Minimizing the Number of Unions

Žarko Ranđelović[a]

Submitted: Dec 25, 2024; Accepted: Jan 29, 2026; Published: Mar 13, 2026

Abstract

For a given number of k-sets, how should we choose them so as to minimize the union-closed family that they generate? Our main aim in this paper is to show that, if $\mathcal{A}$ is a family of k-sets of size $\binom{t}{k}$, and t is sufficiently large, then the union-closed family generated by $\mathcal{A}$ has size at least that generated by the family of all k-sets from a t-set. This proves (for this size of family) a conjecture of Roberts. We also make some related conjectures, and give some other results, including a new proof of the result of Leck, Roberts and Simpson that exactly determines this minimum (for all sizes of the family) when $k = 2$, as well as resolving the conjecture of Roberts when the size of the family is very close to $\binom{t}{k}$.

Mathematics Subject Classifications: 05D05

1 Introduction

Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a family of k-sets where the A_i are distinct. We are interested in minimizing the size of the union-closed family generated by $\mathcal{A}$, namely

$$\langle \mathcal{A} \rangle = \{A_{i_1} \cup A_{i_2} \ldots \cup A_{i_s} | 1 \leqslant s \leqslant n, 1 \leqslant i_1, i_2, \ldots, i_s \leqslant n\}.$$

It is natural to imagine that it is best to take the sets to be 'as close together as possible', so for example all k-sets from a t-set if the size is $\binom{t}{k}$. Note that the union-closed family generated by this family has size $|[t]^{(\geqslant k)}|$, where as usual $[t]$ denotes the set $\{1, 2, \ldots, t\}$ and for any set X we write $X^{(\geqslant k)}$ for the family of all subsets of X of size at least k. Our main aim is to prove this when t is large.

Theorem 1. *Let $\mathcal{A}$ be a family of k-sets of size $\binom{t}{k}$. Then for t sufficiently large we have $|\langle \mathcal{A} \rangle| \geqslant |[t]^{(\geqslant k)}|$.*

This was conjectured by Roberts [1]. In fact, he conjectured a result that should hold for all sizes of $\mathcal{A}$: we discuss this in the final section of the article. In principle the answer

[a]Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia (zarko.randjelovic@turing.mi.sanu.ac.rs).

 https://doi.org/10.37236/13702

could have depended on the size of the ground set, say N, but in fact it does not. The first possible case would be $k = 2$, and for this Leck, Roberts and Simpson showed an exact result. Initial segments of colex are best (colex on $\mathbb{N}^{(k)}$ is the order in which $A < B$ if $\max(A \backslash B) < \max(B \backslash A)$ where $\mathbb{N}^{(k)}$ is the collection of all subsets of $\mathbb{N}$ of size k).

Theorem 2. *(Leck, Roberts and Simpson [2]) Let $\mathcal{A}$ be a family of 2-sets, and let $\mathcal{B}$ be the first $|\mathcal{A}|$ 2-sets in the colex order on $\mathbb{N}^{(2)}$. Then $|\langle \mathcal{A} \rangle| \geqslant |\langle \mathcal{B} \rangle|$.*

We give a new proof of this result showing that the $k = 2$ case can be completely solved using the same main idea from Theorem 1. This approach also identifies all extremal families (see Corollary 8). When $k \geqslant 3$ colex is not best. We discuss this at the end of the article. We will use standard notation for set systems and graphs and their parameters. See Bollobás [3] for general background.

2 Proofs of Results

We will start by giving an overview of the proof of Theorem 1. Let N be the size of the ground set. The idea of the proof is that we will first reduce the ground set to size close to t. To reduce the size of the ground set we first show that it cannot be "too big" and then we remove points belonging to few sets. If N is not close to t then after removing points belonging to few sets we get a set X of size N' none of whose elements belong to too few of the remaining sets. Then we show that most subsets of X are indeed unions. We can show that N' is close to t. Then unless N is close to t starting with X we can add elements from the ground set one by one making many more unions. This will give more unions than the lower bound we are trying to prove. For the close to t case, similar to above, removing elements belonging to few sets we get a set Y such that nearly all subsets of Y are unions, but this time we will have $|Y| \geqslant t$. We will get too many unions unless we have both $|Y| = t$ and Y being the whole ground set which is then the $[t]^{(k)}$ case.

We will first prove a few lemmas. For the following lemmas and the proof of Theorem 1 we will always use the notation that $n = \binom{t}{k}$ where n, t, k are integers such that $t \geqslant k > 1$ and that $\mathcal{A} = \{A_1, \ldots, A_n\}$ is our family of k-sets. We will also define $G = \cup_{i=1}^{n} A_i$. We may view G as the ground set.

The following lemma will be used to show that the ground set size cannot be too big.

Lemma 3. *If $s \in \mathbb{N}$ and $|G| \geqslant sk$ then $|\langle \mathcal{A} \rangle| \geqslant 2^s - 1$.*

Proof. We will prove a more general statement. Suppose that $l \in \mathbb{N}$ and $B_1, \ldots, B_l$ are subsets of $\mathbb{N}$ such that $|B_i| \leqslant k$ for all i. Let $T = \cup_{i=1}^{l} B_i$ and suppose $|T| \geqslant sk$. Finally let

$$\mathcal{F} = \{B_{i_1} \cup B_{i_2} \ldots \cup B_{i_r} | 1 \leqslant r \leqslant l, 1 \leqslant i_1, i_2, \ldots, i_r \leqslant l\}.$$

We will show that $|\mathcal{F}| \geqslant 2^s - 1$. We will prove that there is some $S \subset T$ such that $|S| = s$ and for any non-empty $X \subset S$ there is some $B \in \mathcal{F}$ such that $B \cap S = X$. This will prove the above statement and hence the lemma. To show this we will induct on s.

For $s = 1$ it is trivial as we may choose any $x \in T$ and set $S = \{x\}$.

Suppose it is true for some $s - 1 \geqslant 1$. Now consider the sets $B_1, \ldots, B_l$ and T. For every $x \in T$ we define $T_x = \{i | x \in B_i\}$. Pick an $x \in T$ with the smallest $|T_x|$. Now without loss of generality $x \in B_1$. Let $T' = T \backslash B_1$. We have $|T'| \geqslant (s-1)k$. Now consider the collection of sets

$$\mathcal{H} = \{B_i \cap T' | i \notin T_x\}.$$

Suppose that $y \in T'$. We cannot have $T_x = T_y$ since $y \notin B_1$ but $|T_x| \leqslant |T_y|$ so there must be some $i \in T_y \backslash T_x$. But then $B_i \cap T' \in \mathcal{H}$ so $y \in \cup_{B \in \mathcal{H}} B$. Therefore $\cup_{B \in \mathcal{H}} B = T'$. Now by induction there is some $S' \subset T'$ of size $s - 1$ such that for every non-empty $X \subset S'$ there is some B which is a union of some sets in $\mathcal{H}$ such that $B \cap S' = X$. In other words the projection of a union-closed family containing $\mathcal{H}$ onto S' contains all non-empty subsets of S'. From the definition of $\mathcal{H}$ this means that there is some $B \in \mathcal{F}$ such that $B \cap S' = X$ and $x \notin B$. Now consider all possible unions of sets in the family $\{B_i | i \notin T_x\} \cup \{B_1\}$ and let $S = S' \cup \{x\}$. We get that the projection of $\mathcal{F}$ onto S contains all non-empty subsets of S and $|S| = s$. This completes the induction step and the proof. □

We can see that $|[t]^{(\geqslant k)}| < 2^t - 1$ and so if $|\langle \mathcal{A} \rangle| < |[t]^{(\geqslant k)}|$ then by Lemma 3 we must have $|G| < tk$. This is good as we have restricted the ground set which will make it easier to work with. The next step is showing that we can essentially remove all elements in the ground set that are in too few sets. For any $x \in G$ define $d_x = |\{i | x \in A_i\}|$. Also let

$$\mathcal{A}_x = \{A \subset G | x \in A, |A| \geqslant k, \forall i \ (x \in A_i \Rightarrow A_i \not\subset A)\}.$$

In other words $\mathcal{A}_x$ is the set of all subsets of G of size at least k for whom x is one of the reasons that they are not in $\langle \mathcal{A} \rangle$.

The following lemma will be used to show that after removing elements in too few sets, most subsets of the remaining elements are in fact in $\langle \mathcal{A} \rangle$.

Lemma 4. *Suppose that $s \in \mathbb{N}$ and $x \in G$ such that $d_x \geqslant s\binom{|G|-2}{k-2}$. Then we must have $|\mathcal{A}_x| \leqslant 2^{|G|-s}$.*

Proof. Consider the family $\mathcal{T} = \{B_i | x \in A_i, B_i = A_i \backslash \{x\}\}$. We have that $|\mathcal{T}| = d_x$. We may assume that $G \backslash \{x\} = [p]$ for some p. Now we have $\mathcal{T} \subset [p]^{(k-1)}$. Notice that $\mathcal{A}_x$ consists of all subsets of G containing x that do not have a subset in $\mathcal{T}$. For $0 \leqslant r \leqslant p-1$ we define the *upper shadow* of a family $\mathcal{F} \subset [p]^{(r)}$ to be

$$\partial_+ \mathcal{F} = \{A \cup \{i\} | A \in \mathcal{F}, i \in [p] \backslash A\}.$$

Also define $\partial_+^k \mathcal{F} = \partial_+ \partial_+ \ldots \partial_+ \mathcal{F}$ where ∂_+ is applied k times ($\partial_+^0 \mathcal{F} = \mathcal{F}$). For $N, r \in \mathbb{N}$ define the *lex order* on $[N]^{(r)}$ to be the order in which $A < B$ if $\min(A \backslash B) < \min(B \backslash A)$. Now let $\mathcal{I}$ be the initial segment of lex on $[p]^{(k-1)}$ of size $|\mathcal{T}|$. For $1 \leqslant r \leqslant p$ we define the *lower shadow* of a family $\mathcal{F} \subset [p]^{(r)}$ to be

$$\partial \mathcal{F} = \{A \backslash \{i\} | A \in \mathcal{F}, i \in A\}.$$

Consider the map $g : [p] \to [p]$ given by $g(i) = p + 1 - i$. If $F : \mathcal{P}([p]) \to \mathcal{P}([p])$ is given by $F(A) = g(A^c)$ then F is a bijection that takes any initial segment of lex into a corresponding initial segment of colex. For any $\mathcal{B} \subset [p]^{(r)}$ we have $F(\partial_+(\mathcal{B})) = \partial F(\mathcal{B})$. Applying the Kruskal-Katona theorem (Kruskal [4], Katona [5] and see Bollobás [3]) we see that $|\partial F(\mathcal{T})| \geqslant |\partial F(\mathcal{I})|$ and hence $|\partial_+\mathcal{T}| \geqslant |\partial_+\mathcal{I}|$.

Note that the upper shadow of an initial segment of lex is an initial segment of lex. We can now show by induction that for any $1 \leqslant r \leqslant p-k+1$ we have that $|\partial_+^r\mathcal{T}| \geqslant |\partial_+^r\mathcal{I}|$. We have by above that this is true for $r = 1$. If it is true for $r < p-k+1$ then if $\mathcal{I}'$ is the initial segment of lex of length $|\partial_+^r\mathcal{T}|$ on $[p]^{(k-1+r)}$ then we know that $|\mathcal{I}'| \geqslant |\partial_+^r\mathcal{I}|$. Also since the upper shadow of an initial segment of lex we have that $\partial_+^r\mathcal{I}$ is an initial segment of lex on $[p]^{(k-1+r)}$ and hence $\partial_+^r\mathcal{I} \subset \mathcal{I}'$. Now applying the Kruskal-Katona theorem we have that $|\partial F(\partial_+^r\mathcal{T})| \geqslant |\partial F(\mathcal{I}')|$ and hence

$$|\partial_+^{(r+1)}\mathcal{T}| \geqslant |\partial_+\mathcal{I}'| \geqslant |\partial_+^{(r+1)}\mathcal{I}|.$$

So by induction $|\partial_+^r\mathcal{T}| \geqslant |\partial_+^r\mathcal{I}|$ for all $1 \leqslant r \leqslant p-k+1$. Now if we define the *upset* generated by a family $\mathcal{F} \subset [p]^{(r)}$ to be

$$U_{\mathcal{F}} = \{A \subset [p] | X \subset A \text{ for some } X \in \mathcal{F}\},$$

then by above $|U_{\mathcal{T}}| \geqslant |U_{\mathcal{I}}|$. Notice also that $|\mathcal{A}_x| = |[p]^{(\geqslant k-1)}| - |U_{\mathcal{T}}|$. Since we want an upper bound on $|\mathcal{A}_x|$ we just need to show that $U_{\mathcal{I}}$ contains almost all sets in $[p]^{(\geqslant k-1)}$.

We know that in $[p]$ the number of $(k-1)$-sets containing a fixed element is $\binom{p-1}{k-2}$. We also know that $p = |G| - 1$ and so the number of $(k-1)$-sets containing at least one element from $[s]$ is at most $s\binom{|G|-2}{k-2} \leqslant d_x$ which also means that $s \leqslant p$. Note that in lex these sets are all before sets not containing any elements in $[s]$. so we must have that $\mathcal{I}$ contains all sets with at least one element from $[s]$ which means that the complement of $U_{\mathcal{I}}$ in $[p]^{(\geqslant k-1)}$ is a subset of $\mathcal{P}([p]\backslash[s])$. So we have

$$|\mathcal{A}_x| \leqslant |[p]^{(\geqslant k-1)}| - |U_{\mathcal{I}}| \leqslant |\mathcal{P}([p]\backslash[s])| \leqslant 2^{|G|-s}.$$

□

The following will be a useful fact about getting many new unions.

Lemma 5. *Suppose that $\mathcal{H}$ is a family of sets and A is a set such that $A \not\subset \cup_{S\in\mathcal{H}} S$. Then $|\mathcal{H} \cup \{A \cup S | S \in \mathcal{H}\}| \geqslant (1 + 1/2^{|A|-1})|\mathcal{H}|$.*

Proof. Suppose that $x \in A\backslash \cup_{S\in\mathcal{H}} S$. If two sets S_1, S_2 in $\mathcal{H}$ have $A \cup S_1 = A \cup S_2$ then S_1 and S_2 can only differ on $A \cap (\cup_{S\in\mathcal{H}} S)$ which is a fixed set of size at most $|A| - 1$. So at most $2^{|A|-1}$ different sets in $\mathcal{H}$ can give the same union with A. This means that

$$|\{A \cup S | S \in \mathcal{H}\}| \geqslant \frac{1}{2^{|A|-1}}|\mathcal{H}|.$$

However, no set containing A is in $\mathcal{H}$ since it contains x so we get the desired result. □

We now come to the heart of the proof.

Lemma 6. *For any real $D_k \geqslant 1$ there is a large enough constant B, depending only on k and D_k, such that if $t > B$ and $|G| < t + D_k \log t$ then $|\langle \mathcal{A} \rangle| \geqslant 2^t - \sum_{i=0}^{k-1} \binom{t}{i}$ and equality holds if and only if $\mathcal{A} = X^{(k)}$ where X is a set with $|X| = t$.*

Proof. Given D_k suppose that $t > B$ and $|G| < t + D_k \log t$ where B will be chosen later. Starting with G and all of the A_i we remove elements from G one by one (and hence removing any of the A_i containing those elements) if they are in less than $3\log(2tk)\binom{2t}{k-2}$ of the remaining sets A_i until we get a G' which satisfies that no $x \in G'$ is in less than $3\log(2tk)\binom{2t}{k-2}$ of the remaining sets. Notice that if $|G'| \leqslant t-1$ then at some point in our process we were left with $G'' \subset G$ such that $|G''| = t-1$. We can see that at that point we have removed at most $D_K \log t + 1$ elements. This means that we have not removed that many sets either. In fact we have removed at most $3(D_k \log t + 1)\log(2tk)\binom{2t}{k-2}$ sets. We will make sure that $B > 3k$ to have that $\log(2tk) = \log t + \log(2k) \leqslant 2\log t$, $D_k \log t + 1 \leqslant 2D_k \log t$ and $2t \leqslant 3(t-k)$. Now we just need to make sure that

$$\frac{t-1}{\log^2 t} > 12t^{1/2} \cdot 3^{(k-2)}(k-1)D_k$$

which is possible as $(t-1)/t^{1/2} \geqslant \sqrt{t} - 1 = \exp(\frac{1}{2}\log t) - 1 > \log^3 t/48$ so we ensure that $B > \exp(600 \cdot 3^{(k-2)}(k-1)D_k)$. Now the number of sets removed when we are left with G'' is at most

$$3(D_k \log t + 1)\log(2tk)\binom{2t}{k-2} \leqslant$$
$$\leqslant \frac{12 \cdot 3^{(k-2)} D_k (t-k)^{k-2} \log^2 t}{(k-2)!} < t^{-1/2}\binom{t-1}{k-1}.$$

This means that for large enough B we could have only removed less than $\binom{t-1}{k-1}$ sets and hence when we get G'' we still have more than $\binom{t}{k} - \binom{t-1}{k-1} = \binom{t-1}{k}$ sets remaining. This is a contradiction as $|G''| = t-1$ and hence we have that $|G'| \geqslant t$. Now we will show that if B is large enough we can make sure that the vast majority of subsets of G' are indeed in $\langle \mathcal{A} \rangle$. For any $x \in G'$ define $\mathcal{A}'_x = \mathcal{A}_x \cap \mathcal{P}(G')$. From the proof of Lemma 4, we can see that the lemma is true for any n, not just $n = \binom{t}{k}$. We will apply Lemma 4 on G' and the remaining sets when we are left with G'. For large enough B we have $|G'| \leqslant |G| < 2t$ so if we define $d_{G'}(x) = |\{A_i | x \in A_i, A_i \subset G'\}|$ we have $d_{G'}(x) \geqslant 3\log(2tk)\binom{|G'|-2}{k-2}$ for all $x \in G'$. Now by Lemma 4

$$|\mathcal{A}'_x| \leqslant 2^{|G'| - [3\log(2tk)]} \leqslant 2^{|G'| - 2\log(2tk)} \leqslant \frac{2^{|G'|}}{2tk}.$$

Now since every set in $G'^{(\geqslant k)}$ that is not in $\langle \mathcal{A} \rangle$ must be in some $\mathcal{A}'_x$ we have by the union bound

$$|\langle \mathcal{A} \rangle| \geqslant |G'^{(\geqslant k)}| - |\cup_{x \in G'} \mathcal{A}'_x| \geqslant 2^{|G'|} - \sum_{i=0}^{k-1} \binom{|G'|}{i} - |G'|\frac{2^{|G'|}}{2tk}. \tag{1}$$

Since $|G'| \geqslant t > 3k$ we have $\binom{|G'|}{i} < \binom{|G'|}{i+1}/2$ for $0 \leqslant i \leqslant k-1$ because $|G'| - i > 2(i+1)$. By induction we get that $\binom{|G'|}{i} < \binom{|G'|}{k}/2^{k-i}$. Since exponential beats polynomial, for large enough t we have $|G'|^k < 2^{|G'|}/2^{k+1}$. We will take a large enough B to have that $\sum_{i=0}^{k-1}\binom{|G'|}{i} \leqslant \binom{|G'|}{k} < |G'|^k < 2^{|G'|}/2^{k+1}$. Now if $|G'| > t+1$ by (1) we have

$$|\langle\mathcal{A}\rangle| \geqslant 2^{|G'|} - \frac{2^{|G'|}}{2^{k+1}} - 2^{|G'|}/2 \geqslant 2^{|G'|}/4 \geqslant 2^t$$

so we are done and if $|G'| = t+1$ then $3|G'| < 4t \leqslant 2tk$ so again by (1)

$$|\langle\mathcal{A}\rangle| \geqslant 2^{|G'|} - \frac{2^{|G'|}}{2^{k+1}} - 2^{|G'|}/3 \geqslant (1 - 1/8 - 1/3)2^{|G'|} > 2^{|G'|}/2 = 2^t$$

and we are also done. So we may assume that $|G'| = t$ and that $G' = [t]$. Similar to when we considered G'' we have now removed at most $\binom{t-1}{k-1}/t^{1/2}$ sets and since $n = \binom{t}{k}$ we know that at most $\binom{t-1}{k-1}/t^{1/2}$ sets in $[t]^{(k)}$ are not in $\mathcal{A}$ so for any $x \in [t]$ we have

$$d_{G'}(x) \geqslant \frac{\sqrt{t}-1}{\sqrt{t}}\binom{t-1}{k-1} \geqslant \frac{t-1}{2k}\binom{t-2}{k-2} \geqslant \lceil t^{1/2}\rceil\binom{t-2}{k-2}$$

for large enough B. Now by Lemma 4 on G' we have that

$$|\mathcal{A}'_x| \leqslant 2^{t-t^{1/2}} < \frac{2^t}{t2^{k+1}}$$

for large enough B since $t^{1/2} > \log t/\log 2 + (k+1)$ for large enough t. Now we have that

$$|\cup_{x\in G'}\mathcal{A}'_x| \leqslant \sum_{x\in G'}|\mathcal{A}'_x| \leqslant \frac{2^t}{2^{k+1}}.$$

From before we have $\sum_{i=0}^{k-1}\binom{t}{i} \leqslant 2^t/2^{k+1}$ and thus

$$|\langle\mathcal{A}\rangle \cap [t]^{(\geqslant k)}| \geqslant 2^t - \frac{2^t}{2^{k+1}} - \frac{2^t}{2^{k+1}} = 2^t\left(1 - \frac{1}{2^k}\right).$$

Now if $G \neq G'$ there is some $A_i \in \mathcal{A}, A_i \not\subset G'$ so by Lemma 5

$$|\langle\mathcal{A}\rangle| \geqslant \left(1 + \frac{1}{2^{k-1}}\right)|\langle\mathcal{A}\rangle \cap [t]^{(\geqslant k)}| \geqslant \left(1 + \frac{1}{2^{k-1}}\right)2^t\left(1 - \frac{1}{2^k}\right) > 2^t$$

and we are done. If $G = G'$ we must indeed have that $\mathcal{A} = [t]^{(k)}$ and so $|\langle\mathcal{A}\rangle| = 2^t - \sum_{i=0}^{k-1}\binom{t}{i}$. By the above we only have equality when $\mathcal{A} = X^{(k)}$ for some set X where $|X| = t$. This proves the lemma. □

We now prove our main result. All we need to do is reduce the ground set size enough to apply Lemma 6. To do this we will apply Lemmas 3,4 and 5.

Proof. Proof of Theorem 1. We will show that there is a C_k such that if $t > C_k$ we have $|\langle\mathcal{A}\rangle| \geqslant |[t]^{(\geqslant k)}| = 2^t - \sum_{i=0}^{k-1}\binom{t}{i}$. If $|G| \geqslant tk$ then by Lemma 3 we have $|\langle\mathcal{A}\rangle| \geqslant 2^t - 1$ so we may assume that $|G| < tk$. Now let $s = 2\lceil\log(2tk)\rceil$. Keep removing elements from G one by one (and all of the sets A_i containing them) until no element that is left is in less than $s\binom{tk}{k-2}$ of the remaining sets. Suppose that G' is the set of the remaining elements. We have removed at most $tks\binom{tk}{k-2}$ sets. Since $2k([t/2]+i) \geqslant tk$ for all $i \geqslant 1$ we have that

$$tk\binom{tk}{k-2} \leqslant \frac{(tk)^{k-1}}{(k-2)!} \leqslant (k-1)(2k)^{k-1}\binom{[t/2]+k}{k-1}.$$

We will make sure $C_k > 2k$. Now let $p = 2\lceil\log(2tk)\rceil(k-1)(2k)^{k-1} < 3(2k)^k\log t$. Then we have removed at most $p\binom{[t/2]+k}{k-1}$ sets. But we have also removed at least $\binom{t}{k} - \binom{|G'|}{k}$ sets. If $t > m \geqslant [t/2]+k$ then

$$\binom{t}{k} - \binom{m}{k} = \sum_{l=m}^{t-1}\binom{l}{k-1} \geqslant (t-m)\binom{[t/2]+k}{k-1}.$$

Since exponential beats polynomial take $C_k > 6k$ large enough so that

$$t - [t/2] - k \geqslant t/3 = \exp(\log t)/3 > 3(2k)^k\log t > p.$$

Then if $|G'| \leqslant [t/2]+k$ then we have removed at least

$$\binom{t}{k} - \binom{[t/2]+k}{k} \geqslant (t-[t/2]-k)\binom{[t/2]+k}{k-1} > p\binom{[t/2]+k}{k-1}$$

sets which is impossible. So we must have that $|G'| \geqslant [t/2]+k$ and in fact if $|G'| = t-r$ then we must have that $r \leqslant p < 3(2k)^k\log t$. We will first show that at least $1/4$ of all subsets of G' are actually in $\langle\mathcal{A}\rangle$.

Notice that if $x \in G'$ then $d_{G'}(x) \geqslant s\binom{tk}{k-2} \geqslant s\binom{|G'|-2}{k-2}$. This means we can apply Lemma 4 to G'. Define $\mathcal{A}'_x$ as in Lemma 6. We obtain that for all $x \in G'$ we have

$$|\mathcal{A}'_x| \leqslant 2^{|G'|-s} \leqslant \frac{2^{|G'|}}{2^{2\log(2tk)}} \leqslant \frac{2^{|G'|}}{2tk}.$$

This means that

$$|\cup_{x\in G'}\mathcal{A}'_x| \leqslant \sum_{x\in G'}|\mathcal{A}'_x| \leqslant |G'|\frac{2^{|G'|}}{2tk} \leqslant 2^{|G'|-1}.$$

If $A \subset G'$ and $|A| \geqslant k$ then $A \notin \langle\mathcal{A}\rangle$ if and only if there is some $x \in G'$ such that $A \in A'_x$. This means that

$$|\langle\mathcal{A}\rangle \cap \mathcal{P}(G')| = 2^{|G'|} - |\cup_{x\in G'}\mathcal{A}'_x| - \sum_{i=0}^{k-1}\binom{|G'|}{i} \geqslant$$

$$\geqslant 2^{|G'|} - 2^{|G'|-1} - \sum_{i=0}^{k-1}\binom{|G'|}{i}. \tag{2}$$

We can easily bound the sum of the binomials since if $t > 6k+2$ then we have $|G'| > t/2 > 3k+1$ so just like in the proof of Lemma 6 we will take C_k large enough so that $\sum_{i=0}^{k-1}\binom{|G'|}{i} \leqslant \binom{|G'|}{k} < |G'|^k < 2^{|G'|-k-1} \leqslant 2^{|G'|-2}$ since exponential beats polynomial. Now from (2) we have that $|\langle\mathcal{A}\rangle \cap \mathcal{P}(G')| \geqslant 2^{|G'|-2}$. We will now show that there is a constant D_k dependent on k such that if $|G| \geqslant t + D_k \log t$ then we must have $|\langle\mathcal{A}\rangle| \geqslant 2^t$. Let $\mathcal{H}_0 = \langle\mathcal{A}\rangle \cap \mathcal{P}(G')$ and $G_0 = G'$. First of, if $|G'| \geqslant t+2$ then we are done so suppose that $|G'| < t+2$. We will now show that we can get a lot more unions by adding sets with new elements. If we can pick some $x \in G\backslash G'$ and a set A_i containing x then notice that by taking $\mathcal{H}_1 = \mathcal{H}_0 \cup \{A \cup A_i | A \in \mathcal{H}_0\}$ and $G_1 = G_0 \cup A_i$ we see that by Lemma 5 $|\mathcal{H}_1| \geqslant (1 + 1/2^{k-1})|\mathcal{H}_0|$. This is useful, because we have multiplied the total number of unions by a constant bigger than 1 but dependent on k and we have added at most k new elements. We may keep on going as long as there are new elements and construct $\mathcal{H}_2, \mathcal{H}_3, \ldots, \mathcal{H}_q$ and $G_2, G_3, \ldots, G_q$ with $|H_i| \geqslant (1+1/2^{k-1})|H_{i-1}|$ and $|G_i\backslash G_{i-1}| \leqslant k$ for $1 \leqslant i \leqslant q$ and $G_q = G$. This means that if there are too many elements in G we will get that the total number of unions is too big. First of notice that $|G'| = t - r > t - 3(2k)^k \log t$. Now suppose that $|G| \geqslant t + 2 + (2k + 3k(2k)^k \log t)\log_{1+1/2^{k-1}} 2$. This means that we can repeat the above process of adding a set and less than k new elements at least $q \geqslant (2 + 3(2k)^k \log t)\log_{1+1/2^{k-1}} 2$ times so we have that

$$|\langle\mathcal{A}\rangle| \geqslant |\mathcal{H}_0|\left(1 + \frac{1}{2^{k-1}}\right)^{(2+3(2k)^k \log t)\log_{1+1/2^{k-1}} 2} \geqslant 2^{|G'|-2} \cdot 2^{2+3(2k)^k \log t} \geqslant 2^t.$$

This means that there are constants E_k and $D_k \geqslant 1$ dependent only on k such that if $t > E_k$ and $|G| \geqslant t + D_k \log t$ we have that $|\langle\mathcal{A}\rangle| \geqslant 2^t$. If $|G| < t + D_k \log t$ then since D_k depends only on k using Lemma 6 we get that there is a B dependent only on k such that if $t > B$ and $|G| < t + D_k \log t$ we have the desired result. Taking $C_k = \max(E_k, B)$ proves the theorem. □

We will define $f(n,k)$ to be the minimal size of a union-closed family generated by n k-sets. In the rest of this section we will prove Theorem 2 which solves the case $k = 2$ completely. We will show that if $n \in \mathbb{N}$ and $t \geqslant 2$ is the smallest positive integer such that $n \leqslant \binom{t}{2}$ then $f(n,2) = 2^t - 2^{\binom{t}{2}-n} - t$. We note that $t = \lfloor\sqrt{2n} + 3/2\rfloor$.

Proof. Proof of Theorem 2. As above $n = |\mathcal{A}|$ and let G be the graph with edge set $\mathcal{A} = \{A_1, \ldots, A_n\}$ and vertex set $A_1 \cup A_2 \ldots \cup A_n$. It is trivial to check that the theorem holds for $n \leqslant 3$ so we may assume that $n > 3$. Suppose that $|\langle\mathcal{A}\rangle| \leqslant |\langle\mathcal{B}\rangle|$. Let $t \in \mathbb{N}$ be such that $\binom{t-1}{2} < n \leqslant \binom{t}{2}$. Since $n > 3$ we have that $t \geqslant 4$. We see that $|G| \geqslant t$. Notice that $\mathcal{B} \subset [t]^{(2)}$ so $|\langle\mathcal{B}\rangle| \leqslant 2^t - t - 1 < 2^t - 1$. For any vertex $x \in G$ let $d(x)$ be the degree of x in G. By considering all edges with x we can make at least $2^{d(x)} - 1$ distinct unions. Thus we know that the maximal degree $\Delta(G) < t$. We also may assume that $\operatorname{diam} G \leqslant 2$ since if there are some $x, y \in G$ such that the distance $d(x,y)$ between x and y is at least 3 then we may replace the family $A_1, \ldots, A_n$ with the family obtained from those sets by identifying x and y. This will keep the size of the family to be n and will not increase $|\langle\mathcal{A}\rangle|$ as it will just identify x, y in all of the unions. Denote by $\Gamma(x)$ the set containing all elements adjacent to x in G. We prove the following claim.

Claim 7. *For every $x \in G$ we have $|G| - |\Gamma(x)| < t$.*

Proof. Suppose $x \in G$. Since $\operatorname{diam} G \leqslant 2$ and $|G| > 1$ for every $y \in G \backslash \Gamma(x)$ there is a $z \in \Gamma(x)$ such that $yz \in E(G)$. This means if we consider the family

$$\mathcal{C} = \{A \cap (G \backslash \Gamma(x)) | A \in \langle \mathcal{A} \rangle\}$$

then we have that each singleton subset of $G \backslash \Gamma(x)$ is in $\mathcal{C}$. But $\mathcal{C}$ is closed under unions and also has size at most $|\langle \mathcal{A} \rangle| < 2^t - 1$ meaning that $|G \backslash \Gamma(x)| < t$ which proves the claim. □

With the bound on $\Delta(G)$ this gives $|G| \leqslant 2t - 2$. We will show that $|G| - t$ must be small. Just like in the proof of Theorem 1 instead of counting how many sets are in $\langle \mathcal{A} \rangle$ we will instead count how many are in $\mathcal{F} = \mathcal{P}(G) \backslash \langle \mathcal{A} \rangle$. Observe that if $S \in \mathcal{F}$ and $S \neq \emptyset$ then for some $x \in S$ for all $y \in S \backslash \{x\}$ we have $xy \notin E(G)$. So for each set in $\mathcal{F}$ there is some element which prevents it from being in $\langle \mathcal{A} \rangle$. Similar to our above definition of $\mathcal{A}_x$ let

$$\mathcal{T}_x = \{S \subset G \mid x \in S, \forall y \in S \backslash \{x\} \; xy \notin E(G)\}.$$

Denote by s_x the degree of x in G^c. We have $|\mathcal{T}_x| = 2^{s_x}$ and by the above claim $s_x < t - 1$ for any x. Note that $\mathcal{F} = (\cup_{x \in G} \mathcal{T}_x) \cup \{\emptyset\}$ so we have the bound

$$|\mathcal{F}| \leqslant 1 + \sum_{x \in G} |\mathcal{T}_x| \leqslant 1 + \sum_{x \in G} 2^{s_x}. \tag{3}$$

We also know that $\sum_{x \in G} s_x = 2\left(\binom{|G|}{2} - n\right)$. If $a \geqslant b > 0$ then clearly $2^{a+1} + 2^{b-1} > 2^a + 2^b$. Now consider variables $s'_x \in \mathbb{Z}_{\geqslant 0}$ for $x \in G$. If there are distinct $x, y \in G$ such that $0 < s'_x \leqslant s'_y < t - 1$ then by above replacing s'_x, s'_y with $s'_x - 1, s'_y + 1$ increases $\sum_{x \in G} 2^{s'_x}$. So the biggest the sum $\sum_{x \in G} 2^{s'_x}$ could be subject to the constraints

$$\sum_{x \in G} s'_x = 2\left(\binom{|G|}{2} - n\right) \tag{4}$$

$$s'_x < t - 1 \tag{5}$$

is if all except at most one of the s'_x are either 0 or $t - 2$. Let $2\left(\binom{|G|}{2} - n\right) = d(t-2) + m$ with non-negative integers d, m and $0 \leqslant m < t - 2$. Since the s_x satisfy constraints (4) and (5) we must have $|G| \geqslant d$. By (3) this means that

$$|\mathcal{F}| \leqslant d2^{t-2} + 2^m + |G| - d. \tag{6}$$

Note that $|\langle \mathcal{A} \rangle| = 2^{|G|} - |\mathcal{F}|$ thus

$$2^t - t - 1 \geqslant |\langle \mathcal{A} \rangle| \geqslant 2^{|G|} - d2^{t-2} - 2^m - |G| + d. \tag{7}$$

We now want to bound d in terms of $|G|$ and t. We know that

$$d \leqslant \frac{2\left(\binom{|G|}{2} - n\right)}{t-2} \leqslant \frac{2\left(\binom{|G|}{2} - \binom{t-1}{2}\right)}{t-2} = \frac{(|G|-t+1)(|G|+t-2)}{t-2}$$

$$\leqslant \frac{(|G|-t+1)(3t-4)}{t-2} \leqslant 4(|G|-t+1).$$

We can also get a lower bound on d. Since $d = \left\lfloor \frac{2\left(\binom{|G|}{2}-n\right)}{t-2} \right\rfloor$ we have that

$$d \geqslant \left\lfloor \frac{\binom{|G|}{2} - \binom{t}{2}}{t-2} \right\rfloor = \left\lfloor \frac{(|G|-t)(|G|+t-1)}{2(t-2)} \right\rfloor \geqslant \left\lfloor \frac{(|G|-t)2(t-2)}{2(t-2)} \right\rfloor = |G|-t.$$

Now rearranging the terms in (7) and using the lower bound on d we obtain

$$(d+5)2^{t-2} \geqslant 2^t + 2^m + d2^{t-2} \geqslant 2^{|G|} + 1 + t - |G| + d \geqslant 2^{|G|}.$$

Now let $l = |G| - t + 1$. From the upper bound on d and the previous inequality we obtain $4l + 5 \geqslant d + 5 \geqslant 2^{l+1}$. Note that for $l = 4$ this does not hold and by induction if $2^{l+1} > 4l + 5$ then $2^{l+2} > 2(4l+5) > 4(l+1) + 5$ so we must have $l \leqslant 3$.

Now suppose that $l = 3$. Then $|G| = t+2$. Since $\binom{t+2}{2} - \binom{t}{2} \leqslant \binom{t+2}{2} - n < \binom{t+2}{2} - \binom{t-1}{2}$ we have that $4t + 2 \leqslant d(t-2) + m \leqslant 6t - 2$. If $d > 11$ then $d(t-2) > 11(t-2) \geqslant 6t-2$ since $t \geqslant 4$ but this is impossible so $d \leqslant 11$. Also since $(d+1)(t-2) \geqslant 4t + 2$ we must have $d \geqslant 4$. In both cases we have that $d2^{t-2} + 2^m - d \leqslant 12 \cdot 2^{t-2} - 4$. Now from (6) we have $|\mathcal{F}| \leqslant 12 \cdot 2^{t-2} + t + 2 - 4 = 3 \cdot 2^t + t - 2$. So we have

$$|\langle \mathcal{A} \rangle| = 2^{t+2} - |\mathcal{F}| \geqslant 2^{t+2} - 3 \cdot 2^t - t + 2 = 2^t - t + 2 > 2^t - t - 1 \geqslant |\langle \mathcal{A} \rangle|$$

which is a contradiction.

We move on to the case $l = 2, |G| = t+1$. Since $\binom{t+1}{2} - \binom{t}{2} \leqslant \binom{t+1}{2} - n < \binom{t+1}{2} - \binom{t-1}{2}$ we have that $2t \leqslant d(t-2) + m \leqslant 4t - 4$ so $d \geqslant 2$. If $d \leqslant 3$ or $d = 4, m = 0$ then by (6) we have that $|\mathcal{F}| \leqslant 2^t + t$ so $|\langle A \rangle| > 2^t - t - 1$ which is a contradiction. If $d = 4, m = 2, 4$ then $n \leqslant \binom{t-1}{2} + 2$ so $\mathcal{B}$ is contained in the family consisting of $[t-1]^{(2)}$ and $\{t, 1\}, \{t, 2\}$. Thus $|\langle B \rangle| \leqslant 2^t - t - 2^{t-3}$. But now by (6) we have that $|\mathcal{F}| \leqslant 2^t + 2^{t-3} + t - 3$ and thus $|\langle A \rangle| > 2^t - 2^{t-3} - t \geqslant |\langle B \rangle|$ which is a contradiction. We only have the cases $d > 4$ left but this can only happen for a few cases with small t. Since $t \geqslant 4$ we have that $7(t-2) = 7t - 14 > 4t - 4$ so we must have $d \leqslant 6$. We first prove a few common facts that apply to all these cases.

We first show that all subsets of G of size $t, t+1$ are in $\langle \mathcal{A} \rangle$. Clearly G is in $\langle \mathcal{A} \rangle$, now suppose $S \subset G$ with $|S| = t$. For any $x \in S$ by the claim above there are at most $t-2$ elements in G which are not neighbours of x. This means that $d(x) \geqslant 2$ and hence there is a $y \in S$ such that $xy \in E(G)$. This means that $S \in \langle \mathcal{A} \rangle$ which means all subsets of G of size $t, t+1$ are in $\langle \mathcal{A} \rangle$.

Next we give a lower bound on the number of 3-sets and 4-sets generated by $\mathcal{A}$. Note that any pair of 2-sets in $\mathcal{A}$ generate a 3-set or a 4-set and any set of size 3 or 4 can be generated with a pair of two sets in at most three different ways. This means that there are at least $\left\lceil \frac{n(n-1)}{6} \right\rceil$ distinct sets of size $3, 4$ in $\langle \mathcal{A} \rangle$.

Now we deal with the cases of $d > 4$. If $d = 5$ then $5(t-2) \leqslant 4t-4$ meaning $t \leqslant 6$. If $t = 4$ then m is even and less than two so we must have $m = 0$ and hence $n = 5$. If $t = 5$ then we must have m odd and $m \leqslant 1$ which gives $m = 1$ meaning that $n = 7$. If $t = 6$ then $m = 0$ and $n = 11$. Finally if $d = 6$ we must have $t = 4, m = 0$ which gives $n = 4$. We will show that in each case $|\langle \mathcal{A} \rangle| > |\langle \mathcal{B} \rangle|$ which will give a contradiction.

Case 1: $t = 4, n = 4$ and $\mathcal{B}$ consists of $[3]^{(2)}$ and the set $\{4, 1\}$. Counting sets of size $2, t, t+1$ generated by $\mathcal{A}$ we get that $|\langle \mathcal{A} \rangle| \geqslant 4 + t + 2 = 10$ while $|\langle \mathcal{B} \rangle| = 8$.

Case 2: $t = 4, n = 5$ and $\mathcal{B}$ consists of $[3]^{(2)}$ and the sets $\{4, 1\}$ and $\{4, 2\}$. Counting sets of size $2, t, t+1$ generated by $\mathcal{A}$ we see that $|\langle \mathcal{A} \rangle| \geqslant 5 + t + 2 = 11$ while $|\langle \mathcal{B} \rangle| = 10$.

Case 3: $t = 5, n = 7$ and $\mathcal{B}$ consists of $[4]^{(2)}$ and the set $\{5, 1\}$. Counting sets of size $2, 3, 4, t, t+1$ generated by $\mathcal{A}$ we get that $|\langle \mathcal{A} \rangle| \geqslant 7 + 7 + t + 2 = 21$ while $|\langle \mathcal{B} \rangle| = 19$.

Case 4: $t = 6, n = 11$ and $\mathcal{B}$ consists of $[5]^{(2)}$ and the set $\{6, 1\}$. Here $|\langle \mathcal{B} \rangle| = 42$. Note that similar to above we have that $d(x) \geqslant 2$ for any $x \in G$. Thus if some subset S of G of size 5 is not generated then there must be some $x \in S$ such that $G \backslash S = \Gamma(x)$. In other words $S = G \backslash \Gamma(x)$ for some $x \in G$. Thus at most seven 5-sets are not generated by $\mathcal{A}$. Thus we obtain $|\langle \mathcal{A} \rangle| \geqslant 11 + \left\lceil \frac{11 \cdot 10}{6} \right\rceil + \binom{7}{5} - 7 + 8 = 52$. Thus we have dealt with every possibility in the case of $l = 2$.

Now we deal with the final case when $l = 1$ and $|G| = t$. We may assume that $G = [t]$. We now have that $|\langle \mathcal{A} \rangle|$ is minimized exactly when $|\mathcal{F}|$ is maximized. Now let $r = \binom{t}{2} - n$. We know that $\sum_{i=1}^{t} s_i = 2r$ but if we look at $\mathcal{T}_1, \ldots, \mathcal{T}_t$ we see that at least r sets appear in two of the $\mathcal{T}_i$. This is because all sets corresponding to the edges in G^c appear in two of the $\mathcal{T}_i$. This means that

$$|\mathcal{F}| = |(\cup_{x \in G} \mathcal{T}_x) \cup \{\emptyset\}| \leqslant \sum_{i=1}^{t} 2^{s_i} - r + 1. \tag{8}$$

We consider the sum $X = \sum_{i=1}^{t} 2^{s_i}$ over all possible G^c with r edges. Without loss of generality we may assume that $s_1 \leqslant s_2 \leqslant \ldots \leqslant s_t$. Now suppose that some edge in G^c is not incident with t. Removing this edge we decrease X by at most 2^{s_t} (since two of the s_i decrease by 1. Now we can add an edge incident to t (that is not already present) because $r < \binom{t}{2} - \binom{t-1}{2} = t - 1$. This increases X by at least $2^{s_t} + 1$. So overall we have strictly increased X. This means that any G^c which maximizes X has all edges incident to the highest degree vertex so it is a star. In this case $X = 2^r + r + t - 1$. So we have that

$$|\mathcal{F}| \leqslant 2^r + r + t - 1 - r + 1 = 2^r + t.$$

Now we have that

$$|\langle \mathcal{A} \rangle| \geqslant 2^t - |\mathcal{F}| \geqslant 2^t - 2^r - t.$$

Notice that a set $S \subset [t]$ is not in $\langle \mathcal{B} \rangle$ if and only if $|S| \leqslant 1$ or $S = \{t\} \cup S'$ where $S' \subset \{t-r, t-r+1, \ldots, t-1\}$. This means that $|\langle \mathcal{B} \rangle| = 2^t - t - 1 - (2^r - 1) = 2^t - 2^r - t$. Thus $|\langle \mathcal{A} \rangle| \geqslant |\langle \mathcal{B} \rangle|$. This completes the proof. □

We have now shown that $f(n,2) = 2^t - 2^{\binom{t}{2}-n} - t$. Which families are extremal? From the above proof it is easy to show that the only extremal families are from the colex order (except for the case $n = 2$).

Corollary 8. *Let $\mathcal{A}$ be a family of 2-sets, and let $\mathcal{B}$ be the first $|\mathcal{A}|$ 2-sets in the colex order on $\mathbb{N}^{(2)}$. If $|\langle \mathcal{A} \rangle| = |\langle \mathcal{B} \rangle|$ then either $\mathcal{A}$ is isomorphic to $\mathcal{B}$ or $n = 2$ amd $\mathcal{A}$ contains two disjoint sets.*

Proof. If we have $|\langle \mathcal{A} \rangle| = |\langle \mathcal{B} \rangle|$ and $n \geqslant 4$ then from the above proof we must have $|G| = t$ and $|\mathcal{F}| = 2^r + t$ so G^c must be a star which means that $\mathcal{A}$ is isomorphic to $\mathcal{B}$. It is easy to check that this holds for $n = 1, 3$ as well and the only exception is $n = 2$ where we could also have two disjoint sets in $\mathcal{A}$. □

3 Related results

A conjecture of Roberts [1] asserts that Theorem 1 should be true for all t, not just for t large.

Conjecture 9. (Roberts [1]) Let $k, t \in \mathbb{N}$ where $t \geqslant k$. Let $n = \binom{t}{k}$ and let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a family of n distinct k-sets. Then $|\langle \mathcal{A} \rangle| \geqslant |[t]^{(\geqslant k)}|$.

In a different direction, what happens for values between the binomial coefficients? Suppose that $\binom{t}{k} < n < \binom{t+1}{k}$. It is natural to assume that to minimize $|\langle \mathcal{A} \rangle|$ over all families $\mathcal{A}$ of n distinct k-sets we can pick a family on the ground set $[t+1]$ that contains $[t]^{(k)}$ and some sets containing $t+1$. It is not hard to see that with the same idea as in the proof of Theorem 1 we can get that if $|\langle \mathcal{A} \rangle|$ is minimized then the size of the ground set G must be close to t if t is sufficiently large. From there it is also not hard to see that for most values of n (namely those that are not close to $\binom{t}{k}$), we can apply the idea of Lemma 6 to show that the ground set must have size $t+1$. With these ideas in mind we are able to solve a few special cases when n is very close to $\binom{t}{k}$. If $n = \binom{t}{k} - l$ where $l < \binom{k+1}{2}$ then for sufficiently large t (even larger than we needed for $\binom{t}{k}$) following the same reasoning as before if $|\langle \mathcal{A} \rangle| = f(n,k)$ then we may assume that $\mathcal{A} \subset [t]^{(k)}$.

Now we will show that $|\langle \mathcal{A} \rangle|$ is minimized when $\mathcal{A} = [t]^{k)} \backslash \mathcal{C}_l'$ where $\mathcal{C}_l$ is the initial segment of colex on $[t-1]^{(k-1)}$ of length l and $\mathcal{C}_l' = \{A \cup \{t\} | A \in \mathcal{C}_l\}$. First notice that the only sets in $[t]^{(\geqslant k)}$ that are not in $\langle \mathcal{A} \rangle$ of size at least $k+1$ must have size exactly $k+1$. We then show the following proposition.

Proposition 10. *Let $l < \binom{k+1}{2}$ be a positive integer and $\mathcal{X}$ be a family of k-sets where $|\mathcal{X}| = l$. Let $b_{\mathcal{X}}$ be the number of $(k+1)$-sets with the property that at least k of their k-subsets are in $\mathcal{X}$. Then $l \geqslant b_{\mathcal{X}} k - \binom{b_{\mathcal{X}}}{2}$.*

Proof. Let $C_1, \ldots, C_{b_\mathcal{X}}$ be the $(k+1)$-sets counted in $b_\mathcal{X}$. Then if $i \neq j$ then C_i and C_j have at most one subset of size k in common. Let $\mathcal{X} = \{X_1, \ldots, X_l\}$. The number of pairs (i,j) such that $X_i \subset C_j$ must be at least $b_\mathcal{X} k$ by counting for each j. On the other hand for any $j \neq j'$ there is at most one i such that both (i,j) and (i,j') are counted. So if X_i is counted k_i times then $\binom{k_i}{2}$ is the number of pairs $j \neq j'$ such that $C_j \cap C_{j'} = X_i$ and we have $\sum_{i=1}^{l} \binom{k_i}{2} \leqslant \binom{b_\mathcal{X}}{2}$. Since $k_i \leqslant 1 + \binom{k_i}{2}$ by taking the sum over i the total number of pairs counted is at most $l + \binom{b_\mathcal{X}}{2}$. This gives that $l \geqslant k b_\mathcal{X} - \binom{b_\mathcal{X}}{2}$ as required. □

Going back to our problem notice that since $\mathcal{A}$ is $[t]^{(k)}$ with l sets taken away we have that every set of size at least $k+1$ in $\mathcal{P}([t])$ is in $\langle \mathcal{A} \rangle$ except those $(k+1)$-sets that have at least k of its k-subsets removed. By the above proposition the number of those $(k+1)$-sets is at most the largest $s_l \leqslant k$ such that $l \geqslant s_l k - \binom{s_l}{2} = k + (k-1) + \ldots + (k - s_l + 1)$. Notice that if we remove exactly the sets of $\mathcal{C}'_l$ from $[t]^{(k)}$ then that removes exactly s_l sets of size $k+1$ from $\langle \mathcal{A} \rangle$. The removed sets are $\{t\} \cup [k+1] \backslash \{k+2-i\}$ for $i = 1, \ldots s_l$. This means that it is indeed optimal to take $[t]^{(k)} \backslash \mathcal{C}'_l$ for sufficiently large t.

What about the case when $n = \binom{t}{k} + l$? In this case we can show that if l is fixed then there is a large enough t depending on both k, l such that $|\langle \mathcal{A} \rangle|$ is minimized when $\mathcal{A} = [t]^{(k)} \cup \{Y_1, \ldots, Y_l\}$ where $Y_i = \{1, 2, \ldots, k-2, k-2+i, t+1\}$. Suppose that $|\langle \mathcal{A} \rangle| = f(n,k)$. As before using Lemma 3 to get a bound on the ground set and then using the ideas in Lemma 6 and the proof of Theorem 1 we may assume that $\mathcal{A}$ contains $\binom{t}{k} - o(\binom{t-1}{k-1})$ sets in $[t]^{(k)}$. Now suppose that $X_1, \ldots, X_s$ are all of the sets in $\mathcal{A}$ that are not in $[t]^{(k)}$. Trivially $s \geqslant l$. Further we let $X'_i = X_i \cap [t]$ and $\mathcal{B} = \mathcal{A} \cap [t]^{(k)}$. Notice that if we have a set $B \in \langle \mathcal{B} \rangle$ and there is an i such that $X'_i \subset B$ then if $X = X_i \cup B$ then $X \in \langle \mathcal{A} \rangle$ and $X \cap [t] = B$ but $X \notin [t]$. This gives a lot of sets in $\langle \mathcal{A} \rangle$ in addition to those that are in $\mathcal{P}([t])$.

Let $m_{s,t}$ be the number of sets in the upset generated by an initial segment of lex of length s on $[t]^{(k-1)}$. Let $\delta_{s,t} = m_{s,t}/2^t$. Notice that for $t > s + k$ we have that $\delta_{s,t} = \delta_{s,s+k}$ because the said initial segment of lex only has elements inside $[s+k]$. So define $\delta_s = \delta_{s,s+k}$. Crucially δ_l and δ_{l+1} depend only on l (and of course k) and notice that $\delta_{l+1} > \delta_l$. We see that if $\mathcal{C} = [t]^{(k)} \cup \{Y_1, \ldots, Y_l\}$ we have that

$$|\langle \mathcal{C} \rangle| \sim 2^t (1 + \delta_l) \tag{9}$$

for large t. Notice that for $t > (l+1)k$ the upset generated by $\mathcal{X}' = \{X'_1, \ldots, X'_s\}$ contains at least l sets in $[t]^{(k-1)}$ where it is exactly l if and only if $s = l$ and $|X'_i| = k-1$ for all i.

If the upset generated by $\mathcal{X}'$ has at least $l+1$ sets in $[t]^{(k-1)}$ then applying the Kruskal-Katona theorem similar to the proof of Lemma 4 the upset generated by $\mathcal{X}'$ in $[t]$ has size at least $\delta_{t+1} 2^t$. This means that there are at least $\delta_{l+1} 2^t$ sets in $\langle \mathcal{A} \rangle$ that are not subsets of $[t]$. For large enough t we can guarantee that $|\langle \mathcal{B} \rangle| = (1 - o(1)) 2^t$. Finally from (9) we have that

$$|\langle \mathcal{A} \rangle| \geqslant (1 + \delta_{l+1} - o(1)) 2^t > |\langle \mathcal{C} \rangle| \geqslant f(n,k) = |\langle \mathcal{A} \rangle|$$

which is a contradiction.

We will assume that $t > (l+1)k$. Then by above we must have that $s = l$ and $|X'_i| = k-1$ for all i. This means that $\mathcal{B} = [t]^{(k)}$. Now let a_i be the element in $X_i \backslash X'_i$.

Now we know that each set in the upset generated by $\mathcal{X}'$ gives rise to at least 1 set in $\langle\mathcal{A}\rangle$ that is not contained in $[t]$. So $|\langle\mathcal{A}\rangle| \geqslant |[t]^{(\geqslant k)}| + |\mathcal{U}_{\mathcal{X}'}|$ where $\mathcal{U}_{\mathcal{X}'}$ is the upset generated by $\mathcal{X}'$ in $[t]$. If $Y_i' = Y_i \cap [t]$ and $\mathcal{Y}' = \{Y_1', \ldots, Y_l'\}$ then by using the same idea as in the proof of Lemma 4 we get that $|\mathcal{U}_{\mathcal{X}'}| \geqslant |\mathcal{U}_{\mathcal{Y}'}|$. This means that

$$|\langle\mathcal{A}\rangle| \geqslant |[t]^{(\geqslant k)}| + |\mathcal{U}_{\mathcal{Y}'}| = |\langle [t]^{(k)} \cup \{Y_1, \ldots, Y_l\}\rangle|$$

for sufficiently large t which is what we wanted to show. Notice that also if some $a_i \neq a_j$ then $[t]$ gives rise to both $[t]\cup\{a_i\}, [t]\cup\{a_j\}$ in $\langle\mathcal{A}\rangle$ so we have $|\langle\mathcal{A}\rangle| > |\langle [t]^{(k)}\cup\{Y_1, \ldots, Y_l\}\rangle|$ which is a contradiction. Thus we must have all a_i to be the same.

We also must have that $|\mathcal{U}_{\mathcal{X}'}| = |\mathcal{U}_{\mathcal{Y}'}|$. Using the same idea as in the proof of Lemma 4 we can see that $|\mathcal{U}_{\mathcal{X}'} \cap [t]^{(t-1)}| = |\mathcal{U}_{\mathcal{Y}'} \cap [t]^{(t-1)}| = n - k + 2$. If a set $S \in [t]^{(t-1)}$ is not contained in $\mathcal{U}_{\mathcal{X}'}$ then if $S = [t]\backslash\{a\}$ we have that a is contained in every set of $\mathcal{X}'$. Therefore there are $k-2$ elements in $[t]$ that are contained in every set of $\mathcal{X}'$. Thus $\mathcal{X}'$ must be isomorphic to $\mathcal{Y}'$. This shows that all extremal families are isomorphic to $[t]^{(k)} \cup \{Y_1, \ldots, Y_l\}$ for sufficiently large t.

We have thus found extremal families for some n that are very close to $\binom{t}{k}$ but we need t to be large enough. Still, we do not know the extremal families for most values in between the binomial coefficients or for small values of t.

We will finish with a conjecture for the general result for all n. Colex is not the right order. As noted by Leck, Roberts and Simpson [2] if we consider

$$\mathcal{A} = [4]^{(3)} \cup \{\{1,2,5\}, \{1,3,5\}, \{1,4,5\}\}$$

and let $\mathcal{B}$ be the initial segment of colex on $\mathbb{N}^{(3)}$ of length 7, then $|\langle\mathcal{A}\rangle| = 12$, but $|\langle\mathcal{B}\rangle| = 13$. If $\binom{t}{k} < n < \binom{t+1}{k}$ then as mentioned before for most n if $|\langle\mathcal{A}\rangle| = f(n,k)$ then the ground set of $\mathcal{A}$ has size $t+1$ so we may assume that $\mathcal{A} \subset [t+1]^{(k)}$. If we further assume that $[t]^{(k)} \subset \mathcal{A}$ then the we have $|\langle\mathcal{A}\rangle| = |[t]^{(\geqslant k)}| + |\mathcal{U}_{\mathcal{X}'}|$ where $\mathcal{X}'$ is defined as it is for the $\binom{t}{k} + l$ case and $|\mathcal{X}'| = n - \binom{t}{k}$. As before this is minimized when $\mathcal{X}'$ is the initial segment of lex on $[t]^{(k-1)}$. So something like "colex but lex inside a given maximal element" could be the correct order. Such a mixed ordering has occurred in other problems as well (see Engel and Leck [6] and also Duffus, Howard and Leader [7]). We will define the *max-lex* order on $\mathbb{N}^{(k)}$ to be the order in which $A < B$ if either $\max A < \max B$ or else $\max A = \max B$ and $\min(A\Delta B) \in A$. The following lovely conjecture of Roberts [1] includes Conjecture 3.1 – we strongly believe it is true.

Conjecture 11. (Roberts [1]) Let $k \in \mathbb{N}$ and $\mathcal{A} \subset \mathbb{N}^{(k)}$. If $\mathcal{B}$ is the initial segment of max-lex on $\mathbb{N}^{(k)}$ of size $|\mathcal{A}|$ then $|\langle\mathcal{A}\rangle| \geqslant |\langle\mathcal{B}\rangle|$.

References

[1] I. T. Roberts, Extremal problems and designs on finite sets., Ph.D. thesis, Curtin University, 1999.

[2] U. Leck, I. Roberts, J. Simpson, Minimizing the weight of the union-closure of families of two-sets, *Australas. J. Combin.* 52(1): 67–73, 2012.

[3] B. Bollobás, Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability, Cambridge University Press, USA, 1986.

[4] J. B. Kruskal, The number of simplices in a complex, in: *Mathematical Optimization Techniques*, University of California Press, Berkeley, pp. 251–278, 1963.

[5] G. Katona, A Theorem of Finite Sets, in: *Classic Papers in Combinatorics*, Birkhäuser Boston, Boston, MA, pp. 381–401, 1987.

[6] K. Engel and U. Leck, Optimal antichains and ideals in Macaulay posets, in: *Bolyai Society Mathematical Studies*, Vol. 7, Budapest, pp. 199–222, 1999.

[7] D. Duffus, D. Howard, I. Leader, The width of downsets, *European J.Combin.* 79: 46–59, 2019.