Learning a Hidden Subgraph

Noga Alon * Vera Asodi †

Abstract

We consider the problem of learning a labeled graph from a given family of graphs on n vertices in a model where the only allowed operation is to query whether a set of vertices induces an edge. Questions of this type are motivated by problems in molecular biology. In the deterministic nonadaptive setting, we prove nearly matching upper and lower bounds for the minimum possible number of queries required when the family is the family of all stars of a given size or all cliques of a given size. We further describe some bounds that apply to general graphs.

1 Introduction

Let $\mathcal{H}$ be a family of labeled graphs on the set $V = \{1, 2, \ldots, n\}$, and suppose $\mathcal{H}$ is closed under isomorphism. Given a hidden copy of some $H \in \mathcal{H}$, we have to identify it by asking queries of the following form. For $F \subseteq V$, the query Q_F is: does F contain at least one edge of H? Our objective is to identify H by asking as few queries as possible. We say that a family $\mathcal{F}$ solves the $\mathcal{H}$-problem if for any two distinct members H_1 and H_2 of $\mathcal{H}$, there is at least one $F \in \mathcal{F}$ that contains an edge of one of the graphs H_i and does not contain any edge of the other. Obviously, any such family enables us to learn an unknown member of $\mathcal{H}$ deterministically and non-adaptively, by asking the questions Q_F for each $F \in \mathcal{F}$. Note that for any family $\mathcal{H}$, the set of all pairs of vertices solves the $\mathcal{H}$-problem. Note also that the information theoretic lower bound implies that we need at least $\log |\mathcal{H}|$ queries, where here and throughout the paper, all logarithms are in base 2, unless otherwise specified, and we omit all floor and ceiling signs, when these are not crucial.

There are some families of graphs for which the above problem has been studied, motivated by applications in molecular biology. These include matchings ([1]) and Hamiltonian cycles ([5, 6]). The biological problem is to find, given a set of molecules, pairs that react with each other. Here the vertices correspond to the molecules, the edges to the reactions, and the queries correspond to experiments of putting a set of molecules together in a test tube and determining whether a reaction occurs. The problem of finding a hidden matching is the one encountered by molecular biologists

*Department of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Research supported in part by a USA-Israeli BSF grant, by the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University. Email: nogaa@post.tau.ac.il.

†Department of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Email: veraa@post.tau.ac.il.

when they apply multiplex PCR in order to close the gaps left in a DNA strand after shotgun sequencing. See [1] and its references for more details.

The previous works in this field study the minimum number of queries needed to identify a hidden graph, from various families of graphs. Some of these works consider different query models than the one described above. The authors of [1] study the hidden subgraph problem for the family of matchings. In that paper it is shown that under the deterministic and non-adaptive model, the minimum number of queries that one has to ask in order to identify a hidden matching is $\Theta(n^2)$, that is, one can do better than the trivial algorithm of asking all pairs only by a constant factor. It is also proved that $\Omega(n^2)$ queries are needed in order to find a hidden copy of any bounded-degree graph with a linear size matching. The authors further present randomized non-adaptive algorithms that use $\Theta(n \log n)$ random queries, and deterministic k-round algorithms, that ask $O(n^{1+1/(2(k-1))}\text{polylog}n)$ queries. Grebinski and Kucherov [5, 6] study the family of Hamiltonian cycles. A few query models are discussed in those papers. Besides the model presented above, they consider the additive model, in which the answer to a query is not just "yes" or "no" but the number of edges in the subset. Both models are considered also when the size of the queries is bounded. They present matching lower and upper bounds under each of these models, where some of the upper bounds are achieved by 2-round algorithms, and the other algorithms are fully adaptive. In [7], Grebinski and Kucherov study the problem for low degree graphs, and prove matching lower and upper bounds under the additive non-adaptive model.

In the present paper we consider only the deterministic non-adaptive model, where the answers are only "yes" or "no". The main families considered are families of stars and families of cliques. We study both families of stars or cliques of a given size, and the families of all cliques or all stars. It is shown that the trivial upper bound of $\binom{n}{2}$ is tight up to a $1 + o(1)$-multiplicative term for the families of stars of k edges, for all $n^{\frac{2}{3}} \log^{\frac{2}{3}} n << k \leq n - 2$. For smaller stars, we show that less queries suffice, and give upper and lower bounds on the minimum number of queries needed. These bounds are tight up to some polylog$n$ factor for all sizes of stars, and they are of order k^3, up to the polylog$n$ factors. We show that the problem is easier when the hidden subgraph is a clique. In fact, even for the family of all cliques, the problem can be solved using $O(n \log^2 n)$ queries. We study, as in the case of stars, the problem of a hidden clique of size k, for all values of k. In all cases, we prove upper and lower bounds that are tight up to some polylog$n$ factor, and are of order k^2, up to the polylog$n$ factors. We also consider the case where the family of graphs consists of all the graphs isomorphic to a given general graph G, and give a lower bound that depends on the maximum size of an independent set in G. From this general bound, we obtain a lower bound of $\Omega(\frac{n^2}{\log^2 n})$ for the random graph $G(n, \frac{1}{2})$.

In Section 2 we study the hidden subgraph problem where the hidden graph is a star, in Section 3 we consider the case where the hidden graph is a clique, and in Section 4 we prove a result for general graphs. Section 5 contains some concluding remarks and open problems.

2 Stars

In this section we consider the case where the graphs in $\mathcal{H}$ are stars. Denote by $\mathcal{S}_k$ the family of all graphs on $V = \{1, 2, \ldots, n\}$ that consist of a copy of $K_{1,k}$ and $n-k-1$ isolated vertices. Let $\mathcal{S} = \cup_{k=1}^{n-1} \mathcal{S}_k$. We begin with the following simple claim.

Proposition 2.1 *The minimum size of a family $\mathcal{F}$ that solves the $\mathcal{S}$-problem is exactly $\binom{n}{2}$.*

Proof: We show that any family $\mathcal{F}$ that solves the $\mathcal{S}$-problem must contain all pairs of vertices. Let u and v be two distinct vertices in V. Let S_1 be the star whose center is u and whose leaves are all other vertices of V, and let S_2 be the star whose center is u, and whose leaves are all other vertices of V except for v. Clearly, the answer to a query Q_F where F does not contain u is "no" for both S_1 and S_2, and the answer to a query Q_F with F containing u and another vertex of both stars is "yes" in both cases. Therefore $\mathcal{F}$ must contain the query $\{u, v\}$, or otherwise it cannot distinguish between S_1 and S_2. □

Note that the proof actually shows that even the solution of the $\mathcal{S}_{n-2} \cup \mathcal{S}_{n-1}$-problem requires $\binom{n}{2}$ queries. We now consider the case where the size of the star is known, and prove the following theorem, which gives lower and upper bounds on the minimum size of a family that solves the $\mathcal{S}_k$-problem. These bounds are tight in all cases up to some polylog$n$ factor.

Theorem 2.2 *For all $k \leq n-2$ and $n > 2$, there exists a family of size $min(\lceil \frac{n(n-2)}{2} \rceil, O(k^3 \log n))$ that solves the $\mathcal{S}_k$-problem, and every family that solves the $\mathcal{S}_k$-problem either contains $(1-o(1))\binom{n}{2}$ pairs, or it is of size at least $\Omega(\frac{k^3}{\log^2 n})$. Moreover, if $k \leq \sqrt{n}$, then the size of any family that solves the $\mathcal{S}_k$-problem is at least $\Omega(\frac{k^3 \log n}{\log k})$. For $k = n-1$, the minimum size of such a family is exactly $\lceil \log n \rceil$.*

The best bounds we get, for various values of k, are summarized in Table 1. In the rest of this section we prove these results.

Proposition 2.3 *For all $n > 2$, the minimum size of a family $\mathcal{F}$ that solves the $\mathcal{S}_{n-1}$-problem is exactly $\lceil \log n \rceil$.*

Proof: The family $\mathcal{S}_{n-1}$ is of size n, so we clearly need at least $\lceil \log n \rceil$ queries. To prove that $\lceil \log n \rceil$ queries suffice, we construct the following family that solves the $\mathcal{S}_{n-1}$-problem. Note that we only need to identify the center of the star. Assign distinct vectors of length $\lceil \log n \rceil$ over $\{0, 1\}$ to the vertices in V. For all $1 \leq i \leq \lceil \log n \rceil$, let F_i be the set of all vertices whose i^{th} bit is 1, and let $\mathcal{F} = \{F_i \mid 1 \leq i \leq \lceil \log n \rceil\}$. The answer to a query Q_{F_i} is "yes" if and only if F_i contains the center. Thus, for all i, we can obtain the i^{th} bit of the center from the answer to Q_{F_i}. □

Proposition 2.4 *The minimum size of a family $\mathcal{F}$ that solves the $\mathcal{S}_{n-2}$-problem is 2 for $n = 3$, it is 5 for $n = 4$, and $\lceil \frac{n(n-2)}{2} \rceil$ for all $n \geq 5$*

Table 1: Bounds on the size of a family that solves the $\mathcal{S}_k$-problem.

k	Lower bound	Upper bound
$k \leq \sqrt{n}$	$\Omega(\frac{k^3 \log n}{\log k})$	$O(k^3 \log n)$
$\sqrt{n} < k < \frac{n^{2/3}}{(2\log n)^{1/3}}$	$\Omega(\frac{k^3}{\log^2 n})$	$O(k^3 \log n)$
$\frac{n^{2/3}}{(2\log n)^{1/3}} \leq k \leq O(n^{2/3}\log^{2/3} n)$	$\Omega(\frac{k^3}{\log^2 n})$	$\lceil \frac{n(n-2)}{2} \rceil$
$k = \omega(n^{2/3}\log^{2/3} n)$, $k \leq n-3$	$(1-o(1))\binom{n}{2}$	$\lceil \frac{n(n-2)}{2} \rceil$
$k = n-2$	$\lceil \frac{n(n-2)}{2} \rceil$	$\lceil \frac{n(n-2)}{2} \rceil$
$k = n-1$	$\lceil \log n \rceil$	$\lceil \log n \rceil$

Proof: Let $\mathcal{F}$ be a family that solves the $\mathcal{S}_{n-2}$-problem. Let u, v and w be three distinct vertices in V. Let S_1 be the star whose center is u and in which the isolated vertex is v, and let S_2 be the star whose center is u, and the isolated vertex is w. The only sets that distinguish between S_1 and S_2 are $\{u, v\}$ and $\{u, w\}$, hence $\mathcal{F}$ must contain at least one of them. Thus $\mathcal{F}$ must contain all pairs of vertices but a matching, and hence $|\mathcal{F}| \geq \lceil \frac{n(n-2)}{2} \rceil$. Moreover, it is easy to check directly that for $n = 3$, 2 pairs are necessary and suffice, and so are 5 pairs for $n = 4$.

On the other hand, assume $n \geq 5$, take a maximum matching M on V, and let $\mathcal{F}$ be the family of all pairs of vertices besides those in M. Since $|M| = \lfloor \frac{n}{2} \rfloor$, $|\mathcal{F}| = \lceil \frac{n(n-2)}{2} \rceil$. All the edges but those of M are obtained directly from the queries. Since $n \geq 5$, the center of the star can be identified by these edges, and then the only edge that may be in the graph and was not asked is the edge in M incident with the center, if there is such an edge in M. We can now decide whether this edge exists or not by the size of the star that was found. □

Note that the above upper bound holds for all $\mathcal{S}_k$, where $3 \leq k \leq n-2$.

We now give some general upper and lower bounds on the minimum size of a family that solves the $\mathcal{S}_k$-problem. These bounds are tight up to some polylog$n$ factor. From now on we assume, throughout the section, that n is large.

Proposition 2.5 *For every k, there exists a family $\mathcal{F}$ of size $O(k^3 \log n)$ that solves the $\mathcal{S}_k$-problem.*

Proof: Let $m = ck^3 \log n$ for some absolute constant c, and let $F_1, F_2, ..., F_m$ be m random subsets of V, chosen independently as follows. For every F_i, every $v \in V$ is chosen to be in F_i independently with probability $\frac{1}{k}$. Let S_1 and S_2 be two stars of size k, such that $|E(S_1) \setminus E(S_2)| = |E(S_2) \setminus E(S_1)| = 1$. Let u be the center of S_1 and S_2, let $u_1, ..., u_{k-1}$ be the other common vertices, and let v be the additional vertex of S_1, and w the additional vertex of S_2. F_i distinguishes between S_1 and S_2 if and only if $u \in F_i$, $u_j \notin F_i$ for all $1 \leq j \leq k-1$, and exactly one vertex among v and w is in F_i. Thus

the probability that F_i distinguishes between S_1 and S_2 is

$$\frac{2}{k^2}\left(1-\frac{1}{k}\right)^k = \Omega\left(\frac{1}{k^2}\right).$$

Therefore, the probability that no F_i distinguishes between S_1 and S_2 is

$$\left[1-\Omega\left(\frac{1}{k^2}\right)\right]^m < n^{-(2k+2)}$$

provided c is sufficiently large. For two stars that differ in more edges, this probability is smaller. The number of pairs of stars is smaller than n^{2k+2}, and hence, there is a family $\mathcal{F} = \{F_1, F_2, ..., F_m\}$ that solves the $\mathcal{S}_k$-problem. □

We show that the upper bound given in Proposition 2.5 is tight up to a factor of polylog$n$. More precisely, we show that for every $k \leq n-2$, a family $\mathcal{F}$ that solves the $\mathcal{S}_k$-problem either contains $(1-o(1))\binom{n}{2}$ pairs, or it is of cardinality $\Omega(\frac{k^3}{\text{polylog}n})$.

Proposition 2.6 *For every $k \leq n-2$, if $\mathcal{F}$ is a family that solves the $\mathcal{S}_k$-problem, then $\mathcal{F}$ either contains $(1-o(1))\binom{n}{2}$ pairs, or it is of cardinality at least $\Omega(\frac{k^3}{\log^2 n})$.*

Proof: Let $\mathcal{F}$ be a family that solves the $\mathcal{S}_k$-problem. Then, for every $u \in V$ and $A, B \subseteq V \setminus \{u\}$ such that $|A| = 2$, $|B| = k-1$ and $A \cap B = \emptyset$, there exists a set $F \in \mathcal{F}$ such that $u \in F$, $|F \cap A| = 1$ and $F \cap B = \emptyset$. Indeed, otherwise $\mathcal{F}$ would not distinguish between the two stars whose center is u, which share the vertices of B, and where the additional vertex of one star is one vertex of A, and the additional vertex of the other one is the other vertex of A. Denote by $\mathcal{F}_0$ the family of all sets $F \in \mathcal{F}$ of size 2. Let $m = c \cdot \frac{n \log n}{k}$, and define $\mathcal{F}_1 = \{F \in \mathcal{F} \mid 2 < |F| \leq m\}$, and $\mathcal{F}_2 = \mathcal{F} \setminus (\mathcal{F}_0 \cup \mathcal{F}_1)$. We show that for any constant $\epsilon > 0$, if $|\mathcal{F}_0| \leq (1-\epsilon)\binom{n}{2}$ then $|\mathcal{F}_1 \cup \mathcal{F}_2| > c_1\epsilon^3 \cdot \frac{k^3}{\log^2 n}$ for some constant c_1 that depends only on c. Suppose $|\mathcal{F}_0| \leq (1-\epsilon)\binom{n}{2}$ and $|\mathcal{F}_1 \cup \mathcal{F}_2| \leq c_1\epsilon^3 \cdot \frac{k^3}{\log^2 n}$. For every $u \in V$, denote by V_u the set of vertices $v \in V \setminus \{u\}$ such that $\{u, v\} \notin \mathcal{F}_0$. Let $V' = \{u \in V \mid |V_u| \geq \frac{\epsilon}{2}(n-1)\}$. Since $|\mathcal{F}_0| \leq (1-\epsilon)\binom{n}{2}$, $|V'| \geq \frac{\epsilon}{2}n$. Otherwise, since the pairs of vertices that are not in $\mathcal{F}_0$ are pairs $\{u, v\}$ such that $v \in V_u$, and since $v \in V_u$ if and only if $u \in V_v$, we have

$$\begin{aligned}
|\mathcal{F}_0| &= \binom{n}{2} - \frac{1}{2}\sum_{u \in V}|V_u| \\
&= \binom{n}{2} - \frac{1}{2}\left(\sum_{u \in V'}|V_u| + \sum_{u \in V \setminus V'}|V_u|\right) \\
&> \binom{n}{2} - \frac{1}{2}\left[|V'|(n-1) + |V \setminus V'|\frac{\epsilon}{2}(n-1)\right] \\
&> \binom{n}{2} - \frac{1}{2}\left[\frac{\epsilon}{2}n(n-1) + n\frac{\epsilon}{2}(n-1)\right] \\
&= (1-\epsilon)\binom{n}{2}.
\end{aligned}$$

Choose uniformly a vertex $u \in V'$, and then choose uniformly a subset $A = \{v, w\} \subseteq V_u$. Define $\mathcal{F}_1' = \{F \in \mathcal{F}_1 \mid u \in F, |F \cap A| = 1\}$. For each $F \in \mathcal{F}_1$

$$Pr(F \in \mathcal{F}_1') \leq \frac{|F|(|F|-1)(n-|F|)}{\frac{\epsilon}{2} n \binom{\frac{\epsilon}{2}(n-1)}{2}} \leq \frac{16}{\epsilon^3} \cdot \frac{|F|^2}{n^2}.$$

Therefore,

$$E[|\mathcal{F}_1'|] \leq \frac{16}{\epsilon^3} \sum_{F \in \mathcal{F}_1} \frac{|F|^2}{n^2} \leq \frac{16}{\epsilon^3} |\mathcal{F}_1| \frac{m^2}{n^2},$$

and hence, there is a choice of u and A such that

$$\begin{aligned} |\mathcal{F}_1'| &\leq \frac{16}{\epsilon^3} |\mathcal{F}_1| \frac{m^2}{n^2} \\ &\leq 16 c_1 c^2 \cdot \frac{k^3}{\log^2 n} \cdot \frac{n^2 \log^2 n}{k^2 n^2} \\ &\leq \frac{k}{2} - 1 \end{aligned}$$

provided $c_1 c^2$ is sufficiently small. Thus, there exists a subset $B_1 \subseteq V \setminus (\{u\} \cup A)$ of size $\frac{k}{2} - 1$ that intersects every $F \in \mathcal{F}_1'$. Choose a random subset $B_2 \subseteq V$ of size $\frac{k}{2}$. For every $F \in \mathcal{F}_2$

$$\begin{aligned} Pr(F \cap B_2 = \emptyset) &= \frac{\binom{n-|F|}{\frac{k}{2}}}{\binom{n}{\frac{k}{2}}} \\ &\leq \left(1 - \frac{|F|}{n}\right)^{\frac{k}{2}} \\ &\leq e^{-\frac{km}{2n}} \\ &= n^{-c_2} \end{aligned}$$

for some constant $c_2 = \Theta(c)$. Therefore, if c is sufficiently large, with high probability, $u \notin B_2$, $A \cap B_2 = \emptyset$ and $\forall F \in \mathcal{F}_2 \;\; F \cap B_2 \neq \emptyset$. Denote $B' = B_1 \cup B_2$. $B' \subseteq V \setminus (\{u\} \cup A)$ and $|B'| \leq k-1$. Let B be an arbitrary extension of B' to a subset of $V \setminus (\{u\} \cup A)$ of size $k-1$. Consider the following two stars S_1 and S_2; u is the center of S_1 and S_2, they share the vertices of B, the additional vertex of S_1 is v, and the additional vertex of S_2 is w. Since A was chosen from V_u, the pairs $\{u, v\}$ and $\{u, w\}$ are not in $\mathcal{F}_0$, and thus no set in $\mathcal{F}_0$ can distinguish between S_1 and S_2. Neither can the sets in $\mathcal{F}_1$ that do not contain u, nor those whose intersection with A is not of size 1. All other sets in $\mathcal{F}_1$, i.e. sets $F \in \mathcal{F}_1$ such that $u \in F$, and $|F \cap A| = 1$, and all the sets in $\mathcal{F}_2$, contain a vertex of B, so they cannot distinguish between these two stars either. Thus $\mathcal{F}$ cannot distinguish between S_1 and S_2, contradicting the assumption that it solves the $\mathcal{S}_k$-problem. □

We now prove a better lower bound for $k \leq \sqrt{n}$. This bound is tight up to a factor of $\log k$. For the proof of the this bound, we need a variant of a lemma proved in [8, 4].

Definition 2.7 *Let $\mathcal{A}$ be a family of subsets of a set S. We say that $\mathcal{A}$ is* r-cover-free *if no set in $\mathcal{A}$ is contained in the union of any r other sets in $\mathcal{A}$.*

Lemma 2.8 *Let S be a set of size m, and let $\mathcal{A}$ be a family of n subsets of S. Suppose $\mathcal{A}$ is r-cover-free, where $r \leq 2\sqrt{n}$. Then,*

$$m > \frac{r^2 \log(n - \frac{r}{2})}{10 \log r}.$$

In [8], it is proved that for fixed r and large $n \geq n(r)$, $\frac{\log n}{m} \leq 8 \cdot \frac{\log r}{r^2}$. A simple modification of that proof described below shows that the lemma as stated above holds for every $r \leq 2\sqrt{n}$.

Proof: Let S and $\mathcal{A}$ be as defined in the lemma, and suppose that $m \leq \frac{r^2 \log(n-\frac{r}{2})}{10 \log r}$. As long as $\mathcal{A}$ contains a set A of size greater than $\frac{2m}{r}$, we remove A from $\mathcal{A}$, and its members from S and from all other sets in $\mathcal{A}$. Since $|S| = m$, this process stops after at most $\frac{r}{2}$ steps. Thus, we now have a subset S' of S, and a family $\mathcal{A}'$ of subsets of S', where each subset is of size at most $\frac{2m}{r}$. Denote by m' the size of S', and by n' the size of $\mathcal{A}'$. Clearly, $n' \geq n - \frac{r}{2}$. No set $A \in \mathcal{A}'$ is contained in the union of $\frac{r}{2}$ others, or otherwise, the original set from which A was obtained, would be contained in the union of these $\frac{r}{2}$ sets and the sets that were removed. Thus, every set in $\mathcal{A}'$ has a subset of size at most $\lceil \frac{4m}{r^2} \rceil$ that is not contained in any other set in $\mathcal{A}'$. Otherwise, if there were a set $A \in \mathcal{A}'$ for which the above did not hold, then, since $|A| \leq \frac{2m}{r}$, the set A would have been covered by $\frac{r}{2}$ other sets in $\mathcal{A}'$, which is impossible. Therefore, there are n' distinct sets of size at most $\lceil \frac{4m}{r^2} \rceil$. Thus,

$$n' \leq \binom{m'}{\lceil \frac{4m}{r^2} \rceil}.$$

If $\frac{4m}{r^2} < 1$ then the right hand side of this inequality is m', and thus we have $n' \leq m'$, and hence $n \leq m$, contradicting the assumption that

$$m \leq \frac{r^2 \log(n - \frac{r}{2})}{10 \log r} \leq \frac{4n}{5}(1 + o(1)).$$

Thus, $\frac{4m}{r^2} \geq 1$, and we have

$$n - \frac{r}{2} \leq n' \leq \binom{m'}{\lceil \frac{4m}{r^2} \rceil} \leq \binom{m}{\lceil \frac{4m}{r^2} \rceil} < 2^{\frac{10m \log r}{r^2}},$$

and hence

$$m > \frac{r^2 \log(n - \frac{r}{2})}{10 \log r},$$

contradicting the assumption. □

We use the above lemma to improve the lower bound, for $k \leq \sqrt{n}$.

Proposition 2.9 *For every $k \leq \sqrt{n}$, if $\mathcal{F}$ is a family the solves the $\mathcal{S}_k$-problem, then $|\mathcal{F}| = \Omega(\frac{k^3 \log n}{\log k})$.*

Proof: Let $\mathcal{F}$ be a family that solves the $\mathcal{S}_k$-problem. Choose, randomly, $A, B \subseteq V$, such that $|A| = 2$, $|B| = \frac{k}{2} - 1$, and $A \cap B = \emptyset$. Define $\mathcal{G} = \{F \in \mathcal{F} \mid |F \cap A| = 1, F \cap B = \emptyset\}$. Clearly,

$$\begin{aligned} Pr(F \in \mathcal{G}) &= \frac{|F|(n-|F|)}{\binom{n}{2}} \frac{\binom{n-|F|-1}{\frac{k}{2}-1}}{\binom{n-2}{\frac{k}{2}-1}} \\ &= \frac{2|F|}{n} \frac{\binom{n-|F|}{\frac{k}{2}}}{\binom{n-1}{\frac{k}{2}}} \\ &\leq \frac{2|F|}{n} \left(1 - \frac{|F|-1}{n-1}\right)^{\frac{k}{2}} \\ &\leq \frac{2|F|}{n} e^{-\frac{k|F|}{4n}}. \end{aligned}$$

If $|F| \leq \frac{4n}{k}$ then

$$Pr(F \in \mathcal{G}) \leq \frac{8}{k}.$$

If $|F| > \frac{4n}{k}$, denote $x = \frac{k|F|}{4n}$. Since $x > 1$ we have

$$Pr(F \in \mathcal{G}) \leq \frac{8}{k} x e^{-x} < \frac{8}{ek}.$$

Hence, for all F,

$$Pr(F \in \mathcal{G}) \leq \frac{c}{k}$$

for some constant c, and thus the expected size of $\mathcal{G}$ is $c \cdot \frac{|\mathcal{F}|}{k}$. Therefore, there exists a choice of A and B for which $|\mathcal{G}| \leq c \cdot \frac{|\mathcal{F}|}{k}$. Denote $V' = V \setminus (A \cup B)$, and consider the family $\mathcal{G}' = \{F \cap V' \mid F \in \mathcal{G}\}$. Since $\mathcal{F}$ solves the $\mathcal{S}_k$-problem, for all $u \in V'$, and every $C \subseteq V' \setminus \{u\}$ of size $\frac{k}{2}$, there is a set $F \in \mathcal{G}'$ such that $u \in F$ and $F \cap C = \emptyset$. Otherwise, $\mathcal{F}$ would not distinguish between the two stars whose center is u, that share the $k-1$ vertices of $B \cup C$, and for which the additional vertex of one of them is one element of A, and the additional vertex of the other one is the other element of A. Let $m = |\mathcal{G}'|$, $n' = |V'| = n - \frac{k}{2} - 1$, and let M be the m by n' matrix whose rows are the incidence vectors of the sets in $\mathcal{G}'$. Now let us look at the columns of M as the incidence vectors of subsets of another set, of size m. For every column i, and every set J of $\frac{k}{2}$ columns such that $i \notin J$, there exists a row in which the i^{th} coordinate is 1, and for all $j \in J$, the j^{th} coordinate is 0. Thus, no subset corresponding to a column is contained in the union of $\frac{k}{2}$ subsets correspond to any other $\frac{k}{2}$ columns, and by Lemma 2.8,

$$|\mathcal{G}'| = m > \frac{(\frac{k}{2})^2 \log(n' - \frac{k}{4})}{10 \log \frac{k}{2}} = \Omega\left(\frac{k^2 \log n}{\log k}\right),$$

and hence

$$|\mathcal{F}| \geq \Omega(k|\mathcal{G}|) \geq \Omega(k|\mathcal{G}'|) \geq \Omega\left(\frac{k^3 \log n}{\log k}\right).$$

□

3 Complete Graphs

In this section we consider the case where the hidden graphs are complete graphs. Denote by $\mathcal{C}_k$ the family of all graphs on $V = \{1, 2, \ldots, n\}$, that consist of a copy of K_k, and $n - k$ isolated vertices. Let $\mathcal{C} = \cup_{k=2}^{n} \mathcal{C}_k$.

In the following theorem, we prove lower and upper bounds on the minimum size of a family that solves the $\mathcal{C}$-problem.

Theorem 3.1 *Any family that solves the $\mathcal{C}$-problem is of size at least $\Omega(n \log n)$, and there exists a family of size $O(n \log^2 n)$ that solves the $\mathcal{C}$-problem.*

Proposition 3.2 *The minimum size of a family $\mathcal{F}$ that solves the $\mathcal{C}$-problem is at least $\Omega(n \log n)$.*

Proof: Let $\mathcal{F}$ be a family that solves the $\mathcal{C}$-problem. Let $u \in V$, and $V_1 = V \setminus \{u\}$. In order to distinguish between the complete graph on V_1 and the complete graph on $V_1 \cup \{u\}$, $\mathcal{F}$ must contain a query $F_1(u) = \{u, v_1\}$ for some $v_1 \in V_1$. Now let $V_2 = V_1 \setminus \{v_1\}$. In order to distinguish between the complete graph on V_2 and the complete graph on $V_2 \cup \{u\}$, $\mathcal{F}$ must contain a query $F_2(u)$ such that $u \in F_2(u)$ and $|F_2(u) \cap V_2| = 1$. Denote by v_2 the vertex in $F_2(u) \cap V_2$. We can continue in this way and define for all $1 \leq i \leq n-2$ a set $V_i = V_{i-1} \setminus \{v_{i-1}\}$, and find a set $F_i(u) \in \mathcal{F}$ that distinguishes between the complete graph on V_i and the complete graph on $V_i \cup \{u\}$. Then $u \in F_i(u)$, and $|F_i(u) \cap V_i| = 1$. Denote by v_i the vertex in $F_i(u) \cap V_i$. For all $1 \leq i \leq n-2$, $|V_i| = n - i$, and since $|F_i(u) \cap V_i| = 1$, $|F_i(u)| \leq i + 1$. Furthermore, all the sets $F_i(u)$ for $1 \leq i \leq n-2$ are distinct, since the vertices v_i are distinct, and $v_i \in F_i(u)$, but for all $j < i$, $v_i \notin F_j(u)$. $\mathcal{F}$ contains these sets $F_i(u)$ for all $u \in V$. For every vertex $u \in V$, and all $1 \leq i \leq n-2$, assign a weight to the pair (u, i), defined by $w(u, i) = \frac{1}{|F_i(u)|}$. For a set $F \in \mathcal{F}$, there are at most $|F|$ vertices u (the vertices in F) such that $F = F_i(u)$ for some i. Thus the total weight corresponding to a set $F \in \mathcal{F}$ is at most 1, that is,

$$\sum_{(u,i):F_i(u)=F} w(u, i) \leq |F| \cdot \frac{1}{|F|} = 1.$$

Therefore,

$$\begin{aligned}
|\mathcal{F}| &\geq \sum_{F \in \mathcal{F}} \sum_{(u,i):F_i(u)=F} w(u, i) \\
&= \sum_{u \in V} \sum_{i=1}^{n-2} w(u, i) \\
&= \sum_{u \in V} \sum_{i=1}^{n-2} \frac{1}{|F_i(u)|} \\
&\geq \sum_{u \in V} \sum_{i=1}^{n-2} \frac{1}{i+1} \\
&= \Omega(n \log n).
\end{aligned}$$

□

Proposition 3.3 *There exists a family $\mathcal{F}$ of cardinality $O(n \log^2 n)$ that solves the $\mathcal{C}$-problem.*

Proof: We construct the family $\mathcal{F}$ recursively as follows. First, the set V is in $\mathcal{F}$. Now partition V into two halves V_1 and V_2 and find the part of the clique in each half. The clique is the union of the cliques found in V_1 and V_2. This works as long as the part of the clique in each V_i is of size 0 or of size at least 2. But if the part of the clique in V_i is of size 1, then the answer to Q_{V_i} is "no", and we need some additional queries to find this vertex. Suppose that the clique has one vertex in V_1. We show that we can find this vertex by the following queries. Assign distinct vectors of length $\lceil \log |V_1| \rceil$ over $\{0,1\}$ to the vertices in V_1. For all $1 \leq i \leq \lceil \log |V_1| \rceil$, $j \in \{0,1\}$ and $u \in V_2$, we have the following set $F(i,j,u) = \{v \in V_1 \mid \textit{the } i^{th} \textit{ bit of } v \textit{ is } j\} \cup \{u\}$ in $\mathcal{F}$. If the answer to Q_V is "yes" and the answer to Q_{V_1} is no, then there is at least one vertex u of the clique in V_2. If there are no vertices of the clique in V_1 then the answers to all $Q_{F(i,j,u)}$ are "no". Otherwise, there is precisely one vertex v of the clique in V_1. The answer to $Q_{F(i,j,u)}$ is "yes" if and only if u is in the clique, and the i^{th} bit of v is j. Since there is at least one vertex of the clique in V_2, we can obtain v from these queries. We should have similar queries for the case that V_2 contains one vertex of the clique. Denote by $f(n)$ the number of queries needed for n vertices. Then, by the above discussion,

$$f(n) \leq 4 \cdot \frac{n}{2} \cdot \log \frac{n}{2} + 2f(\frac{n}{2}) + 1 = O(n \log^2 n).$$

□

We now give upper and lower bounds for cliques of a given size. These results are tight up to a factor of polylog$n$ for all admissible sizes.

Theorem 3.4 *For every k, there exists a family $\mathcal{F}$ of size $O(k^2 \log n)$ that solves the $\mathcal{C}_k$-problem, and every family that solves the $\mathcal{C}_k$-problem either contains $\Omega(n)$ pairs, or it is of size at least $\Omega(\frac{k^2}{\log n})$. Moreover, for all $k \leq n^{\frac{1}{3}}$, the size of any family that solves the $\mathcal{C}_k$-problem is at least $\Omega(\frac{k^2 \log n}{\log k})$, and for all $k \leq \sqrt{n}$ it is at least $\Omega(k^2)$. In addition, for all s, there exists a family of size $(s+1)\lceil \frac{n}{2} \rceil$ that solves the $\mathcal{C}_{n-s}$-problem.*

The best bounds we have, for various values of k, are summarized in Table 2. In the rest of this section we prove these results.

Proposition 3.5 *For every k, there exists a family $\mathcal{F}$ of size $O(k^2 \log n)$ that solves the $\mathcal{C}_k$-problem.*

Proof: Let $m = ck^2 \log n$ for some absolute constant c, and let $F_1, F_2, ..., F_m$ be m random subsets of V, chosen independently as follows. For every F_i, every $v \in V$ is chosen to be in F_i independently with probability $\frac{1}{k}$. Let C_1 and C_2 be two complete graphs of size k such that $|V(C_1) \setminus V(C_2)| = |V(C_2) \setminus V(C_1)| = 1$. Let $v_1, ..., v_{k-1}$ be the common vertices of C_1 and C_2, and let u_i be the additional vertex of C_i for $i = 1, 2$. F_i distinguishes between C_1 and C_2 if and only if exactly one

Table 2: Bounds on the size of a family that solves the $\mathcal{C}_k$-problem.

k	Lower bound	Upper bound
$k \leq n^{\frac{1}{3}}$	$\Omega(\frac{k^2 \log n}{\log k})$	$O(k^2 \log n)$
$n^{\frac{1}{3}} < k \leq \sqrt{n}$	$\Omega(k^2)$	$O(k^2 \log n)$
$\sqrt{n} < k < \sqrt{n \log n}$	$\Omega(\frac{k^2}{\log n})$	$O(k^2 \log n)$
$\sqrt{n \log n} \leq k \leq n - \log^2 n$	$\Omega(n)$	$O(n \log^2 n)$
$k = n - s,\ s < \log^2 n$	$\Omega(n)$	$(s+1)\lceil \frac{n}{2} \rceil$

vertex among u_1 and u_2, and exactly one vertex among $v_1, ..., v_{k-1}$ are in F_i. Thus the probability that F_i distinguishes between C_1 and C_2 is

$$\frac{2}{k} \cdot \frac{k-1}{k} \left(1 - \frac{1}{k}\right)^{k-1} = \Omega\left(\frac{1}{k}\right).$$

Therefore, the probability that no F_i distinguishes between C_1 and C_2 is

$$\left[1 - \Omega\left(\frac{1}{k}\right)\right]^m \leq n^{-2k}$$

for an appropriate value of c. For two cliques that differ in more vertices, this probability is smaller. The number of pairs of cliques is smaller than n^{2k}, and hence, there is a family $\mathcal{F} = \{F_1, F_2, ..., F_m\}$ that solves the $\mathcal{C}_k$-problem. □

Proposition 3.6 *For every k, if $\mathcal{F}$ is a family the solves the $\mathcal{S}_k$-problem, then $\mathcal{F}$ either contains $\Omega(n)$ pairs, or it is of cardinality at least $\Omega(\frac{k^2}{\log n})$.*

Proof: Clearly, we may assume that $k^2 > \log n$, since otherwise there is nothing to prove. Let $\mathcal{F}$ be a family that solves the $\mathcal{C}_k$-problem. Then, for all $A, B \subseteq V$ such that $|A| = 2$, $|B| = k-1$ and $A \cap B = \emptyset$, there exists a set $F \in \mathcal{F}$ such that $|F \cap A| = 1$ and $|F \cap B| = 1$. Indeed, otherwise $\mathcal{F}$ would not distinguish between the complete graph on B and one vertex of A, and the complete graph on B and the other vertex of A. Denote by $\mathcal{F}_0$ the family of all sets $F \in \mathcal{F}$ of size 2. Let $m = c \cdot \frac{n \log n}{k}$, and define $\mathcal{F}_1 = \{F \in \mathcal{F} \mid 2 < |F| \leq m\}$, and $\mathcal{F}_2 = \mathcal{F} \setminus (\mathcal{F}_0 \cup \mathcal{F}_1)$. We show that if, say, $|\mathcal{F}_0| \leq \frac{1}{10}n$ then $|\mathcal{F}_1 \cup \mathcal{F}_2| > c_1 \cdot \frac{k^2}{\log n}$ for some constant c_1 that depends only on c. Suppose $|\mathcal{F}_0| \leq \frac{1}{10}n$ and $|\mathcal{F}_1 \cup \mathcal{F}_2| \leq c_1 \cdot \frac{k^2}{\log n}$. Choose uniformly a subset $A = \{u, v\} \subseteq V$, and define $\mathcal{F}_1' = \{F \in \mathcal{F}_1 \mid |F \cap A| = 1\}$. For each $F \in \mathcal{F}_1$

$$Pr(F \in \mathcal{F}_1') = 2 \cdot \frac{|F|}{n} \cdot \frac{n - |F|}{n-1} \leq 2 \cdot \frac{|F|}{n}$$

Therefore,

$$E[|\mathcal{F}_1'|] \leq 2 \sum_{F \in \mathcal{F}_1} \frac{|F|}{n} \leq 2|\mathcal{F}_1|\frac{m}{n}.$$

By Markov's inequality, the probability that $|\mathcal{F}_1'| > 4|\mathcal{F}_1|\frac{m}{n}$ is at most $\frac{1}{2}$. Since $|\mathcal{F}_0| \leq \frac{1}{10}n$, the probability that there is a set $F \in \mathcal{F}_0$ such that $F \cap A \neq \emptyset$ is less than $\frac{2}{5}$. Thus, there exists a choice of A such that for all $F \in \mathcal{F}_0$, $F \cap A = \emptyset$ and $|\mathcal{F}_1'| \leq 4|\mathcal{F}_1|\frac{m}{n}$. For such a choice of A, and appropriate values of c and c_1,

$$\begin{aligned} |\mathcal{F}_1'| &\leq 4|\mathcal{F}_1| \cdot \frac{m}{n} \\ &\leq 4c_1 c \cdot \frac{k^2}{\log n} \cdot \frac{n \log n}{kn} \\ &\leq \frac{k-1}{4}. \end{aligned}$$

Thus, there exists a subset $B_1 \subseteq V \setminus A$ of size $\frac{k-1}{2}$ such that for every $F \in \mathcal{F}_1'$, $|F \cap B_1| \geq 2$. Choose a random subset $B_2 \in V$ of size $\frac{k-1}{2}$. For all $F \in \mathcal{F}_2$

$$\begin{aligned} Pr(|F \cap B_2| \leq 1) &= \frac{\binom{n-|F|}{\frac{k-1}{2}}}{\binom{n}{\frac{k-1}{2}}} + \frac{k-1}{2} \cdot \frac{|F|}{n} \cdot \frac{\binom{n-|F|}{\frac{k-1}{2}-1}}{\binom{n-1}{\frac{k-1}{2}-1}} \\ &\leq \left(1 - \frac{|F|}{n}\right)^{\frac{k-1}{2}} + \frac{k-1}{2} \cdot \frac{|F|}{n} \cdot \left(1 - \frac{|F|-1}{n-1}\right)^{\frac{k-1}{2}-1} \\ &\leq e^{-\frac{k-1}{2} \cdot \frac{m}{n}} + \frac{k-1}{2} \cdot \frac{n}{n} \cdot e^{-(\frac{k-1}{2}-1)\frac{m-1}{n-1}} \\ &\leq n^{-c_2} \end{aligned}$$

for some constant $c_2 = \Theta(c)$. Therefore, if c is sufficiently large, then, with high probability, $A \cap B_2 = \emptyset$ and for every $F \in \mathcal{F}_2$, $|F \cap B_2| \geq 2$. Denote $B' = B_1 \cup B_2$. $B' \subseteq V \setminus A$ and $|B'| \leq k-1$. Let B be an arbitrary extension of B' to a subset of $V \setminus A$ of size $k-1$. Let C_1 be the complete graph on $B \cup \{u\}$, and let C_2 be the complete graph on $B \cup \{v\}$. Since for every $F \in \mathcal{F}_0$ $u, v \notin F$, no set in $\mathcal{F}_0$ can distinguish between C_1 and C_2. Neither can sets in $\mathcal{F}_1$ that contain both u and v, or that contain none of them. All other sets in $\mathcal{F}_1$, i.e. sets that contain exactly one vertex among u and v, and all the sets in $\mathcal{F}_2$, contain at least two vertices of B, so they cannot distinguish between these two cliques either. Thus $\mathcal{F}$ cannot distinguish between C_1 and C_2, contradicting the assumption that it solves the $\mathcal{C}_k$-problem. □

We now prove a better lower bound for $k \leq n^{\frac{1}{3}}$. This bound is tight up to a factor of $\log k$.

Lemma 3.7 *Let S be a set of size m, and let $\mathcal{A}$ be a family of n subsets of S. Suppose that there are no distinct $A, B_1, ..., B_r, C_1, ..., C_r \in \mathcal{A}$ for which*

$$A \subseteq \bigcup_{i=1}^{r} B_i$$

and

$$A \subseteq \bigcup_{i=1}^{r} C_i,$$

where $r \leq n^{\frac{1}{3}}$. *Then* $m = \Omega(\frac{r^2 \log n}{\log r})$.

Proof: Let $\mathcal{B} = \emptyset$. As long as there exist $A, B_1, ..., B_r \in \mathcal{A}$ such that

$$A \subseteq \bigcup_{i=1}^{r} B_i,$$

remove $A, B_1, ..., B_r$ from $\mathcal{A}$, and add A to $\mathcal{B}$. Let $\mathcal{A}'$ be the family obtained from $\mathcal{A}$ at the end of this process, and denote the size of $\mathcal{B}$ by l. Then, $|\mathcal{A}'| = n - l(r+1)$, and both $\mathcal{A}'$ and $\mathcal{B}$ are r-cover-free. $\mathcal{A}'$ is clearly r-cover-free, or otherwise the above process would not stop. $\mathcal{B}$ is also r-cover-free, because if there were $A, C_1, ..., C_r \in \mathcal{B}$ such that

$$A \subseteq \bigcup_{i=1}^{r} C_i,$$

then, there are also $B_1, ..., B_r \in \mathcal{A}$, that were removed from $\mathcal{A}$ together with A, such that

$$A \subseteq \bigcup_{i=1}^{r} B_i,$$

contradicting the assumption. If $l \geq \frac{n^{\frac{2}{3}}}{4}$, then, since $r \leq n^{\frac{1}{3}}$, we have by Lemma 2.8,

$$m > \frac{r^2 \log(l - \frac{r}{2})}{10 \log r} = \Omega\left(\frac{r^2 \log n}{\log r}\right).$$

Otherwise $l < \frac{n^{\frac{2}{3}}}{4}$, and thus, since $r < n^{\frac{1}{3}}$, $|\mathcal{A}'| = n - l(r+1) > \frac{n}{2}$. Hence, by Lemma 2.8,

$$m > \frac{r^2 \log(\frac{n}{2} - \frac{r}{2})}{10 \log r} = \Omega\left(\frac{r^2 \log n}{\log r}\right).$$

□

Proposition 3.8 *For every* $k \leq n^{\frac{1}{3}}$, *if* $\mathcal{F}$ *is a family that solves the* $\mathcal{C}_k$*-problem, then* $|\mathcal{F}| = \Omega(\frac{k^2 \log n}{\log k})$.

Proof: Let $\mathcal{F}$ be a family that solves the $\mathcal{C}_k$-problem. Define $m = |\mathcal{F}|$, and let M be the m by n matrix whose rows are the incidence vectors of the sets in $\mathcal{F}$. Consider the columns of M as the incidence vectors of subsets of another set, of size m. For $1 \leq i \leq n$, let G_i be the subset corresponding to the i^{th} column of M. Define the family $\mathcal{G}$ as follows. $\mathcal{G} = \{G_{2i-1} \cup G_{2i} \mid 1 \leq i \leq \frac{n}{2}\}$. We claim that there are no distinct sets $A, B_1, ..., B_{\frac{k-1}{4}}, C_1, ..., C_{\frac{k-1}{4}} \in \mathcal{G}$, such that

$$A \subseteq \bigcup_{i=1}^{\frac{k-1}{4}} B_i \tag{1}$$

and

$$A \subseteq \bigcup_{i=1}^{\frac{k-1}{4}} C_i. \tag{2}$$

Suppose there were such sets. A is the union of two subsets corresponding to two distinct columns of M. Let u and v be the vertices corresponding to these columns. Similarly, let $w_1, ..., w_{k-1}$ be the vertices corresponding to $B_1, ..., B_{\frac{k-1}{4}}, C_1, ..., C_{\frac{k-1}{4}}$. The members of A are the queries that contain u or v. Since (1) and (2) hold, each such query contains at least two vertices from $w_1, ..., w_{k-1}$. Thus, no query distinguishes between the complete graph on $u, w_1, ..., w_{k-1}$ and the complete graph on $v, w_1, ..., w_{k-1}$. Hence, there are no such sets in $\mathcal{G}$, and therefore, by Lemma 3.7, with $r = \frac{k-1}{4}$ and $\mathcal{A} = \mathcal{G}$,

$$|\mathcal{F}| = m = \Omega\left(\frac{k^2 \log n}{\log k}\right).$$

□

We now prove that for all $n^{\Omega(1)} \leq k \leq \sqrt{n}$, any family that solves the $\mathcal{C}_k$-problem is of size at least $\Omega(k^2)$.

Definition 3.9 *Let A be a subset of a set S, and let $\mathcal{A}$ be a family of subsets of S. We say that A is* covered twice *by $\mathcal{A}$ if for all $a \in A$, there are at least two sets in $\mathcal{A}$ that contain a.*

Lemma 3.10 *Let S be a set of size m, and let $\mathcal{A}$ be a family of n subsets of S. Suppose that no set in $\mathcal{A}$ is covered twice by any other r sets in $\mathcal{A}$, where $n^{\Omega(1)} \leq r \leq \sqrt{n}$. Then $m = \Omega(r^2)$.*

Proof: Suppose $m \leq \epsilon r^2$, for some small constant $\epsilon > 0$. We show that if ϵ is sufficiently small, then there is a set $A \in \mathcal{A}$ that is covered twice by some other r sets in $\mathcal{A}$. As long as there exists $a \in S$ that belongs to one or two sets in $\mathcal{A}$, remove these sets from $\mathcal{A}$. After removing these sets, a belongs to no set in $\mathcal{A}$. Therefore, this process stops after at most m steps, and then every $a \in S$ belongs to zero or at least three sets. Let $\mathcal{A}'$ be the family of the remaining sets, and denote its size by n'. Thus $n' \geq n - 2m \geq n - 2\epsilon r^2 \geq (1 - 2\epsilon)n$. If there exists a set $A \in \mathcal{A}'$ such that $|A| \leq \frac{r}{2}$, then it is covered twice by a family of at most r sets in $\mathcal{A}' \setminus \{A\}$, consisting of two arbitrarily chosen sets that contain each member of A. Suppose now that every set $A \in \mathcal{A}'$ is of size greater than $\frac{r}{2}$. Choose randomly $\frac{r}{2}$ sets $B_1, ..., B_{\frac{r}{2}} \in \mathcal{A}'$. Let C be the set of all $a \in S$ that belong to at most one set from $B_1, ..., B_{\frac{r}{2}}$. Now choose randomly another set $A \in \mathcal{A}'$. If $|A \cap C| \leq \frac{r}{4}$, then for all $a \in A \cap C$, choose two sets in $\mathcal{A}' \setminus \{A\}$ that contain a. These sets, together with $B_1, ..., B_{\frac{r}{2}}$, form a family of at most r sets that cover A twice. We now show that $E[|A \cap C|] \leq \frac{r}{5}$, and hence there exists a choice of $B_1, ..., B_{\frac{r}{2}}$ and $A \neq B_1, ..., B_{\frac{r}{2}}$ for which $|A \cap C| \leq \frac{r}{4}$. Therefore A is covered twice by r other sets, contradicting the assumption. Let $a \in S$, and let k be the number of sets in $\mathcal{A}'$ that contain a. The probability that $a \in A \cap C$ is at most

$$\frac{k}{n'}\left[\frac{\binom{n'-k}{\frac{r}{2}}}{\binom{n'}{\frac{r}{2}}} + \frac{k\binom{n'-k}{\frac{r}{2}-1}}{\binom{n'}{\frac{r}{2}}}\right] = \frac{k}{n'}\left[\frac{\binom{n'-k}{\frac{r}{2}}}{\binom{n'}{\frac{r}{2}}} + \frac{kr}{2n'}\frac{\binom{n'-k}{\frac{r}{2}-1}}{\binom{n'-1}{\frac{r}{2}-1}}\right]$$

$$\leq \frac{k}{n'}\left(1-\frac{k}{n'}\right)^{\frac{r}{2}} + \frac{k^2 r}{2n'^2}\left(1-\frac{k-1}{n'-1}\right)^{\frac{r}{2}-1}$$
$$\leq \frac{k}{n'}e^{-\frac{kr}{2n'}} + \frac{k^2 r}{2n'^2}e^{-\frac{kr}{4n'}}.$$

We now show that this probability is at most $\frac{c}{r}$ for some constant c. Let us first consider the term $\frac{k}{n'}e^{-\frac{kr}{2n'}}$. If $k \leq \frac{2n'}{r}$ then this term is at most $\frac{2}{r}$. If $k > \frac{2n'}{r}$, denote $x = \frac{kr}{2n'}$. Since $x > 1$ we have $\frac{k}{n'}e^{-\frac{kr}{2n'}} = \frac{2}{r}xe^{-x} < \frac{2}{er}$. Consider now the term $\frac{k^2 r}{2n'^2}e^{-\frac{kr}{4n'}}$. If $k \leq \frac{8n'}{r}$ then this term is at most $\frac{32}{r}$. If $k > \frac{8n'}{r}$ then denote $x = \frac{kr}{4n'}$. Then $x > 2$, and $\frac{k^2 r}{2n'^2}e^{-\frac{kr}{4n'}} = \frac{8}{r}x^2e^{-x}$. It is easy to check that x^2e^{-x} is decreasing for all $x > 2$, and hence $\frac{8}{r}x^2e^{-x} < \frac{32}{e^2 r}$.

Thus the probability that $a \in A \cap C$ is at most $\frac{c}{r}$ for some constant c. Therefore, we have

$$E[|A \cap C|] \leq \frac{cm}{r} \leq \frac{c\epsilon r^2}{r} \leq \frac{r}{5}$$

provided ϵ is sufficiently small, completing the proof of the lemma. □

Proposition 3.11 *For every $n^{\Omega(1)} \leq k \leq \sqrt{n}$, if $\mathcal{F}$ is a family that solves the $\mathcal{C}_k$-problem, then $|\mathcal{F}| = \Omega(k^2)$.*

Proof: For $n^{\Omega(1)} \leq k \leq n^{1/3}$ the result follows from Proposition 3.8. We thus assume that $k > n^{1/3}$. Let $\mathcal{F}$ be a family that solves the $\mathcal{C}_k$-problem. Define $m = |\mathcal{F}|$, and let M be the m by n matrix whose rows are the incidence vectors of the sets in $\mathcal{F}$. Consider the columns of M as the incidence vectors of subsets of another set, of size m. For $1 \leq i \leq n$, let G_i be the subset corresponding to the i^{th} column of M. Define $\mathcal{G} = \{G_{2i-1} \cup G_{2i} \mid 1 \leq i \leq \frac{n}{2}\}$. We claim that there are no distinct sets $A, B_1, ..., B_{\frac{k-1}{2}} \in \mathcal{G}$, such that A is covered twice by $B_1, ..., B_{\frac{k-1}{2}}$. Suppose there were such sets. A is the union of two subsets corresponding to two distinct columns of M. Let u and v be the corresponding vertices. Similarly, let $w_1, ..., w_{k-1}$ be the vertices corresponding to $B_1, ..., B_{\frac{k-1}{2}}$. The members of A are the queries that contain u or v. Since A is covered twice by $B_1, ..., B_{\frac{k-1}{2}}$, each such query contains at least two vertices from $w_1, ..., w_{k-1}$. Thus, no query distinguishes between the complete graph on $u, w_1, ..., w_{k-1}$ and the complete graph on $v, w_1, ..., w_{k-1}$. Hence, there are no such sets in $\mathcal{G}$, and therefore, by Lemma 3.10,

$$|\mathcal{F}| = m = \Omega(k^2).$$

□

We conclude the section with a simple upper bound, which improves our estimate for cliques that contain almost all the vertices.

Proposition 3.12 *For every s, there exists a family of size at most*

$$(s+1)\lceil\frac{n}{2}\rceil$$

that solves the $\mathcal{C}_{n-s}$-problem.

Proof: For each $u \in V$, ask $s+1$ pairs that contain u. u is in the clique if and only if the answer to at least one of these queries is "yes". □

4 General Graphs

In this section we consider families that contain all the graphs on V isomorphic to a graph G. Denote by $\mathcal{H}_G$ the family of all graphs isomorphic to G.

Theorem 4.1 *Let $G = (V, E)$ be a graph on n vertices, and suppose that there are three vertices $u, v, w \in V$, such that for every two of them, the sets of their neighbours except these vertices themselves are distinct, i.e. $N(u) \setminus \{v\} \neq N(v) \setminus \{u\}$, $N(u) \setminus \{w\} \neq N(w) \setminus \{u\}$, and $N(v) \setminus \{w\} \neq N(w) \setminus \{v\}$. Then, the size of any family that solves the $\mathcal{H}_G$-problem is at least $\Omega(\frac{n^2}{\alpha^2(G)})$, where $\alpha(G)$ is the maximum size of an independent set in G.*

Proof: For any two vertices $x, y \in V$, denote by $A(x, y)$ the set of vertices $z \in V \setminus \{x, y\}$ such that z is a neighbour of both x and y, or of none of them. We show that there are two vertices among u, v, and w, for which the size of this set is at least $\frac{1}{3}n - 1$. Suppose that $A(u, v) < \frac{1}{3}n - 1$. Then, $V \setminus (A(u, v) \cup \{u, v, w\}) > \frac{2}{3}n - 2$, and each one of these vertices is a neighbour of exactly one vertex among u and v. Thus, each one of these vertices is in $A(u, w)$ or in $A(v, w)$, and hence at least one of these sets is of size at least $\frac{1}{3}n - 1$. Assume, without loss of generality, that $|A(u, v)| \geq \frac{1}{3}n - 1$.

Let $\mathcal{F}$ be a family that solves the $\mathcal{H}_G$-problem, and let $\alpha = \alpha(G)$. Assume that $|\mathcal{F}| < \frac{n^2}{12\alpha^2}$. Every set $F \in \mathcal{F}$ is of size at most α, or otherwise the answer to Q_F is "yes" (and is known in advance). For every $x \in V$, denote by $f(x)$ the number of sets $F \in \mathcal{F}$ such that $x \in F$.

$$\sum_{x \in V} f(x) = \sum_{F \in \mathcal{F}} |F| \leq \alpha |\mathcal{F}| < \frac{n^2}{12\alpha}. \tag{3}$$

Let $V' = \{x \in V \mid f(x) < \frac{n}{6\alpha}\}$. Then $|V'| \geq \frac{n}{2}$, since otherwise

$$\sum_{x \in V} f(x) \geq \sum_{x \in (V \setminus V')} f(x) \geq \frac{n}{2} \cdot \frac{n}{6\alpha} = \frac{n^2}{12\alpha},$$

contradicting (3). For $x \in V'$, the number of vertices $z \in V$ such that there exists a set $F \in \mathcal{F}$ that contains both x and z is at most

$$\sum_{F : x \in F} |F| \leq f(x)\alpha < \frac{n}{6}.$$

Let $x, y \in V'$, and let A be the set of all vertices $z \in V$ such that there exists a set $F \in \mathcal{F}$ that contains x or y, and z.

$$|A| \leq \sum_{F : x \in F} |F| + \sum_{F : y \in F} |F| < \frac{n}{3}.$$

Let G_1 be a graph isomorphic to G, where u is mapped to x, v is mapped to y, and only vertices from $A(u, v)$ are mapped into A. Let G_2 be the graph in which u is mapped to y, v is mapped to x, and the rest of it is identical to G_1. The only queries that could distinguish between G_1 and G_2 are queries Q_F where F contains x or y, but then all the other vertices in F are in $A(u, v)$, and thus, the answer to Q_F is the same for G_1 and G_2. Therefore, $\mathcal{F}$ cannot distinguish between G_1 and G_2, contradicting the assumption that it solves the $\mathcal{H}_G$-problem. □

Corollary 4.2 *Let $G = G(n, \frac{1}{2})$ be the random graph on n vertices. Then, almost surely, any family that solves the $\mathcal{H}_G$-problem is of size at least $\Omega(\frac{n^2}{\log^2 n})$.*

Proof: The corollary follows from Theorem 4.1, since, almost surely, $\alpha(G) = O(\log n)$ (see, for example, [3] or [2]), and since obviously, there are, almost surely, three vertices u, v and w with distinct sets of neighbours, as defined in the theorem. □

5 Concluding Remarks

- It will be interesting to close the polylogarithmic gaps between the upper and the lower bounds proved in this paper.
- Another intriguing challenge is to obtain a general way to estimate, for every graph G, the number of queries needed to identify a hidden graph isomorphic to G. In particular, the problem of characterizing all graphs for which the trivial upper bound of $O(n^2)$ is best possible seems interesting. Our results enable us to prove an $\Omega(n^2)$ lower bound for the number of queries required to identify a hidden copy of any graph with at least one isolated vertex, containing a vertex of degree 1 which is adjacent to a vertex of high degree. We omit the details.
- The problems considered here can be investigated when more than one round is allowed, and in case the algorithms are fully adaptive.

References

[1] N. Alon, R. Beigel, S. Kasif, S. Rudich, B. Sudakov, Learning a Hidden Matching, Proceedings of the 43rd IEEE FOCS, 2002, 197-206. Also: SIAM J. Computing 33 (2004), 487-501.

[2] N. Alon and J. H. Spencer, **The Probabilistic Method**, Second Edition, Wiley, New York, 2000.

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

[4] A. G. Dyachkov and V. V. Rykov, Bounds on the Length of Disjunctive Codes, *Problemy Peredachi Informatsii* Vol. 18, no. 3 (1982), 158-166.

[5] V. Grebinski and G. Kucherov, Optimal Query Bounds for Reconstructing a Hamiltonian Cycle in Complete Graphs, Proc. 5th Israeli Symposium on Theoretical Computer Science (1997), 166-173.

[6] V. Grebinski and G. Kucherov, Reconstructing a Hamiltonian Cycle by Querying the Graph: Application to DNA Physical Mapping, *Discrete Applied Math.* 88 (1998), 147-165.

[7] V. Grebinski and G. Kucherov, Optimal Reconstruction of Graphs under the Additive Model, *Algorithmica* 28(1) (2000), 104-124.

[8] M. Ruszinkó, On the Upper Bound of the size of the r-cover-free families, *Journal of Combinatorial Theory* Series A vol. 66, no. 2, May 1994, 302-310.