Discrete Mathematics 309 (2009) 3083–3091

Contents lists available at ScienceDirect

Discrete Mathematics

journal homepage: www.elsevier.com/locate/disc

Pruning processes and a new characterization of convex geometries

Federico Ardila [a], Elitza Maneva [b]

[a] Department of Mathematics, San Francisco State University, 94132 San Francisco, CA, USA

[b] IBM Almaden Research Center, 95120 San José, CA, USA

ARTICLE INFO

Article history:
Received 2 February 2008
Accepted 13 August 2008
Available online 20 September 2008

Keywords:
Antimatroid
Convex geometry
k-SAT
Pruning process
Removal process

ABSTRACT

We provide a new characterization of convex geometries via a multivariate version of an identity that was originally proved, in a special case arising from the k-SAT problem, by Maneva, Mossel and Wainwright. We thus highlight the connection between various characterizations of convex geometries and a family of removal processes studied in the literature on random structures.

1. Introduction

This article studies a general class of procedures in which the elements of a set are removed one at a time according to a given rule. We refer to such a procedure as a *removal process*. If every element which is removable at some stage of the process remains removable at any later stage, we call this a *pruning process*. The subsets that one can reach through a pruning process have the elegant combinatorial structure of a *convex geometry*. Our first goal is to highlight the role of convex geometries in the literature on random structures, where many pruning processes have been studied without exploiting their connection to these objects. Our second contribution is a proof that a generalization of a polynomial identity, first obtained for a specific removal process in [17], provides a new characterization of pruning processes and of convex geometries. To prove this result we also show how a convex geometry is equivalent to a particular kind of interval partition of the Boolean lattice.

Two equivalent families of combinatorial objects, known as *convex geometries* and *antimatroids*, were defined in the 1980s [8,11]. The fact that these objects can be characterized via pruning processes has been known since then. Some examples of pruning processes considered at that time are the removal of vertices of the convex hull of a set of points in $\mathbb{R}^n$, the removal of the leaves of a tree, and the removal of minimal elements of a poset. More recently various pruning processes have been studied in the literature on random structures, and referred to also as peeling, stripping, whitening, coarsening, identifying, etc. A typical example is the removal of vertices of degree less than k in the process of finding the k-core of a random (hyper)graph.

In [17], a surprising identity was proved to hold for a particular removal process which arises in the context of the k-SAT problem. In this paper, we answer the question posed by Mossel [23] of characterizing the combinatorial structures that satisfy (the multivariate version of) that identity: they are precisely the convex geometries or equivalently the pruning processes. That is the content of our main result, Theorem 3.1 and Corollary 3.2. It says that any pruning process has the following two properties, and that in fact either of these two properties characterizes pruning processes among removal processes.

E-mail addresses: federico@math.sfsu.edu (F. Ardila), enmaneva@us.ibm.com (E. Maneva).

0012-365X/$ – see front matter
doi:10.1016/j.disc.2008.08.020

- Suppose there is a subset S of elements that we do not wish to remove. Then there is a unique minimal set $\tau(S)$ achievable by the pruning process which contains S.
- Suppose to each element e corresponds a weight p_e. For every set S reachable by the pruning process, whose set of removable elements is $U \subseteq S$, define the weight of S to be $\prod_{e \notin S} p_e \prod_{e \in U}(1 - p_e)$. Then the sum of the weights of all reachable sets is 1.

Equivalently, after appropriate rewording, either of these properties characterize convex geometries among set systems.

Outline. In Section 2 we define precisely the equivalent concepts of convex geometry, antimatroid, and pruning process. We describe a few different ways of looking at these objects: as set systems, as languages generated by a set of circuits or pruning rules, and as lattices. These different representations bring forward the differences between various pruning processes that have been studied. Section 3 is devoted to our new characterization of convex geometries. Finally, in Section 4 we apply our results to the k-SAT problem and the distribution considered in [17], thereby generalizing Theorem 6 of [17].

Related work. Antimatroids and convex geometries were first identified in the context of lattice theory by Dilworth [7]. Since then they have appeared in a variety of combinatorial situations. Two particularly important treatments are Edelman and Jamison's convexity approach [8], and Korte and Lovász's greedoid approach [11]. Two good introductions to the subject are [3,12].

Two notable examples of pruning processes on random structures appear in the analysis of identifiable vertices in random hypergraphs [6], and of k-cores in random hypergraphs [22,25]. Additionally, in the analysis of satisfiability problems such processes have appeared repeatedly — most notably for the pure-literal rule algorithm for k-SAT [5,21,22,26], and the study of the clustering of solutions for XOR-SAT [20] and k-SAT [1,18].

Pruning processes appear also in practical applications; for example in error correcting codes such as LDPC codes [9,16] and LT codes [14] over the erasure channel. A unified analysis of the pruning processes in error-correcting codes and the pure-literal rule is provided in [15].

Our work and in particular the implication $1 \rightarrow 3$ of our main Theorem 3.1 is related to previous work of Aivaliotis, Gordon, and Graveman [2] and Gordon [10]. For more information on this connection, see Section 3.1.

A word on terminology. The objects of this paper have been studied under several different names. In particular, other authors have referred to pruning processes as *shelling processes*. We prefer to avoid this name, which may lead to confusion with the other, better established notion of shelling in combinatorics. The term "pruning" is more accurate, since a pruning process is equivalent to the successive removal of outermost elements of a convex geometry. A good example to keep in mind throughout the paper is the process of pruning of a tree by successively removing its leaves.

2. Convex geometries and antimatroids

Convex geometries and antimatroids are equivalent families of combinatorial objects. Convex geometries provide a combinatorial abstraction of the notion of convexity. Antimatroids describe pruning processes, where we remove elements from (or add elements to) a set one at a time, and once an element becomes available for removal, it remains available until it is removed. There are many equivalent definitions of these objects and a vast underlying theory [3,12]. We now present four points of view which we will use.

2.1. Convex sets and closure operations

A *convex geometry* is a pair $(E, \mathcal{N})$ where E is a set and $\mathcal{N} \subseteq 2^E$ is a collection of subsets of E satisfying:

(N1) $E \in \mathcal{N}$.
(N2) If $A, B \in \mathcal{N}$ then $A \cap B \in \mathcal{N}$.
(N3) For every $A \in \mathcal{N}$ with $A \neq E$ there is an $x \notin A$ such that $A \cup x \in \mathcal{N}$.

The sets in $\mathcal{N}$ are called *closed* or *convex*. It is sometimes convenient to think of $\mathcal{N}$ as a poset ordered by containment; this is a lattice. We can then think of (N3) as a property of *accessibility from the top*: every closed set can be obtained from E by removing one element at a time in such a way that every intermediate set in the process is also closed.

The *closure* of $A \subseteq E$ is defined to be

$$\tau(A) = \bigcap_{\{C \in \mathcal{N} : C \supseteq A\}} C,$$

which is the minimum closed set containing A. It is easy to see that τ is, in fact, a *closure operator*; that is, for all A we have $A \subseteq \tau(A)$ and $\tau(\tau(A)) = \tau(A)$, and for all $A \subseteq B$ we have $\tau(A) \subseteq \tau(B)$. Also, a set A is closed if and only if $\tau(A) = A$.

Example 2.1. For a given graph, consider the subgraphs that can be obtained by successively removing leaves (nodes of degree 1). The vertex sets of these subgraphs are the closed sets of a convex geometry. The minimal closed set is the 2-core of the graph. Fig. 1 shows a specific graph and its lattice of closed sets; for example, the closure of the set $\{a, f, g\}$ is the set $\{a, c, e, f, g\}$.

An *extreme point* of a set A is an element $a \in A$ which is not in the closure of $A - a$. The set $\mathrm{ex}(A)$ of extreme points of A is the unique minimal set whose closure is A.

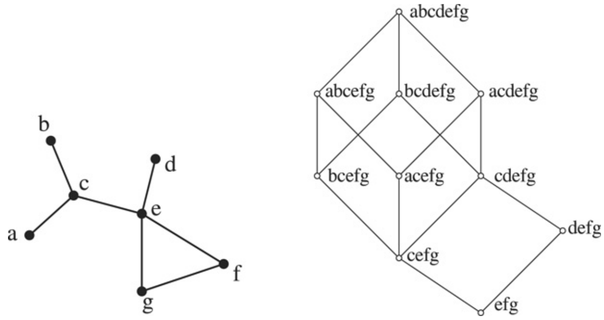

Fig. 1. The set of configurations reachable by the process that successively removes leaves from a graph. The bottom configuration is the 2-core of the graph.

2.2. Antimatroids and pruning processes

Next we define antimatroids, which are equivalent to convex geometries. Let E be a set, whose elements we regard as letters. A word over the alphabet E is called *simple* if it contains no repeated letters; let E_s^* be the set of simple words over E.

An *antimatroid* is a pair $(E, \mathcal{L})$ where E is a set and $\mathcal{L} \subseteq E_s^*$ is a set of simple words satisfying:

(L1) If $\alpha\beta \in \mathcal{L}$ then $\alpha \in \mathcal{L}$; that is, any beginning section of a word of $\mathcal{L}$ is in $\mathcal{L}$.
(L2) If $\alpha, \beta \in \mathcal{L}$ and $|\alpha| > |\beta|$, then α contains a letter x such that $\beta x \in \mathcal{L}$.
(L3) If $\alpha, \beta \in \mathcal{L}$ and $x \in E$ are such that $\alpha x, \alpha\beta \in \mathcal{L}$ and $x \notin \beta$, then $\alpha\beta x \in \mathcal{L}$.

Axiom (L1) says that $\mathcal{L}$ is left hereditary, (L2) is an exchange axiom, and (L3) states that, as we build up a word of $\mathcal{L}$ from left to right, any letter which can be added to the word at a certain stage can still be added at any later stage.

The supports of the words in $\mathcal{L}$ are called the *feasible* subsets of E. The feasible subsets determine $\mathcal{L}$: a word is in $\mathcal{L}$ if and only if every initial segment of it is a feasible set. The following theorem provides a one-to-one correspondence between antimatroids and convex geometries.

Theorem 2.2 ([12, Theorem III.1.3]). *Let E be a finite set and $\mathcal{F}$ be a collection of subsets of E. Then $\mathcal{F}$ is the collection of feasible sets of an antimatroid if and only if $\mathcal{F}^* = \{E - F : F \in \mathcal{F}\}$ is the collection of closed sets of a convex geometry.*

By the above theorem, the feasible sets of the antimatroid corresponding to our example of a convex geometry can be read on the descending paths from the top of the lattice of closed sets. A letter corresponds to every edge — this is the element that is removed. Any set of removed elements is the complement of a convex set, thus it is a feasible set. We will see below that the words of the antimatroid can also be read on the descending paths.

An alternative characterization of antimatroids starts by defining a set $H(x) \subseteq 2^{E-x}$ for every $x \in E$, which is a collection of *alternative precedences* for x; each precedence is a set not containing x. Let $\mathcal{L}$ be the set of words on the alphabet E such that x can only appear after at least one of its precedences has appeared:

$$\mathcal{L} = \{x_1 \dots x_k : \text{ for all } i \text{ there is a set } A \in H(x_i) \text{ with } A \subseteq \{x_1, \dots, x_{i-1}\}\}.$$

In the example of Fig. 1 the alternative precedences for c are $\{a, b\}$, $\{a, e\}$, and $\{b, e\}$. In general, for the process of removing leaves to obtain the 2-core of a graph, a vertex of degree d becomes removable as soon as $d - 1$ of its neighbors are removed. Thus its precedences are all the $(d - 1)$-subsets of its set of d neighbors.

A *removal process* is a procedure in which the elements of a set are removed one at a time according to a given rule. If every element which is removable at some stage of the process remains removable at any later stage, we call this a *pruning process*.

We are now in a position to explain the correspondence between pruning processes and antimatroids. Given an antimatroid $\mathcal{L}$ on ground set E, we can consider each word w of $\mathcal{L}$ as a removal sequence, which instructs us to remove the elements of w from left to right. These removal sequences describe a pruning process: if an element x is removable at a certain stage described by word $w \in \mathcal{L}$ (that is, if $wx \in \mathcal{L}$), then one of the alternative precedences of x appears in w. At any later stage the removal sequence w' will contain w as a prefix, and therefore will contain that alternative precedence for x as well.

Conversely, suppose we are given a pruning process on a set E. For each element x let the alternative precedences of x be the subsets $A \subseteq E$ such that x is removable in $E - A$. Clearly the antimatroid determined by these sets of alternative precedences consists of the removal sequences in our pruning process.

Example 2.1 gives rise to an antimatroid whose words are the *pruning sequences* of leaves that one can successively remove from the graph. These correspond to the descending paths from the top of the lattice of closed sets; the word indicates the elements that are being removed as we walk down. For example the leftmost path in the lattice of Fig. 1 gives the word *dabc*, which corresponds to a valid order of successively removing leaves from the graph.

2.3. Circuits and paths

The circuit description of a convex geometry is a good way to reveal the pruning process which generates it. A *rooted set* is a set with a designated element called the root. Any collection of rooted sets gives rise to a convex geometry as follows. Suppose $\mathcal{C}$ is a family of rooted subsets of E; label them (A_i, a_i) where $a_i \in A_i$ and $A_i \subseteq E$. Call a subset S of E *full* if $A_i - a_i \subseteq S$ implies $a_i \in S$ for each i; that is, if the root of a set is never the only element of the rooted set missing from S. Say a full set S is *accessible from E* if there exists a sequence of full subsets $E = S_0 \supset S_1 \supset \cdots \supset S_k = S$ with $|S_i - S_{i+1}| = 1$ for each i. The following is, in a different language, Lemma 3.7 of [12].

Proposition 2.3 ([12]). *Let $\mathcal{C}$ be a collection of rooted sets in E. The collection $\mathcal{N}(\mathcal{C})$ of full subsets of E which are accessible from E is the collection of closed sets of a convex geometry on E.*

The set of rooted sets can be interpreted as pruning rules. Let for every $e \in E$, $\mathcal{C}_e = \{C \setminus \{e\} : (C, e) \in \mathcal{C}\}$. Then the corresponding pruning process is the one in which an element e is removable if and only if at least one element has been removed from each set in $\mathcal{C}_e$.

Conversely, from a convex geometry it is possible to recover a set of rooted sets that generates it. A *free set* is one of the form $\mathrm{ex}(A)$. A *circuit* is a minimal set which is not free, and one can check that each circuit C has a unique element a which is in the closure of the remaining ones. This element is called the *root* of the circuit, and (C, a) is called a *rooted circuit*.

Even though we will not need this fact, let us point out that the collection of full sets of a family of rooted sets has a nice structure.

Proposition 2.4. *A collection $\mathcal{F} \subseteq 2^E$ is the collection of full sets determined by a family of rooted sets if and only if it contains E and is closed under intersection.*

Proof. First, it is easy to see that the collection of full sets determined by a family of rooted sets $\mathcal{C}$ contains E and is closed under intersection.

Next, suppose $\mathcal{F}$ is a collection of subsets of E that contains E and is closed under intersection. We start with $\mathcal{C}$ being the complete set of rooted sets on E. For every $F \in \mathcal{F}$ remove from $\mathcal{C}$ all rooted sets $(A \cup a, a)$, where $A \subseteq F$ and $a \notin F$. We claim that $\mathcal{F}$ is the collection of full sets of $\mathcal{C}$.

It is immediate by the construction that every $F \in \mathcal{F}$ is a full set. It remains to show that there are no other full sets. Suppose $D \subseteq E$ is a full set for $\mathcal{C}$. That means that all rooted sets $(A \cup a, a)$ with $A \subseteq D$ and $a \notin D$ have been removed. In particular $(D \cup a, a)$ has been removed for all $a \notin D$. This implies that for every $a \notin D$ there exists $F_a \in \mathcal{F}$ such that $D \subseteq F_a$ and $a \notin F_a$. Since $D = \cap F_a$ and $\mathcal{F}$ is closed under intersection, D is in $\mathcal{F}$. □

So far in this paper, rooted sets have played the role of circuits in a convex geometry. It is worth pointing out, however, that one can consider rooted sets as *paths* which generate a convex geometry in a different way, as follows. Suppose $\mathcal{P}$ is a family of rooted subsets of E which we now call *paths*; label them (P_i, p_i), where $p_i \in P_i$ and $P_i \subseteq E$. Let a subset S of E be *path-full* if for every $e \in S$ there exists $(P, e) \in \mathcal{P}$ such that $P \subseteq S$. Let it be *path-closed* if it is path-full, and accessible from E by a sequence $E = S_0 \supset S_1 \supset \cdots \supset S_k = S$ of path-full subsets with $|S_i - S_{i+1}| = 1$ for each i. Then the path-closed sets are the closed sets of a convex geometry, and every convex geometry arises in this way from a set of paths.

In this context, the rooted sets can again be interpreted as pruning rules. For every $e \in E$ let $\mathcal{P}_e = \{P \setminus \{e\} : (P, e) \in \mathcal{P}\}$. Then the corresponding pruning process is the one in which an element e is removable if and only if there is a set $P \in \mathcal{P}_e$ such that every element of P has been removed.

The pruning processes in the literature on random structures are generated by rules for removing elements which can usually be represented in a natural way through circuits or paths. While both points of view are equivalent, sometimes one is more natural than the other. For example, in finding the k-core of a graph [25], a vertex becomes removable when at least one element has been removed from every k-subset of its neighbors (circuit rule). On the other hand, in the case of identifiable vertices in hypergraphs [6] a vertex is removable if, in at least one of the hyperedges in which it appears, every other vertex has been removed (path rule).

In the rest of this paper, all the rooted sets that appear play the role of circuits.

2.4. Lattices

Finally, we can also think of convex geometries as meet-distributive lattices. A lattice L is *meet-distributive* if for any element $x \neq \widehat{0}$ the interval $[m(x), x]$ is a Boolean lattice, where $m(x)$ is the meet of the elements covered by x. Meet-distributive lattices are precisely the posets of closed sets of convex geometries [3, Prop. 8.7.5].

One might wonder whether something more specific can be said about the convex geometries that arise from a set of circuits each of size at most k. This question will be particularly natural in Section 4, where convex geometries are applied to the k-SAT problem for a fixed value of k.

For $k = 2$ the situation is very nice. Recall that a lattice L is *distributive* if $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ for any x, y, z in L.

Proposition 2.5 ([12, Cor. 3.10]). *Let $\mathcal{C}$ be a set of rooted sets of size 2. The convex geometry $\mathcal{N}(\mathcal{C})$ generated by these sets is a distributive lattice. Conversely, every distributive lattice arises in this way.*

For higher values of k, if we start with a collection $\mathcal{C}$ of rooted circuits of size k, the resulting convex geometry $\mathcal{N}$ generally has additional rooted circuits of different sizes. For the case of $k = 2$, all circuits of the generated convex geometry have size 2. However, for example, the convex geometry defined by the rooted 3-sets (abc, b) and (bde, d) also has $(acde, d)$ as a circuit.

If we have a bound on the size of the circuits of a convex geometry, we can make the following statement. A lattice L is *k-distributive* if $x \wedge (y_0 \vee \cdots \vee y_k) = (x \wedge y_0) \vee \cdots \vee (x \wedge y_k)$ for any $x, y_0, \ldots, y_k$ in L.

Proposition 2.6 ([13, Cor. 4.3.]). *Let $k \geq 3$ be an integer. If all circuits of a convex geometry have size at most k, then its poset of closed sets is a $(k-1)$-distributive lattice. Not every $(k-1)$-distributive lattice arises in this way.*

3. Convex geometries as interval partitions of Boolean lattices

In this section we describe our new characterization of convex geometries. We show that convex geometries on a set E are characterized by the fact that they induce a certain partition of the Boolean lattice 2^E. This partition is encoded in a polynomial identity which, as we will later see, generalizes Theorem 4.2 from [17].

For any collection $\mathcal{N} \subseteq 2^E$ of subsets of E, and a set A in $\mathcal{N}$, say that an element $a \in A$ is *excludable from A* if $A - a$ is in $\mathcal{N}$. Let $\mathrm{ex}(A)$ be the set of excludable elements of A. When $\mathcal{N}$ is the collection of closed sets of a convex geometry, $\mathrm{ex}(A)$ is the set of extreme points of A.

Theorem 3.1. *Let $\mathcal{N} \subseteq 2^E$ be a collection of subsets of a non-empty set E. The following statements are equivalent:*

1. *$\mathcal{N}$ is the collection of closed sets of a convex geometry.*
2. *As A ranges over $\mathcal{N}$ the intervals $[\mathrm{ex}(A), A]$ partition the Boolean lattice 2^E; that is, for every $D \subseteq E$ there is a unique $A \in \mathcal{N}$ such that $\mathrm{ex}(A) \subseteq D \subseteq A$.*
3. *For any collection of p_i and q_i for $i \in E$ such that $p_i + q_i = 1$ for all i, we have*
$$\sum_{A \in \mathcal{N}} \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j = 1.$$

Proof. 1. *implies* 2. Notice that if $A \in \mathcal{N}$ is such that $\mathrm{ex}(A) \subseteq D \subseteq A$, then we have that $A = \tau(\mathrm{ex}(A)) \subseteq \tau(D) \subseteq \tau(A) = A$; so the only possible choice for A is $A = \tau(D)$. It remains to notice that, since $\mathrm{ex}(A)$ is the unique minimal set such that $A = \tau(\mathrm{ex}(A))$, and $A = \tau(D)$, it follows that $D \supseteq \mathrm{ex}(A)$, and therefore $D \in [\mathrm{ex}(A), A]$.
2. *implies* 1. We define the map $\phi : 2^E \to \mathcal{N}$, as follows: for every $D \subseteq E$, let $\phi(D)$ be the unique element $A \in \mathcal{N}$ such that $\mathrm{ex}(A) \subseteq D \subseteq A$. We need to show axioms (N1)–(N3) of a convex geometry $(E, \mathcal{N})$: E is in $\mathcal{N}$, $\mathcal{N}$ is closed under intersection, and every $A \in \mathcal{N}$ is accessible from E.

Axiom (N1) obviously holds, because $\phi(E) = E$ is in $\mathcal{N}$. To show (N3), we show that every set $A \in \mathcal{N}$ is accessible from any superset $B \supseteq A$ that also belongs to $\mathcal{N}$. It suffices to prove that there exists an element of $B \setminus A$ that is excludable from B. If that were not the case, then $\mathrm{ex}(B) \subseteq A$, and both of the intervals $[\mathrm{ex}(A), A]$ and $[\mathrm{ex}(B), B]$ would contain A, a contradiction.

Finally, we need to prove (N2), which states that $\mathcal{N}$ is closed under intersection. First we prove the following statement:

If $B \in \mathcal{N}$ and $A \subseteq B$ then $\phi(A) \subseteq B$.

Suppose that we remove one element at a time from B in any arbitrary way, with the restriction that the intermediate sets in the process must all be in $\mathcal{N}$ and contain A. We keep doing this until we cannot continue anymore; suppose the set we obtain is C; by construction, $B \supseteq C \supseteq A$. That means that every element excludable from C is in A, so $\mathrm{ex}(C) \subseteq A \subseteq C$. Thus $C = \phi(A)$ and we obtain the desired statement.

Now suppose A_1 and A_2 are in $\mathcal{N}$. From $A_1 \cap A_2 \subseteq A_1$ we obtain that $\phi(A_1 \cap A_2) \subseteq A_1$. Similarly $\phi(A_1 \cap A_2) \subseteq A_2$, so $\phi(A_1 \cap A_2) \subseteq A_1 \cap A_2$. But the reverse inclusion holds by definition, so we must have equality. It follows that $A_1 \cap A_2$ is in $\mathcal{N}$.
2. *implies* 3. Observe that
$$\sum_{D \subseteq E} \prod_{i \notin D} p_i \prod_{j \in D} q_j = \prod_{h \in E} (p_h + q_h) = 1.$$
Therefore, since for every D there is a unique A such that $\mathrm{ex}(A) \subseteq D \subseteq A$, it suffices to prove that for every $A \in \mathcal{N}$:
$$\prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j = \sum_{D \in [\mathrm{ex}(A), A]} \prod_{i \notin D} p_i \prod_{j \in D} q_j.$$
This is easily seen to be true because:
$$\begin{aligned}
\sum_{D \in [\mathrm{ex}(A), A]} \prod_{i \notin D} p_i \prod_{j \in D} q_j &= \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j \sum_{R \subseteq A \setminus \mathrm{ex}(A)} \left(\prod_{i \in A \setminus (\mathrm{ex}(A) \cup R)} p_i \prod_{j \in R} q_j \right) \\
&= \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j \prod_{h \in A \setminus \mathrm{ex}(A)} (p_h + q_h) \\
&= \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j.
\end{aligned}$$

3. *implies* 2. Consider any set $D \subseteq E$, and let $p_a = 0$ if $a \in D$ and $p_a = 1$ otherwise. The equality becomes:

$$\begin{aligned} 1 &= \sum_{A \in \mathcal{N}} \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j \\ &= \sum_{A \in \mathcal{N} : \mathrm{ex}(A) \subseteq D \subseteq A} 1 \end{aligned}$$

Therefore there is exactly one set $A \in \mathcal{N}$ for which $\mathrm{ex}(A) \subseteq D \subseteq A$. □

The following corollary gives a characterization of pruning processes among removal processes:

Corollary 3.2. *A removal process on a set E is a pruning process if and only if, for each subset S of E, there is a unique minimal set $\tau(S)$ containing S which is achievable by the removal process.*

Proof. As outlined in Section 2.2, a pruning process gives rise to a convex geometry, and in that case $\tau(S)$ is just the convex closure of S. For the other direction, let $\mathcal{N}$ consist of the sets achievable by the removal process; it suffices to show that $\mathcal{N}$ is the collection of closed sets of a convex geometry. We will show that property 2 of Theorem 3.1 holds. For any set $S \subseteq E$, it holds that $\mathrm{ex}(\tau(S)) \subseteq S \subseteq \tau(S)$, because if there is an excludable element of $\tau(S)$ that is not in S, then $\tau(S)$ would not be the minimal set containing S. Furthermore, any set $T \in \mathcal{N}$ for which $\mathrm{ex}(T) \subseteq S \subseteq T$ is a minimal set containing S because all of its excludable elements are in S. Since there is a unique such set, it is $\tau(S)$. □

We conclude this section by offering a probabilistic interpretation of property 3 of Theorem 3.1.

3.1. A probabilistic interpretation

Let $(E, \mathcal{N})$ be a convex geometry, and fix $0 \le p_e, q_e \le 1$ with $p_e + q_e = 1$ for each element e of E. Define a probability distribution π_1 on the subsets of E by independently deleting element e with probability p_e and keeping it with probability q_e:

$$\Pr_{\pi_1}(A) = \prod_{i \notin A} p_i \prod_{j \in A} q_j, \quad A \subseteq E.$$

Define a probability distribution π_2 on the convex sets of E by:

$$\Pr_{\pi_2}(A) = \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j, \quad A \in \mathcal{N}.$$

The implication $1 \to 3$ of Theorem 3.1 tells us that π_2 is, indeed, a probability distribution. Furthermore, in order to sample from π_2, it suffices to sample from π_1 and compute the closure of the obtained set.

Theorem 3.3. *Let $f(A)$ be a function defined on the subsets A of E which depends only on $\tau(A)$. Then the expected value of f when we sample from the distribution π_1 on all subsets of E equals the expected value of f when we sample from the distribution π_2 on the convex sets of E.*

Proof. Assuming that $f(D) = f(\tau(D))$, the identity

$$\sum_{D \subseteq E} f(D) \prod_{i \notin D} p_i \prod_{j \in D} q_j = \sum_{A \in \mathcal{N}} f(A) \prod_{i \notin A} p_i \prod_{j \in \mathrm{ex}(A)} q_j. \tag{1}$$

can be established in exactly the same way as implication $1 \to 3$ of Theorem 3.1. □

We note that Aivaliotis, Gordon, and Graveman [2] and Gordon [10] studied the problem of choosing a random subset of a convex geometry under the distribution π_1. They related the expected rank of this random subset to the Tutte polynomial of the antimatroid. In particular, they discovered (1) in a special case which is no simpler than the general case.

The fact that π_2 is a probability distribution on $\mathcal{N}$ is not explicitly stated in [2] or [10], and neither is the probabilistic interpretation of the right hand side of (1). However, these two results follow very easily from that work. Our theorem that the probabilistic property of Theorem 3.3 (or the weaker condition 3 of Theorem 3.1) characterizes convex geometries is new.

4. Convex geometries in the k-SAT problem

Let F be a Boolean formula such as

$$F = (\bar{x}_1 \vee \bar{x}_2 \vee x_3) \wedge (x_2 \vee \bar{x}_3 \vee \bar{x}_4).$$

We can assume that F is written in *conjunctive normal form* as a conjunction of certain clauses C in the variables V and their negations. The *Boolean satisfiability problem (SAT)* is to determine whether there is some assignment of TRUE (1) and

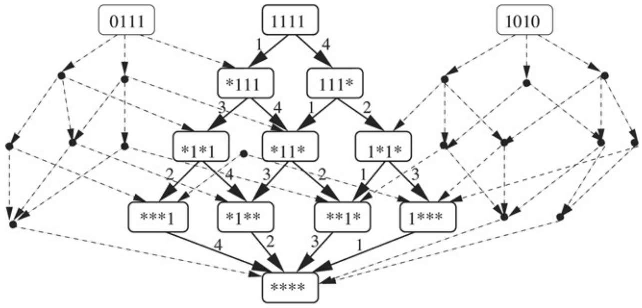

Fig. 2. Some of the valid partial assignments for the formula $(\bar{x}_1 \vee \bar{x}_2 \vee x_3) \wedge (x_2 \vee \bar{x}_3 \vee \bar{x}_4)$. Highlighted are the assignments below the satisfying assignment (1, 1, 1, 1). Edges are labeled with the index of the variable whose value differs in the two adjacent assignments.

FALSE (0) to the variables which makes the entire formula true. The *k-SAT problem* is the same problem when restricted to formulas with clauses of a fixed size k. For $k = 2$ there is a polynomial time algorithm for deciding satisfiability; however for $k \geq 3$ the problem is NP-complete.

In their analysis of the Survey Propagation algorithm [19,4] for 3-SAT, Maneva et al. [17] discovered a polynomial identity that holds for any SAT problem and any satisfying assignment. To define this identity first we need to introduce the concept of *partial assignments*, where to each variable is assigned one of the values 0, 1, or $*$; the value $*$ indicates that a variable is unassigned and free to take either value. Say that a partial assignment $\boldsymbol{x}$ is *invalid* for a clause C if plugging $\boldsymbol{x}$ into C gives either $0 \vee 0 \vee \cdots \vee 0$ (which makes the clause invalid) or $0 \vee \cdots \vee 0 \vee * \vee 0 \vee \cdots \vee 0$ (where the $*$ is not free to take either value). A partial assignment $\boldsymbol{x}$ is *valid* for a formula if it is valid for all its clauses. For example, some valid partial assignments for the formula $F = (\bar{x}_1 \vee \bar{x}_2 \vee x_3) \wedge (x_2 \vee \bar{x}_3 \vee \bar{x}_4)$ are $(1, 1, 1, 1)$, $(*, 1, *, *)$ and $(1, *, *, 1)$, and some invalid partial assignments are $(1, 1, 0, *)$ and $(*, *, 1, 1)$.

Definition 4.1. Given a Boolean formula F, the poset $P(F)$ of valid partial assignments is defined by decreeing that $\boldsymbol{a}$ covers $\boldsymbol{b}$ if $\boldsymbol{b}$ is obtained from $\boldsymbol{a}$ by switching a 0 or 1 to a $*$.

Fig. 2 shows part of the poset of valid partial assignments for the formula F above. Note that in this example, somewhat surprisingly, $(1, 1, 1, 1)$ is not greater than $(1, *, *, 1)$ in $P(F)$, because to stay valid one must switch x_2 and x_3 from 1 to $*$ simultaneously: $(1, *, 1, 1)$ and $(1, 1, *, 1)$ are invalid.

Definition 4.1 suggests that, given a valid partial assignment $\boldsymbol{a}$, we call a coordinate i either:

(a) a *star* $*$,
(b) *unconstrained* if $a_i \in \{0, 1\}$ and setting $a_i = *$ keeps the assignment valid, or
(c) *constrained* if $a_i \in \{0, 1\}$ and setting $a_i = *$ gives an invalid assignment; that is, if a_i is the only satisfying variable in some clause of F.

Let $S(\boldsymbol{a})$, $U(\boldsymbol{a})$, $C(\boldsymbol{a})$, and $N(\boldsymbol{a})$ be the sets of star, unconstrained, constrained, and numerical variables of $\boldsymbol{a}$, respectively; so $V = S(\boldsymbol{a}) \cup N(\boldsymbol{a})$ and $N(\boldsymbol{a}) = U(\boldsymbol{a}) \cup C(\boldsymbol{a})$.

Maneva et al. [17] defined the *weight* of a partial assignment $\boldsymbol{a}$ to be

$$W(\boldsymbol{a}) = p^{|S(\boldsymbol{a})|} q^{|U(\boldsymbol{a})|},$$

where p and q are parameters in the interval $[0, 1]$. They considered the probability distribution which assigns to $\boldsymbol{a}$ a probability proportional to $W(\boldsymbol{a})$ for every valid partial assignment $\boldsymbol{a}$. The survey propagation algorithm was then proved to be equivalent to applying the belief propagation marginalization heuristic [24] to this distribution with suitably chosen p and q. This distribution has the following property, which should not look surprising in view of Theorem 3.1:

Theorem 4.2 (*[17]*). *For any satisfying assignment $\boldsymbol{a}$ of a Boolean formula F and $p + q = 1$,*

$$\sum_{\boldsymbol{b} \leq \boldsymbol{a}} p^{|S(\boldsymbol{b})|} q^{|U(\boldsymbol{b})|} = 1,$$

summing over all valid partial assignments $\boldsymbol{b}$ which are less than $\boldsymbol{a}$ in $P(F)$; that is, summing over the subposet $P(F)_{\leq \boldsymbol{a}}$.

Thus the probability distribution on partial assignments with $p > 0$ may be regarded as a "smoother" version of the uniform distribution over satisfying assignments, which corresponds to the case $p = 0$: if we choose a valid partial assignment $\boldsymbol{b}$ at random, then the probability of being under $\boldsymbol{a}$ is the same for any satisfying assignment $\boldsymbol{a}$. Another consequence of the above theorem is that, if the total weight of all valid partial assignments is less than 1, then the formula has no satisfying assignment. In recent work of Sinclair and the second author [18], this fact was used in conjunction with the first-moment method to bound the probability of satisfiability of a random SAT formula with clauses of sizes 2 and 3.

Consider the following experiment:
1. in a valid assignment $\boldsymbol{a}$, change a random unconstrained variable to $*$, and
2. repeat until there are no unconstrained variables.

This procedure has been referred to as "peeling", "whitening", "coarsening" and "pruning". We now recognize it as a pruning process on the set of variables which have numerical values in $\boldsymbol{a}$. At each stage, we are allowed to remove an unconstrained variable; notice that if a variable becomes unconstrained, it remains unconstrained throughout this process.

This experiment is equivalent to taking a random path from $\boldsymbol{x}$ down the partial order $P(F)$, by choosing at each step a random partial assignment that is covered by the current one. For a fixed choice of $\boldsymbol{a}$, any such path terminates at the same partial assignment, which is known as a "core". (Note, however, that different $\boldsymbol{a}$ may lead to different core assignments.)

Achlioptas and Ricci-Tersenghi [1] examined the above pruning process and proved that, for $k \geq 9$ and a formula chosen from a particular distribution of interest, there is a high probability that the process will terminate before removing all variables. This is not known to hold for $k = 3$.

With the above description of the removal process, our next result follows easily.

Theorem 4.3. *Let F be a SAT formula with variables V, and let $\boldsymbol{a}$ be a valid (possibly partial) assignment for F. Let*

$$\mathcal{N} = \{N(\boldsymbol{b}) : \boldsymbol{b} \text{ is a valid partial assignment such that } \boldsymbol{b} \leq \boldsymbol{a}\}.$$

Then $(N(\boldsymbol{a}), \mathcal{N})$ is a convex geometry. Conversely, every convex geometry arises in this way from a valid assignment for a SAT formula.

Proof. We show that this statement is equivalent to Proposition 2.3. Consider the clauses of F with a unique satisfying variable in $\boldsymbol{a}$, which give $0 \vee \cdots \vee 0 \vee 1 \vee 0 \vee \cdots \vee 0$ when we plug $\boldsymbol{a}$ into them. If C is the set of variables in such a clause (which must be a subset of $N(\boldsymbol{a})$) and v is the unique satisfying variable, form a rooted set (C, v). Then $(N(\boldsymbol{a}), \mathcal{N})$ is clearly the convex geometry generated by these rooted sets. Conversely, given a convex geometry, one can encode its rooted sets into the clauses of a SAT formula with a valid assignment. □

The convex geometry corresponding to assignment $(1, 1, 1, 1)$ in Fig. 2 is the collection of sets of assigned variables in assignments lying below $(1, 1, 1, 1)$:

$$\{\{1, 2, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3\}, \{2, 4\}, \{2, 3\}, \{1, 3\}, \{4\}, \{2\}, \{3\}, \{1\}, \emptyset\}$$

and the words of the corresponding antimatroid can be read out by going down the directed edges; the feasible sets are:

$$\{\emptyset, \{1\}, \{4\}, \{1, 3\}, \{1, 4\}, \{2, 4\}, \{1, 2, 3\}, \{1, 3, 4\}, \{1, 2, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}.$$

In the particular case of 2-SAT, the convex geometry is very special. By Proposition 2.5, the poset $P(F)_{\leq a}$ of valid partial assignments is a distributive lattice.

Notice that a SAT formula F generally has several different valid partial assignments, and each assignment $\boldsymbol{a}$ gives rise to a convex geometry $G(F, \boldsymbol{a})$. These different convex geometries fit together nicely, as seen in Fig. 2. If $G(F, \boldsymbol{a})$ and $G(F, \boldsymbol{b})$ have a non-empty intersection, then their intersection is the convex geometry $G(F, \boldsymbol{c})$ for the unique element $\boldsymbol{c}$ with maximal $N(\boldsymbol{c})$ for which $c_i = *$ if $a_i \neq b_i$, and $c_i = a_i = b_i$ otherwise.

The machinery that we have built up now provides a more illustrative multivariate version of Maneva, Mossel, and Wainwright's Theorem 4.2 on the probability distribution determined by a SAT problem F and a valid assignment $\boldsymbol{a}$. More importantly, in view of Theorem 3.1, it tells us that the identity of Theorem 4.2 holds precisely **because** a SAT problem gives rise to a convex geometry. Therefore convex geometries are really the context in which this identity should be understood.

Theorem 4.4. *For a valid partial assignment $\boldsymbol{b}$ of a Boolean formula F with variables V, let $S(\boldsymbol{b})$, $U(\boldsymbol{b})$ and $C(\boldsymbol{b})$ denote the sets of star, unconstrained, and constrained variables of $\boldsymbol{b}$ in F, respectively. Let p_i and q_i be such that $p_i + q_i = 1$ for all $i \in V$. Then, for any valid assignment $\boldsymbol{a}$ of a Boolean formula F,*

$$\sum_{\boldsymbol{b} \leq \boldsymbol{a}} \prod_{i \in S(\boldsymbol{b})} p_i \prod_{j \in U(\boldsymbol{b})} q_j = 1$$

summing over all valid assignments $\boldsymbol{b}$ which are less than $\boldsymbol{a}$ in $P(F)$.

Proof. The result follows directly from Theorems 3.1 and 4.3. □

Acknowledgments

We would like to thank Elchanan Mossel for posing the question of characterizing the combinatorial objects satisfying Theorem 4.2, and Laci Lovász and Martin Wainwright for helpful discussions. We would also like to thank the referee for very useful suggestions for improving the exposition.

References

[1] D. Achlioptas, F. Ricci-Tersenghi, On the solution-space geometry of random constraint satisfaction problems, in: Proc. 38th ACM Symp. Theory of Computing, 2006.

[2] M. Aivaliotis, G. Gordon, W. Graveman, When bad things happen to good trees, J. Graph Theory 37 (2001) 79–99.
[3] A. Björner, G. Ziegler, Introduction to greedoids, in: N. White (Ed.), Matroid Applications, Cambridge University Press, 1992, pp. 284–354.
[4] A. Braunstein, M. Mézard, R. Zecchina, Survey propagation: An algorithm for satisfiability, Random Structures and Algorithms 27 (2005) 201–226.
[5] A.Z. Broder, A.M. Frieze, E. Upfal, On the satisfiability and maximum satisfiability of random 3-cnf formulas, in: Proc. 4th ACM-SIAM Symp. Discrete Algorithms, 1993, pp. 322–330.
[6] R.W.R. Darling, J.R. Norris, Structure of large random hypergraphs, Ann. Appl. Probab. 15 (2005) 125–152.
[7] R.P. Dilworth, Lattices with unique irreducible decompositions, Ann. Math. 41 (1940) 771–777.
[8] P.H. Edelman, R. Jamison, The theory of convex geometries, Geom. Dedicata 19 (1985) 247–274.
[9] R.G. Gallager, Low-Density Parity Check Codes, MIT Press, Cambridge, MA, 1963.
[10] G. Gordon, Expected rank in antimatroids, Adv. in Appl. Math. 32 (2004) 299–318.
[11] B. Korte, L. Lovász, Shelling structures, convexity and a happy end, in: B. Bolobás (Ed.), Graph theory and combinatorics, Proceedings of the Cambridge Combinatorial Conference in Honor of Paul Erdös, Academic Press, London, UK, 1984, pp. 219–232.
[12] B. Korte, L. Lovász, R. Schrader, Greedoids, Springer-Verlag, Berlin, Heidelberg, 1991.
[13] L. Libkin, n-distributivity, dimension and Carathéodory's theorem, Algebra Universalis 34 (1995) 72–95.
[14] M. Luby, LT codes, in: Proc. 42nd IEEE Symp. Foundations of Computer Science, 2002, pp. 271–282.
[15] M. Luby, M. Mitzenmacher, M.A. Shokrollahi, Analysis of random processes via AND-OR tree evaluation, in: Proc. 9th ACM-SIAM Symp. Discrete Algorithms, 1998, pp. 364–373.
[16] M. Luby, M. Mitzenmacher, M.A. Shokrollahi, D.A. Spielman, V. Stemann, Practical loss-resilient codes, in: Proc. 29th ACM Symp. Theory of Computing, 1997, pp. 150–159.
[17] E. Maneva, E. Mossel, M.J. Wainwright, A new look at Survey Propagation and its generalizations, J. ACM 55 (2007) 2–41. Extended abstract appeared in Proc. 16th ACM-SIAM Symp. on Discrete Algorithms, 2005, pp.1089–1098.
[18] E. Maneva, A. Sinclair, On the satisfiability threshold and clustering of solutions of random 3-SAT formulas, Technical Report 0710.0805, ArXiv, 2007, Theoret. Comput. Sci. (in press).
[19] M. Mézard, G. Parisi, R. Zecchina, Analytic and algorithmic solution of random satisfiability problems, Science 297 (2002) 812 (Scienceexpress published on-line 27-June-2002; 10.1126/science.1073287).
[20] M. Mézard, F. Ricci-Tersenghi, R. Zecchina, Two solutions to diluted p-spin models and XORSAT problems, J. Stat. Phys. 111 (2003) 505.
[21] M. Mitzenmacher, Tight thresholds for the pure literal rule, Technical Report 1997-011, DEC/SRC, 1997.
[22] M. Molloy, The pure literal rule threshold and cores in random hypergraphs, in: Proc. 15th ACM-SIAM Symp. Discrete Algorithms, 2004, pp. 672–681.
[23] E. Mossel, Personal communication, 2004.
[24] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, Palo Alto, CA, 1988.
[25] B. Pittel, J. Spencer, N. Wormald, Sudden emergence of a giant k-core in a random graph, J. of Combin. Theory, Ser. B 67 (1996) 111–151.
[26] J.W. Rosenthal, J.M. Plotkin, J. Franco, The probability of pure literals, J. Comput. Logic 9 (1999) 501–513.