiently large c satisfying

$$\frac{c}{2} - c_3 - c_2 - \frac{c}{5} = \frac{3}{10}\cdot c - c_3 - c_2 > c_1,$$

then we have

$$n_i\cdot p(n) + \left(\frac{c}{2} - c_3 - c_2 - \frac{c}{5}\right)\sqrt{n\cdot p(n)\cdot(1-p(n))} > n_i\cdot p(n) + c_1\sqrt{n\cdot p(n)\cdot(1-p(n))}.$$

Then by Theorem 4.4 these vertices won't be removed at this step.

– If the vertex removed in this step is from $\mathbf{L} \cup \mathbf{S}$, since all the edges between $\mathbf{K} \setminus (\mathbf{L} \cup \mathbf{S})$ and $\mathbf{L} \cup \mathbf{S}$ have already been ignored, degree of $v \in \mathbf{K} \setminus (\mathbf{L} \cup \mathbf{S})$ remains unchanged.
– Otherwise the vertex removed in this step belongs to $\mathbf{F}$. Suppose it is the j-th vertex removed from $\mathbf{F}$. Obviously we have $j \leq i + 1$.
  Fix an abitrary vertex $v \in \mathbf{K} \setminus (\mathbf{L} \cup \mathbf{S})$, then we know $\mathbf{s}_j(v) \leq c_2 \sqrt{n \cdot p(n) \cdot (1 - p(n))}$.
  By Definition 4.5, the number of its neighbours in $\mathbf{F}$ that has been removed in the first $i + 1$ steps is $\mathbf{s}_j(v) + j \cdot p(n)$. Hence by equation (6) its degree is at least

$$
\begin{aligned}
& n \cdot p(n) + \left(\frac{c}{2} - c_3 - \frac{c}{5}\right) \sqrt{n \cdot p(n) \cdot (1 - p(n))} - (\mathbf{s}_j(v) + j \cdot p(n)) \\
&= (n - j) \cdot p(n) + \left(\frac{c}{2} - c_3 - \frac{c}{5}\right) \sqrt{n \cdot p(n) \cdot (1 - p(n))} - \mathbf{s}_j(v) \\
&\geq (n - (i + 1)) \cdot p(n) + \left(\frac{c}{2} - c_3 - c_2 - \frac{c}{5}\right) \sqrt{n \cdot p(n) \cdot (1 - p(n))} \\
&= n_{i+1} \cdot p(n) + \left(\frac{c}{2} - c_3 - c_2 - \frac{c}{5}\right) \sqrt{n \cdot p(n) \cdot (1 - p(n))}
\end{aligned}
$$

Therefore in G_{i+1} such vertex still has degree at least

$$
n_{i+1} \cdot p(n) + \left(\frac{c}{2} - c_3 - c_2 - \frac{c}{5}\right) \sqrt{n \cdot p(n) \cdot (1 - p(n))}.
$$

Note that equation (4) shows that $c_1 = \Theta(\sqrt{c})$, therefore we can always choose a sufficiently large c to satisfy requirements for c and c_1.

Therefore with high probability we have $k_{\hat{r}} = |\mathbf{K} \setminus (\mathbf{L} \cup \mathbf{S})| \geq \frac{4}{5} \cdot k(n)$. □

Proof: (of the main Theorem 3.1) The expected size of a maximum clique in $\mathbf{G}'$ is $O(t(n)) = o(\log n) = o\left(\sqrt{n \cdot p(n) \cdot (1 - p(n))}\right)$. Thus by Markov inequality almost surely there is no clique of size $\sqrt{n \cdot p(n) \cdot (1 - p(n))}$ in $\mathbf{G}'$, thus no clique of size $\sqrt{n \cdot p(n) \cdot (1 - p(n))}$ in $\mathbf{F}$. Therefore the removal phase won't stop before $f_{\hat{r}} = \frac{1}{c} \cdot k(n) \geq \sqrt{n \cdot p(n) \cdot (1 - p(n))}$. Then by the above theorem, the algorithm works. The total failure probability is $\frac{1}{10} + \frac{1}{10} + o(1) < \frac{1}{3}$. □