

Dense Subgraphs in Random Graphs

Paul Balister^{*}, Béla Bollobás[†], Julian Sahasrabudhe[‡], Alexander Veremyev[§]

January 19, 2019

Abstract

For a constant $\gamma \in [0, 1]$ and a graph G , let $\omega_\gamma(G)$ be the largest integer k for which there exists a k -vertex subgraph of G with at least $\gamma \binom{k}{2}$ edges. We show that if $0 < p < \gamma < 1$ then $\omega_\gamma(G_{n,p})$ is concentrated on a set of two integers. More precisely, with $\alpha(\gamma, p) = \gamma \log \frac{\gamma}{p} + (1 - \gamma) \log \frac{1 - \gamma}{1 - p}$, we show that $\omega_\gamma(G_{n,p})$ is one of the two integers closest to $\frac{2}{\alpha(\gamma, p)} (\log n - \log \log n + \log \frac{e\alpha(\gamma, p)}{2}) + \frac{1}{2}$, with high probability. While this situation parallels that of cliques in random graphs, a new technique is required to handle the more complicated ways in which these “quasi-cliques” may overlap.

1 Introduction

Let $G = (V(G), E(G))$ be a simple, undirected graph where $V(G)$ denotes the set of *vertices* of G (sometimes called *nodes*) and $E(G)$ denotes the set of *edges*. A graph G is said to be *complete* if all possible edges are present: if $\{i, j\} \in E(G)$ for all $i, j \in V(G)$, $i \neq j$. For a subset $S \subseteq V(G)$, we denote by $G[S]$ the subgraph of G *induced by* S : the graph with vertex set S and edge set $\{\{i, j\} : i, j \in S\} \cap E(G)$. A *clique* C is a subset of $V(G)$ for which $G[C]$ is a complete graph [25].

Cliques are an indispensable concept in the theory of graphs and have been extensively studied in various contexts, reaching back to the 1930s with the celebrated results of Ramsey [31] and Turán [36]. In random graphs, cliques have also been a central topic of study with

^{*}Department of Mathematical Sciences, University of Memphis, Memphis TN 38152, USA. Partially supported by NSF grant DMS 1600742

[†]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB3 0WB, UK *and* Department of Mathematical Sciences, University of Memphis, Memphis TN 38152, USA *and* London Institute for Mathematical Sciences, 35a South St., Mayfair, London W1K 2XF, UK. Partially supported by NSF grant DMS 1600742 and MULTIPLEX grant no. 317532.

[‡]Instituto Nacional de Matemática Pura e Aplicada (IMPA), Estr. Dona Castorina, 110 Jardim Botânico, Rio de Janeiro RJ 22460-320, Brasil *and* Peterhouse, University of Cambridge, Cambridge CB2 1RD, UK

[§]Industrial Engineering & Management Systems, University of Central Florida, Orlando, Florida, USA. Supported, in part, by the AFRL Mathematical Modeling and Optimization Institute

its roots in the pioneering work of Erdős [11] on the probabilistic method. The term *clique*, however, was introduced more recently in sociometry [25] to model cohesive subgroups of tightly knit elements in a graph [8]. For example, in social networks, where vertices correspond to “actors” and edges indicate relationships between actors [39], a clique represents a group of people any two of which have a certain kind of relationship (friendship, acquaintance, etc.) with each other [27]. Some of the earliest work on cliques, in the context of sociometry, is presented in [25, 24, 14].

However, in some applications the clique is not the correct concept; often we do not care that *all* edges are present in a particular subset, but only that the set is very “well connected”, in some appropriate sense. Consequently, a number of relaxations of the notion of “clique” have appeared in the literature in recent years [30, 22].

One of the most popular and widely used clique relaxation models is the γ -*quasi-clique*, where $\gamma \in [0, 1]$ is a parameter [2]. In particular, for $\gamma \in [0, 1]$, we say that a subset $S \subseteq V(G)$ of a graph G is a γ -quasi-clique if the graph $G[S]$, induced by S , has at least $\gamma \binom{|S|}{2}$ edges.

This concept was first defined by Abello, Pardalos & Resende [1] who were interested in quasi-cliques in graphs representing telecommunications data. Later, the idea of “dense clusters” (a more general concept which includes γ -quasi-cliques) was studied in the context of molecular interaction networks described by Hartwell, Leland, Hopfield, John, Leibler, Stanislas, Murray and Andrew [15] and further analyzed by Spirin and Mirny [34]. They reported that dense subgraphs in molecular interaction networks correspond to meaningful modules or building blocks of molecular networks such as protein complexes or dynamic functional units. The problem of finding large dense subgraphs have also appeared in a number of other domains including biology [3, 9, 18, 4], social network analysis [10, 23, 39], finance [5, 19, 32, 33] and data mining [28, 35].

Given the myriad of instances for which the notion is useful, one would like to efficiently compute solutions to basic questions about quasi-cliques in a given graph: for example, “what is the largest γ -quasi clique in a given graph G ”? However, it comes as no surprise that the computational problem of finding the largest quasi-clique in a given graph (along with many other such questions) is a hard computational problem in general [29, 17] – similar to the sister problem of finding large cliques in graphs [21, 16]. Moreover, the literature on *exact* computational methods for this class of problems is extremely sparse and mostly focuses on the development and application of heuristic methods. The currently available exact methods allow one to explicitly compute the maximum γ -quasi-clique size only in relatively small and sparse graphs. Hence, studying the typical behavior of quasi-cliques in the binomial random graph might be useful for evaluating the performance and scalability of heuristic methods.

To this end, we study the order of the largest γ -quasi-clique in the binomial random graph, a project initiated in a paper of Veremyev, Boginski, Krokhmal and Jeffcoat [37], where the upper bound on the largest γ -quasi-clique is presented. For a graph G , we let $\omega_\gamma(G)$ be the size of the largest subset of vertices of G that induces a γ -quasi clique. Of course, $\omega_1(G)$ is the classical “clique number” of G , often denoted by $\omega(G)$.

We strengthen the result of the aforementioned paper and prove that $\omega_\gamma(G_{n,p})$ is concentrated on two explicitly determined points, with high probability as $n \rightarrow \infty$, provided

$0 < p < \gamma < 1$ are fixed real numbers. See Section 2 for a more careful statement of this result.

Although these bounds are asymptotic, computational experiments suggest that they are quite accurate even for relatively small ($n = 50, 100$) graphs generated using the $G_{n,p}$ model. For the results of these experiments, see Section 5.

It has recently been drawn to our attention that the behavior of a closely related graph parameter has been studied by Kang and McDiarmid [20], the so-called *t-dependence number* of a graph G , which is defined as the largest subset of vertices of G that induces a graph with maximum degree t . A little thought reveals that the complement of a t -dependent set on k vertices is a γ -quasi-clique with $\gamma = 1 - t/(k - 1)$. One consequence of their results is that the largest βn -dependent set in $G_{n,p}$ is equal to $(\kappa(\beta, p) + o_n(1)) \log n$, where $0 < p < \beta < 1$ are fixed and $\kappa(\cdot, \cdot)$ is an implicitly defined function. This implies (after some minor adjustments) the existence of the asymptotic constant in our main result. Later, Fountoulakis, Kang and McDiarmid [12] went on to give more precise results in the binomial random graph for much smaller values of t , a regime which roughly corresponds to very dense quasi-cliques.

2 Notation and statement of the main result

As usual write $[n]$ for the set $\{1, \dots, n\}$ and $G_{n,p}$ for the binomial random graph on vertex set $[n]$ with edge probability $p \in (0, 1)$. We use the notation $O_n(1)$ to denote a quantity that is bounded by a constant as n tends to infinity and we use $o_n(1)$ to denote a quantity that tends to zero as n tends to infinity. We say that a sequence of events E_n holds *with high probability* (henceforth whp) if $\mathbb{P}(E_n) = 1 - o_n(1)$. For a graph G we let $e(G)$ denote the number of edges in the graph.

A complete subgraph on k vertices will be called a *k-clique*, and we define the *clique number* $\omega(G)$ of a graph G to be the largest integer k for which G contains a k -clique. The study of the clique number of $G_{n,p}$ was first carefully considered by Matula [26], who noticed that the clique number of $G_{n,p}$ is concentrated on a small set of values. These results were later strengthened by Grimmett and McDiarmid [13] and then Bollobás and Erdős [7], who showed that for fixed $0 \leq p \leq 1$ the clique number takes one of only *two values*, whp (See also Theorem 11.1 in [6]). We prove that a similar phenomena persists for γ -quasi-cliques. However, a significant difficulty arises when controlling the concentration of the count of γ -quasi cliques directly. We tackle this issue by instead controlling a closely related random variable, which is more naturally handled.

We call n -vertex graph a γ -quasi-clique if $e(G) \geq \gamma \binom{n}{2}$. For a graph G , we define $\omega_\gamma(G)$ to be the largest integer k for which there exists a γ -quasi-clique subgraph of order k . For $0 < p < \gamma < 1$, we show that $\omega_\gamma(G_{n,p})$ is concentrated on two points whp as $n \rightarrow \infty$.

Theorem 1. *Let $0 < p < \gamma < 1$ and $\varepsilon > 0$ be fixed and define*

$$\alpha(\gamma, p) := \gamma \log \frac{\gamma}{p} + (1 - \gamma) \log \frac{1 - \gamma}{1 - p}.$$

Then

$$\omega_\gamma(G_{n,p}) - \frac{2}{\alpha(\gamma,p)} \left(\log n - \log \log n + \log \frac{e\alpha(\gamma,p)}{2} \right) \in (-\varepsilon, 1 + \varepsilon),$$

whp. In particular, $\omega_\gamma(G_{n,p})$ is one of the two integers closest to

$$\frac{2}{\alpha(\gamma,p)} \left(\log n - \log \log n + \log \frac{e\alpha(\gamma,p)}{2} \right) + \frac{1}{2},$$

whp.

As usual, the binary entropy function for $\gamma \in (0, 1)$ is

$$h(\gamma) := \gamma \log \frac{1}{\gamma} + (1 - \gamma) \log \frac{1}{1 - \gamma}.$$

We use the following consequence of Stirling's formula. If $\gamma \in (0, 1)$ is fixed, we have

$$\binom{n}{\gamma n + O_n(1)} = e^{nh(\gamma) - \frac{1}{2} \log(n\gamma(1-\gamma)) + O_n(1)}. \quad (1)$$

We first set out to give an upper bound for $\omega_\gamma(G_{n,p})$ which holds with high probability. Let $X_k = X_{k,\gamma}(G_{n,p})$ be the random variable which counts the number of subgraphs of $G_{n,p}$ that are γ -quasi-cliques on k vertices. We easily obtain an upper bound on $\omega_\gamma(G_{n,p})$ by bounding $\mathbb{E}X_k$. In preparation, we state a basic fact about binomial random variables.

Lemma 2. *Let $0 < p < \gamma < 1$ be fixed and $N \rightarrow \infty$. We have*

$$\mathbb{P}(\text{Bin}(N, p) = \lceil \gamma N \rceil) = e^{-N\alpha(\gamma,p) + O(\log N)},$$

and

$$\mathbb{P}(\text{Bin}(N, p) \geq \gamma N) = e^{-N\alpha(\gamma,p) + O(\log N)}.$$

Proof. We have

$$\begin{aligned} \mathbb{P}(\text{Bin}(N, p) = \lceil \gamma N \rceil) &= \binom{N}{\lceil \gamma N \rceil} p^{\lceil \gamma N \rceil} (1-p)^{\lfloor (1-\gamma)N \rfloor} \\ &= e^{Nh(\gamma) - \frac{1}{2} \log(N\gamma(1-\gamma)) + N(\gamma \log p + (1-\gamma) \log(1-p)) + O(\log \frac{p}{1-p})} \\ &= e^{-N\alpha(\gamma,p) + O(\log N)}. \end{aligned}$$

The second result follows as, for $r \geq \gamma N > pN$, $\mathbb{P}(\text{Bin}(N, p) = r)$ is decreasing in r , and hence

$$\mathbb{P}(\text{Bin}(N, p) = \lceil \gamma N \rceil) \leq \mathbb{P}(\text{Bin}(N, p) \geq \gamma N) \leq (N+1)\mathbb{P}(\text{Bin}(N, p) = \lceil \gamma N \rceil).$$

□

We now may establish an upper bound on $\omega_\gamma(G_{n,p})$, that holds whp, thus proving one of the inequalities implicit in the statement of Theorem 1. In the following sections, we go on to show that the distribution of quasi-cliques (actually a subclass of these quasi-cliques) is sufficiently concentrated to prove Theorem 1.

Lemma 3. *Let $0 < p < \gamma < 1$ and $\varepsilon > 0$ be fixed. Then as $n \rightarrow \infty$*

$$\omega_\gamma(G_{n,p}) < \frac{2}{\alpha(\gamma, p)}(\log n - \log \log n + \log \frac{e \cdot \alpha(\gamma, p)}{2}) + 1 + \varepsilon,$$

whp.

Proof. With $X_k = X_{k,\gamma}(G_{n,p})$ and $S = \binom{k}{2}$ we have

$$\begin{aligned} \mathbb{E}X_k &= \binom{n}{k} \mathbb{P}(\text{Bin}(S, p) \geq \gamma S) \\ &\leq \frac{n^k}{k!} e^{-S\alpha(\gamma, p) + O(\log S)} \\ &= e^{k(\log n - \frac{\alpha(\gamma, p)(k-1)}{2} - \log(k/e) + o_k(1))} \end{aligned}$$

Let $\kappa = \frac{2}{\alpha(\gamma, p)}(\log n - \log \log n + \log \frac{e \cdot \alpha(\gamma, p)}{2}) + 1 + \varepsilon$. If $k = \lceil \kappa \rceil$ then

$$\log n - \frac{\alpha(\gamma, p)(k-1)}{2} - \log(k/e) + o_k(1) < -\frac{\varepsilon \cdot \alpha(\gamma, p)}{2} + o_k(1)$$

is negative for large enough n , and hence the expectation must tend to zero. Thus we have $\mathbb{P}(X_k > 0) \leq \mathbb{E}X_k = o_n(1)$. The existence of a γ -quasi-clique on $j > k$ vertices implies, by a simple averaging argument, that there exists a γ -quasi-clique subgraph on k vertices. Thus if $X_k = 0$ then $X_j = 0$ for all $j > k$. Hence $\omega_\gamma(G_{n,p}) < \kappa$ with high probability. \square

3 γ -flat subgraphs

To show that $G_{n,p}$ contains a γ -quasi-clique of order roughly $\frac{2}{\alpha(\gamma, p)} \log n$ whp, we count a slightly restricted class of subgraphs. The advantage of working with this restricted class is that the second moment of their count is controlled more naturally. Roughly speaking, we say that a γ -quasi-clique G is γ -flat if every induced subgraph of G is close to being a γ -quasi clique.

To make this definition precise, we need a few definitions. First, for a graph G and a subset A of the vertex set of G , let us define $e(A)$ to be the number of edges with both end-points in A .

Now, for $\gamma \in (0, 1)$ and $\ell \in [k]$, we define $S = \binom{k}{2}$, $T = \binom{\ell}{2}$, and set

$$D_k(\ell) = \min(T, S - T) \ell^{-1/2} \log k.$$

Call an k -vertex graph G γ -flat if $e(G) = \lceil \gamma \binom{k}{2} \rceil$ and for all $A \subseteq V(G)$ with $\ell = |A| \in [2, k-1]$, we have $e(A) \leq \gamma \binom{\ell}{2} + D_k(\ell)$. We note that $\min(T, S - T)$ is clearly an upper bound on $e(A) - \gamma \binom{\ell}{2}$ when $e(G) = \lceil \gamma \binom{k}{2} \rceil$, so this is only a restriction on $e(A)$ when $|A| = \ell > (\log k)^2$.

We shall show that if a subset of k vertices in $G_{n,p}$ has $\lceil \gamma \binom{k}{2} \rceil$ edges then it is reasonably likely that it will also be γ -flat, and hence the two notions are “typically” interchangeable. For positive integers n, m , $0 \leq m \leq \binom{n}{2}$, we define the *Erdős-Rényi random graph* $G(n, m)$ as the uniform probability space that is supported on all n vertex graphs with exactly m edges.

Lemma 4. *Let $G = G(k, \lceil \gamma \binom{k}{2} \rceil)$ and let γ be fixed and $k \rightarrow \infty$. Then G is γ -flat with high probability.*

Proof. Let $G = G(k, \lceil \gamma \binom{k}{2} \rceil)$ be realized on the vertex set $[k]$ and fix a subset $A \subseteq [k]$ with $\ell = |A| \in [2, k-1]$. We shall show that

$$\binom{k}{\ell} \mathbb{P}(e(A) \geq \gamma \binom{\ell}{2} + D_k(\ell)) \leq k^{-2}. \quad (2)$$

Set $S = \binom{k}{2}$, $T = \binom{\ell}{2}$, $R = S - T$, and put $C(L) = \mathbb{P}(e(A) = L)$. Note that

$$C(L) = \binom{T}{L} \binom{R}{\lceil \gamma S \rceil - L} \binom{S}{\lceil \gamma S \rceil}^{-1}$$

and for $0 \leq L < T$, we have

$$Q(L) := \frac{C(L+1)}{C(L)} = \frac{T-L}{L+1} \left(\frac{\lceil \gamma S \rceil - L}{R - \lceil \gamma S \rceil + L + 1} \right). \quad (3)$$

From (3) we see that $Q(L)$ is strictly decreasing as L increases. Let $L = \lceil \gamma T \rceil + r \leq T$. Then if $r \geq 0$,

$$\begin{aligned} Q(L) &\leq Q(\gamma T + r) \leq \left(\frac{(1-\gamma)T - r}{\gamma T + r + 1} \right) \left(\frac{\gamma R + 1 - r}{(1-\gamma)R + r} \right) \\ &= \left(\frac{1 - \frac{r}{(1-\gamma)T}}{1 + \frac{r+1}{\gamma T}} \right) \left(\frac{1 - \frac{r-1}{\gamma R}}{1 + \frac{r}{(1-\gamma)R}} \right) \\ &\leq \min \left\{ 1 - \frac{r}{(1-\gamma)T}, 1 - \frac{r-1}{\gamma R} \right\} \leq e^{-\frac{c(r-1)}{\min(R, T)}}, \end{aligned}$$

where $c = 1/\max(\gamma, 1-\gamma) > 0$ is a constant. Hence

$$C(L) = C(\lceil \gamma T \rceil + r) \leq C(\lceil \gamma T \rceil + 1) \prod_{s=1}^r e^{-\frac{c(s-1)}{\min(R, T)}} \leq e^{-\frac{cr(r-1)}{2\min(R, T)}}, \quad (4)$$

where we have used the (trivial) fact that $C(\lceil \gamma T \rceil + 1) \leq 1$. Now $T\ell^{-1/2} \log k \geq \frac{1}{2} \log k$ and $R\ell^{-1/2} \log k \geq (k-1)^{1/2} \log k$, so $D_\ell(k) \rightarrow \infty$ uniformly in ℓ as $k \rightarrow \infty$. Thus for large k we have $cr(r-1)/(2\min(R, T)) \geq c' \min(R, T)\ell^{-1}(\log k)^2$ for some $c' > 0$ when $r > D_k(\ell) - 1$. Hence

$$\begin{aligned} \mathbb{P}(e(A) \geq \gamma T + D_k(\ell)) &= \sum_{\gamma T + D_k(\ell) \leq L \leq T} C(L) \\ &\leq \ell^2 e^{-c' \min(R, T)\ell^{-1}(\log k)^2} \end{aligned} \quad (5)$$

for large enough k .

Consider the case when $R < T$. Then $R = (k - \ell)(k + \ell - 1)/2 > \ell(k - \ell)/2$ and so $c' \min(R, T)\ell^{-1}(\log k)^2 > 5(k - \ell) \log k \geq (k - \ell) \log k + 4 \log k$ for large enough k . Now $\binom{k}{\ell} = \binom{k}{k-\ell} \leq k^{k-\ell}$, so

$$\binom{k}{\ell} \mathbb{P}(e(A) \geq \gamma \binom{\ell}{2} + D_k(\ell)) \leq \binom{k}{\ell} k^2 e^{-(k-\ell) \log k - 4 \log k} \leq k^{-2},$$

as required. Now suppose $R \geq T$. Then $c' \min(R, T)\ell^{-1}(\log k)^2 > 3\ell \log k \geq \ell \log k + 4 \log k$ when k is large enough. Now $\binom{k}{\ell} \leq k^\ell$, so

$$\binom{k}{\ell} \mathbb{P}(e(A) \geq \gamma \binom{\ell}{2} + D_k(\ell)) \leq \binom{k}{\ell} k^2 e^{-\ell \log k - 4 \log k} \leq k^{-2},$$

as required. Hence (2) holds for all $\ell \in [2, k-1]$.

Now, for $2 \leq \ell \leq k-1$, let Y_ℓ be the random variable counting the number of subsets A of order ℓ which induce more than $\gamma \binom{\ell}{2} + D_k(\ell)$ edges. By (2) we have

$$\mathbb{P}(Y_\ell > 0) \leq \mathbb{E}(Y_\ell) \leq \binom{k}{\ell} \mathbb{P}(e(A) \geq \gamma \binom{\ell}{2} + D_k(\ell)) \leq k^{-2},$$

for large enough k . So the probability that $Y_\ell > 0$ for any of the $< k$ choices for ℓ is at most $k^{-1} = o_k(1)$. \square

Let $Z_k = Z_{k,n}$ be the random variable counting the number of copies of γ -flat subgraphs of order k in $G_{n,p}$, with p fixed and $n \rightarrow \infty$. We now easily bound $\mathbb{E}Z_k$, by using Lemma 4, to relate it to the quantity $\mathbb{E}X_k$.

Lemma 5. *Let $\varepsilon > 0$ and $k \leq \frac{2}{\alpha(\gamma, p)}(\log n - \log \log n + \log \frac{e \cdot \alpha(\gamma, p)}{2}) + 1 - \varepsilon$ with $k \rightarrow \infty$ as $n \rightarrow \infty$. Then $\mathbb{E}Z_k \rightarrow \infty$.*

Proof. We apply Lemma 4 to deduce that

$$\mathbb{E}Z_k = \binom{n}{k} \mathbb{P}(G(k, p) \text{ is } \gamma\text{-flat}) \geq (1 + o_k(1)) \binom{n}{k} \mathbb{P}(e(G(k, p)) = \lceil \gamma S \rceil).$$

Now $k = O(\log n)$ by assumption, so $\binom{n}{k} = \frac{n^k}{k!}(1 - O(k^2/n)) = (1 + o_k(1))\frac{n^k}{k!}$. Hence

$$\begin{aligned}\mathbb{E}Z_k &= (1 + o_k(1))\frac{n^k}{k!}\mathbb{P}(\text{Bin}(S, p) = \lceil \gamma S \rceil) \\ &= \frac{n^k}{k!}e^{-S\alpha(\gamma, p) + O(\log S)} \\ &= e^{k(\log n - (k-1)\alpha(\gamma, p)/2 - \log(k/e) + o_k(1))}.\end{aligned}$$

However, the exponent in the last line tends to infinity when $k \rightarrow \infty$ and $k \leq \frac{2}{\alpha(\gamma, p)}(\log n - \log \log n + \log \frac{e\alpha(\gamma, p)}{2}) + 1 - \varepsilon$. \square

In the next section we turn to estimate the variance of Z_k .

4 The second moment

To prove our lower bound on $\omega_\gamma(G_{n,p})$, we count the number of γ -flat subsets of order k in $G_{n,p}$, where k is roughly $\frac{2}{\alpha(\gamma, p)} \log n$. For $k \in [n]$, recall that Z_k is the random variable which counts the number of γ -flat subsets of $G(n, p)$. To apply Chebyshev's inequality, we aim to estimate the fraction

$$F = \frac{\text{Var} Z_k}{(\mathbb{E} Z_k)^2} = \frac{\mathbb{E} Z_k^2 - (\mathbb{E} Z_k)^2}{(\mathbb{E} Z_k)^2}. \quad (6)$$

In particular, we shall show $F = o(1)$, as both k and n tend to infinity. Let $A, B \subseteq [n]$ with $|A| = |B| = k$ and $|A \cap B| = \ell$. We think of $\ell \in [2, k-1]$ and treat the degenerate cases $\ell \in \{0, 1, k\}$ separately. Put $S = \binom{k}{2}$, $T = \binom{\ell}{2}$, $R = S - T$ and let $g_\ell(L)$ denote the probability that $e(A) = \lceil \gamma S \rceil$, $e(B) = \lceil \gamma S \rceil$ and $e(A \cap B) = L$. We note that

$$g_\ell(L) = \binom{T}{L} \binom{R}{\lceil \gamma S \rceil - L}^2 p^{2\lceil \gamma S \rceil - L} (1-p)^{2\lfloor (1-\gamma)S \rfloor - T + L}$$

and consider the ratio

$$\begin{aligned}R_\ell(L) &= \frac{g_\ell(L)}{\mathbb{P}(e(A) = \lceil \gamma S \rceil)^2} \\ &= \binom{T}{L} \binom{R}{\lceil \gamma S \rceil - L}^2 \binom{S}{\lceil \gamma S \rceil}^{-2} p^{-L} (1-p)^{L-T}.\end{aligned} \quad (7)$$

The following lemma gives us a suitable way of estimating the quantity $R_\ell(L)$, for our purposes. For the remainder of the section, we maintain the assumption that $0 < p < \gamma \leq 1$ and that $k \rightarrow \infty$.

Lemma 6. *Let $2 \leq \ell \leq k-1$, $r \geq 0$ be an integer and set $\lambda = 2 \cdot \frac{\gamma}{1-\gamma} \frac{1-p}{p}$. Then*

$$R_\ell(\lfloor \gamma T \rfloor + r) \leq \lambda^r e^{T\alpha(\gamma, p) + O_k(1)} \quad (8)$$

and

$$R_\ell(\lfloor \gamma T \rfloor - r) \leq R_\ell(\lfloor \gamma T \rfloor). \quad (9)$$

Proof. We first bound $R_\ell(\lfloor \gamma T \rfloor)$. Note that $R \geq k - 1$ and hence $R^2 T = R^2(S - R) \geq S^2$ for $2 \leq \ell \leq k - 1$. Since $S = \binom{k}{2} \rightarrow \infty$ and γ is fixed, we may bound line (7) by using equation (1), to obtain

$$\begin{aligned} R_\ell(\lfloor \gamma T \rfloor) &= e^{Th(\gamma) + 2(S-T)h(\gamma) - 2Sh(\gamma) + \frac{1}{2} \log \frac{S^2}{\gamma(1-\gamma)R^2 T} + O_k(1)} p^{-\lfloor \gamma T \rfloor} (1-p)^{-\lceil (1-\gamma)T \rceil} \\ &\leq e^{-Th(\gamma) + O_k(1)} p^{-\lfloor \gamma T \rfloor} (1-p)^{-\lceil (1-\gamma)T \rceil} \\ &= e^{T\alpha(\gamma, p) + O_k(1)}. \end{aligned}$$

Now put $C(L) = R_\ell(L + 1)/R_\ell(L)$ and observe that $C(L)$ can be written as

$$\frac{1-p}{p} \cdot \frac{T-L}{L+1} \left(\frac{\lceil \gamma S \rceil - L}{R - \lceil \gamma S \rceil + L + 1} \right)^2.$$

From this expression, we see that $C(L)$ strictly decreases as L increases and therefore

$$\begin{aligned} C(\lfloor \gamma T \rfloor + r) &\leq C(\lfloor \gamma T \rfloor) \\ &\leq \frac{1-p}{p} \cdot \frac{\gamma}{1-\gamma} \left(1 + \frac{1}{\gamma R} \right). \end{aligned}$$

Now note that since $\ell < k$, by assumption, we have that $R \geq k - 1$ and thus R tends to infinity with k . Hence, for large k ,

$$C(\lfloor \gamma T \rfloor + r) \leq 2 \frac{1-p}{p} \frac{\gamma}{1-\gamma} = \lambda.$$

We now apply this inequality r times to obtain

$$R_\ell(\lfloor \gamma T \rfloor + r) \leq \lambda^r R_\ell(\lfloor \gamma T \rfloor),$$

which holds for k sufficiently large, but independently of r . This proves the inequality (8).

To prove the inequality (9) we note that $C(L)$ is strictly decreasing and

$$\begin{aligned} C(\lfloor \gamma T \rfloor - 1) &\geq \frac{1-p}{p} \frac{\gamma}{1-\gamma} \cdot \left(1 + \frac{1}{(1-\gamma)T} \right) \left(1 + \frac{1}{\gamma T} \right) \\ &\geq \frac{1-p}{p} \frac{\gamma}{1-\gamma} > 1. \end{aligned}$$

Thus $R_\ell(\lfloor \gamma T \rfloor - r) \leq R_\ell(\lfloor \gamma T \rfloor)$. □

We are now in a position to show that $F = o(1)$ as n and k tend to infinity.

Lemma 7. *Let $k \leq \frac{2}{\alpha(\gamma, p)}(\log n - \log \log n + \log \frac{e\alpha(\gamma, p)}{2}) + 1 - \varepsilon$. We have*

$$\mathbb{E}Z_k^2 = (1 + o_k(1))(\mathbb{E}Z_k)^2.$$

Proof. We consider the fraction F , from equation (6). We keep with the convention that $S = \binom{k}{2}$, $T = \binom{\ell}{2}$, and $R = S - T$. Let E_A and E_B denote the events that A , resp. B , induces a γ -flat subgraph. Let E'_A , resp. E'_B , denote the event that A , resp. B , induce exactly $\lceil \gamma S \rceil$ edges. Note that $E_A \subseteq E'_A$ and $E_B \subseteq E'_B$. Now write $t(\ell) = t_{n,k}(\ell) = \binom{k}{\ell} \binom{n-k}{k-\ell} \binom{n}{k}^{-1}$. We now turn to bound F . We may expand Z_k as a sum of indicators

$$Z_k = \sum_{A \subset V(G), |A|=k} \mathbf{1}(E_A)$$

Hence $\mathbb{E}Z_k = \binom{n}{k} \mathbb{P}(E_A)$ thus

$$F = \frac{\mathbb{E}Z_k^2 - (\mathbb{E}Z_k)^2}{\mathbb{E}Z_k} = \sum_{A,B} \binom{n}{k}^{-2} \frac{\mathbb{P}(E_A \cap E_B) - \mathbb{P}(E_A)\mathbb{P}(E_B)}{\mathbb{P}(E_A)^2}.$$

We now divide the sum with respect to $|A \cap B| = \ell$ to obtain

$$\begin{aligned} F &= \sum_{\ell=0}^k \binom{n}{k} \binom{k}{\ell} \binom{n-k}{k-\ell} \binom{n}{k}^{-2} \frac{\mathbb{P}(E_A \cap E_B) - \mathbb{P}(E_A)\mathbb{P}(E_B)}{\mathbb{P}(E_A)\mathbb{P}(E_B)} \\ &= \sum_{\ell=2}^{k-1} t(\ell) \cdot \frac{\mathbb{P}(E_A \cap E_B) - \mathbb{P}(E_A)^2}{\mathbb{P}(E_A)^2} + o_n(1), \end{aligned} \quad (10)$$

where we have eliminated the first two terms in the above sum as E_A and E_B are independent events when $|A \cap B| \leq 1$. We have also eliminated the last term in the sum, i.e. when $E_A = E_B$. This is justified, as this term is at most $((\binom{n}{k} \mathbb{P}(E_A)))^{-1} = (\mathbb{E}Z_k)^{-1} = o_k(1)$, by Lemma 5. Let us denote the ℓ th term in the sum at (10) as $F(\ell)$.

Lemma 4 implies that

$$\frac{\mathbb{P}(E_A \cap E_B) - \mathbb{P}(E_A)^2}{\mathbb{P}(E_A)^2} \leq (1 + o_k(1)) \frac{\mathbb{P}(E_A \cap E_B)}{\mathbb{P}(E'_A)^2}.$$

For $\ell \in [2, k-1]$, our “flatness condition” on subsets of A applies and hence

$$\begin{aligned} \mathbb{P}(E_A \cap E_B) / \mathbb{P}(E'_A)^2 &= \mathbb{P}(E'_A)^{-2} \sum_{0 \leq L \leq \gamma T + D_k(\ell)} \mathbb{P}(E_A \cap E_B \mid e(A \cap B) = L) \mathbb{P}(e(A \cap B) = L) \\ &\leq \mathbb{P}(E'_A)^{-2} \sum_{0 \leq L \leq \gamma T + D_k(\ell)} \mathbb{P}(E'_A \cap E'_B \mid e(A \cap B) = L) \mathbb{P}(e(A \cap B) = L) \\ &= \sum_{0 \leq L \leq \gamma T + D_k(\ell)} R_\ell(L) \\ &= \sum_{0 \leq L < \gamma T} R_\ell(L) + \sum_{\gamma T \leq L \leq \gamma T + D_k(\ell)} R_\ell(L) \\ &\leq T \lambda^{D_k(\ell)} e^{T\alpha(\gamma, p) + O_k(1)}. \end{aligned}$$

This last inequality follows from applying the inequality (8) (from Lemma 6) to each term in the right sum and applying the inequality (9) (again from Lemma 6) to the left sum. So we may bound the ℓ th term in the sum (10) as

$$F(\ell) \leq t(\ell) T \lambda^{D_k(\ell)} e^{T\alpha(\gamma,p)+O_k(1)}.$$

We first consider the case when $R < T$. Write $\delta := k - \ell$. Now

$$t(\ell) = \binom{k}{\delta} \binom{n-k}{\delta} \binom{n}{k}^{-1} \leq (kn)^\delta \binom{n}{k}^{-1}$$

and $\mathbb{E}Z_k = \binom{n}{k} e^{-S\alpha(\gamma,p)+O(\log k)} \rightarrow \infty$. Also $D_k(\ell) = R\ell^{-1/2} \log k = o_k(R)$ as $R < T$ implies $\ell \geq k/2$. Thus

$$F(\ell) \mathbb{E}Z_k \leq e^{\delta \log(kn) - R\alpha(\gamma,p) + o_k(R)}$$

But $R = \delta(k + \ell - 1)/2 \geq 2k\delta/3$ and $k\alpha(\gamma,p) \sim 2 \log n$. Thus $F(\ell) \leq (\mathbb{E}Z_k)^{-1} e^{-(\frac{1}{3} - o_k(1))\delta \log n}$. In particular, $\sum_{\ell: R < T} F(\ell) = o(1)$.

Now consider the case when $R \geq T$. In this case we use the bound $t(\ell) \leq (1 + o_k(1))(k^2/n)^\ell$ to deduce that

$$F(\ell) \leq e^{T\alpha(\gamma,p) + \ell \log(k^2/n) + O_k(T\ell^{-1/2} \log k) + O_k(\log k)} = e^{\ell((\ell-1)\alpha(\gamma,p)/2 - \log(n) + O_k(k^{1/2} \log k))}.$$

Now $k = O(\log n)$ and $\ell < 3k/4$. Thus $(\ell-1)\alpha(\gamma,p)/2 \leq (3/4 + o_k(1)) \log n$. Hence $F(\ell) \leq e^{-(1/4 - o(1)) \log n}$ and so $\sum_{\ell: R \geq T} F(\ell) = o(1)$. □

After these preparations, it is only a small step to finish the proof of Theorem 1.

Proof of Theorem 1. Let $\varepsilon > 0$ be given. The upper bound on $\omega_\gamma(G_{n,p})$ follows from Lemma 3. For the lower bound, assume $k \leq \frac{2}{\alpha(\gamma,p)}(\log n - \log \log n + \log \frac{e\alpha(\gamma,p)}{2}) + 1 - \varepsilon$. From Lemma 5 we know that $\mathbb{E}Z_k \rightarrow \infty$, so for sufficiently large n we have $\mathbb{E}Z_k > 0$ and thus we may apply Chebyshev's inequality to show that the quantity $\mathbb{P}(X_k = 0)$ is small. We have

$$\mathbb{P}(X_k = 0) \leq \mathbb{P}(Z_k = 0) \leq \mathbb{P}(|Z_k - \mathbb{E}Z_k| \geq \mathbb{E}Z_k) \leq \text{Var}(Z_k)/\mathbb{E}(Z_k)^2 = F = o(1),$$

where we have used the fact that every γ -flat set is a γ -quasi-clique for the first inequality. The third inequality is Chebyshev's inequality and the bound on F is the content of Lemma 7. □

5 Computational Experiments

Here, we note the bounds obtained from Theorem 1 are actually quite accurate in practice, even for relatively small values of n . To illustrate, we performed a small set of computational

γ	ω_{min}^γ	ω_{max}^γ	ω_{avg}^γ	ω_{th}^γ	γ	ω_{min}^γ	ω_{max}^γ	ω_{avg}^γ	ω_{th}^γ	γ	ω_{min}^γ	ω_{max}^γ	ω_{avg}^γ	ω_{th}^γ
$n = 50$														
$p = 0.20$					$p = 0.15$					$p = 0.10$				
0.9	4	5	4.95	5.72	0.9	3	5	4.12	5.06	0.9	3	5	3.27	4.39
0.8	5	7	6.01	6.92	0.8	4	6	5.19	6.03	0.8	3	5	4.28	5.14
0.7	6	8	7.2	8.44	0.7	5	8	6.02	7.26	0.7	3	6	5.05	6.09
0.6	8	11	9.48	10.41	0.6	6	10	7.62	8.87	0.6	5	8	6.15	7.34
0.5	10	15	12.58	12.64	0.5	8	12	9.85	10.99	0.5	6	10	7.8	9.05
$n = 100$														
$p = 0.15$					$p = 0.10$					$p = 0.05$				
0.9	4	5	4.98	5.82	0.9	3	6	4.41	4.99	0.8	3	5	4.1	4.73
0.85	4	6	5.6	6.4	0.8	5	7	5.23	5.92	0.6	5	7	5.72	6.65
0.8	6	7	6.21	7.04	0.7	5	8	6.11	7.12	0.4	7	11	9.06	10.56
0.75	6	8	6.95	7.78	0.6	7	10	7.74	8.75	0.3	11	16	12.77	14.44

Table 1: Largest quasi-cliques in graphs generated according to $G_{n,p}$ model. For each n, p , the minimum ω_{min}^γ , maximum ω_{max}^γ and average ω_{avg}^γ cardinalities of the largest quasi-cliques identified in 100 instances are reported. These values are compared against the values given by the formula for ω_{th}^γ at (11).

experiments for graphs of size $n = 50$ and $n = 100$ and different values of p . For each pair n, p we generated 100 instances of graphs sampled according to the corresponding $G_{n,p}$ model. We have also selected various values of γ ranging from 0.3 to 0.9.

For each γ, n, p in Table 1 we report the minimum ω_{min}^γ , maximum ω_{max}^γ and average ω_{avg}^γ cardinalities of the largest γ -quasi-cliques and compare this to ω_{th}^γ , the “theoretical” value obtained from the formula in Theorem 1. That is,

$$\omega_{th}^\gamma(n) = \frac{2}{\alpha(\gamma, p)} \left(\log n - \log \log n + \log \frac{e\alpha(\gamma, p)}{2} \right) + \frac{1}{2}. \quad (11)$$

Observe that the obtained formula provides an accurate estimate of γ -quasi-clique number $\omega_\gamma(G)$ in graph instances generated according to the binomial random graph $G_{n,p}$, even for relatively small values of n .

To identify the largest γ -quasi-clique in these experiments, we used the so-called feasibility check version of formulation **F4** in [38] (or **AlgF4**). Previous experimental work has suggested this algorithm to be the best performing on instances generated from $G_{n,p}$.

References

- [1] J. Abello, P.M. Pardalos, and M.G.C. Resende. On maximum clique problems in very large graphs. In J. Abello and J. Vitter, editors, *External Memory Algorithms and Visualization*, pages 119–130. American Mathematical Society, Boston, 1999.

- [2] J. Abello, M. G. C. Resende, and S. Sudarsky. Massive quasi-clique detection. In S. Rajsbaum, editor, *LATIN 2002: Theoretical Informatics*, pages 598–612, London, 2002. Springer-Verlag.
- [3] G. D. Bader and C. W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.
- [4] S. Bastkowski, V. Moulton, A. Spillner, and T. Wu. The minimum evolution problem is hard: a link between tree inference and graph clustering problems. *Bioinformatics*, page 623, 2015.
- [5] V. Boginski, S. Butenko, and P. M. Pardalos. Statistical analysis of financial networks. *Computational Statistics & Data Analysis*, 48(2):431–443, 2005.
- [6] B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.
- [7] B. Bollobás and P. Erdős. Cliques in random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 80:419–427, 1976.
- [8] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo. The maximum clique problem. In *Handbook of Combinatorial Optimization*, volume 4, pages 1–74. Kluwer Academic Publishers, 1999.
- [9] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9):2443–2450, May 2003.
- [10] M. A. Crenson. Social networks and political processes in urban neighborhoods. *American Journal of Political Science*, 22(3):578–594, 1978.
- [11] P. Erdős. Some remarks on the theory of graphs. *Bull. Amer. Math. Soc.*, 53:292–294, 1947.
- [12] Nikolaos Fountoulakis, Ross J. Kang, and Colin McDiarmid. Largest sparse subgraphs of random graphs. *European Journal of Combinatorics*, 35:232 – 244, 2014. Selected Papers of EuroComb’11.
- [13] G.R. Grimmett and C.J.H. McDiarmid. On colouring random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 77:313–324, 1975.
- [14] F. Harary and I. C. Ross. A procedure for clique detection using the group matrix. *Sociometry*, 20:205–215, 1957.
- [15] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.

- [16] J. Håstad. Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Mathematica*, 182(1):105–142, 1999.
- [17] Klaus Holzapfel, Sven Kosub, Moritz G Maaß, and Hanjo Täubig. The complexity of detecting fixed-density clusters. *Discrete Applied Mathematics*, 154(11):1547–1562, 2006.
- [18] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21(suppl 1):i213–i221, 2005.
- [19] W Q. Huang, X T. Zhuang, and S. Yao. A network analysis of the Chinese stock market. *Physica A: Statistical Mechanics and its Applications*, 388(14):2956 – 2964, 2009.
- [20] Ross J Kang and Colin McDiarmid. The t-improper chromatic number of random graphs. *Combinatorics, Probability and Computing*, 19(1):87–98, 2010.
- [21] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103, New York, New York, USA, 1972. Plenum.
- [22] C. Komusiewicz. Multivariate algorithmics for finding cohesive subnetworks. *Algorithms*, 9(1):21, 2016.
- [23] P. Lee and L.V.S. Lakshmanan. Query-driven maximum quasi-clique search. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 522–530. SIAM, 2016.
- [24] R. D. Luce. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15(2):169–190, 1950.
- [25] R.D. Luce and A.D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.
- [26] D.W. Matula. *Combinatory Mathematics and its Applications*. Chapel Hill, North Carolina, 1972.
- [27] R. J. Mokken. Cliques, clubs and clans. *Quality and Quantity*, 13(2):161–173, 1979.
- [28] T. Nguyen, H. W. Lauw, and P. Tsaparas. Micro-review synthesis for multi-entity summarization. *Data Mining and Knowledge Discovery*, pages 1–29, 2017.
- [29] J. Pattillo, A. Veremyev, S. Butenko, and V. Boginski. On the maximum quasi-clique problem. *Discrete Applied Mathematics*, 161(1–2):244 – 257, 2013.
- [30] J. Pattillo, N. Youssef, and S. Butenko. On clique relaxation models in network analysis. *European Journal of Operational Research*, 226(1):9–18, 2013.

- [31] F. P. Ramsey. On a problem of formal logic. *Proc. London Math. Soc.*, 30:264–286, 1930.
- [32] D. Saban, F. Bonomo, and N. E. Stier-Moses. Analysis and models of bilateral investment treaties using a social networks approach. *Physica A: Statistical Mechanics and its Applications*, 389(17):3661–3673, 2010.
- [33] K. Sim, J. Li, V. Gopalkrishnan, and G. Liu. Mining maximal quasi-bicliques to co-cluster stocks and financial ratios for value investment. In *Proceedings of the Sixth International Conference on Data Mining, ICDM '06*, pages 1059–1063, Washington, DC, USA, 2006. IEEE Computer Society.
- [34] V. Spirin and L.A. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.
- [35] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 104–112. ACM, 2013.
- [36] P. Turán. On a extremal problem in graph theory. *Matematikai és Fizikai Lapok*, 48:436–452, 1941.
- [37] A. Veremyev, V. Boginski, P. A. Krokhmal, and D. E. Jeffcoat. Dense percolation in large-scale mean-field random networks is provably “explosive”. *PloS one*, 7(12):e51883, 2012.
- [38] A. Veremyev, O. A. Prokopyev, S. Butenko, and E. L. Pasiliao. Exact mip-based approaches for finding maximum quasi-cliques and dense subgraphs. *Computational Optimization and Applications*, 64(1):177–214, 2016.
- [39] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, New York, 1994.