Birthdays, Broadcasts, and Boolean Algebras:
Probabilistic Boolean Algebras and Applications

Alexa Sharp

Cornell University, Ithaca NY 14853, USA,
asharp@cs.cornell.edu

Abstract. In the area of extremal finite set theory there are many combinatorial results concerning the selection of m k-element sets. This type of set selection can also be viewed as a boolean algebra. In this paper we consider a probabilistic construction of this boolean algebra, concentrating on the structure and properties such an algebra may form, particularly the structure of the algebra's atoms. The results are then applied to a generalization of the popular birthday problem, where the event of interest is now whether all selected sets have a unique element; we find an upper bound on the probability of this event. We also extend the definition of the generalized birthday problem to model content protection protocols. While these protocols are widely used in digital media rights management, they are insufficiently analyzed due to a lack of such an underlying model. We focus on the event that revoking the rights of multiple pirate users inadvertently causes the rights of other, authorized users to be unjustly revoked; we give an exact formula for the probability of this event.

1 Introduction

In this paper we consider a probabilistic view of a boolean algebra formed by the selection of m k-element sets selected independently and uniformly at random from a set X. The combinatorial (non-probabilistic) version of this problem has been extensively studied in extremal finite set theory [1–3]. These problems are generally concerned with finding the maximum number of sets that can be selected and maintain a given property. In contrast, we fix the number of sets and investigate, in a probabilistic fashion, the structure of the resulting boolean algebra, in particular the structure of atoms. We use the results to analyze the following applications:

A Generalized Birthday Problem. The *birthday problem*, as introduced by mathematician Richard von Mises in 1939, asks for the minimum number of people required so that the probability of two having the same birthday is greater than one half. The assumptions are a year of length 365 days, with each day equally likely as a birthday, and birthdays independent from person to person. It is well known that the answer is 23 [4].

Many versions of this problem have been studied. (See Diaconis and Mosteller [5] for a discussion of these methods.) The natural generalization we are concerned with is based on the following model: we select m k-sets uniformly and independently at random with replacement from N elements. We are interested in the probability that each set has a unique element not shared with any other set, or conversely, the probability that at least one of the m sets has each of its elements also selected by some other set. In extremal finite set theory, this question asks for the probability that our m sets form an $(m-1)$-cover-free family.

Rights Revocation in Content Protection Protocols. There has been an interest recently in content protection technology and digital rights management, stimulated by the growing amount of digital media now available to consumers and the prospect of home entertainment and broadcast networks. The goal of content protection is to allow authorized users access to all of their own licensed content, but to prevent unauthorized users from accessing or distributing this same content.

At a high-level, this works in the following way: On the manufacturing side, each manufactured device (PDA, MP3 player, TV, cable receiver) is assigned a combination of keys taken from a large set of keys. On the content owners' side, the digitized content is encrypted such that only authorized devices can access the protected subject matter. When a device tries to play some content, the device key is used to determine whether the device is indeed authorized to play the content. A device is authorized to play some content if at least one of its assigned keys is valid.

Problems arise when unauthorized users and devices enter the picture. When an unauthorized device is detected, we first determine which keys are used by the pirated device, and then we revoke these keys to prevent further illegal access of the protected content. Unfortunately, the unauthorized device's keys can overlap other devices' keys; it is conceivable we may accidentally invalidate the rights of a legitimate device by revoking the rights of multiple unauthorized devices with keys completely overlapping the keys of said authorized device. To guarantee the non-occurrence of such undesirable behaviour, we would require all authorized devices to hold at least one key not contained in the combined union of the unauthorized devices' keys. In contrast, the generalized birthday problem can be viewed as requiring all devices, both authorized and unauthorized, to hold at least one key not contained in the combined union of all other devices' keys. In this sense, the birthday problem can be viewed as m instances of the rights revocation problem where in instance i all but device A_i is pirated.

This application was motivated by a talk given by IBM's Tushar Chandra on their xCP technology for digital media rights management [8]. Although revocation is a core component of this technology, it appeared at the time to lack a formal model. Unfortunately the details of xCP are proprietary and not readily accessible. Regardless, we believe the presented model is general enough to encompass whatever details the technology may incorporate and to be useful in

other related applications.

In this paper we first consider the generalized birthday problem, for which we obtain an upper bound on the probability of the m selected sets forming an $(m-1)$-cover-free family, along with an exact formula for the expected number of sets with unique elements. We then modify the birthday problem even further to obtain a model for the rights revocation problem discussed above. With this new model we are able to find the probability distribution of the number of unjustly invalidated devices; this can then be used by manufacturers to decide how many keys to create and how many to assign to each device in order to keep the probability of invalidation error below an acceptable minimum. Along the way we use multiple intermediate results pertaining to the properties of boolean algebras, in particular to the structure of atoms, which may be of independent interest.

2 The Underlying Model

A *boolean algebra* is a set A together with binary operations $+$ and $\cdot$, a unary operation $-$, and elements 0, 1 of A such that the following laws hold: commutative and associative laws for addition and multiplication, distributive laws both for multiplication over addition and for addition over multiplication, and the following special laws:

$$\begin{array}{ll} x + (x \cdot y) = x & x \cdot (x + y) = x \\ x + (-x) = 1 & x \cdot (-x) = 0 \end{array}$$

One of the most common boolean algebras is a collection $\mathcal{A} = A_1, \ldots, A_m$ of subsets of a set X closed under the operations of union, intersection and complementation with respect to X, with members $\emptyset$ and X. In fact, Stone's Representation Theorem states that every boolean algebra is isomorphic to such a boolean algebra of sets [9]. In our case, we are only considering finite boolean algebras.

Of particular interest are the *atoms* of our boolean algebra. The atoms are defined to be $B_1 \cap B_2 \cap \cdots \cap B_m$, where $B_i \in \{A_i, \overline{A_i}\}$, i.e. the minimal non-zero elements.

Finally, let us define the notion of an r-cover free family. Let X be an N-set and let $\mathcal{A}$ be a set of subsets of X. $(X, \mathcal{A})$ is called a *r-cover-free family* provided that, for any subset $A_i \in \mathcal{A}$ and any other r subsets $B_1, \ldots, B_r \in \mathcal{A}$, we have $A_i \not\subseteq \cup_{j=1}^{r} B_j$. In this paper we consider the case of $(m-1)$-cover-free families, where $m = |\mathcal{A}|$. In particular, we are interested in families of sets such that no set is contained in the union of all the others.

3 The Generalized Birthday Problem

The birthday problem asks for the minimum number of people necessary such that the probability of at least one repeated birthday amongst the group is

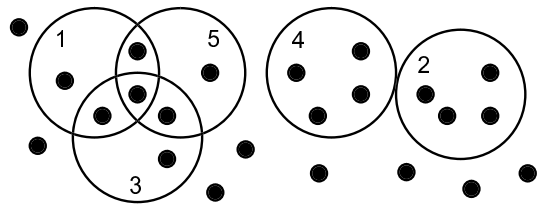

Fig. 1. There are 10 atoms in this case with $m = 5$ and $k = 4$.

greater than one half. In particular, we could ask for the probability that at least two people out of m share the same birthday, or the probability that none share a birthday. A natural generalization would be to ask for the probability that each set has a unique element not shared with any other set, when we pick m k-element sets uniformly and independently at random with replacement, from N elements. Equivalently, we could ask for the probability that at least one of the m sets has each of its elements contained in at least one of the other $m-1$ sets. These two formulations are asking for the probability that we either do or do not have an $(m-1)$-cover-free family, respectively. Note that the $k = 1$ case is precisely the original birthday problem, and the $k = 2$ case can be viewed as asking m people for their parents' birthdays and finding the probability that each person has a parent's birthday not shared by anyone else's parents.

In order to answer these questions, we focus on the expected number of sets out of m with unique elements. When $k = 1$ this is analogous to finding the expected number of unique birthdays in a group of m people.

Theorem 1. *After m set selections*

(a) $E[\textit{num. sets with unique elements}] = m \cdot \sum_{w=1}^{k} (-1)^{w+1} \binom{N}{w} \frac{\binom{N-w}{k-w}}{\binom{N}{k}} \left[\frac{\binom{N-w}{k}}{\binom{N}{k}} \right]^{m-1}$

(b) $Pr[\textit{all } m \textit{ sets have a unique element}] \leq \sum_{w=1}^{k} (-1)^{w+1} \binom{N}{w} \frac{\binom{N-w}{k-w}}{\binom{N}{k}} \left[\frac{\binom{N-w}{k}}{\binom{N}{k}} \right]^{m-1}$.

Proof. For (a) we use lemma 7 for the probability that any given set A_j has a unique element, yielding

$$\begin{aligned} E[\text{num. sets with unique elements}] &= m \cdot Pr[\text{set } A_j \text{ has a unique element}] \\ &= m \cdot \sum_{w=1}^{k} (-1)^{w+1} \binom{N}{w} \frac{\binom{N-w}{k-w}}{\binom{N}{k}} \left[\frac{\binom{N-w}{k}}{\binom{N}{k}} \right]^{m-1} \end{aligned}$$

We then use Markov's inequality to obtain the bound of (b) on the probability of such an event occuring.

$$Pr[\text{all } m \text{ sets have a unique element}] \leq Pr[\text{set } A_j \text{ has a unique element}] \qquad \square$$

If N is large then the event that any two sets have a unique element are by and large independent, and we would have the following result:

Lemma 1. *After m sets selections, and for large N*

$$Pr[\textit{all sets have a unique element}] \approx (Pr[\textit{set } A_j \textit{ has a unique element}])^m .$$

4 Rights Revocation in Content Protection Protocols

The introduction includes a discussion on the importance of digital rights and the issue of unauthorized users. As an illustration of the problem, consider figure 2, where it would be unacceptable to revoke the keys in the sets 1, 2, and 3, as all of the keys in set 4 would be unjustly invalidated leaving no key for step 4 to use to decrypt the content. But it is acceptable to revoke the keys in the sets 1, 2, and 4, or all of sets 1, 2, 3, and 4.

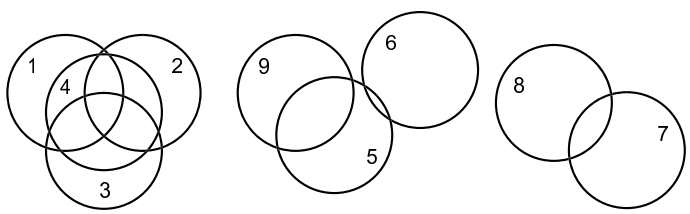

Fig. 2. Revoking sets 1, 2 and 3 invalidates set 4 even if set 4's device is authorized.

How does this relate to the atoms problem we originally described? Suppose we have a large set of N keys from which manufacturers select the unique sets of keys assigned to each device. Each device is assigned a random set of k keys, such as the 9 sets in figure 2. We would like to know the likelihood of accidentally invalidating the rights of an authorized user by revoking the rights of multiple unauthorized users such as the previously described scenario of invalidating sets 1, 2, and 3 but not set 4.

More formally, the problem is stated as follows: We select m k-sets independently and uniformly at random from a set of N elements. We then start revoking some t of the m sets, where the t sets are selected for revocation independently and uniformly at random. We ask for the probability that we inadvertently invalidate an authorized set.

Before we answer this question, we point out that, unlike the birthday problem, it is allowable for some of our m k-sets to have no unique element, such as set 4 in figure 2. This is fine so long as we do not revoke all the sets covering such a set and not the set itself, as this is precisely the undesirable situation we are trying to avoid.

Theorem 2. *For m set selections and t set revocations*

(a) $Pr[c \text{ sets unjustly revoked}] = \binom{m-t}{c} p^c (1-p)^{m-t-c}$
(b) $E[\text{number of unjustly revoked sets}] = (m-t) \cdot p$,

where $p = \sum_{s \geq k} \binom{N}{s} \sum_{i=0}^{s-k} (-1)^i \binom{s}{i} \left(\frac{\binom{s-i}{k}}{\binom{N}{k}} \right)^t \frac{\binom{s}{k}}{\binom{N}{k}}$.

Proof. Since the sets are both selected and revoked uniformly and independently at random, the process can be viewed as revoking t sets all at once and then considering the remaining $m-t$ sets individually. If any of the $m-t$ remaining sets are repeated, i.e. have no unique element, with respect to only the t revoked sets then we have invalidated a set we shouldn't have. In particular, we do not have to concern ourselves with repeats in the $m-t$ authorized sets. Furthermore, we may ignore repeated sets in the t revoked sets because all t sets are being invalidated, and invalidating a set twice is acceptable. These two facts allow us to ignore repeats in the first t set selections, and to treat all remaining $m-t$ selections exactly the same.

Viewed in this way, we essentially have $m-t$ independent and uniform set selections, and each has the same probability of being completely contained in the t revoked sets. In particular, we have a sequence of Bernoulli trials where the probability p of success on the ith trial is the probability that set A_i is repeated with respect to the t revoked sets. From corollary 3,

$$\begin{aligned} p &= Pr[t+1^{st} \text{ set has no unique elements}] \\ &= \sum_{s \geq k} \binom{N}{s} \sum_{i=0}^{s-k} (-1)^i \binom{s}{i} \left(\frac{\binom{s-i}{k}}{\binom{N}{k}} \right)^t \frac{\binom{s}{k}}{\binom{N}{k}}. \end{aligned}$$

We want to find the probability of c successes:

$$Pr[c \text{ sets unjustly revoked}] = \binom{m-t}{c} p^c (1-p)^{m-t-c}$$

The estimated number of such sets is

$$E[\text{number of unjustly revoked sets}] = (m-t) \cdot p \qquad \square$$

Corollary 1. *For m set selections and t set revocations, the probability that no set is unjustly revoked is*

$$\left[1 - \sum_{s \geq k} \binom{N}{s} \sum_{i=0}^{s-k} (-1)^i \binom{s}{i} \left(\frac{\binom{s-i}{k}}{\binom{N}{k}} \right)^t \frac{\binom{s}{k}}{\binom{N}{k}} \right]^{m-t}.$$

5 Properties of Elements

Suppose you have a set X of N elements, from which you select m k-sets independently and uniformly at random. In this section we will explore some probabilistic

properties of the elements after such set selections. Some of these conclusions will be used in subsequent results, while others remain independent of the applications and results presented in other sections and are included for the sake of completeness.

Lemma 2. *Consider an element $x \in X$. After m set selections, the probability that x is in exactly d of the m sets is $\binom{m}{d}\left(\frac{k}{N}\right)^d\left(\frac{N-k}{N}\right)^{m-d}$ and the expected number of sets containing x is $m \cdot \frac{k}{N}$.*

Proof. We view the m set selections as a series of m Bernoulli trials, where a success on the ith trial occurs when the set A_i contains the element x, and a failure when the set does not. Accordingly, the probability of success p is just $\frac{k}{N}$, and the probability of failure q is just $1-p$. Within this framework it is easy to see that the probability an element x is in exactly d of the m sets is just the probability of d successes in m Bernoulli trials, which corresponds to the dth binomial coefficient of $(p+q)^m$. Consequently

$$Pr[x \text{ is in exactly } d \text{ of the } m \text{ sets}] = \binom{m}{d}\left(\frac{k}{N}\right)^d\left(\frac{N-k}{N}\right)^{m-d}.$$

Similarly, we are able to calculate the expected number of sets out of m that contain the element x. We just use a common fact of Bernoulli trials; the expected number of successes in m trials is mp. □

Lemma 3. *After m set selections, the expected number of unselected elements of X remaining is $N\left(\frac{N-k}{N}\right)^m$.*

Proof. Using lemma 2 where x is an arbitrary element, we have

$$\begin{aligned} E[\text{num. unselected elements after } m \text{ selections}] &= N \cdot Pr[\text{element } x \text{ in no sets}] \\ &= N\left(\frac{N-k}{N}\right)^m \end{aligned}$$ □

Now instead of finding the expected value, we can find the probability distribution of number of selected elements.

Lemma 4. *After m set selections, the probability that exactly s elements total are selected is*

$$\binom{N}{s}\sum_{i=0}^{s-k}(-1)^i\binom{s}{i}\left(\frac{\binom{s-i}{k}}{\binom{N}{k}}\right)^m.$$

Proof. Consider a set S of s elements. The inclusion-exclusion principle implies

$$Pr[\text{exactly } S \text{ selected after } m \text{ time steps}] = \sum_{i=0}^{s-k}(-1)^i\binom{s}{i}\left(\frac{\binom{s-i}{k}}{\binom{N}{k}}\right)^m.$$

Along with the fact that there are $\binom{N}{s}$ ways of selecting the set S, we have

$$Pr[\text{exactly } s \text{ elts selected after } m \text{ timesteps}] = \binom{N}{s}\sum_{i=0}^{s-k}(-1)^i\binom{s}{i}\left(\frac{\binom{s-i}{k}}{\binom{N}{k}}\right)^m$$ □

Lemma 5. *Suppose we have selected $m-1$ sets of size k, and we are about to select the mth k-set. The expected number of previously selected elements also selected on the mth time step is $k \cdot \left(1 - (\frac{N-k}{n})^{m-1}\right)$.*

Proof. If we knew there were s total elements selected after the $m-1^{st}$ set selection, then we would expect the mth set to select $\frac{k}{N} \cdot s$ repeated elements. By linearity of expectations and lemma 3, we have

$$\begin{aligned} &E[\text{num repeated elts on } m^{th} \text{ selection}] \\ &\quad = \tfrac{k}{N} E[\text{num elements selected by } m-1^{st} \text{ timestep}] \\ &\quad = k \cdot \left(1 - (\tfrac{N-k}{n})^{m-1}\right) \end{aligned}$$

□

Corollary 2. *The expected number of new (as of yet unselected) elements selected by the mth set is $k\left(\frac{N-k}{n}\right)^{m-1}$.*

Although more complex, we can determine the probability distribution for the number of repeated elements, instead of the expected value. For a simple bound we can use Markov's inequality and obtain

$$Pr[\text{num repeated elts on } m^{th} \text{ selection} \geq c] \leq \frac{k}{c}\left(1 - (\frac{N-k}{N})^{m-1}\right)$$

However, using the probability distribution on total number of element selected (lemma 4) we can get an exact (more complex) formula.

Lemma 6. *The probability of having exactly c repeated elements on the mth set selection is*

$$\sum_{s \geq k} \frac{\binom{s}{c}\binom{N-s}{k-c}}{\binom{N}{k}} \cdot Pr[s \text{ elts selected after } m-1 \text{ selections}].$$

The proof follows from the definition of conditional probability. The following corollary is of particular interest.

Corollary 3. *The probability that the mth set has no unique element is just the probability that it selects all k elements from previously selected elements:*

$$Pr[m^{th} \text{ set repeated}] = \sum_{s \geq k} \frac{\binom{s}{k}}{\binom{N}{k}} \binom{N}{s} \sum_{i=0}^{s-k} (-1)^i \binom{s}{i} \left(\frac{\binom{s-i}{k}}{\binom{N}{k}}\right)^{m-1}.$$

As mentioned in the generalized birthday problem, we would like to determine the probability that an arbitrary set A_j has at least one unique element not selected by the other $m-1$ sets.

Lemma 7. *The probability that an arbitrary set A_j has at least 1 unique element after m set selections is*

$$\sum_{w=1}^{k} (-1)^{w+1} \binom{N}{w} \frac{\binom{N-w}{k-w}}{\binom{N}{k}} \left[\frac{\binom{N-w}{k}}{\binom{N}{k}}\right]^{m-1}.$$

Proof. Consider a set W of w elements. The probability that some set A_j has these w elements unique to it is

$$Pr[\text{elements } W \text{ are unique to set } A_j] = \frac{\binom{N-w}{k-w}}{\binom{N}{k}} \left[\frac{\binom{N-w}{k}}{\binom{N}{k}} \right]^{m-1}$$

The first multiplicand is the probability that set A_j selects the w unique elements, the second exponentiated multiplicand is the probability that the remaining $m-1$ sets never select any of these w unique elements. Therefore, using the inclusion-exclusion principle, the probability that set A_j has at least 1 unique element is

$$\sum_{w=1}^{k} (-1)^{w+1} \binom{N}{w} \frac{\binom{N-w}{k-w}}{\binom{N}{k}} \left[\frac{\binom{N-w}{k}}{\binom{N}{k}} \right]^{m-1} \qquad \square$$

For any element y, let $Atom_m(y)$ denote the elements in y's atom after m set selections. Consider the subset $X^r = \{x_1, x_2, \ldots x_r\} \subseteq X$. The following lemma states the probabilities of various placements of $x_1, x_2, \ldots, x_r$ into atoms.

Lemma 8. *For $X^r = \{x_1, x_2, \ldots, x_r\} \subseteq X$, $z \in X^r$ and $y \notin X^r$, we have*

(a) $Pr[x_1, x_2, \ldots, x_r \text{ in same atom}] = \left[\frac{\binom{N-r}{k-r} + \binom{N-r}{k}}{\binom{N}{k}} \right]^m$

(b) $Pr[x_1, \ldots, x_{r+d} \text{ in same atom} \mid x_1, \ldots, x_r \text{ in same atom}] = \left[\frac{\binom{N-(r+d)}{k} + \binom{N-(r+d)}{k-(r+d)}}{\binom{N-r}{k} + \binom{N-r}{k-r}} \right]^m$

(c) $Pr[\forall x_i \in X^r \,.\, x_i \notin Atom_m(y)] = 1 - \sum_{d=1}^{r} (-1)^{d-1} \binom{r}{d} \left[\frac{\binom{N-d-1}{k} + \binom{N-d-1}{k-d-1}}{\binom{N}{k}} \right]^m$

(d) $Pr[\forall x_i \in \overline{X^r} . \, x_i \notin Atom_m(z) \mid X^r \subseteq Atom_m(z)] = 1 - \sum_{d=1}^{N-r} (-1)^{d-1} \binom{N-r}{d} \left[\frac{\binom{N-(r+d)}{k} + \binom{N-(r+d)}{k-(r+d)}}{\binom{N-r}{k} + \binom{N-r}{k-r}} \right]^m$

Proof. Let P_i be the probability that the ith set splits $x_1, \ldots, x_r$ into different atoms, i.e. set A_i selects a proper subset of X^r, guaranteeing the selected and unselected elements to be in different atoms. Since the m sets are selected independently,

$$\begin{aligned}
Pr[x_1, x_2, \ldots, x_r \text{ in same atom}] &= Pr[\text{no set splits } x_1, x_2, \ldots, x_r] \\
&= Pr[\text{set } A_1 \text{ doesn't split them}] \cdots Pr[\text{set } A_m \text{ doesn't split them}] \\
&= \overline{P_1} \cdot \overline{P_2} \cdots \overline{P_m} = (\overline{P_i})^m \qquad \text{for an arbitrary set } A_i \\
&= (Pr[\text{set } A_i \text{ contains all of } x_1, \ldots, x_r \text{ or contains none of them}])^m \\
&= \left[\frac{\binom{N-r}{k-r} + \binom{N-r}{k}}{\binom{N}{k}} \right]^m
\end{aligned}$$

which we will denote by p_m^r. A similar technique is used to obtain the result of (b); the denominator counts the number of k-sets that do not split $x_1, \ldots, x_r$, whereas the numerator counts the number of k-sets that do not split $x_1, \ldots, x_{r+d}$. We denote the probability of part (b) by $p_m^{r,d}$. For (c), consider the probability that $x_1, x_2, \ldots, x_r$ are all in different atoms than y after m set selections:

$$
\begin{aligned}
Pr[\forall x_i \in X^r \,.\, x_i \notin Atom_m(y)] &= 1 - Pr[\exists x_i \in X^r \,.\, x_i \in Atom_m(y)] \\
&\overset{incl-excl}{=} 1 - \textstyle\sum_{d=1}^{r} (-1)^{d-1} Pr[\exists Y \subseteq X^r \,.\, |Y| = d \wedge Y \subseteq Atom_m(y)] \\
&= 1 - \textstyle\sum_{d=1}^{r} (-1)^{d-1} \binom{r}{d} p_m^{1,d} \\
&= 1 - \textstyle\sum_{d=1}^{r} (-1)^{d-1} \binom{r}{d} \left[\frac{\binom{N-(d+1)}{k} + \binom{N-(d+1)}{k-(d+1)}}{\binom{N-1}{k} + \binom{N-1}{k-1}} \right]^m \\
&= 1 - \textstyle\sum_{d=1}^{r} (-1)^{d-1} \binom{r}{d} \left[\frac{\binom{N-d-1}{k} + \binom{N-d-1}{k-d-1}}{\binom{N}{k}} \right]^m
\end{aligned}
$$

At last, part (d) considers the probability that $\overline{X^r} = \{x_{r+1}, x_{r+2}, \ldots, x_N\}$ are in different atoms than an element z after m set selections, given that the remaining elements $x_1, x_2, \ldots, x_r$ are in the same atom as z, for $z \in X^r$.

$$
\begin{aligned}
&Pr[\ \forall\ x_i \in \overline{X^r}.\ x_i \notin Atom_m(z) \mid X^r \subseteq Atom_m(z)] \\
&\quad = 1 - Pr[\exists x_i \in \overline{X^r} \,.\, x_i \in Atom_m(z) \mid X^r \subseteq Atom_m(z)] \\
&\quad = 1 - \sum_{d=1}^{N-r} (-1)^{d-1} Pr[\exists Y \subseteq \overline{X^r} \,.\, |Y| = d \wedge Y \subseteq Atom_m(z) \mid X^r \subseteq Atom_m(z)] \\
&\quad = 1 - \sum_{d=1}^{N-r} (-1)^{d-1} \binom{N-r}{d} p_m^{r,d} \\
&\quad = 1 - \sum_{d=1}^{N-r} (-1)^{d-1} \binom{N-r}{d} \left[\frac{\binom{N-(r+d)}{k} + \binom{N-(r+d)}{k-(r+d)}}{\binom{N-r}{k} + \binom{N-r}{k-r}} \right]^m
\end{aligned}
$$

which we denote by q_m^r. □

6 Properties of Atoms

Now we will use results of the previous section to investigate the structure of atoms created by the m selected sets. In particular, we consider the number of atoms and their sizes.

Lemma 9. *Consider an element $x \in X$. The probability that x's atom is of size exactly r after m set selections is*

$$
\binom{N-1}{r-1} \left[1 - \sum_{s=1}^{N-r} (-1)^{s-1} \binom{N-r}{s} \left(\frac{\binom{N-s-r}{k} + \binom{n-s-r}{k-s-r}}{\binom{N-r}{k} + \binom{N-r}{k-r}} \right)^m \right] \left[\frac{\binom{N-r}{k} + \binom{N-r}{k-r}}{\binom{N}{k}} \right]^m .
$$

Proof. First notice that

$$
Pr[|Atom_m(x)| = r] = \binom{N-1}{r-1} Pr[\text{exactly these } r-1 \text{ elements are in } x\text{'s atom}]
$$

where the choose term is the number of ways we can select the other $r-1$ elements of x's atom, and we multiply this by the probability that the $r-1$ elements in

question are indeed in x's atom. Without loss of generality, let $x = x_1$ and let the other $r-1$ elements selected for x's atoms be $x_2, \ldots, x_r$. Then the remaining $N-r$ elements not in x's atom are $x_{r+1}, \ldots, x_N$. The above equation becomes

$$\begin{aligned}
&\binom{N-1}{r-1} Pr[x_2, \ldots, x_r \in Atom_m(x) \wedge x_{r+1}, \ldots, x_N \notin Atom_m(x)] \\
&= \binom{N-1}{r-1} Pr[x_{r+1}, \ldots, x_N \notin Atom_m(x) \mid x_2, \ldots, x_r \in Atom_m(x)] \cdot \\
&\qquad Pr[x_2, \ldots, x_r \in Atom_m(x)] \\
&= \binom{N-1}{r-1} q_m^r \cdot p_m^r
\end{aligned}$$

where the first equality follows from the definition of conditional probability and the remaining follow from lemma 8. □

We now have the probability distribution of a specific element's atom size. From this we reconstruct the probability distribution of atom sizes.

Corollary 4. *After m set selections, the expected number of atoms of size r is $\frac{N}{r} Pr[|Atom_m(x)| = r]$.*

Proof. After m set selections, the expected number of elements in atoms of size r is just $N \cdot Pr[|Atom_m(x)| = r]$. Thus to obtain the expected number of atoms of size r, we need only divide the expected number of elements in atoms of size r by r. □

The expected number of atoms of a given size now allows us to derive a formula for the expected total number of atoms.

Lemma 10. *After m set selections, the expected number of atoms total is*

$$\sum_{r=1}^{N} \frac{N}{r} Pr[|Atom_m(x)| = r]$$

Proof. Using linearity of expectation

$$\begin{aligned}
E[\text{number of atoms after } m \text{ selections}] &= E[\sum_{r=1}^{N} \text{number of atoms of size } r] \\
&= \sum_{r=1}^{N} E[\text{number of atoms of size } r] \\
&= \sum_{r=1}^{N} \frac{N}{r} Pr[|Atom_m(x)| = r]
\end{aligned}$$

□

A quantity naturally desired is the number of atoms consisting of selected elements. For this purpose, recall there is one 'larger' atom containing precisely all the unselected elements so far, and the remaining 'smaller' atoms comprising solely of elements selected so far. If we assume that $N >> k$ (in particular, if $N > k(m+1)$) then there is exactly 1 large atom of size between $N-k$

and $N - km$ of unselected elements, and multiple smaller atoms of selected elements with sizes ranging between 1 and k. Thus armed with this assumption, we are guaranteed to have one large atom of unselected elements. In this case, the expected number of smaller atoms should just be one less than the expected number of total atoms. In particular,

$$E[\text{number of smaller atoms after } m \text{ selections}] = \sum_{r=1}^{N} \frac{N}{r} Pr[|Atom_m(x)| = r] - 1$$

7 Final Remarks

Naturally, many of the quantities discussed in this paper could use with improved, simplified, or exact bounds. Other than these endeavours, one direction for further work would be to look at different versions of the problems presented; other generalizations of the birthday problem could be formulated and analyzed in similar ways, and the digital rights problem could be considered when the devices are revoked one at a time, trying to avoid the undesirable behaviour at each time step, rather than only at the end of the revocation phase. Another direction to investigate would be to look at things from the boolean algebra view and try to formulate other probabilistic treatments of boolean algebra.

References

1. P. Erdos, P. Frankl, and Z. Furedi. Families of Finite Sets in Which No Set is Covered By the Union of r Others. *Israel Journal of Mathematics*, 51:79-89, 1985.
2. N. Alon. Probabilistic Methods in Extremal Finite Set Theory. *Proc. Conference on Extremal Problems for Finite Sets*, Bolyai Soc. Math. Stud. 3, pp.39-57, 1994.
3. A. D'yachkov, P. Vilenkin, A. Macula, and D. Torney. Families of finite sets in which no intersection of l sets is covered by the union of s others. *J. Combin. Theory*, 99:195-208, 2002.
4. W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed.* New York: Wiley, pp.31-33, 1968.
5. P. Diaconis and F. Mosteller. Methods for Studying Coincidences. *Journal of the American Statistical Association*, 84:853-861, 1989.
6. B. Chor, A. Fiat, M. Naor, and B. Pinkas. Tracing Traitors. *IEEE Transactions on Information Technology*, Vol.46, No.3, 2000.
7. D. Naor, M. Naor, and J. Lotspiech. Revocation and Tracing Schemes for Stateless Receivers. Advances in Cryptology - Crypto 2001, Lecture Notes in Computer Science, vol. 2139, pp.41-62, 2001.
8. xCP: eXtensible Content Protection. Broadcast Paper, IBM Research Division Almaden Research Center.
9. M. Stone. The Theory of Representations for Boolean Algebras. *Trans. Amer. Math. Soc.* 40 (1936), 37–111.