Towards a polynomial form of Rödl’s theorem

Jacob Fox
Stanford University,
Stanford, CA 94305, USA

Alex Scott
Oxford University,
Oxford, UK

Paul Seymour
Princeton University,
Princeton, NJ 08544, USA

July 4, 2022; revised August 4, 2022

1Supported by a Packard Fellowship and by NSF grants DMS-1953990 and DMS-2154129.
2Supported by AFOSR grants A9550-19-1-0187 and FA9550-22-1-0234, and by NSF grants DMS-1800053 and DMS-2154169.
Abstract

A theorem of Rödl says that for every graph $H$, and every $\varepsilon > 0$, there exists $\delta > 0$ such that if $G$ is a graph that has no induced subgraph isomorphic to $H$, then there exists $X \subseteq V(G)$ with $|X| \geq \delta |G|$ such that one of $G[X], \overline{G}[X]$ has maximum degree at most $\varepsilon |X|$. But for fixed $H$, how does $\delta$ depend on $\varepsilon$? If the dependence is polynomial for a graph $H$, then $H$ satisfies the Erdős-Hajnal conjecture; and it has been conjectured that the dependence is always polynomial. Here we prove this conjecture in an assortment of special cases, for instance when $H$ can be obtained from subgraphs of the four-vertex path $P_4$ by repeated vertex-substitution.

We also show that when $H$ is $P_4$ itself, the dependence is linear, and indeed for every cograph $G$ and every $\varepsilon > 0$, there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|$ such that one of $G[X], \overline{G}[X]$ has maximum degree at most $\varepsilon^2 |G|$ (and hence at most $\varepsilon |X|$). We sharpen this as much as we can. Finally, we prove a similar, weaker result for the class of comparability graphs and for every hereditary class with the “strong Erdős-Hajnal property”. 
1 Introduction

Some terminology and notation: $G[X]$ denotes the induced subgraph with vertex set $X$ of a graph $G$; $|G|$ denotes the number of vertices of $G$; $\overline{G}$ is the complement graph of $G$; $P_4$ denotes the path with four vertices; a graph is $H$-free if it has no induced subgraph isomorphic to $H$; and a cograph is a $P_4$-free graph.

A very useful theorem of V. Rödl [14] says:

1.1 For every graph $H$ and every $\varepsilon > 0$, there exists $\delta > 0$ such that for every $H$-free graph $G$, there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon|G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon|X|(|X| - 1)$ edges.

How does $\delta$ depend on $\varepsilon$, for a given graph $H$? B. Sudakov and the first author [10] studied this, and proposed the conjecture (conjecture 7.1 in their paper) that the dependence is polynomial, or more exactly:

1.2 Conjecture: For every graph $H$ there exists $c > 0$ such that for every $\varepsilon$ with $0 < \varepsilon \leq 1/2$ and every $H$-free graph $G$, there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon|G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon|X|(|X| - 1)$ edges.

Our first result, proved in the next section, shows that 1.2 holds in a particularly nice form when $H = P_4$. We will show:

1.3 For every $\varepsilon > 0$ and every cograph $G$, there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon|G|$ such that one of $G[X], \overline{G}[X]$ has maximum degree at most $\varepsilon^2|G| \leq \varepsilon|X|$.

This is not tight, and with more work we can find sharper bounds, discussed later.

But which other graphs $H$ satisfy 1.2? It is shown in [10] that every graph $H$ satisfying 1.2 also satisfies the Erdős-Hajnal conjecture [8, 9], and the latter is only known to hold for a rather paltry set of graphs $H$: the "bull" (which is the graph obtained from a triangle by adding two leaves adjacent to different vertices of the triangle), the cycle $C_5$ of length five, and graphs that can be obtained from induced subgraphs of these by repeated vertex-substitution. But can we at least show that these few graphs $H$ satisfy 1.2? No, not yet; indeed, we have no proof that either the bull or $C_5$ satisfies 1.2. But in this paper we make a start. We will show:

1.4 If $H$ can be built by repeated vertex-substitution starting from subgraphs of $P_4$ then $H$ satisfies 1.2.

Since we just used it twice, we had better define “vertex-substitution” before we go on. Let $H_1, H_2$ be graphs, let $v \in V(H_1)$, and let $N$ be the set of all neighbours of $v$ in $H_1$. Let $H$ be obtained from the disjoint union of $H_1 \setminus \{v\}$ and $H_2$ by making every vertex of $H_2$ adjacent to every vertex in $N$. Then $H$ is obtained by substituting $H_2$ for the vertex $v$ of $H_1$, and this operation is called vertex-substitution.

We have not been able to show that the class of all graphs $H$ that satisfy 1.2 is closed under vertex-substitution; but the class that satisfy a stronger property is closed under vertex-substitution, and we will show that $P_4$ and its subgraphs have this stronger property, and 1.4 follows from this. The “stronger property” is as follows. Let us say a copy of $H$ in $G$ is an isomorphism from $H$ to an induced subgraph of $G$. We say a graph $H$ is a virus if there exist $c, d > 0$ such that for every graph $G$ and every $\varepsilon$ with $0 < \varepsilon \leq 1/2$, either
there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges; or

- there are at least $\varepsilon d|G||H|$ copies of $H$ in $G$.

(So either large areas of the graph are under tight lock-down, or there are a huge number of copies of the virus.) We will show the following two results, which together imply 1.4:

1.5 If $H_1, H_2$ are viruses and $H$ is obtained by substituting $H_2$ for a vertex of $H_1$, then $H$ is a virus.

1.6 Every graph on at most four vertices is a virus.

A proper hereditary class is a class of graphs closed under isomorphism and under taking induced subgraphs, and not the class of all graphs. The conjecture 1.2 can be reformulated as:

1.7 Conjecture: If $F$ is a proper hereditary class, there exists $c > 0$ such that for every $\varepsilon$ with $0 < \varepsilon \leq 1/2$ and every $G \in F$, there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges.

So far we have only considered hereditary classes defined by excluding one induced subgraph, but the question also makes sense for other hereditary classes. We will show that 1.7 holds for the class of comparability graphs and for classes of graphs with the “strong Erdős-Hajnal property”, which we define later. There is also a “viral” version of the property when more than one graph is excluded, and we will show that for every path (or caterpillar) $T$, the set $\{T, T\}$ is viral.

2 Dense and sparse sets in cographs

In this section we prove 1.3. We remark first that the result is almost tight. For instance, let $m \geq 2$ be an integer, take $\varepsilon$ slightly smaller than $1/m$ and $n$ a very large integer, and let $G$ be the cograph consisting of the disjoint union of $m$ complete graphs each with $n$ vertices. One can show that if $X \subseteq V(G)$ and one of $G[X], \overline{G}[X]$ has maximum degree at most $\varepsilon |X|$, then $|X| \leq |G|/m$. Indeed, by the pigeonhole principle, any $X \subseteq V(G)$ contains a clique with at least $|X|/m$ vertices and hence $G[X]$ has maximum degree at least $|X|/m - 1$. If $|X| > |G|/m$, then $G$ has a vertex not in this clique and the degree of this vertex in $\overline{G}[X]$ is at least $|X|/m$. One might hope that the result really is tight, but it is not, at least when $\varepsilon \geq 1/2$, as we shall see in the next section.

Cographs are well understood. There is a theorem discovered independently by several authors (see [7]), that:

2.1 If $G$ is a cograph with $|G| \geq 2$ then one of $G, \overline{G}$ is disconnected.

We will use 2.1 to prove 1.3 by induction on $|G|$. Applying it directly does not seem to work, and to use induction we will use a strengthening of 1.3, the following (1.3 follows by setting $x = y = \varepsilon$):

2.2 If $G$ is a cograph then, for all $x, y$ with $0 \leq x, y \leq 1$, either:

- there exists $X \subseteq V(G)$ with $|X| \geq x|G|$ such that $G[X]$ has maximum degree at most $xy|G|$; or

- there exists $Y \subseteq V(G)$ with $|Y| \geq y|G|$ such that $\overline{G}[Y]$ has maximum degree at most $xy|G|$.  

2
Proof. If $|G| \leq 1$ the result is true, so we assume that $|G| \geq 2$ and the result holds for all cographs with fewer vertices, and for all choices of $x, y \in [0, 1]$. By 2.1, taking complements if necessary, we may assume that $G$ is not connected; let $G_1, G_2$ be two non-null subgraphs of $G$, with union $G$ and with $V(G_1) \cap V(G_2) = \emptyset$. Now are are given $x, y \in [0, 1]$. If $x = 0$ then the first bullet holds, taking $X = \emptyset$, so we assume that $x > 0$.

If for $i = 1, 2$ there exists $X_i \subseteq V(G_i)$ with $|X_i| \geq x|G_i|$ such that $G[X_i]$ has maximum degree at most $xy|G|$, then $|X_1 \cup X_2| \geq x|G|$ and $G[X_1 \cup X_2]$ has maximum degree at most $xy|G|$, and the first bullet of the theorem holds. We may therefore assume that:

(1) There does not exist $X_1 \subseteq V(G_1)$ with $|X_1| \geq x|G_1|$ such that $G[X_1]$ has maximum degree at most $xy|G|$.

Let $|G_1| = c|G|$. Since $|G_1| \geq x|G_1|$, there exists $X_1 \subseteq V(G_1)$ with $|X_1| = \lfloor x|G_1| \rfloor$, and hence $G[X_1]$ has maximum degree at most $|X_1| - 1 \leq x|G_1| = xc|G|$; and therefore $xc|G| > xy|G|$ by (1), and so $y < c$ (because $x > 0$). Let $y' = yc; then 0 \leq y' \leq 1$. By (1), it follows from the inductive hypothesis, applied to $G_1, x, y'$, that there exists $Y \subseteq V(G_1)$ with $|Y| \geq y'|G_1|$ such that $G[Y]$ has maximum degree at most $xy'|G_1|$. Since $y'|G_1| = y|G|$ and $G_1[Y] = G[Y]$, the second bullet of the theorem holds. This proves 2.2.

This has several useful consequences, and here is one, a strengthening of 1.3:

2.3 Let $G$ be a cograph, and let $0 \leq \varepsilon \leq 1$. Then there exists $X, Y \subseteq V(G)$, such that $G[X, \overline{G}[Y]$ both have maximum degree at most $\varepsilon|G|$, and with $|X| \cdot |Y| \geq |G|^2$.

Proof. Let $I$ be the set of $x \in [0, 1]$ such that for some $X \subseteq V(G)$, $|X| \geq x|G|$ and $G[X]$ has maximum degree at most $\varepsilon|G|$; and let $J$ be the set of $x \in [0, 1]$ such that for some $Y \subseteq V(G)$, $|Y| \geq \varepsilon|G|$ and $\overline{G}[Y]$ has maximum degree at most $\varepsilon|G|$. By 2.2, $I \cup J = [0, 1]$. Since $I, J$ are nonempty closed sets (because $G$ is finite), it follows that $I \cap J \neq \emptyset$. This proves 2.3.

Let us say $G$ is good if for all $x, y$ with $0 \leq x, y \leq 1$, either:

- there exists $X \subseteq V(G)$ with $|X| \geq x|G|$ such that $G[X]$ has maximum degree at most $xy|G|$; or
- there exists $Y \subseteq V(G)$ with $|Y| \geq y|G|$ such that $\overline{G}[Y]$ has maximum degree at most $xy|G|$.

Thus, complements of good graphs are good; 2.2 says that all cographs are good; and its proof shows that goodness is preserved under taking disjoint unions. Which other graphs are good? This is still open, but Tung Nguyen (private communication) has shown that:

- all forests are good;
- the bull is not good;
- a cycle of length at least five is good if and only if its length is a multiple of six; and
- goodness is not preserved under vertex-substitution; indeed, substituting a two-vertex graph for a vertex of a good graph does not always preserve goodness.
3 A tighter bound for cographs

Let us say two disjoint subsets $A, B$ are complete to each other if every vertex in $A$ is adjacent to every vertex in $B$, and anticomplete if there are no edges between $A, B$. The result 1.3 is neat, and it seemed plausible that it would be tight, but it is not. For $\varepsilon \in [0, 1]$, let $\delta_\varepsilon$ be the supremum of all $\delta$ such that for every cograph $G$, there exists $X \subseteq V(G)$ such that $|X| \geq \delta|G|$ and one of $G[X], \overline{G[X]}$ has maximum degree at most $\varepsilon\delta|G|$. Thus 1.3 says that $\delta_\varepsilon \geq \varepsilon$, but we will show the following, which implies that $\delta_\varepsilon \geq 1/(2 - \varepsilon) > \varepsilon$ when $1/2 \leq \varepsilon < 1$:

**3.1 For every non-null cograph $G$ and every $\varepsilon$ with $1/2 \leq \varepsilon < 1$, there is a set $X \subseteq V(G)$ with $|X| > \delta|G|$ such that one of $G[X], \overline{G[X]}$ has maximum degree at most $\varepsilon\delta|G|$, where $\delta = 1/(2 - \varepsilon)$.**

**Proof.** Let $G$ be a non-null cograph, and let $1/2 \leq \varepsilon < 1$. Let $\delta = 1/(2 - \varepsilon)$ and $d = \varepsilon/(2 - \varepsilon)$; we must show that there is a set $X \subseteq V(G)$ with $|X| > \delta|G|$ such that one of $G[X], \overline{G[X]}$ has maximum degree at most $d|G|$.

We partition $V(G)$ into sets $X_1, \ldots, X_k$ as follows. Suppose that $i \geq 1$ and we have defined $X_1, \ldots, X_{i-1}$, such that $V(G) \neq X_1 \cup \cdots \cup X_{i-1}$. Let $Y = V(G) \setminus (X_1 \cup \cdots \cup X_{i-1})$. If $|Y| = 1$, let $X_i = Y$ and $k = i$. Now we assume that $|Y| > 1$, and define $X_i$ as follows. By 2.1, one of $G[Y], \overline{G[Y]}$ is not connected. Let $X_i$ be a subset of $Y$ that is the vertex set of a component of one of $G[Y], \overline{G[Y]}$, chosen with $|X_i|$ minimum. Thus $|X_i| \leq |Y|/2$, and in particular $|X_i| \leq (1/2)|X_i|$.

(1) We may assume that $|X_i| \leq \delta|G|/2$ for $1 \leq i \leq k - 1$.

Suppose that some $|X_i| > \delta|G|/2$, and let $Y = V(G) \setminus (X_1 \cup \cdots \cup X_i)$. Choose $A \subseteq X_i$ with $|A| = |\delta|G|/2 + 1$. As we saw, $|Y| \geq |X_i|$, and so there exists $B \subseteq Y$ with $|B| = |A|$. Now the set $A \cap B$ has cardinality more than $\delta|G|$. Moreover, from the construction, $X_i$ is either complete or anticomplete to $Y$, and by taking complements if necessary, we may assume the former. But then every vertex in $A \cap B$ has no neighbours in $B$ and has at most $|A| - 1 \leq \delta|G|/2 \leq \varepsilon\delta|G|$ neighbours in $A$, and the same for $B$; and so setting $X = A \cup B$ satisfies the theorem. This proves (1).

We may assume that $|G| \geq 2$ and so $k \geq 2$. If $d|G| \geq |G| - 1$, then the theorem is satisfied with $X = V(G)$ (because $\delta < 1$ and every vertex has at most $d|G|$ neighbours in $G$). So we may assume that $d|G| < |G| - 1$. Choose $h$ with $0 \leq h \leq k - 1$, minimum such that $|X_{h+1} \cup \cdots \cup X_k| \leq d|G| + 1$. (This is possible since the condition is satisfied when $h = k - 1$). Since $|G| > d|G| + 1$ it follows that $h \geq 1$. By moving to the complement if necessary, we may assume that there $X_h, Y$ are anticomplete, where $Y = X_{h+1} \cup \cdots \cup X_k$. Let $I$ be the set of all $i \in \{1, \ldots, h\}$ such that $X_i, Y$ are anticomplete, and let $J$ be the set of all $i \in \{1, \ldots, h\}$ such that $X_i, Y$ are complete. Thus $h \in I$. Moreover, all the sets $X_i (i \in I)$ are pairwise anticomplete, and the sets $X_i (i \in J)$ are pairwise complete.

Choose $Z \subseteq X_h$ such that $|Y \cup Z| = |d|G| + 1|$ (this is possible since $X_h \cup Y > d|G| + 1$ from the minimality of $h$). Let $A$ be the union of $Y$ and the sets $X_i (i \in I)$. Since each of the sets $X_i (i \in I)$ and $Y$ have cardinality at most $d|G| + 1$ by (1), and there are no edges between them, it follows that $G[A]$ has maximum degree at most $d|G|$. Similarly, let $B$ be the union of $Y \cup Z$ and the sets $X_i (i \in J)$; then since these sets all have cardinality at most $d|G| + 1$, and there are no edges of $\overline{G}$ between any two of them, it follows that $\overline{G[B]}$ has maximum degree at most $d|G|$. But $|A| + |B| = |G| + |Y| + |Z|$, and so one of $|A|, |B|$ has cardinality at least $(|G| + |Y| + |Z|)/2$. To
complete the proof it suffices to show that \(|G| + |Y| + |Z|)/2 \geq \delta |G|\). Certainly \(|Y \cup Z| > d|G|\); and hence

\(|G| + |Y| + |Z|)/2 > (1 + d)|G|/2 = (1 + \varepsilon/(2 - \varepsilon))|G|/2 = \delta |G|.

This proves 3.1.

3.1 says that \(|X| > \delta |G|\), and hence \(|X| \geq |\delta |G| + 1|\). Next we show that this is tight.

3.2 Let \(1/2 \leq \varepsilon \leq 1\) and \(\delta = 1/(2 - \varepsilon)\). For each even integer \(2n\), there is a cograph \(G\) with \(2n\) vertices such that if \(X \subseteq V(G)\) and one of \(G[X], \overline{G}[X]\) has maximum degree at most \(\varepsilon \delta |G|\), then \(|X| \leq \delta |G| + 1\) (and hence \(|X| \leq |\delta |G| + 1|\).

**Proof.** Let \(G\) be the “half-graph” with vertex set \(\{a_1, \ldots, a_n, b_1, \ldots, b_n\}\), in which \(\{a_1, \ldots, a_n\}\) is a stable set, \(\{b_1, \ldots, b_n\}\) is a clique, and \(a_i, b_j\) are adjacent if and only if \(i \leq j\). This graph is a cograph; choose \(X \subseteq V(G)\) such that \(G[X]\) has maximum degree at most \(\varepsilon \delta |G|\), with \(|X|\) maximum, and subject to that with \(|X \cap A|\) maximum. Since \(|X| > |G|/2\), it contains a vertex \(b \in B\); and so for each \(a \in A\), since \(b\) dominates \(a\), and we cannot trade \(b\) for \(a\), it follows that \(a \in X\) and so \(A \subseteq X\).

Let \(|X \cap B| = i\) say; then there is a vertex in \(X \cap B\) with \(i\) neighbours in \(A\) and adjacent to all other vertices in \(X \cap B\), and since its degree in \(G[X]\) is at most \(\varepsilon \delta |G|\), we deduce that \(2i - 1 \leq \varepsilon \delta |G|\). So \(|X \cap B| \leq (\varepsilon \delta |G| + 1)/2\), and hence \(|X| \leq |G|/2 + (\varepsilon \delta |G| + 1)/2 = \delta |G| + 1/2\). Similarly (the graph is not quite self-complementary), if \(X \subseteq V(G)\) and \(\overline{G}[X]\) has maximum degree at most \(\varepsilon \delta |G|\), it follows that \(|X| \leq \delta |G| + 1\). This proves 3.2.

What can we say about the values of \(\delta \varepsilon\) through the remainder of the range of \(\varepsilon\)? As far as we know, though it seems unlikely, it may be that \(\delta \varepsilon = \varepsilon\) for all \(\varepsilon \leq 1/2\). To try to settle this, we focussed on \(\delta_{1/3}\), but all we found is an example that shows that \(\delta_{1/3} \leq 3/8\), and more generally \(\delta \varepsilon \leq 1/(4 - 4\varepsilon)\) when \(1/3 \leq \varepsilon < 2/5\), the following. Suppose that \(1/3 \leq \varepsilon \leq 2/3\) and \(\delta > 1/(4 - 4\varepsilon)\); then for sufficiently large integer \(n\), we can choose four disjoint sets \(A, B, C, D\) with

\(|A| < \delta n - 100,\)
\(|B| < (1 - \varepsilon)\delta n - 100,\)
\(|C| < (1 - 2\varepsilon)\delta n - 100,\) and
\(|D| < (1 - \varepsilon)\delta n - 100,\)

and with union of cardinality \(n\). Make \(B\) a clique, and \(A, C, D\) stable sets, and make \(C\) complete to \(D\), and \(A\) complete to \(B \cup C \cup D\). One can check (we omit the details) that there is no \(X \subseteq V(G)\) with \(|X| \geq \delta |G|\) such that one of \(G[X], \overline{G}[X]\) has maximum degree at most \(\varepsilon \delta |G|\).

4 Comparability graphs and other hereditary classes

A graph is a *comparability graph* if there is a partial order on the element set \(V(G)\) such that for all distinct \(u, v \in V(G)\), \(u, v\) are adjacent in \(G\) if and only if one of \(u, v\) is less than the other in the partial order. We will show that the class of comparability graphs satisfies 1.7, because of the following:

4.1 If \(G\) is a comparability graph, then for every \(\varepsilon\) with \(0 \leq \varepsilon \leq 1\), either
there exists $X \subseteq V(G)$ with $|X| \geq \frac{\varepsilon^2}{16(1+\varepsilon)} |G|$ such that $G[X]$ has at most $\varepsilon|X|(|X|-1)$ edges; or

there exists $X \subseteq V(G)$ with $|X| \geq \frac{\varepsilon}{4(1+\varepsilon)} |G|$ such that $G[X]$ has at most $\varepsilon|X|^2$ edges.

**Proof.** We may assume that $\varepsilon \neq 0$. Let $G$ be a comparability graph and let $\prec$ be the corresponding partial order on $V(G)$. Let $t = [\varepsilon^3 |G|/(16(1+\varepsilon))]$. Define another partial order $\prec^*$ on $V(G)$ where $a \prec^* b$ if there are at least $t$ elements $c$ in $V(G)$ such that $a < c < b$. By Mirsky’s theorem (the dual of Dilworth’s theorem), either there is a sequence

$$a_1 \prec^* a_2 \prec^* a_3 \prec^* \cdots$$

with at least $1/\varepsilon + 1$ terms, or we can partition $V(G)$ into fewer than $1/\varepsilon + 1$ sets that are antichains under $\prec^*$. In the former case, any $t$ of the intermediate elements between $a_i$ and $a_{i+1}$ form a part of a complete $\lceil 1/\varepsilon \rceil$-partite subgraph of $G$ with parts of size $t$, and which therefore has at least $t/\varepsilon \geq \varepsilon^2 |G|/(16(1+\varepsilon))$ vertices and has edge density at least $1-\varepsilon$, completing this case. So we may assume that there exists $A \subseteq V(G)$ with $|A| \geq |G|/(1+1/\varepsilon)$ that is an antichain under the partial order $\prec^*$.

Next we obtain an upper bound on the number of triples $a < c < b$ with $a, b, c \in A$. For fixed $a, b$, there are at most $t-1$ choices of $c$, and so there are at most $(t-1)|A|^2/2$ such triples. Let $r = ((t-1)|A|)^{1/2}$, and let $A_0$ be the set of $c \in A$ such that there are at least $r$ elements $a \in A$ with $a < c$ and at least $r$ elements $b \in A$ with $b > c$. So $|A_0| \leq r^{-2}(t-1)|A|^2/2 = |A|/2$. We can partition $A \setminus A_0$ into two subsets $A_1$ and $A_2$, where for every element in $A_1$ there are fewer than $r$ elements in $A$ below it, and for every element in $A_2$ there are fewer than $r$ elements in $A$ above it. Without loss of generality, $|A_1| \geq |A_2|$, so $|A_1| \geq |A|/4 \geq |G|/(4+4/\varepsilon)$.

Since $|G| \leq (4+4/\varepsilon)|A_1|$ and $|A| \leq 4|A_1|$, it follows that $|G| \cdot |A| \leq 16(1+1/\varepsilon)|A_1|^2$. Consequently

$$r^2 = (t-1)|A| \leq (\varepsilon^3/(16(1+\varepsilon)))|G| \cdot |A| \leq (\varepsilon^3/(16(1+\varepsilon)))(16(1+1/\varepsilon)|A_1|^2 = \varepsilon^2 |A_1|^2,$$

and hence $r \leq \varepsilon |A_1|$. It follows that the number of edges with both ends in $A_1$ is at most $|A_1|r \leq \varepsilon |A_1|^2$. Since $|A_1| \geq |G|/(4+4/\varepsilon)$, the second bullet of the theorem holds. This proves 4.1.

Thus, we have proved 1.7 for cographs and for comparability graphs; but there are many other hereditary classes for which we could try to prove it. For instance, does it hold for the class of perfect forests?

Here is another type of hereditary class that we can show satisfies 1.7. A pure pair in a graph $G$ is a pair of disjoint subsets $A, B$ of $G$ such that either every vertex in $A$ is adjacent to every vertex in $B$ (that is, $A$ is complete to $B$), or there are no edges between $A, B$ (that is, $A$ is anticomplete to $B$). Let us say a hereditary class $\mathcal{F}$ has the strong Erdős-Hajnal property if there exists $\delta > 0$, such that every graph $G \in \mathcal{F}$ with $|G| \geq 2$ admits a pure pair $A, B$ with $|A|, |B| \geq \delta |G|$. Such classes exist: the following two results were shown in [4, 5].

4.2 Let $\mathcal{F}$ be a hereditary class of graphs such that some forest is not in $\mathcal{F}$, and the complement of some forest is not in $\mathcal{F}$. Then $\mathcal{F}$ has the strong Erdős-Hajnal property.

4.3 Let $\mathcal{F}$ be a hereditary class of graphs such that for some graphs $H_1, H_2$, there is no subdivision of $H_1$ that belongs to $\mathcal{F}$, and there is no subdivision of $H_2$ whose complement belongs to $\mathcal{F}$. Then $\mathcal{F}$ has the strong Erdős-Hajnal property.
An argument of Sophie Spirkl (private communication) shows the following:

**4.4 Let** \( F \)** be a hereditary class with the strong Erdős-Hajnal property. Then \( F \) satisfies 1.7.

**Proof.** Let \( \delta \) be as in the definition of the strong Erdős-Hajnal property. Let \( c = 5 \log(1/\delta) \) (logarithms in this paper are always to base two), and let \( 0 < \varepsilon \leq 1/2 \). We will show that for every \( G \in F \), there exists \( X \subseteq V(G) \) with \( |X| \geq \varepsilon^c |G| \) such that one of \( G[X], \overline{G}[X] \) has at most \( \varepsilon |X|(|X| - 1) \) edges. Let \( t = \lceil 2 + 2 \log(1/\varepsilon) \rceil \). Suppose first that \( |G| \leq \delta^{-t} \); then
\[
\varepsilon^c |G| \leq \varepsilon^c \delta^{-t} \leq \varepsilon^c \delta^{-3-2 \log(1/\varepsilon)},
\]
and so
\[
\log(\varepsilon^c |G|) \leq -(5 \log(1/\delta))(\log(1/\varepsilon)) + (3 + 2 \log(1/\varepsilon)) \log(1/\delta) \leq 0
\]
since \( \varepsilon \leq 1/2 \). Consequently \( \varepsilon^c |G| \leq 1 \) and the result holds trivially.

Thus we may assume that \( |G| \geq \delta^{-t} \). For each integer \( s \) with \( 0 \leq s \leq t \), we define inductively a set \( A_s \) of pairwise disjoint subsets of \( V(G) \), with the following properties:

- \( |A_s| = 2^s \);
- each member of \( A_s \) has cardinality \( \delta^s |G| \);
- \( A,B \) is a pure pair for all distinct \( A,B \in A_s \); and
- the pattern graph of \( A_s \) is a cograph

where the pattern graph of \( A_s \) means the graph \( P_s \) with vertex set \( A_s \) where distinct \( A,B \in A_s \) are adjacent in \( P_s \) if and only if \( A,B \in A_s \) are complete to \( B \).

The inductive definition is as follows: let \( A_0 = \{ V(G) \} \). Now let \( 1 \leq s \leq t \) and suppose that \( A_{s-1} \) is defined. Since \( \delta^s |G| \geq 1 \), it follows that \( \delta^{s-1} |G| > 1 \). Consequently \( |A| \geq 2 \) for each \( A \in A_{t-1} \), and therefore there is a pure pair \( C_A, D_A \subseteq A \) with \( |C_A|, |D_A| \geq \delta |A| \geq \delta^s |G| \); and hence we may choose \( C_A, D_A \) with cardinality \( \delta^s |G| \). Let \( A_s \) be the union of the sets \( \{ C_A, D_A \} \) over all \( A \in A_{s-1} \). It is easy to see that the four bullets above are satisfied.

Since the pattern graph \( P_t \) of \( A_t \) is a cograph, by 1.3 (with \( \varepsilon \) replaced by \( \varepsilon/2 \)) there exists \( |R| \geq (\varepsilon/2)|P_t| = \varepsilon 2^{t-1} \) such that one of \( P_t[R], \overline{P_t}[R] \) has maximum degree at most (\( \varepsilon/2)^2 |P_t| = \varepsilon^2 2^{t-2} \), and from the symmetry we may assume the first. Let \( X \subseteq V(G) \) be the union of the sets \( A \in R \). Thus
\[
|X| \geq \varepsilon 2^{t-1} |\delta^t |G| | \geq (\varepsilon/2)(2\delta^t |G|).
\]

Since
\[
\log((2/\varepsilon)(2\delta)^{-t}) = 1 + \log(1/\varepsilon) + t(\log(1/\delta) - 1) \\
\leq 2 \log(1/\varepsilon) + (3 + 2 \log(1/\varepsilon))(\log(1/\delta) - 1) \\
= (3 + 2 \log(1/\varepsilon)) \log(1/\delta) - 3 \\
\leq 5 \log(1/\varepsilon) \log(1/\delta) = c \log(1/\varepsilon)
\]

it follows that
\[
|X| \geq (\varepsilon/2)(2\delta^t |G|) \geq \varepsilon^c |G|.
\]

Moreover, every vertex in \( X \) has degree in \( G[X] \) at most \((1 + 2^t - 2^t \varepsilon^2)|\delta^t |G| | - 1 \). Since \( 1 + 2^t - 2^t \varepsilon^2 \leq 2^{t-1} \varepsilon^2 \) (since \( 2^{t-2} \varepsilon^2 \geq 1 \)) and \( |\delta^t |G| | \leq 2\delta^t |G| \) (since \( |\delta^t |G| | \geq 1 \)) it follows that \( G[X] \) has maximum degree at most \( \varepsilon^2 (2\delta^t |G|) - 1 \leq 2\varepsilon(|X| - 1) \), and so has at most \( \varepsilon |X|(|X| - 1) \) edges. This proves 4.4. 

\[\square\]
5 Viruses and vertex-substitution

So far we have been considering classes of graphs in which there is no induced subgraph of a certain type, but it seems better in some respects just to assume that there are not “many” copies of the subgraph, and now we turn to that. First, let us prove 1.5, which we restate:

5.1 If $H_1, H_2$ are viruses and $H$ is obtained by substituting $H_2$ for a vertex of $H_1$, then $H$ is a virus.

**Proof.** Let $H$ be obtained by substituting $H_2$ for a vertex $v$ say of $H_1$. For $i = 1, 2$, since $H_i$ is a virus, there exist $c_i, d_i$ as in the definition of “virus”. Let $c = \max(c_1, c_2 + d_1 + 1)$, and $d = (|H_2| + 1)(d_1 + 1) + d_2$. To show that $H$ is a virus, we will show that:

1. For every graph $G$ and all $\varepsilon$ with $0 < \varepsilon \leq 1/2$, either
   a. there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon^n |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges; or
   b. there are at least $\varepsilon^d |G|^{|H|}$ copies of $H$ in $G$.

Since $\varepsilon^n |G| \geq \varepsilon^n |G|$, we may assume that there is no $X \subseteq V(G)$ with $|X| \geq \varepsilon^n |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges, since otherwise the first bullet of (1) holds. Consequently, from the choice of $c_1, d_1$, there are at least $\varepsilon^d |G|^{|H_1|}$ copies of $H_1$ in $G$. For each copy $\phi$ of $H_1 \setminus \{v\}$ in $G$, let $N(\phi)$ be the set of all vertices $u \in V(G)$ such that extending $\phi$ by mapping $v$ to $u$ gives a copy of $H_1$ in $G$, and let $n(\phi) = |N(\phi)|$. Let $\Phi$ be the set of all copies of $H_1 \setminus \{v\}$ in $G$; then

$$\sum_{\phi \in \Phi} n(\phi) \geq \varepsilon^d |G|^{|H_1|}.$$

Let $\Psi$ be the set of all $\phi \in \Phi$ such that $n(\phi) \geq \varepsilon^{d_1 + 1} |G|$. Since

$$\sum_{\phi \in \Phi \setminus \Psi} n(\phi) \leq \sum_{\phi \in \Phi \setminus \Psi} \varepsilon^{d_1 + 1} |G| \leq |G|^{|H_1| - 1} \varepsilon^{d_1 + 1} |G|,$$

it follows that

$$\sum_{\phi \in \Psi} n(\phi) \geq \varepsilon^{d_1} |G|^{|H_1|}(1 - \varepsilon) \geq \varepsilon^{d_1 + 1} |G|^{|H_1|}.$$

Since $n(\phi) \leq |G|$, we deduce that $|\Psi| \geq \varepsilon^{d_1 + 1} |G|^{|H_1| - 1}$.

Let $\phi \in \Psi$. Thus $|N(\phi)| = n(\phi) \geq \varepsilon^{d_1 + 1} |G|$. From the choice of $c_2, d_2$, either there exists $X \subseteq N(\phi)$ with $|X| \geq \varepsilon^{c_2} |N(\phi)|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges, or there are $\varepsilon^{d_2} |N(\phi)|^{H_2}$ copies of $H_2$ in $G[N(\phi)]$. In the first case, since $\varepsilon^{c_2} |N(\phi)| \geq \varepsilon^{c_2} \varepsilon^{d_1 + 1} |G| \geq \varepsilon^n |G|$, the first bullet of (1) holds; so we may assume that there are at least

$$\varepsilon^{d_2} |N(\phi)|^{H_2} \geq \varepsilon^{d_2} \varepsilon^{(d_1 + 1)|H_2|} |G|^{H_2}$$

copies of $H_2$ in $G[N(\phi)]$, and hence each $\phi \in \Psi$ can be extended to at least $\varepsilon^{d_2} \varepsilon^{(d_1 + 1)|H_2|} |G|^{H_2}$ copies of $H$. Since $|\Psi| \geq \varepsilon^{d_1 + 1} |G|^{|H_1| - 1}$, there are at least

$$\varepsilon^{d_1 + 1} |G|^{|H_1| - 1} \geq \varepsilon^{d_2} \varepsilon^{(d_1 + 1)|H_2|} |G|^{H_2} = \varepsilon^{d_1 + 1 + d_2 + (d_1 + 1)|H_2|} |G|^{H} \geq \varepsilon^d |G|^{|H|}$$
copies of $H$ in $G$, and hence the second bullet of (1) holds. This proves (1), and hence shows that $H$ is a virus, and proves 5.1.\[\square\]
Let us deduce 1.6, which we restate:

5.2 Every graph on at most four vertices is a virus.

Proof. Every graph on at most two vertices is a virus from the definition. Every graph on three or four vertices, apart from $P_4$, can be obtained through vertex-substitution from smaller graphs. So from 1.5 it suffices to prove that $P_4$ is a virus. The proof uses both 1.3 and a polynomial bound in the induced graph removal lemma for $P_4$. The induced graph removal lemma (see [3, 6, 11]) says that for each graph $H$ and $0 < \varepsilon \leq 1/2$ there exists $\delta > 0$ such that every graph $G$ with at most $\delta |G|^2$ copies of $H$ can be made $H$-free by adding or deleting at most $\varepsilon |G|^2$ edges. For $H = P_4$, Alon and the first author [1] proved a polynomial bound, that is, we can take $\delta = \varepsilon^d$, where $d$ is an absolute constant.

To show that $P_4$ is a virus, let $d$ be as above, and let $0 < \varepsilon \leq 1/2$. We will show that either

- there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|/2 \geq \varepsilon^2 |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges; or

- there are at least $(\varepsilon/2)^3 d |G|^4 \geq \varepsilon^6 d |G|^4$ copies of $P_4$ in $G$.

Let us first dispose of a trivial case, when $\varepsilon |G|/2 \leq 3$. Then the first bullet holds if $|G| \geq 6$ (because then $G$ or $\overline{G}$ has a triangle), and also if $|G| \leq 5$ (because then $\varepsilon |G|/2 \leq 2$). So we may assume that $\varepsilon |G|/2 > 3$. From the choice of $d$ (with $\varepsilon$ replaced by $(\varepsilon/2)^3$), either $G$ contains at least $(\varepsilon/2)^3 d |G|^4$ copies of $P_4$ (in which case we are done), or we can obtain a $P_4$-free graph $G'$ with the same vertex set as $G$ by adding or deleting at most $(\varepsilon/2)^3 |G|^2$ edges from $G$. In the latter case, by 1.3, there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|/2$ such that one of $G'[X], \overline{G'}[X]$ has maximum degree at most $(\varepsilon/2)^3 |G|$. Then one of $G[X], \overline{G}[X]$ has at most $$(\varepsilon/2)^3 |G|^2 + (\varepsilon/2)^2 |G||X|/2 \leq \frac{3}{4} \varepsilon |X|^2 \leq \varepsilon |X|(|X| - 1)$$ edges (since $\varepsilon |G|/2 > 3$). This proves 5.2.

6 Viral sets

There is a natural extension of the concept of a virus to finite sets of excluded induced subgraphs. Let us say a set $\mathcal{H}$ of graphs is viral if there exist $c, d > 0$ such that for every graph $G$ and every $\varepsilon$ with $0 < \varepsilon \leq 1/2$, either

- there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon^c |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges; or

- for some $H \in \mathcal{H}$, there are at least $\varepsilon^d |G| |H|$ copies of $H$ in $G$.

(This makes sense if $\mathcal{H}$ is infinite, but we will use it only for finite sets.) We call $c, d$ viral exponents for $\mathcal{H}$. We will have more to say about viral sets, but first, what about an analogue of 1.5? It is easy to see that if $\mathcal{F}, \mathcal{H}$ are finite viral sets, and for each $F \in \mathcal{F}$ and $H \in \mathcal{H}$, $J(F, H)$ is a graph obtained by substituting $H$ for some vertex of $F$, then the set of graphs $\{J(F, H) : F \in \mathcal{F}, H \in \mathcal{H}\}$ is viral; but one can be more delicate. For instance, we have:
6.1 Let $H = \{H_1, \ldots, H_k\}$ be viral, and for $1 \leq i \leq k$ let $J_i$ be obtained by substituting a copy of $H_i$ for some vertex of $H_i$. Then $\{J_1, \ldots, J_k\}$ is viral.

This follows by repeated application of the following:

6.2 Let $H = \{H_1, \ldots, H_k\}$ be viral, and let $J_1$ be obtained by substituting a copy of $H_1$ for some vertex of $H_1$. Then $\{J_1, H_2, \ldots, H_k\}$ is viral.

**Proof.** Let $c, d > 0$ be as in the definition of a viral set, for $H$. Let $c' = c + d + 1$ and

$$d' = 2d + 1 + (d + 1) \max(|H_1|, \ldots, |H_k|) + \log(k).$$

Let $G$ be a graph. We must show that:

1. One of the following holds:
   - there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges; or
   - for some $i \in \{2, \ldots, k\}$, there are at least $\varepsilon^d |G|^{\varepsilon |H_i|}$ copies of $H_i$ in $G$; or
   - there are at least $\varepsilon^d |G|^{|J_1|}$ copies of $J_1$ in $G$.

From the choice of $c, d$, and since $c' \geq c$, we may assume that for some $i \in \{2, \ldots, k\}$, there are at least $\varepsilon^d |G|^{\varepsilon |H_i|}$ copies of $H_i$ in $G$; and since $d' \geq d$, we may assume that $i = 1$. Let $J_1$ be obtained from $H_1$ by substituting a copy of $H_1$ for some vertex $v_1$ of $H_1$; and for $2 \leq j \leq k$, let $J_j$ be the graph obtained by substituting $H_j$ for the vertex $v_1$ of $H_1$. Let $\Phi$ be the set of all copies of $H_1 \setminus \{v_1\}$ in $G$. 

For each $\phi \in \Phi$, let $N(\phi)$ be the set of all $v \in V(G)$ such that mapping $v_1$ to $v$ extends $\phi$ to a copy of $H_1$; and let $\Psi$ be the set of all $\phi \in \Phi$ such that $|n(\phi)| \geq \varepsilon^{d+1}|G|$. As in the proof of 5.1, it follows that $|\Psi| \geq \varepsilon^{d+1}|G|^{\varepsilon |H_1| - 1}$.

2. We may assume that there exists $j \in \{1, \ldots, k\}$ such that there are at least $\varepsilon^{d'} |G|^{|J_j|}$ copies of $J_j$ in $G$.

Let $\phi \in \Psi$. From the choice of $c, d$, either there exists $X \subseteq N(\phi)$ with $|X| \geq \varepsilon |N(\phi)|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges, or for some $j \in \{1, \ldots, k\}$ there are $\varepsilon^d |N(\phi)|^{\varepsilon |H_j|}$ copies of $H_j$ in $G[N(\phi)]$. In the first case, since $\varepsilon^d |N(\phi)| \geq \varepsilon^{d+1} |G| \geq \varepsilon^d |G|$, the theorem holds; so we may assume that for each $\phi \in \Psi$, there exists $j \in \{1, \ldots, k\}$ such that $\phi$ extends to at least

$$\varepsilon^d |N(\phi)|^{\varepsilon |H_j|} \geq \varepsilon^d |G|^{|J_j|} |G|^{\varepsilon |H_j|} = \varepsilon^{d+(d+1)|H_j|} |G|^{|H_j|}$$

copies of $J_j$.

There are at least $|\Psi|/k \geq \varepsilon^{d+1}|G|^{\varepsilon |H_1| - 1}/k$ copies $\phi \in \Psi$ with the same value of $j$; and so there exists $j \in \{1, \ldots, k\}$ such that there are at least

$$\varepsilon^{d+1}|G|^{\varepsilon |H_1| - 1} \varepsilon^{d+(d+1)|H_j|} |G|^{|H_j|}/k = \varepsilon^{2d+1+(d+1)|H_j|} |G|^{|J_j|}/k \geq \varepsilon^{d'} |G|^{|J_j|},$$

copies of $J_j$ in $G$ (the last because $\varepsilon \log k \leq (1/2) \log k = 1/k$). This proves (2).
Now there are two cases, $j = 1$ and $j \geq 2$. If $j = 1$, then the third bullet of (1) holds. If $j > 1$, then since each copy of $H_j$ in $G$ only extends to at most $|G|^{\lceil |H_j| - 1 \rceil}$ copies of $J_j$, there are at least

$$\varepsilon^d |G|^{\lceil |J_j| / |G|^{\lceil |H_j| - 1 \rceil}} = \varepsilon^d |G|^{\lceil |H_j| - 1 \rceil}$$

copies of $H_j$ in $G$ and the second bullet of (1) holds. This proves (1), and hence proves 6.2 and 6.1.

Here is another useful lemma for manipulating viral sets:

**6.3** Let $\mathcal{H}, \mathcal{H}'$ be sets of graphs, and suppose that each member of $\mathcal{H}$ has an induced subgraph isomorphic to a member of $\mathcal{H}'$. If $\mathcal{H}$ is viral then $\mathcal{H}'$ is viral.

**Proof.** Let $c, d$ be viral exponents for $\mathcal{H}$. We will show that $c, d$ are also viral exponents for $\mathcal{H}'$. Let $G$ be a graph and let $0 < \varepsilon \leq 1/2$. It follows that either

- there exists $X \subseteq V(G)$ with $|X| \geq \varepsilon |G|$ such that one of $G[X], \overline{G}[X]$ has at most $\varepsilon |X|(|X| - 1)$ edges; or
- for some $H \in \mathcal{H}$, there are at least $\varepsilon d |G|^{\lceil |H| \rceil}$ copies of $H$ in $G$.

If the first holds, we are done, so we assume that the second holds. Choose $H' \in \mathcal{H}'$ isomorphic to an induced subgraph of $H$. Each copy of $H$ in $G$ is an extension of a copy of $H'$ in $G$, and each copy of $H'$ in $G$ extends to at most $|G|^{\lceil |H| - |H'| \rceil}$ copies of $H$; so there are at least $\varepsilon d |G|^{\lceil |H'| \rceil}$ copies of $H'$ in $G$, and again we are done. This proves 6.3.

## 7 Sparse pairs of sets

There is also a viral version of pure pairs and 4.4 that we found useful. Let us say that disjoint subsets $A, B$ of the vertex set of a graph are $c$-sparse to each other if there are at most $c|A| \cdot |B|$ edges of $G$ between $A, B$; and and $c$-dense to each other if they are $c$-sparse to each other in $\overline{G}$. Such pairs of sets will provide an analogue of pure pairs that work better with the viral version of 1.2. First, we need a lemma that allows us to tidy up a sparse pair.

**7.1** Let $0 < c, d < 1$, and let $A, B$ be disjoint subsets of $V(G)$, both of cardinality at least $d|G|$, and $c$-sparse to each other. Then there exist $A' \subseteq A$ and $B' \subseteq B$, both of cardinality at least $d|G|/2$, such that each vertex of $A'$ has at most $4cd|G|$ neighbours in $B'$ and vice versa.

**Proof.** If $d|G| \leq 1$, the result is clear (since $c < 1$), so we assume that $d|G| > 1$. By averaging over all subsets $A_1$ of $A$ with cardinality $[d|G|]$ it follows that there exists $A_1 \subseteq A$ with $|A_1| = [d|G|]$ such that $A_1, B$ are $c$-sparse to each other; and similarly there exists $B_1 \subseteq B$ with $|B_1| = [d|G|]$ such that $A_1, B_1$ are $c$-sparse to each other. Since there are at most $c|A_1| \cdot |B_1|$ edges between $A_1, B_1$, at most $|A_1|/2$ vertices in $A_1$ have at least $2c|B_1|$ neighbours in $B_1$, and so there exists $A_2 \subseteq A_1$ with $|A_2| \geq |A_1|/2$ such that every vertex in $A_2$ has fewer than $2c|B_1|$ neighbours in $B_1$. Since $|A_1|/2 \geq d|G|/2$, there exists $A' \subseteq A_2$ with $|A'| = [d|G|/2]$ such that every vertex in $A'$ has fewer than $2c|B_1|$ neighbours in $B_1$. Similarly there exists $B' \subseteq B_1$ with $|B'| = [d|G|/2]$ such that every
vertex in \( B' \) has fewer than \( 2|A_1| \) neighbours in \( A_1 \). Now \(|B_1| \leq 2d|G|\) since \(|B_1| = [d|G|]\) and \(d|G| \geq 1\). Hence every vertex in \( A' \) has fewer than \( 4cd|G| \) neighbours in \( B_1 \); and similarly every vertex in \( B' \) has fewer than \( 4cd|G| \) neighbours in \( A_1 \). But then \( A', B' \) satisfy the theorem. This proves 7.1.

One way to prove that a set of graphs is viral is to apply the following, a relative of 4.4:

### 7.2 Let \( \mathcal{H} \) be a finite set of graphs. Suppose that there exist \( a, b > 0 \) such that for every graph \( G \) with \(|G| \geq 2\) and every \( \epsilon \) with \( 0 \leq \epsilon \leq 1/2 \), either:

- there exist disjoint \( A, B \subseteq V(G) \), both with cardinality at least \( a|G| \), that are \( \epsilon \)-sparse or \( \epsilon \)-dense to each other; or
- for some \( H \in \mathcal{H} \), there are at least \( \epsilon^b|G|^{|H|} \) copies of \( H \) in \( G \).

Then \( \mathcal{H} \) is viral.

**Proof.** We may assume that \( a \leq 1/4 \). Choose \( p \) such that \( a = (1/2)^p \). Let \( c = 5p + 2 \) and let \( d \) be the maximum of \( b(5p + 6) + 5(p + 1)|H| \) over all \( H \in \mathcal{H} \). Let \( G \) be a graph and let \( 0 \leq \epsilon \leq 1/2 \). We will show that:

1. **One of the following holds:**
   - there exists \( X \subseteq V(G) \) with \(|X| \geq \epsilon^c|G|\) such that one of \( G[X], \overline{G}[X] \) has at most \( \epsilon|X|(|X| - 1) \) edges; or
   - for some \( H \in \mathcal{H} \), there are at least \( \epsilon^d|G|^{|H|} \) copies of \( H \) in \( G \).

   Choose \( t \) minimum such that \( (1/2)^t \leq \epsilon^4 \). So \( (1/2)^t \geq \epsilon^4/2 \geq \epsilon^5 \). Since \( c \geq 5p + 2 \), and \( 2^t \leq \epsilon^{-5} \), it follows that
   
   \[
   \epsilon^c \leq \epsilon^{5(p+1)} \leq (1/2)^{(p+1)t}.
   \]

   Moreover, we may assume that \( \epsilon^c|G| > 2 \), since otherwise the first bullet above holds; so \( (1/2)^{(p+1)t}|G| > 2 \), and therefore \( (a/2)^t|G| > 2 \).

   Define \( \epsilon_1 = \epsilon^{5p+6} \). For each integer \( s \) with \( 0 \leq s \leq t \), we define inductively a set \( \mathcal{A}_s \) of \( 2^s \) pairwise disjoint subsets of \( V(G) \), each of cardinality \( \lfloor (a/2)^s|G| \rfloor \), and a cograph \( P_s \) with vertex set \( \mathcal{A}_s \), with the following requirement, where \( f_s \) is defined by
   
   \[
   f_s = 4\epsilon_1 \left( 1 + (a/2) + (a/2)^2 + \cdots + (a/2)^{s-1} \right) |G|.
   \]

   Let us say a pair \( u, v \) of vertices is \( s \)-contrary if there exist distinct \( A, B \in \mathcal{A}_s \) with \( u \in A \) and \( v \in B \), and \( u, v \) are adjacent in \( G \) if and only if \( A, B \) are \( A, B \) are nonadjacent in \( P_s \). Then for each \( A \in \mathcal{A}_s \) and each \( u \in A \), we require that there are at most \( f_s \) vertices \( v \) such that \( u, v \) are \( s \)-contrary.

   The inductive definition is as follows. Define \( \mathcal{A}_0 = \{V(G)\} \), and let \( P_0 \) be the graph with vertex set \( \mathcal{A}_0 \) (and therefore with only one vertex). Now let \( 1 \leq s \leq t \), and suppose that \( \mathcal{A}_{s-1} \) and \( P_{s-1} \) are defined. Let \( A \in \mathcal{A}_{s-1} \). Thus \(|A| \geq (a/2)^t|G| > 2 \), and so, from the hypothesis applied to \( G[A] \), with \( \epsilon \) replaced by \( \epsilon_1 \), either

   - there exist disjoint \( C, D \subseteq A \), both with cardinality at least \( a|A| \), that are \( \epsilon_1 \)-sparse or \( \epsilon_1 \)-dense to each other; or
We will show that the last expression in the above is at most $2\varepsilon$ because

$$|G|^{|H|} \geq \varepsilon^b(a/2)^{s-1}|H| \geq \varepsilon^d|G|^{|H|},$$

the last because

$$\varepsilon^b(a/2)^{|H|} = \varepsilon^{b(5p+6)}(1/2)^{(p+1)|H|} \geq \varepsilon^{b(5p+6)(5p+1)|H|} \geq \varepsilon^{(5p+6)(5p+1)|H|} \geq \varepsilon^d.$$  

Thus we may assume that there exist disjoint $C,D \subseteq A$, both with cardinality at least $a|A|$, that are $\varepsilon_1$-sparse or $\varepsilon_1$-dense to each other. Since $|A| \geq (a/2)^{s-1}|G|$, 7.1 (applied in the complement if $C,D$ are $\varepsilon_1$-dense) implies that there exist $C_A \subseteq C$ and $D_A \subseteq D$, both with cardinality at least $(a/2)^s|G|$, such that either every vertex of $C_A$ has at most $4\varepsilon(a/2)^{s-1}|G|$ neighbours in $B_A$ and vice versa, or the same holds in $\overline{G}$. By replacing $C_A,D_A$ by subsets if necessary, we may assume that $|C_A|,|D_A| = [(a/2)^s|G|]$. Let $\mathcal{H}_s$ be the union of the sets $\{C_A,D_A\}$ over all $A \in \mathcal{A}_{s-1}$, and define $P_s$ in the natural way (that is, $C,A,D$ are adjacent in $P_s$ if and only if every vertex in $C_A$ has at most $4\varepsilon(a/2)^{s-1}|G|$ neighbours in $D_A$ and vice versa, and for distinct $A,A' \in \mathcal{A}_{s-1}$, if $A,A'$ are adjacent in $P_{s-1}$ then $C_A,D_A$ are both adjacent to both of $C_{A'},D_{A'}$ in $P_s$, and otherwise the four pairs are nonadjacent in $P_s$.)

To complete the inductive definition, we must check that for each $A \in \mathcal{A}_{s-1}$, and each $u \in C_A \cap D_A$, there are at most $f_s$ vertices $v$ such that $u,v$ are s-contrary. To show this we may assume from the symmetry that $u \in C_A$ and $C_A,D_A$ are nonadjacent in $P_s$. Now $u,v$ are s-contrary if and only if either $u,v$ are $(s-1)$-contrary, or $v \in D_A$ and $u,v$ are adjacent in $G$. There are only $f_{s-1}$ vertices $v$ such that $u,v$ are $(s-1)$-contrary, and there are only $4\varepsilon(a/2)^{s-1}|G|$ vertices $v \in D_A$ such that $u,v$ are adjacent in $G$. Since $f_s = f_{s-1} + 4k\varepsilon(a/2)^{s-1}|G|$, this verifies the conditions of the inductive definition and so completes the definition.

From 1.3 applied to $P_t$ (with $\varepsilon$ replaced by $\varepsilon/2$), there exists $|R| \geq \varepsilon/2|P_t|$ such that one of $P_t[R],\overline{P}_t[R]$ has maximum degree at most $(\varepsilon/2)^2|P_t| = (\varepsilon/2)^2t$, and from the symmetry we may assume the first. Let $X \subseteq V(\overline{G})$ be the union of the sets $A \in R$. Thus

$$|X| \geq (\varepsilon/2)^2(a/2)^t|G| = (\varepsilon/2)a^t|G|.$$  

Since $(1/2)^t \geq \varepsilon^2/8 \geq \varepsilon^5$, it follows that

$$(\varepsilon/2)a^t \geq \varepsilon^2(1/2)^pt \geq \varepsilon^2(\varepsilon^5)p = \varepsilon^{5p+2} = \varepsilon^c,$$

and so $|X| \geq \varepsilon^c|G|$.

Moreover, let $m = \lceil (a/2)^t|G| \rceil$. Since

$$f_t \leq 4\varepsilon|G|/(1-a/2) = 4\varepsilon^{5p+6}|G|/(1-a/2) \leq \varepsilon^{5p+3}|G|$$

(because $\varepsilon \leq 1/2$ and $1 - a/2 \geq 1/2$), it follows that every vertex in $X$ has degree in $G[X]$ at most

$$(1 + (\varepsilon/2)^22^t)m + f_t - 1 \leq m + (\varepsilon/2)^22^tm + \varepsilon^{5p+3}|G| - 1.$$

We will show that the last expression in the above is at most $2\varepsilon(|X| - 1)$, and since $|X| \geq (\varepsilon/2)^2tm$, it suffices to show that

$$m + (\varepsilon/2)^22^tm + \varepsilon^{5p+3}|G| - 1 \leq \varepsilon^22^tm - 2\varepsilon.$$  

13
To show this, observe that
\[
\begin{align*}
m &\leq \varepsilon^4 2^t m \leq \varepsilon^2 2^t m / 4 \\
(\varepsilon/2)^2 2^t m &\leq \varepsilon^2 2^t m / 4 \\
\varepsilon 5p + 3|G| &\leq \varepsilon^2 (1/2)^p |G| / 2 \leq \varepsilon^2 2^t (a/2)^i |G| / 2 \leq \varepsilon^2 2^t m / 2 \\
-1 &\leq -2\varepsilon
\end{align*}
\]
and the claim follows by adding. This proves 7.2.

Because of the result of 4.2, it follows from 4.4 that for every two forests \(T_1, T_2\), the hereditary class of graphs that are both \(T_1\)-free and \(T_2\)-free satisfies 1.7. But if we use 7.2 we can sometimes obtain a stronger theorem. For instance, we will use 7.2 to show that for every two paths \(P, P'\), the set \(\{P, \overline{P}\}\) is viral. (We have not yet been able to do the same for a general pair of forests.) The proof breaks into several steps. First, we need the following strengthening of Rödl’s theorem, due to Nikiforov [13]:

7.3 For all \(\varepsilon > 0\) and every graph \(H\), there exist \(\gamma, \delta > 0\) such that if \(G\) is a graph containing fewer than \(\gamma |G|^i H\) copies of \(H\), then there exists \(D \subseteq V(G)\) with \(|D| \geq \delta |G|\) such that one of \(G[D], \overline{G[D]}\) has at most \(\varepsilon |D|(|D| - 1)\) edges.

This can be formulated equivalently (we omit the proof of equivalence, which is easy) as:

7.4 For all \(\varepsilon > 0\) and every graph \(H\), there exist \(\gamma, \delta > 0\) such that if \(G\) is a graph containing fewer than \(\gamma |G|^i H\) copies of \(H\), then there exists \(D \subseteq V(G)\) with \(|D| \geq \delta |G|\) such that one of \(G[D], \overline{G[D]}\) has maximum degree at most \(\varepsilon |D|\).

Second, let us show:

7.5 Let \(G\) be a graph in which every vertex has degree at most \(|G|/8\). Let \(A \subseteq V(G)\) with \(|G|/4 \leq |A| \leq |G|/2\), and let \(B = V(G) \setminus A\). Let \(0 \leq \varepsilon \leq 1/2\) with \(1/\varepsilon \leq |G|/16\). Then either:

- there exist \(X \subseteq A\) with \(|X| \geq |G|/8\), and \(Y \subseteq B\) with \(|Y| \geq |G|/4\), such that every vertex in \(X\) has at most \(4\varepsilon|Y|\) neighbours in \(Y\); or

- there is a set \(\mathcal{F}\) of pairwise disjoint subsets of \(A\), with \(|\mathcal{F}| \geq \varepsilon |G|/8\), such that for each \(F \in \mathcal{F}\) there are at least \(|B|/4\) and at most \(|B|/2\) vertices in \(B\) with a neighbour in \(F\).

**Proof.** For \(X \subseteq A\), we denote by \(N_B(X)\) the set of vertices in \(B\) with a neighbour in \(X\). Let \(\mathcal{F}\) be a set of pairwise disjoint subsets of \(A\), such that \(|F| \leq 1/\varepsilon\) and \(|B|/4 \leq |N_B(F)| \leq |B|/2\) for each \(F \in \mathcal{F}\). Choose \(\mathcal{F}\) maximal; then we may assume that \(|\mathcal{F}| \leq \varepsilon |G|/8\), because otherwise the second bullet holds. Let \(Z\) be the union of the members of \(\mathcal{F}\); thus \(|Z| \leq |G|/8\). Choose \(C \subseteq A \setminus Z\) such that
\[
\varepsilon |G| \cdot |C| \leq |N_B(C)| \leq |B|/2,
\]
where \(N_B(C)\) denotes the set of vertices in \(B\) with a neighbour in \(C\). (This is possible, since we may take \(C = \emptyset\).) It follows that
\[
\varepsilon |G| \cdot |C| \leq |B|/2 \leq 3|G|/8,
\]

14
and so \( \varepsilon|C| \leq 3/(8\varepsilon) \). If \( |N_B(C)| \geq |B|/4 \), we may add \( C \) to \( F \), contrary to the maximality of \( F \); so \( |N_B(C)| < |B|/4 \).

Suppose that there exists \( v \in A \setminus (Z \cup C) \) that has at least \( \varepsilon|G| \) neighbours in \( B \setminus N_B(C) \). It has at most \( |G|/8 \) neighbours, since every vertex of \( G \) has degree at most \( |G|/8 \). Let \( C' = C \cup \{v\} \); then

\[
\varepsilon|G| \cdot |C'| = \varepsilon|G| \cdot |C| + \varepsilon|G| \leq |N_B(C')| \leq |N_B(C)| + |G|/4 \leq |B|/4 + |G|/8 \leq |B|/2,
\]

contrary to the maximality of \( C \). Thus there is no such \( v \), and therefore every vertex in \( A \setminus (F \cup C) \) has fewer than \( \varepsilon|G| \) neighbours in \( B \setminus N_B(C) \). Since

\[
|A \setminus (F \cup C)| \geq |G|/4 - |G|/8 - 1/\varepsilon \geq |G|/16
\]

(because \( 1/\varepsilon \leq |G|/16 \)), and

\[
|B \setminus N_B(C)| \geq 3|B|/4 \geq |G|/4,
\]

and \( \varepsilon|G| \leq 4\varepsilon|B \setminus N_B(C)| \), the first bullet of the theorem holds. This proves 7.5.

This is used to show:

**7.6** Let \( k \geq 1 \) be an integer. Let \( G \) be a graph in which every vertex has degree at most \( 2^{-k-1}|G| \). Let \( A \subseteq V(G) \) with \( |G|/4 \leq |A| \leq |G|/2 \), and let \( 0 \leq \varepsilon \leq 1/2 \) with \( 1/\varepsilon \leq 2^{-k-4}|G| \). Then either:

- there exist disjoint \( X,Y \subseteq V(G) \) with \( |X| \geq 2^{-k-1}|G| \) and \( |Y| \geq 2^{-k}|G| \), such that every vertex in \( X \) has at most \( 4\varepsilon|Y| \) neighbours in \( Y \); or

- there are at least \( (\varepsilon/8)^k|G|^k \) copies of \( P_k \) in \( G \), each with first vertex in \( A \) and no other vertex in \( A \).

**Proof.** We proceed by induction on \( k \). If \( k = 1 \), the second bullet holds, since \( |A| \geq |G|/4 \geq (\varepsilon/8)|G| \); so we may assume that \( k \geq 2 \) and the result holds for \( k - 1 \). We may assume that the first bullet of 7.5 does not hold, since otherwise the first bullet of the theorem holds. So by 7.5, there exists \( F \) as in the second bullet of 7.5. Let \( B = V(G) \setminus A \), let \( F \in F \), and let \( A' \) be the set of vertices in \( B \) with a neighbour in \( F \). Thus \( |B|/4 \leq |A'| \leq |B|/2 \). Every vertex of \( G[B] \) has degree at most \( 2^{-k-1}|G| \leq 2^{-k}|B| \), and \( 1/\varepsilon \leq 2^{-k-4}|G| \leq 2^{-k-3}|B| \). From the inductive hypothesis on \( k \), applied to \( A' \) and \( G[B] \), either:

- there exist disjoint \( X,Y \subseteq B \) with \( |X| \geq 2^{-k}|G| \) and \( |Y| \geq 2^{1-k}|G| \), such that every vertex in \( X \) has at most \( 4\varepsilon|Y| \) neighbours in \( Y \); or

- there are at least \( (\varepsilon/8)^{k-1}|G|^{k-1} \) copies of \( P_{k-1} \) in \( G[B] \), each with first vertex in \( A' \) and no other vertex in \( A' \).

If the first holds, then the first bullet of the theorem holds; so we may assume the second holds. Since each of these copies of \( P_{k-1} \) extends to (at least one) copy of \( P_k \) in \( G \) with first vertex in \( F \) and no other vertex in \( A \), we deduce that there are at least \( (\varepsilon/8)^{k-1}|G|^{k-1} \) copies of \( P_k \) in \( G \) with first vertex in \( F \) and no other vertex in \( A \). Since there are at least \( \varepsilon|G|/8 \) choice of \( F \in F \), and they are pairwise disjoint (and therefore no copies of \( P_k \) are double-counted), there are at least \( (\varepsilon/8)^k|G|^k \) copies of \( P_k \) in \( G \) with first vertex in \( A \) and no other vertex in \( A \). This completes the induction, and so proves 7.6. \[ \square \]
We deduce

7.7 \( \{P, \overline{P}\} \) is viral for every two paths \( P, P' \).

**Proof.** By 6.3, it suffices to show that \( \{P_k, \overline{P}_k\} \) is viral for each integer \( k \); and for this, because of 7.2, it suffices to show that there exist \( a, b > 0 \) such that:

(1) For every graph \( G \) with \( |G| \geq 2 \), and every \( \varepsilon \) with \( 0 \leq \varepsilon \leq 1/2 \), either:

- there exist disjoint \( X, Y \subseteq V(G) \), both with cardinality at least \( a|G| \), that are \( \varepsilon \)-sparse or \( \varepsilon \)-dense to each other; or
- for some \( H \in \{P_k, \overline{P}_k\} \), there are at least \( \varepsilon^b|G|^k \) copies of \( H \) in \( G \).

By 7.4, there exist \( \gamma, \delta > 0 \) such that if \( G \) is a graph containing fewer than \( \gamma|G|^k \) induced labelled copies of \( P_k \), then there exists \( D \subseteq V(G) \) with \( |D| \geq \delta|G| \) such that one of \( G[D], \overline{G}[D] \) has maximum degree at most \( 2^{-k-1}|D| \). Choose \( a' \) such that if a graph \( G \) with \( |G| \geq 2 \) is both \( P_k \)-free and \( \overline{P}_k \)-free then it has a pure pair with both sets of cardinality at least \( a'|G| \). Let \( a = \min(a', 2^{-k-1}, \delta/2) \). Choose \( b \) so large that \( 2^b \geq \gamma^{-1} \), and \((1/2)^b-6k \leq \delta^k \) and \((1/2)^b-k \leq 2^{-k-6} \delta^k \). We will show that for every graph \( G \) and every \( \varepsilon \) with \( 0 \leq \varepsilon \leq 1/2 \), the two bullets above are satisfied.

Thus, let \( G \) be a graph with \( |G| \geq 2 \), and let \( 0 \leq \varepsilon \leq 1/2 \). If \( a|G| \leq 1 \) then the second bullet above holds, so we may assume that \( |G| > 1/a \). If \( \varepsilon^b|G|^k < 1 \) then (1) holds since \( a \leq a' \), so we may assume that \( \varepsilon^b|G|^k \geq 1 \). Since \( \varepsilon^b \leq \gamma \) we may assume that there exists \( D \subseteq V(G) \) with \( |D| \geq \delta|G| \) such that one of \( G[D], \overline{G}[D] \) has maximum degree at most \( 2^{-k-1}\delta|G| \), and by moving to the complement if necessary, we may assume that \( G[D] \) has maximum degree at most \( 2^{-k-1}\delta|G| \). Choose \( A \subseteq D \) with \( |D|/4 \leq |A| \leq |D|/2 \) (this is possible since \( |D| \geq 2 \)).

Since \( \varepsilon^{b/k}|G| \geq 1 \), and \( \varepsilon^{b/k-1} < 2^{-k-6} \delta^k \), it follows that

\[
1 \leq \varepsilon^{b/k-1}|G| \varepsilon \leq 2^{-k-6} \delta|G|,
\]

and so \( 4/\varepsilon \leq 2^{-k-2} \delta|G| \). Since \( 0 \leq \varepsilon/4 \leq 1/2 \) and \( 4/\varepsilon \leq 2^{-k-4} \delta|G| \leq 2^{-k-4}|D| \), we may apply 7.6 to \( G[D] \), with \( \varepsilon \) replaced by \( \varepsilon/4 \), and deduce (since \( |D| \geq \delta|G| \geq 2 \), because \( |G| \geq 1/a \geq 2/\delta \)) that either:

- there exist disjoint \( X, Y \subseteq V(G) \) with \( |X| \geq 2^{-k-1}|D| \) and \( |Y| \geq 2^{-k}|D| \), such that every vertex in \( X \) has at most \( \varepsilon|Y| \) neighbours in \( Y \); or
- there are at least \( (\varepsilon/32)^k|D|^k \) copies of \( P_k \) in \( G[D] \), each with first vertex in \( A \) and no other vertex in \( A \).

In the first case, \( X, Y \) are \( \varepsilon \)-sparse to each other, and since \( 2^{-k-1}|D| \geq a|D| \), the first bullet of (1) holds. In the second case, since \( (\varepsilon/32)^k|D|^k \geq \varepsilon^b|G|^k \), the second bullet of (1) holds. This proves 7.7.
8 Caterpillar viruses

In this section we extend 7.7 from paths to caterpillars. The proof we gave of 7.7 was a development of the proof by Bousquet, Lagoutte, and Thomassé [2] of the special case of 4.2 when the two forests are both paths; and now the proof we will give of our caterpillar result is a development of the proof by Liebenau, Pilipczuk, Seymour and Spirkl [12] of the special case of 4.2 when the two forests are both caterpillars. We have not yet been able to prove the viral extension of 4.2 in full generality.

A caterpillar is a tree \( H \), such that some path \( S \) of \( H \) contains all vertices of degree at least two.

We will show:

8.1 \( \{H, H'\} \) is viral for every two caterpillars \( H, H' \).

A rooted tree is a pair \((H, h)\) where \( H \) is a tree and \( h \in V(H) \), and a rooted forest is a set of pairwise vertex-disjoint rooted trees. If \( F = \{(H_1, h_1), (H_2, h_2), \ldots, (H_k, h_k)\} \) is a rooted forest, then \( H_1 \cup \cdots \cup H_k \) is a forest, which we denote by \( F^* \). If \( F \) is a rooted forest, we define \( V(F) = V(F^*) \) and call it the vertex set of \( F \). We say a rooted forest \( F \) captures a tree \( T \) if some member \((H, h)\) \( \in F \) has an induced subgraph isomorphic to \( T \).

Let \( F \) be a rooted forest, and let \((A, u), (B, v)\) be distinct members of \( F \). Let \( C \) be the tree obtained from \( A \cup B \) by adding the edge \( uv \). We denote by \( F(u \to v) \) the rooted forest

\[
(F \setminus \{(A, u), (B, v)\}) \cup \{(C, v)\}.
\]

Thus \( F(u \to v) \) is a rooted forest with the same vertex set, with one more edge and one fewer member.

A rooted caterpillar is a rooted tree \((H, h)\) where \( H \) is a caterpillar, and there is a path \( S \) of \( H \) with one end \( h \) that contains all vertices of degree at least two; and the minimal such path \( S \) is called the spine. We call \( h \) the head.

For \( \ell \geq 2 \), a caterpillar is \( \ell \)-uniform if every vertex has degree one or \( \ell \), and there are \( \ell \) vertices of degree \( \ell \). For each \( \ell \), such a caterpillar is unique up to isomorphism, and we denote it by \( M_\ell \). Every caterpillar is an induced subgraph of \( M_\ell \) for all sufficiently large \( \ell \), and so by 6.3 it suffices to prove 8.1 (for all \( \ell \geq 2 \)) when \( H = H' = M_\ell \). Thus, let us fix some integer \( \ell \geq 2 \).

Here is a game. A game position is a rooted forest \( F \), and the initial position \( F_1 \) contains \( N \) rooted trees, each with exactly one vertex. (We will fix \( N \) later.) There are two players, A and B, and \( N + 1 \) rounds; and for \( 1 \leq i \leq N \), the \( i \)th round proceeds as follows, where \( F_i \) is the game position at the start of the \( i \)th round:

- **Player A** selects some \((A, u)\) \( \in F_i \).
- **Player B** selects some \((B, v)\) \( \in F_i \) different from \((A, u)\).
- The new game position \( F_{i+1} \) is \( F_i(u \to v) \).

Thus for each \( i \), \( F_i^{*+1} \) is a supergraph of \( F_i^* \), with the same vertex set and with one more edge and one fewer component; so \( F_N^* \) is a tree. Player A wins if \( F_N^* \) (or equivalently, some \( F_i^* \)) has an induced subgraph isomorphic to \( M_\ell \).

We need a theorem that is implicit in [12]:

17
8.2 If $N = 2^{\ell(\ell+1)-1}$ then player A has a winning strategy.

Proof. Let us say a rooted caterpillar $(H, h)$ is a baby if

- its spine has at most $\ell$ vertices;
- every vertex of the spine has degree exactly $\ell$, except possibly for the head;
- the head has degree at most $\ell$; and strictly less if the spine has fewer than $\ell$ vertices.

The value of a rooted tree $(H, h)$ is the maximum of $2^{|J|-1}$ over all rooted subtrees $(J, h)$ of $(H, h)$ such that $(J, h)$ is a baby. Thus, the value of $(H, h)$ is at most $2^{\ell(\ell+1)-1}$, and if $(H, h)$ has value $2^{\ell(\ell+1)-1}$ then $H$ contains $M_\ell$ as an induced subgraph (indeed, with the same head).

Let us say a member of $F_i$ is expectant if its head has degree $\ell - 1$. The strategy we describe also has the property that for each $i$, either

- $F_i$ captures $M_\ell$; or
- each member of $F_i$ has maximum degree at most $\ell$, and its head has degree at most $\ell - 1$; and
- at most one member of $F_i$ is expectant.

The strategy is as follows (we call this the proper strategy). If $F_i$ captures $M_\ell$, player A choose $(A, u) \in F_i$ arbitrarily. If there is no such member, there are two cases, depending whether some member of $F_i$ is expectant, or not. If $(A, u)$ is the unique expectant member of $F_i$, player A should choose $(A, u)$. If there is no such member, player A should choose the member $(A, u)$ of $F_i$ with lowest value.

To see that this works, observe that, with $(B, v)$ as in the description of a round, and letting $(C, v)$ denote the new member of $F_{i+1}$, if $F_i$ does not capture $M_\ell$, then:

- No member of $F_{i+1}$ is expectant except possibly $(C, v)$; and $C$ has maximum degree at most $\ell$, and its root has degree at most $\ell - 1$.
- $(B, v)$ is not expectant, so the value of $(C, v)$ is at least twice that of $(B, v)$, and therefore if the value of $(A, u)$ is at most that of $(B, v)$ then the value of $F_{i+1}$ is at least that of $F_i$.
- Finally, suppose that the value of $(A, u)$ is more than that of $(B, v)$. It follows that $(A, u)$ is expectant: let $(H, u)$ be the largest baby in $(A, u)$. If the spine of $(H, u)$ has fewer than $\ell$ vertices, then $(C, v)$ has value at least twice that of $(A, u)$, and so the value of $F_{i+1}$ is at least that of $F_i$. If the spine of $(H, u)$ has exactly $\ell$ vertices, then adding the edge $uv$ to $(H, u)$ gives $M_\ell$, and so $C$ contains $M_\ell$, and therefore $F_{i+1}$ captures $M_\ell$.

This proves 8.2.
The initial position $F_1$ is prescribed, but for $i > 0$ there are many possibilities for $F_i$, depending on the choices of player B, even if player A follows the proper strategy. Let us say a rooted forest $F$ is credible if there is a game in which player A follows the proper strategy and $F$ is one of the game positions. The key properties of credibility are:

- the rooted forest with $N = 2^{\ell(\ell+1)-1}$ members, each with one vertex, is credible;
- if $F$ is credible and has at least two members, then there is a member $(A,u)$ such that for all other members $(B,v)$, the rooted forest $F(u \rightarrow v)$ is credible;
- if $F$ is credible and has only one member, then $F$ captures $M_i$.

If $F$ is a rooted forest, we define $R(F)$ to be the set of roots of the members of $F$. If $u,v \in V(F)$, we say $v$ is the parent of $u$ if $u,v$ are adjacent in $F^*$, and therefore belong to the same member $(H,h)$ of $F$, and the path of $H$ between $u,h$ contains $v$.

Now let $G$ be a graph, and let $B$ be a set of pairwise disjoint, nonempty subsets of $V(G)$; we call $B$ a blockade. (In previous papers the same word was used for a sequence of subsets rather than a set, but here the order will not matter.) The members of $B$ are called blocks. We are concerned with rooted forests with vertex set $B$; thus, each block of $B$ is a vertex of the forest.

Let $F$ be a rooted forest, with vertex set $B$. A map $\phi$ with domain $B$, that maps each $B \in B$ to a nonempty subset $\phi(B) \subseteq B$, is called an appearance of $F$ if it satisfies the following:

- for all distinct $B,B' \in B$, if $B,B'$ are nonadjacent in $F^*$ and not both in $R(F)$, then there are no edges of $G$ between $\phi(B)$ and $\phi(B')$;
- for all distinct $B,B' \in B$, if $B$ is the parent of $B'$ in $F$, then every vertex of $\phi(B)$ has a neighbour in $\phi(B')$.

Two appearances $\phi, \phi'$ of $F$ are diverse if there exists $v \in V(F) \setminus R(F)$ such that $\phi(v) \cap \phi'(v) = \emptyset$.

We would like to show that there are many copies of $M_i$ in $G$. Our method to do so is to start with a blockade with $N = 2^{\ell(\ell+1)-1}$ blocks (all large), and show that there is a rooted tree $(H,h)$ such that $H$ contains $M_i$, and there are many appearances of $\{(H,h)\}$ that are pairwise diverse. To prove this, we will show by induction that for $i = N,N-1,\ldots,1$, there is a credible rooted forest with $i$ members and with many appearances that are pairwise diverse.

Let $\phi$ be an appearance of a rooted forest $F$. An extrusion of $\phi$ is a map $\phi'$, also with domain $B$, such that $\phi'(B) = \phi(B)$ for each $B \in V(F) \setminus R(F)$, and $\phi'(B) \subseteq \phi(B)$ for each $B \in R(F)$. We need the sets $\phi(B)$ to be large when $B \in R(F)$. Let us say an appearance $\phi$ of $F$ is $\tau$-large, where $\tau \geq 0$, if $|\phi(B)| \geq \tau$ for every $B \in R(F)$.

We begin with:

**8.3** Let $B$ be a blockade in a graph $G$, with $|B| = N = 2^{\ell(\ell+1)-1}$. Suppose that every vertex of $G$ has degree at most $\theta|G|$. Let $F$ be a rooted forest with vertex set $B$, with at least two members, and let $\phi$ be a $\tau$-large appearance of $F$, where $\tau \geq 3\theta|G|$ is an integer. Let $\varepsilon > 0$ with $\varepsilon \leq 1$, such that $\varepsilon \tau \geq 3/2$. Suppose that $B_1 \in R(F)$ is such that for every $B_2 \in R(F)$ different from $B_1$, the rooted forest $F(B_1 \rightarrow B_2)$ is credible. Then there exists $B_2 \in R(F)$ different from $B_1$ such that either:

- there exists $P \subseteq B_1$ and $Q \subseteq B_2$ with $|P| \geq \tau/3$ and $|Q| \geq 2\tau/3$ such that $P,Q$ are $\varepsilon$-sparse to each other;
there are at least $2\varepsilon\tau/(9N)$ appearances of $F(B_1 \rightarrow B_2)$, all extrusions of $\phi$, and all $(\tau/3)$-large and pairwise diverse.

**Proof.** Choose $k$ maximum such that there are $k$ subsets $A_1, \ldots, A_k$ of $\phi(B_1)$ with the following properties:

- $A_1, \ldots, A_k$ are pairwise disjoint, and each has cardinality at most $3/(2\varepsilon)$;
- for $1 \leq i \leq k$, there exists $B_i \in R(F)$ with $B_i \neq B_1$ such that at least $\tau/3$ vertices of $\phi(B_i)$ have a neighbour in $A_i$, and for each $B \in R(F)$ with $B \neq B_1, B_2$, at least $\tau/3$ vertices of $\phi(B)$ have no neighbour in $A_i$.

Suppose first that $k \geq 2\varepsilon\tau/9$. By hypothesis, for $1 \leq i \leq k$ there exists $B_i$ as in the second bullet above, and since there are at most $N$ choices for $B_2$, there are at least $k/N$ values of $i$ such that the corresponding $B_i$'s are all equal. Thus we may assume that there exists $B_2 \in R(F)$ with $B_2 \neq B_1$ such that for each integer $i$ with $1 \leq i \leq k/N$, at least $\tau/3$ vertices of $\phi(B_i)$ have a neighbour in $A_i$, and for each $B \in R(F)$ with $B \neq B_1, B_2$, at least $\tau/3$ vertices of $\phi(B)$ have no neighbour in $A_i$. Let $1 \leq i \leq k/N$ be an integer. For each $B \in B$, define $\phi_i(B)$ as follows:

- if $B \notin R(F)$ then $\phi_i(B) = \phi(B)$;
- $\phi_i(B_1) = A_i$;
- $\phi_i(B_2)$ is the set of vertices in $\phi(B_2)$ with a neighbour in $A_i$;
- for each $B \in R(F)$ with $B \neq B_1, B_2$, $\phi_i(B)$ is the set of vertices in $\phi(B)$ with no neighbour in $A_i$.

Then each $\phi_i$ is an extrusion of $\phi$, and an $(\tau/3)$-large appearance of the rooted forest $F(B_1 \rightarrow B_2)$, and they are pairwise diverse (since the sets $A_1, A_2, \ldots$ are pairwise disjoint). Moreover, $F(B_1 \rightarrow B_2)$ is credible, from the choice of $B_1$. Since $k/N \geq 2\varepsilon\tau/(9N)$, the second bullet of the theorem holds.

Thus we may assume that $k \leq 2\varepsilon\tau/9$. Let $A = A_1 \cup \cdots \cup A_k$. It follows that $|A| \leq \tau/3$. Choose $B_2 \in F$ different from $B_1$, choose $Y \subseteq B_2$ with $|Y| = \tau$, and choose $X \subseteq \phi(B_1) \setminus A$ maximal such that $|N_Y(X)| \geq 2\varepsilon\tau|X|/3$, where $N_Y(X)$ denotes the set of vertices in $Y$ with a neighbour in $X$. It follows that $|X| \leq 3/(2\varepsilon)$.

Suppose that $|N_Y(X)| \geq \tau/3$. Then we may choose $A_{k+1} \subseteq X$ minimal such that for some $B \in R(F) \setminus \{B_1\}$, at least $\tau/3$ vertices of $\phi(B)$ have a neighbour in $A_{k+1}$. It follows from the minimality of $A_{k+1}$ that for every $B' \in R(F) \setminus \{B_1\}$, at most $\tau/3 + \theta|G|$ vertices of $\phi(B')$ have a neighbour in $A_{k+1}$, and since $\tau/3 + \theta|G| \leq 2\tau/3$ and $|\phi(B')| \geq \tau$, at least $\tau/3$ vertices in $\phi(B')$ have no neighbour in $A_{k+1}$. This contradicts the maximality of $k$, and so proves that $|N_Y(X)| < \tau/3$.

Since $|N_Y(X)| \geq 2\varepsilon\tau|X|/3$, it follows that $|X| < 1/(2\varepsilon)$. Let $P = \phi(B_1) \setminus (A \cup X)$ and $Q = \phi(Y) \setminus N_Y(X)$. Thus $|P| \geq \tau - (\tau/3) - 1/(2\varepsilon) \geq \tau/3$, and $|Q| \geq 2\tau/3$. Moreover, every vertex in $P$ has fewer than $2\varepsilon\tau/3 \leq \varepsilon|Q|$ neighbours in $Q$, from the maximality of $X$, and so $P, Q$ are $\varepsilon$-sparse to each other. Hence the first bullet of the theorem holds. This proves 8.3.  

We deduce:
8.4 Let $\tau > 0$ be an integer, and let $B$ be a blockade in a graph $G$, with $|B| = N = 2^{(t+1)-1}$, such that every block has cardinality $\tau 3^{N-1}$. Suppose that every vertex of $G$ has degree at most $\theta |G|$, where $\tau \geq \theta |G|$. Let $\varepsilon > 0$ with $\varepsilon \leq 1$, with $\varepsilon \tau \geq 1/2$. Then either

- there exist distinct $B_1, B_2 \in B$ and $P \subseteq B_1$ and $Q \subseteq B_2$ with $|P| \geq \tau/3$ and $|Q| \geq 2\tau/3$ such that $P, Q$ are $\varepsilon$-sparse to each other; or
- there is a rooted tree $(H, h)$ such that $\{(H, h)\}$ is credible, and there are at least $(2\varepsilon \tau/(9N^2))^{|H|}$ appearances of $\{(H, h)\}$, pairwise diverse. Consequently there are at least

$$
\frac{2}{(9N^2)^N} 3^{-N^2} \varepsilon^N N \ell |M_\ell|
$$

copies of $M_\ell$ in $G$.

Proof. We assume the first bullet is false, and will prove the following statement by induction on $t$:

(1) For $0 \leq t \leq N - 1$, there is a credible rooted forest $F$ with $N - t$ members, such that there is a set of at least $(2\varepsilon \tau/(9N^2))^t$ appearances of $F$, each $\tau 3^{N-t-1}$-large and pairwise diverse.

The claim is true for $t = 0$, since the rooted forest with $N$ members, each with only one vertex, is credible, and admits a $\tau 3^N$-large appearance $\phi$ where $\phi(B) = B$ for each $B \in B$. Suppose that $1 \leq t \leq N - 1$, and the claim holds for $t - 1$. Let $F$ be a credible rooted forest with $N - t + 1$ members, such that there is a set $\Phi$ of at least $(2\varepsilon \tau/(9N^2))^{t-1}$ appearances of $F$, each $\tau 3^{N-t}$-large and pairwise diverse. Since $F$ is credible, there exists $B_1 \in R(F)$ such that for all $B_2 \in R(F)$ different from $B_1$, the rooted forest $F(B_1 \rightarrow B_2)$ is credible.

Since the first bullet of the theorem is false, and $\tau 3^{N-t} \geq 3\theta |G|$, and $\varepsilon \tau 3^{N-t} \geq 3/2$, it follows from 8.3 (with $\tau$ replaced by $\tau 3^{N-t}$) that for each $\phi \in \Phi$, there exists $B_2 \in R(F)$ different from $B_1$ such that there are at least $2\varepsilon \tau/(9N)$ appearances of $F(B_1 \rightarrow B_2)$, all extrusions of $\phi$, and all $(\tau/3)$-large and pairwise diverse. Since there are at most $N$ choices of $B_2$, there exists $\Phi' \subseteq \Phi$ with $|\Phi'| \geq \Phi/N$ such that the $B_2$'s are the same for all $\phi \in \Phi'$. Consequently there exists $B_2 \in R(F)$ different from $B_1$ such that for all $\phi \in \Phi'$, there are at least $2\varepsilon \tau/(9N)$ appearances of $F(B_1 \rightarrow B_2)$, all extrusions of $\phi$, and all $3^{N-t-1}$-large and pairwise diverse. Since $|\Phi'| \geq (2\varepsilon \tau/(N(9N^2)))^{t-1}$, we have altogether at least

$$
(2\varepsilon \tau/(N(9N^2)))^{t-1}(2\varepsilon \tau/(9N)) = (2\varepsilon \tau/(9N^2))^t
$$
appearances of $F(B_1 \rightarrow B_2)$, all $3^{N-t-1}$-large and pairwise diverse. (To see the last, any two extrusions of the same member of $\Phi'$ are diverse from the construction, and extrusions of different members of $\Phi'$ are diverse since the members of $\Phi'$ are themselves diverse.) This proves (1).

From (1) with $t = N - 1$, there is a credible rooted forest $F = \{(H, h)\}$, such that there is a set $\Phi$ of at least $(2\varepsilon \tau/(9N^2))^{N-1}$ appearances of $F$, each $\tau$-large and pairwise diverse. Since $V(H) = B$, we may assume that $h = B_1$ say. Let $\phi \in \Phi$, and let $v \in \phi(B_1)$. We claim that:

(2) There is a copy $\psi$ of $H$ in $G$ such that $\psi(B) \in \phi(B)$ for each $B \in B$, and $\psi(B_1) = v$.

Let $V(H) = B = \{B_1, \ldots, B_N\}$, numbered such that for $2 \leq j \leq N$, the parent (in $(H,h)$) of
$B_j$ is some $B_i$ where $i < j$. Define $\psi(B_1) = v$, and for $2 \leq j \leq N$ define $\psi(B_j)$ inductively as follows: let $B_i$ be the parent of $B_j$ in $(H, h)$, and choose $\psi(B_j) \in \phi(B_j)$ adjacent to $\psi(B_i)$ (this is possible from the definition of an appearance). It also follows from the definition of an appearance that $\psi$ is a copy of $H$. This proves (2).

For each $\phi \in \Phi$ and each $v \in \phi(B_1)$, we define $\psi(\phi, v)$ to be some $\psi$ as in (1). We claim that all these copies of $(H, h)$ are distinct: because suppose that $\psi(\phi, v) \neq \psi(\phi', v')$ for some $(\phi, v) \neq (\phi', v')$. It follows that $v = v'$, and so $\phi, \phi'$ are diverse, a contradiction. Thus there are at least

$$(2\varepsilon/9N^2)^{N-1} = (2\varepsilon/9N^2)^{N-1}N$$

distinct copies of $H$ in $G$, each mapping $B$ to a vertex of $B$ for each $B \in \mathcal{B}$. Since each contains a copy of $M_\ell$, and each copy of $M_\ell$ can be extended to only $(\tau 3^{N-1})^{N-|M_\ell|}$ copies of $H$ that map each $B \in \mathcal{B}$ to a vertex of $B$ (since each block has size $\tau 3^{N-1}$), we deduce that there are at least

$$(2\varepsilon/9N^2)^{N-1} \tau 3^{N-1}|M_\ell-N| \geq (2/9N^2)3^{N-2}N^2\varepsilon^{N}|M_\ell|$$

such copies of $M_\ell$. This proves 8.4.

Now we deduce the main result of this section, which we restate:

8.5 \{H, H'\} is viral for every two caterpillars $H, H'$.

**Proof.** As we discussed earlier, it suffices to prove 8.1 when $H = H' = M_\ell$. Let $m = |M_\ell| = \ell(\ell+1)$. By 7.2, it suffices to show that there exist $a, b > 0$ such that:

1. For every graph $G$ with $|G| \geq 2$ and every $\varepsilon$ with $0 \leq \varepsilon \leq 1/2$, either:
   - there exist disjoint $A, B \subseteq V(G)$, both with cardinality at least $a|G|$, that are $\varepsilon$-sparse or $\varepsilon$-dense to each other; or
   - for some $H \in \{M_\ell, \overline{M}_\ell\}$, there are at least $\varepsilon^b|G|^m$ copies of $H$ in $G$.

Let $\theta = 1/(2N^3N^{-1})$. By 7.4, there exist $\gamma, \delta > 0$ such that if $G$ is a graph containing fewer than $\gamma|G|^m$ copies of $M_\ell$, then there exists $D \subseteq V(G)$ with $|D| \geq \delta|G|$ such that one of $G[D], \overline{G}[D]$ has maximum degree at most $\theta|D|$. By 4.2, we may choose $a'$ such that if a graph $G$ with $|G| \geq 2$ is both $M_\ell$-free and $\overline{M}_\ell$-free then it has a pure pair with both sets of cardinality at least $a'|G|$. Let $a = \min(a', \theta\delta/3)$, and choose $b$ so large that

$$2^{-b} \leq \gamma$$

$$(1/2)^{b/m} \leq \theta\delta$$

$$b \geq m,$$ and

$$(1/2)^{b-N} \leq (2/9N^2)^N3^{-N}(\theta\delta)^m.$$ 

We will show that (1) holds. Thus, let $G$ be a graph with $|G| \geq 2$, and let $0 \leq \varepsilon \leq 1/2$. We may assume that $a|G| > 1$, for otherwise the first bullet of (1) holds; and so $|G| > 1/a \geq 3/(\theta\delta)$. Also we may assume that $\varepsilon^b|G|^m > 1$, since otherwise either the second bullet of (1) holds, or $G$ admits a
pure pair with both sets of cardinality at least \(a'|G| \geq a|G|\), and the first bullet of (1) holds. We may assume that there are fewer than \(\varepsilon b|G|^m\) copies of \(M_\ell\) in \(G\), and hence fewer than \(\gamma|G|^m\) copies of \(M_\ell\), since \(\varepsilon \leq 2^{-b} \leq \gamma\). Hence there exists \(D \subseteq V(G)\) with \(|D| \geq \delta|G|\) such that one of \(G[D], \overline{G[D]}\) has maximum degree at most \(\theta|D|\), and by taking complements if necessary, we may assume the first.

Let \(\tau = [2\theta|D|]\). Since \(|D| \geq N(3^{N-1}\tau)\), there is a blockade \(B\) in \(G\) with \(N\) blocks, each a subset of \(D\) with cardinality \(3^{N-1}\tau\). Note that \(\tau \geq \theta|D|\), since \(\varepsilon \geq \theta\delta|G| \geq 1\). Since \(\varepsilon b|G|^m > 1\), it follows that \(\varepsilon b/m|G| \geq 1\), and consequently

\[
\varepsilon \varepsilon \geq \varepsilon \theta \delta |G| \geq \varepsilon (1/2)^{b/m}|G| \geq \varepsilon b/m|G|/2 \geq 1/2.
\]

Hence we may apply 8.4 to \(G[D]\) and \(B\). We deduce that either

\[\begin{itemize}
\item there exist distinct \(B_1, B_2 \in B\) and \(P \subseteq B_1\) and \(Q \subseteq B_2\) with \(|P| \geq \tau/3\) and \(|Q| \geq 2\tau/3\) such that \(P, Q\) are \(\varepsilon\)-sparse to each other; or
\item there are at least \((2/(9N^2))^N 3^{-N} \varepsilon^N \tau^m\) copies of \(M_\ell\) in \(G\).
\end{itemize}\]

In the first case, since \(\tau/3 \geq \theta|G|/3 \geq a|G|\), the first bullet of (1) holds; and in the second case, since

\[
(2/(9N^2))^N 3^{-N} \varepsilon^N \tau^m \geq (2/(9N^2))^N 3^{-N} \theta^N |G|^m \geq \varepsilon b |G|^m,
\]

(from the choice of \(b\) and since \(\varepsilon \leq 1/2\)), the second bullet of (1) holds. This proves (1) and hence proves 8.1. \(\square\)

Acknowledgement

Thanks to Maria Chudnovsky, Tung Nguyen and Sophie Spirkl for stimulating discussions on the topic of this paper.

References


