Small complete arcs in projective planes

J. H. Kim * V. H. Vu †

April 24, 2002

Abstract

In the late 1950's, B. Segre introduced the fundamental notion of arcs and complete arcs [48, 49]. An arc in a finite projective plane is a set of points with no three on a line and it is complete if cannot be extended without violating this property. Given a projective plane $\mathcal{P}$, determining $n(\mathcal{P})$, the size of its smallest complete arc, has been a major open question in finite geometry for several decades. Assume that $\mathcal{P}$ has order q, it was shown by Lunelli and Sce [41], more than 40 years ago, that $n(\mathcal{P}) \geq \sqrt{2q}$. Apart from this bound, practically nothing was known about $n(\mathcal{P})$, except for the case $\mathcal{P}$ is the Galois plane. For this case, the best upper bound, prior to this paper, was $O(q^{3/4})$ obtained by Szőnyi using the properties of the Galois field $GF(q)$.

In this paper, we prove that $n(\mathcal{P}) \leq \sqrt{q}\log^c q$ for any projective plane $\mathcal{P}$ of order q, where c is a universal constant. Together with Lunelli-Sce's lower bound, our result determines $n(\mathcal{P})$ up to a polylogarithmic factor. Our proof uses a probabilistic method known as the dynamic random contruction or Rödl's nibble. The proof also gives a quick randomized algorithm which produces a small complete arc with high probability.

The key ingredient of our proof is a new concentration result, which applies for non-Lipschitz functions and is of independent interest.

1 INTRODUCTION

A projective plane of order q consists of a set of q^2+q+1 points and a set of q^2+q+1 lines, where each line contains exactly $q+1$ points and two distinct points lie on exactly one line. It is easy to deduce from the definition that each point is contained in exactly $q+1$ lines and two distinct lines have exactly one common point. Finite projective planes are fundamental object in combinatorics and several related areas such as coding theory; for more details about projective planes we refer to [25] and [32]. Throughout the paper, we assume that our projective plane has order q.

The most important example of a projective plane is, perhaps, the Galois plane $PG(2,q)$, constructed as follows. Let V be the vector space of dimension three over the Galois Field $GF(q)$, where q is a prime power. The points and lines of the projective plane $PG(2,q)$ are the 1 and 2 dimensional subspaces of V, respectively, with the natural inclusion. For large q, there are many planes of order q not isomorphic to $PG(2,q)$. On the other hand, it is not known whether there exists a projective plane of order not a prime power.

*Microsoft Research, Microsoft Corporation, Redmond, WA 98052; jehkim@microsoft.com

†Microsoft Research, Microsoft Corporation, Redmond, WA 98052; vanhavu@microsoft.com.

This work was partially carried out while both authors were at AT&T Bell Laboratories.

In the late 1950's, B. Segre ([48, 49]) introduced the notions of arcs and complete arcs. An *arc* in a plane is a set of points with no three on a line and maximal arcs under the set inclusion are called *complete arcs*. A line containing two points of an arc is called a *secant*. By definition, an arc is complete if and only if its secants cover the whole plane. Segre [48] asked the following fundamental question

In a plane $\mathcal{P}$ of order q, how many points can a complete arc have ?

Since Segre's introduction, this question has become one of the main research topics in finite geometry. Especially, the problems of finding the maximum and minimum possible sizes of complete arcs and characterizing those complete arcs have attracted much attention.

For the maximum size, it is fairly trivial that an arc cannot have more than $q+2$ points (the reader may consider it an easy exercise). Segre himself proved that for odd q, an arc of $PG(2,q)$ has at most $q+1$ points, and the maximum is attained if and only if the arc is a conic, which is basically the set of points (x,y,z) satisfying $xz = y^2$. (Of course, with respect to the definition of $PG(2,q)$ given above, a point (x,y,z) actually means the one dimensional subspace generated by (x,y,z).) For even q, the maximum is $q+2$, but the characterization of all such arcs is still not completed [10]. The second largest cardinalities have been studied too (see e.g. [27, 28, 29, 10, 53] and references therein).

The other direction, the minimum possible size, seems to be more interesting from the combinatorial point of view since it is a mini-max question. Given a plane $\mathcal{P}$ of order q, we denote by $n(\mathcal{P})$ the size of a smallest complete arc in $\mathcal{P}$. The main goal of this paper is to give a nearly sharp estimate on $n(\mathcal{P})$. But before presenting our result, let us give a brief review of the long and rich history of this fundamental problem.

The first lower bound on $n(\mathcal{P})$ was obtained by Lunelli and Sce [41] about 40 years ago, shortly after the introduction of arcs. This bound is $(2q)^{1/2}$ and its proof is quite simple. Observe that the union of the secants of a complete arc should cover all q^2+q+1 points, and each secant covers $q+1$ points, it follows that a complete arc must have at least $(q^2+q+1)/(q+1) \geq q$ secants. To have q secants, the arc should have at least $(2q)^{1/2}$ points. Notice that this proof does not make use of the crucial property that an arc does not contain three co-linear points. Therefore, the bound is true for all sets whose secants cover the whole plane. Sets with this property are called *saturated* and have been studied by various authors [40, 7, 57].

The only improvements over Lunelli and Sce's bound in the last 40 years that we know of are the results of Blokhuis [9] and Ball [6], who improved it to $(3q)^{1/2}$ for $PG(2,q)$, where q is prime and the square of a prime, respectively. It is not known whether this improved bound holds for every Galois plane, although Blokhuis conjectured that this should be the case [11]. Fisher [21] (see also [11]) used computer simulations to generate complete arcs in many planes of small orders, and conjectured that the average size of a complete arc is about $(3q\log q)^{1/2}$. In relation to this conjecture, Szőnyi [56] proved that there exist certain planes, called André planes, of order q which contain complete arcs of size at most $cq^{1/2}\log^2 q$ for some constant c. Although this bound is quite close to the lower bound of Lunelli and Sce, Szőnyi's proof has a somewhat different nature. Instead of proving the existence of small complete arcs in a given plane, he constructed a plane together with a small complete arc in it.

The above mentioned results strongly support the conjecture that $n(\mathcal{P})$ should be fairly

close to $q^{1/2}$. Indeed, de Resmini ([11]) made a very precise conjecture for $PG(2,q)$: $n(PG(2,q)) = \lfloor 3q^{1/2} - 2 \rfloor$ if q is even and $n(PG(2,q)) = \lfloor 3q^{1/2} - 3 \rfloor$ if q is odd. Small complete arcs were also studied in relation with blocking sets and for more details we refer to [8, 13, 14, 15].

There have been a large number of attempts to construct small complete arcs in order to obtain a good upper bound for $n(\mathcal{P})$. Somewhat surprisingly, this problem is hard even at the very beginning. While an arc cannot have more than $q+2$ points, it is already difficult to construct a complete arc having ϵq points, for a small constant ϵ. Most of the attempts focus on the Galois plane $PG(2,q)$ and in the remaining of this paragraph and the next paragraph we restrict ourselves to this plane. The first important result was a construction by Abatangelo [1], which gives a complete arc of size roughly $q/3$. The main idea of this construction followed a suggestion of Segre and Lombardo-Radice [50, 42]: choose a special small set of points from a proper algebraic curve (which itself is an arc but of large size) and next show that the secants of the chosen set cover most of the points of the plane, then extend the set if necessary to be complete. This idea was also applied in other papers including [17, 19, 20, 67] to construct arcs of somewhat larger sizes.

To show that the secants cover almost the whole plane, Abatangelo applied a deep theorem of Weil in algebraic geometry, which estimates the number of solutions of equations of a certain type. Using a similar idea, Korchmáros [39] improved the bound to $q/4$. Bounds better than $\Omega(q)$ requires more sophisticated algebraic techniques which have been developed in a sequence of papers [58, 59, 67, 47, 51] (see also [53, 54, 55] for surveys). In [52], Szőnyi proved that $n(PG(2,q)) \leq cq^{3/4}$, where c is a constant not depend on q. Szőnyi's bound was the best upper bound known prior to this paper.

The main back-draft of the algebraic technique is that it can hardly be applied without the presence of the Galois field $GF(q)$. Consequently, very little has been known about complete arcs of planes other than $PG(2,q)$. For instance, for a general plane, no upper bound substantially better than $(q+1)/2$ (see [26]) had been proved prior to this paper. Taking into account the trivial fact that an arc cannot have more than $q+2$ points, this rather weak bound underlines the difficulty of the problem.

In this paper, we obtain a major improvement concerning the upper bound of $n(\mathcal{P})$ for a general plane. Our main theorem is the following

Theorem 1.1 *There are positive constants c and M such that the following holds. In every projective plane of order $q \geq M$, there is a complete arc of size at most $q^{1/2}\log^c q$.*

Theorem 1.1, together with Lunelli-Sce's lower bound, determines $n(\mathcal{P})$ up to a poly-logarithmic factor. As our arc is created in a random manner, the result also gives some support to Fisher's conjecture.

The proof of Theorem 1.1 brings a very useful by-product. To prove it, we have discovered a new and powerful concentration result (see Section 4), which, since the completion of this paper, has become a topic in itself. This result and its variants have found many applications in diverse areas, ranging from the theory of random graphs to additive number theory, leading to remarkable improvements in several old problems. For instance, a variant of this result plays an essential role in the solution of Nathanson's conjecture on the existence of thin Waring's basis [62]. The interested reader is referred to [60] for a recent survey.

In order to obtain Theorem 1.1, we actually prove a stronger result

Theorem 1.2 *There are absolute constants c and M such that in any projective plane of order $q \geq M$, one can find an arc with $\Theta(q^{1/2}\log^{1/2} q)$ points whose secants cover all but $q^{1/2}\log^{c} q$ points of the plane.*

Throughout the paper, we assume that q is sufficiently large, whenever needed. The asymptotic notations $\Theta, O, ...$, ect are used under the assumption that $q \to \infty$.

It is easy to derive Theorem 1.1 from Theorem 1.2. Let us remark here that our proof of Theorem 1.2 also provides an efficient randomized algorithm which produces the desired arc with probability close to 1. This algorithms runs in $polylog(q)$ steps, where each step consist of $O(q^4)$ basic operations. Theorem 1.2 also implies the following corollary.

Corollary 1.3 *There are positive constants c_1, c_2 and M such that in any projective plane of order $q \geq M$, one can find a complete arc of size between $\frac{1}{c_1} q^{1/2}\log^{1/2} q$ and $q^{1/2}\log^{c_2} q$.*

The proof of Theorem 1.2 uses the dynamic random contruction (or Rödl's nibble method), whose description is given in the next section. Section 3 presents our main lemma and the proof of Theorem 1.2 via this lemma. In Section 4, we provide the key tools to prove the main lemma. These includes the new concentration result mentioned earlier. The proof of the main lemma follows in Section 5. The last section is for several remarks and open questions.

2 DYNAMIC RANDOM CONSTRUCTION

2.1 The nibble method

As already mentioned, the desired arc in Theorem 1.2 is produced by a randomized algorithm with polynomial number of basic operations. One basic operation consists of either checking the incidence of a point and a line, or deleting a point from a line. This algorithm is based on the nibble method, a powerful and sophisticated tool from probabilistic combinatorics. In the following, we give a brief introduction to this method.

When one wants to construct an object with certain structural constrains such as packings, covers, graphs without certain small subgraphs and arcs in a plane, random greedy construction is considered a natural way to generate it: Randomly order all possible elements of the desired object and *select* each of them one by one in the order if and only if it together with already selected ones cause no conflict, i.e. no violation to the given constrains. Here we mean by "select" that we choose and permanently add it to the desired object being constructed. We may discard at each step all elements that cause any conflict with already selected ones and then randomly select a non-discarded one. This is an equivalent construction and will be called the random greedy construction (RGC) which stands for random greedy construction. For example, the RGC of a complete arc is the following. Initially, the arc being constructed is empty. At each step, discard all points contained in any secant of already selected points and select one non-discarded point uniformly at random. Then the set of all selected points is a complete arc. In many cases, it is believed that the RGC yields an almost optimal desired object. However, proving it seems to be

very hard. For the case of complete arcs, we believe that the resulting complete arc is of size at most $q^{1/2+\epsilon}$ but have no idea to prove it.

The nibble method, or dynamic random construction using nibble (DRC), is an approximated version of RGC. DRC has been initiated by a seminal paper of Ajtai, Komlós and Szemerédi [2] to construct a large independent set of a triangle-free graph and become well-known to combinatorialists by Rödl [46] who used the construction to settle the Erdős-Hanani conjecture regarding Steiner systems. It has been developed and become more sophisticated and powerful to solve intriguing combinatorial problems regarding packings and edge-colorings of hypergraphs or multigraphs ([44, 23, 33, 36, 4, 5, 24]), chromatic numbers of sparse graphs ([34, 30, 31, 64, 65]), Ramsey numbers ([35]), and some general graph coloring problems ([43, 45]).

Rather than selecting one element at each step, DRC randomly and independently chooses elements, not selected yet, with certain probability so that a bunch of elements are chosen together. This is called a nibble. The size of a nibble is the number of chosen elements or sometimes its expectation. Since the set of chosen elements may violate the constrains, we take a subset of it satisfying the constrains. Elements of this subset are called selected in the above meaning. Though the way constructing this subset varies depending on problems and/or for the sake of simplicity, chosen elements contribute no conflicts with previously and currently chosen ones are usually selected. We discard each unchosen element that may cause any new conflict if it were added to chosen elements regardless what the selected elements actually are. Since not all chosen elements are selected, some elements are unnecessarily discarded but the set of remaining, i.e. non-discarded unchosen, elements is defined with respect to randomly and independently chosen elements so that the structure of the set might be well-understood. For the next step, corresponding new constrains are to be imposed. (For the example of a complete arc, initial constrains are that no three points are in a line. After choosing a bunch of points, we must add a new constraint that no two points are in a line containing a selected point.)

If the size of a nibble is too big so that many of the chosen elements contribute at least one conflict, then it would be hard to predict the structure of selected elements and/or too many elements would unnecessarily be discarded. Thus the size needs to be small enough that most chosen elements do not cause any conflict. Consequently, only few elements would be unnecessarily discarded. For example, if we choose $\theta q^{1/2}$ random points from a plane of order q, a simple computation yields that each chosen point causes a conflict with probability at most $(q+1)\binom{q}{2}(\theta q^{-3/2})^2 \leq \theta^2$. Thus as long as θ is small enough, most chosen points do not cause any conflict. In the case that each nibble is of size one, DRC would exactly be RGC. In general, DRC is believed a good approximation of RGC as long as nibble sizes are small enough.

How to nibble a good arc. Our algorithm to construct a small complete arc of a given plane works roughly as follows. Initially Ω_0 and S_0 both denote the set of all points of the plane and A_0, which will be extended to the desired arc, is empty. Generally, Ω_i is essentially the set of points which are not covered by the secants of the current arc A_i, and S_i is a subset of Ω_i. At the i^{th} step, we choose a random subset B_i of S_i by picking each point in S_i with the same probability p_i, independently. Add an appropriate subset of B_i to A_i so that the new set, say A_{i+1}, is still an arc. Now Ω_{i+1} is obtained from Ω_i by deleting all the points covered by the secants of A_{i+1}. To obtain S_{i+1}, we delete from S_i not only

the points covered by the secants of A_{i+1}, but a few more points, chosen randomly. The purpose of this additional deletion is to keep certain structural properties of the S_i's.

We repeat the process until all but $q^{1/2}\log^c q$ points are covered by the secants of the current arc (see Theorem 1.2).

The implementation of this idea, nevertheless, turns out to be technical. The detailed algorithm follows in the next subsection.

2.2 The algorithm

Our input is a plane $\mathcal{P}$ of order q. Initially, Ω_0 and S_0 both consist of all points of the plane, and A_0 is empty. We will keep track of two running parameters a_i and b_i where $a_0 = 0$ and $b_0 = 1$. In general $a_i = |A_i|q^{-1/2}$ and b_i is roughly $|S_i|/|S_0|$. Set $\theta = \log^{-2} q$.

Provided that after the first i steps, Ω_i, S_i and A_i have been constructed with points covered by secants of A_i in neither Ω_i nor S_i, we execute the following three operations at Step $i+1$:

Choose: Choose each point v in S_i with probability $p_i = \theta(b_i q^{3/2})^{-1}$. Let B_i be the set of chosen points. A point x in B_i is "good" if it does not cause any conflict in $A_i \cup B_i$, i.e. no two other points of $A_i \cup B_i$ are co-linear with it. Since no points of S_i are covered by secants of A_i, x is good if and only if

- There are no $y \in B_i$ and $z \in A_i$ so that x, y, z are co-linear.
- There are no $y, z \in B_i$ so that x, y, z are co-linear.

Let M_i be the set of all "good" points and let the new arc be $A_{i+1} = A_i \cup M_i$.

Delete: Delete from Ω_i all the points covered by secants of A_{i+1} or in B_i. Since Ω_i has no points covered by secants of A_i, a point v in Ω_i is deleted if and only if one of the following events occurs:

- v is in B_i.
- There are $x \in M_i$ and $y \in A_i$ such that x, y, v are co-linear.
- There are x and y in M_i such that x, y, v are co-linear.

Let D_i be the set of all deleted points in this operation. For each $v \in \Omega_i$, denote $P_i(v) = Pr(v \in D_i)$ and let P_i^u and P_i^l be the maximum and minimum over $P_i(v)$'s, respectively (u and l stand for upper and lower bounds, respectively).

Compensate: To define S_{i+1}, we delete from S_i all points in D_i and independently remove each point v in S_i with probability

$$P_i^{com}(v) := (P_i^u - P_i(v))/(1 - P_i(v)).$$

Denote by C_i the set of removed points. For the next step set

$$\Omega_{i+1} = \Omega_i \backslash D_i,\ S_{i+1} = S_i \backslash (D_i \cup C_i),\ A_{i+1} = A_i \cup M_i,$$

and

$$a_{i+1} = |A_{i+1}|q^{-1/2},\ b_{i+1} = b_i(1 - P_i^u)\ .$$

Throughout the paper, we say that after step $i+1$, a point v is surviving if $v \in S_{i+1}$ and undeleted if $v \in \Omega_{i+1}$. Clearly, every surviving point is undeleted.

Stop: The algorithm stops after the completion of step N, where N is the first number such that $b_N \leq q^{-3/2}\log^c q$, for some constant c (later, we will set $c = 300$).

When the algorithm halts, we obtain an arc A_N. To prove Theorem 1.2, we shall show that this arc has $\Theta(q^{1/2}\log^{1/2} q)$ points and its secants cover all but $O(q^{1/2}\log^c q)$ points of the plane. Due to the algorithm, the set of points uncovered by the secants of A_N is a subset of $\Omega_N \bigcup(\cup_{i=1}^N B_i \backslash M_i)$.

To achieve our goal, we need to analyze the algorithm in two phases, depending on whether $b_i \geq q^{-1}\log^{c_1} q$ or not, where c_1 is some other constant significantly smaller than c (later we will set $c_1 = 100$). In each phase, we consider a number of parameters such as the number of undeleted points on a line, the size of the current arc etc., and prove that they behave essentially as their expectations predict. Technically speaking, we show that these parameters are strongly concentrated around their means. This is the main task in the proof and requires the new concentration result presented in Section 4. Once we could take control all these parameters, the desired properties of A_N follows via an elementary (but still not so obvious) calculation (see subsection 3.2 and 3.3).

The reader may notice that the algorithm follows closely the description of the dynamic random construction method given in the previous subsection. The only new move is the "Compensation" operation. It is clear from the algorithm that the probability that a point is deleted depends on its (geometrical) situation. Therefore, the deleting probability $P_i(v)$'s are not necessarily the same for all v's. On the other hand, our purpose is to make S_{i+1} a random-like subset of S_i, so we want that each point in S_i has the same chance to stay in S_{i+1}, or each point in S_i is thrown out with the same probability. The compensation probability $P_i^{com}(v)$ is introduced for this purpose.

2.3 Notation

Throughout the paper, $[xyz]$ (sometimes $[x, y, z]$ in order to avoid confusion caused by sub-indices) will mean that the three points x, y, z are co-linear. This notion will be used in summations, for instance, $\sum_{j,j',[xjj']} t_j t_{j'}$ is summing the products $t_j t_{j'}$'s over all pairs j, j' such that j, j' and x are co-linear. The unique line containing the two points x and y is denoted by (xy). For each point $v \in \Omega_i$, $A_i(v)$ is the set of surviving points (excluding v itself) in the lines connecting v with a point in A_i. Formally,

$$A_i(v) = \{x \in S_i \backslash v | \exists u \in A_i, [vxu]\} .$$

Initially, $A_0(v) = \emptyset$ for all v. In general, for any set of points $X = \{v_1, ..., v_k\}$, $A_i(X) = A_i(v_1, .., v_k) := \cap_{j=1}^k A_i(v_j)$. We may define the same things with respect to B_i; $B_i(v)$ is the set of surviving points on the lines connecting v with a point in B_i. The set $B_i(X)$ is similarly defined. Sometimes, we need to consider undeleted points instead of surviving points (Ω_i instead of S_i). We set

$$A_i'(v) := \{x \in \Omega_i \backslash v | \exists u \in A_i, [vxu]\} .$$

Next, we denote by $T_i(v)$ the set of (unordered) pairs of surviving points which are co-linear with v. Formally,

$$T_i(v) = \{\{x, y\} | x, y \in S_i \setminus v, [vxy]\} .$$

In the beginning, $T_0(v)$ contains $(q+1)\binom{q}{2}$ pairs for all v.

For a line l, $S_i(l)$ denotes the set of surviving points in l, i.e., $S_i(l) = S_i \cap l$. Whenever $S_i(l)$ is used, we assume that l is not a secant (otherwise $S(l)$ is empty). Furthermore, we set $S_i(l,v) = S_i(l) \cap A_i(v)$ and $S_i(l,u,v) = S_i(l) \cap A_i(u) \cap A_i(v)$ for a line l and any two points u, v. Similarly, set $\Omega_i(l) = \Omega_i \cap l$, $\Omega_i(l,v) = \Omega_i(l) \cap A'_i(v)$ and $\Omega_i(l,u,v) = \Omega_i(l) \cap A'_i(u) \cap A'_i(v)$.

For an event $\mathcal{A}$, $\mathbf{1}_{\mathcal{A}}$ denotes the characteristic indicator of $\mathcal{A}$: $\mathbf{1}_{\mathcal{A}} = 1$ if $\mathcal{A}$ holds and 0 otherwise. All logarithms have the natural base.

3 MAIN LEMMA

In this section, we describe our main lemma and prove Theorem 1.2 modulo this lemma. We start with the description of the main lemma. A patient reader will have a better understanding about the properties described in the main lemma, after reading the subsequent subsections.

3.1 Main Lemma

Our main lemma asserts that with probability close to 1, certain properties hold at every step of the algorithm. To analyze the algorithm, we need to split it into two phases, depending on the size of S_i. In each step of Phase 1, we need to consider three primary properties and seven secondary properties. In each step of Phase 2, we need to maintain five properties.

Before presenting these properties, let us recall that $\theta = \log^{-2} q$. Furthermore, let $b'_i = \prod_{j=0}^{i-1}(1 - P_j^l)$. Recall that $1 - P_j^l$ is an upper bound on the probability that a point from Ω_{j-1} remains in Ω_j, so b'_i can be interpreted as the upper bound for the chance that a fix point in Ω_0 remains in Ω_i. We set $c = 300$, $c_1 = 100$. These parameters are far from being best possible but we make no attempt to optimize them.

Phase 1. This phase consists of all steps where the parameter $b_i q$ of the input is at least $\log^{c_1} q$ (see the description of the algorithm). This roughly means that in this phase, each line has at least $\log^{c_1} q$ surviving points (see Property **(2)**). We want the following three primary and seven secondary properties hold for the outputs of any step in this phase.

Primary Properties.

(1) $\theta q^{1/2}(1 - o(1)) \le |M_i| \le \theta q^{1/2}(1 + o(1))$ and $|B_i| \le 2\theta q^{1/2}$.

(2) $b_{i+1} q(1 - (i+1)\log^{-13} q) \le |S_{i+1}(l)| \le b_{i+1} q(1 + (i+1)\log^{-13} q)$

(3) $|\Omega_{i+1}(l)| \le b'_{i+1} q(1 + (i+1)\log^{-13} q)$.

Secondary Properties

For all points $u, w, v, z \in \Omega_i$ and all lines l, where $l \cap \Omega_{i+1} \neq \emptyset$

(4) $|S_{i+1}(l,v)| \le 8(i+1)a_{i+1}b_{i+1}q^{1/2} + (i+1)\log^{40} q$

(5) $|S_{i+1}(l,u,v)| \le (i+1)\log^4 q$

(6) $|A_{i+1}(u,v)| \le (i+1)b_{i+1}q + (i+1)\log^{40} q$

(7) $|A_{i+1}(u,v,w)| \le ib_{i+1}q^{1/2} + (i+1)\log^{10} q$

(8) $|A_{i+1}(u,v,w,z)| \le (i+1)\log^6 q$

(9) $|\Omega_{i+1}(l,v)| \le 8(i+1)a_{i+1}b'_{i+1}q^{1/2} + (i+1)\log^{40} q$

(10) $|\Omega_{i+1}(l,u,v)| \le (i+1)\log^4 q$.

Taking into account the relations

$$\begin{aligned} |T_{i+1}(v)| &= \sum_{v\in l}\binom{|S_{i+1}(l)|-1}{2},\ |A_{i+1}(v)| = \sum_{l=(av),a\in A_{i+1}} (|S_{i+1}(l)|-1) \\ |S_{i+1}| &= \sum_{l}|S_{i+1}(l)|/(q+1),\ |\Omega_{i+1}| = \sum_{l}|\Omega_{i+1}(l)|/(q+1), \end{aligned}$$

Property **(2)** implies the following four properties **(11)** $\frac{1}{2}b_{i+1}^2q^3(1-3(i+1)\log^{-13}q) \le |T_{i+1}(v)| \le \frac{1}{2}b_{i+1}^2q^3(1+3(i+1)\log^{-13}q)$.

(12) $a_{i+1}b_{i+1}q^{3/2}(1-2(i+1)\log^{-13}q) \le |A_{i+1}(v)| \le a_{i+1}b_{i+1}q^{3/2}(1+2(i+1)\log^{-13}q)$

(13) $(1-\log^{-10}q)^{i+1}b_{i+1}q^2 \le |S_{i+1}| \le (1+\log^{-10}q)^{i+1}b_{i+1}q^2$

(14) $|\Omega_{i+1}| \le (1+\log^{-10}q)^{i+1}q^2b'_{i+1}$.

In the second phase, we want to keep these four properties (with a somewhat different error terms for **(12)**) together with Property **(1)**.

Phase 2. In this phase, we consider all steps $i+1$ whose input satisfies

$$q^{-3/2}\log^c q \le b_i \le q^{-1}\log^{c_1} q.$$

By Property **(2)** of the first phase, the second phase starts when each line has roughly $\log^{c_1} q$. The properties we are interested in are

(1) $\theta q^{1/2}(1-o(1)) \le |M_i| \le \theta q^{1/2}$ and $|B_i| \le 2\theta q^{1/2}$.

(2) $\frac{1}{2}b_{i+1}^2q^3(1-3(i+1)\log^{-13}q)) \le |T_{i+1}(v)| \le \frac{1}{2}b_{i+1}^2q^3(1+3(i+1)\log^{-13}q)$

(3) $a_{i+1}b_{i+1}q^{3/2}(1-O(i\theta^2)) \le |A_{i+1}(v)| \le a_{i+1}b_{i+1}q^{3/2}(1+O(i\theta^2))$

(4) $(1-\log^{-10}q)^{i+1}b_{i+1}q^2 \le |S_{i+1}| \le (1+\log^{-10}q)^{i+1}b_{i+1}q^2$

(5) $|\Omega_{i+1}| \le (1+\log^{-10}q)^{i+1}q^2b'_{i+1}$

for all points $v \in \Omega_{i+1}$.

We say that our algorithm runs *perfectly* up to a step j if the required properties hold for the outputs of steps $1,2,\ldots,j$. If all properties hold at every step in both phases, we say that the algorithm runs *entirely perfectly* .

The following remark may give the reader a better understanding of the above properties.

Remark.

• Before the $(i+1)^{th}$ step, S_i has roughly b_iq^2 points. In this step, each point from S_i is selected into B_i with probability $p_i = \theta(b_iq^{3/2})^{-1}$, therefore B_i has roughly $\theta q^{1/2}$ points. Most of the points of B_i remain in M_i as the number of points causing conflicts is small. Thus, one can expect M_i also to have around $\theta q^{1/2}$ points. This is quantified in the fist property of both phases.

• The main terms in the left and right hand sides of Property **(2)** of Phase 1 and Properties **(2-4)** of Phase 2 are exactly the expectations of the corresponding quantities. Thus, these properties state that the quantities in question are strongly concentrated around their means.

• As the reader will see in the next subsections, the proof of Theorem 1.2 (assuming the Main Lemma) does not require all properties, but only few of them. For instance, none of the secondary properties will be used. However, the primary properties, by themselves, cannot be proved using induction. The secondary properties are thus introduced so that

together with the primary properties they form a sufficiently strong induction hypothesis which we are able to prove. The reader will have a clearer picture about this point and the relation between the properties after reading the next two subsections, especially Remark 3.10.

Lemma 3.1 *(Main Lemma) Our algorithm runs entirely perfectly with probability close to 1.*

In the next two subsections, we show that if the algorithm runs entirely perfectly, then the final arc A_N satisfies the statement of Theorem 1.2.

3.2 Some other lemmas

In this subsection, we present some estimates which are needed in the proof of Theorem 1.2. The key fact we want to deduce is the following: If the algorithm runs entirely perfectly, then the gap $P_i^u - P_i^l$ is sufficiently small at every step. As $|\Omega_0|\prod_{i=0}^{N-1}(1-P_i^l)$ is (essentially) an upper bound for $|\Omega_N|$ and $|\Omega_0|\prod_{i=0}^{N-1}(1-P_i^u)$ is (essentially) a lower bound for $|S_N|$, appropriate bounds on $P_i^u - P_i^l$'s imply that $|\Omega_N|/|S_N| = O(1)$. On the other hand, by the description of the stopping time and Property **(4)** of the second phase, $|S_N| = O(q^{1/2}\log^c q)$. So, this way we can obtain the desired bound $O(q^{1/2}\log^c q)$ for $|\Omega_N|$.

The first two lemmas provide a upper bound on P_i^u and a lower bound on P_i^l, respectively. These bounds also explain the necessity of the estimates on $|A_i(v)|$ and $|T_i(v)|$ in the main lemma.

Lemma 3.2 *We have*

$$P_i^u \le p_i + p_i \max_v |A_i(v)| + p_i^2 \max_v |T_i(v)|,$$

where the maximum is taken over the set Ω_i.

Proof. By the description of M_i and the deletion operation, it is clear that for all $v \in \Omega_i$:

$$P_i(v) \le Pr(v \in B_i) + Pr(B_i \cap A_i(v) \neq \emptyset) + Pr(\exists l \text{ containing } v \text{ s.t. } |B_i \cap (l \backslash v)| \ge 2).$$

Let P_1, P_2, P_3 denote the first, second and third terms of the right hand side, respectively. It is obvious that $P_1 \le p_i$ (the strict inequality is possible as v may not be in S_i). To show $P_2 \le p_i \max|A_i(v)|$, consider

$$\begin{aligned} P_2 &= 1 - Pr(B_i \cap A_i(v) = \emptyset) \\ &= 1 - (1-p_i)^{A_i(v)} \\ &\le 1 - (1 - p_i|A_i(v)|) \\ &= p_i|A_i(v)|. \end{aligned}$$

To bound P_3, note that

$$P_3 \le p_i^2 \sum_{v \in l} \binom{|l\backslash v|}{2} \le p_i^2 \max_v |T_i(v)|.$$

This completes the proof. □

Lemma 3.3 *We have*

$$P_i^l \geq p_i \min_v |A_i(v)| - 2p_i^2 \max_v |A_i(v)|^2 - p_i^3 \max_v |A_i(v)| \max_v |T_i(v)|,$$

where the maximum and minimum are taken over the set Ω_i.

Proof. Notice that

$$P_i(v) \geq Pr(M_i \cap A_i(v) \neq \emptyset) = 1 - Pr(M_i \cap A_i(v) = \emptyset).$$

In order to upper bound $Pr(M_i \cap A_i(v) = \emptyset)$, consider

$$\begin{aligned} Pr(M_i \cap A_i(v) = \emptyset) &= Pr(B_i \cap A_i(v) = \emptyset) + Pr(\{B_i \cap A_i(v) \neq \emptyset\} \wedge \{M_i \cap A_i(v) = \emptyset\}) \\ &= P_4 + P_5, \end{aligned}$$

where P_4 and P_5 are, respectively, the first and second terms of the right hand side. Repeating the computation in the previous proof we have

$$\begin{aligned} P_4 = (1-p_i)^{|A_i(v)|} &\leq 1 - p_i |A_i(v)| + p_i^2 \binom{|A_i(v)|}{2} \\ &\leq 1 - p_i \min_v |A_i(v)| + p_i^2 \max_v \binom{|A_i(v)|}{2}, \end{aligned}$$

and

$$\begin{aligned} P_5 &\leq \sum_{x \in A_i(v)} Pr(\{x \in B_i\} \wedge \{x \notin M_i\}) \\ &\leq |A_i(v)| \max_{x \in A_i(v)} Pr(\{x \in B_i\} \wedge \{x \notin M_i\}) \\ &\leq |A_i(v)| p_i \max_{x \in A_i(v)} \Big(Pr(B_i \cap A_i(x) \neq \emptyset) + Pr(\exists l \text{ containing } x \text{ s.t. } |B_i \cap (l \backslash x)| \geq 2) \Big) \\ &\leq |A_i(v)| p_i \max_{x \in A_i(v)} \Big(p_i |A_i(x)| + p_i^2 |T_i(x)| \Big). \end{aligned}$$

Lemma 3.3 now follows from the bounds of P_4 and P_5. The constant 2 can be replaced by $3/2$ but we do not bother. □

Remark 3.4 *It will be useful to keep in mind the following asymptotics (under the assumption that the algorithm runs perfectly up to the* i^{th} *step).*

$$\begin{aligned} p_i |A_i(v)| &\sim \theta (b_i q^{3/2})^{-1} a_i b_i q^{3/2} = \theta a_i \sim i\theta^2 \\ p_i |A_i(v)|^2 &\sim i^2 \theta^5 \\ p_i^3 |A_i(v)||T_i(v)| &\sim \frac{1}{2} \theta^3 a_i \sim \frac{1}{2} i \theta^4 \\ p_i^2 |T_i(v)| &\sim \frac{1}{2} \theta^2. \end{aligned}$$

Here $A \sim B$ *means* A/B *is almost one. Usually, we only need* $0.9 \leq A/B \leq 1.1$ *for the computation. In the first asymptotic we used* $p_i = \theta (b_i q^{3/2})^{-1}$.

The next lemma gives an estimate on the running time N, assuming that the algorithm runs entirely perfectly .

Lemma 3.5 *If the algorithm runs entirely perfectly , then the running time* $N = \Theta(\theta^{-1}\log^{1/2} q)$.

Proof. By the definition of N the following holds:

$$\prod_{i=0}^{N-1}(1-P_i^u) = b_N \le q^{-3/2}\log^c q \le b_{N-1} = \prod_{i=0}^{N-2}(1-P_i^u).$$

Now assume, for contradiction, that $N \ge L = 10\theta^{-1}\log^{1/2} q$ (we can assume L is integer, for the sake of convenience). It follows that

$$b_L \ge b_{N-1} \ge q^{-3/2}\log^c q. \tag{1}$$

On the other hand,

$$b_L = \prod_{i=1}^{L-1}(1-P_i^u) \le \prod_{i=1}^{L-1}(1-P_i^l).$$

Taking logarithmic, it follows from (1) that

$$-(\frac{3}{2}+o(1))\log q \le \sum_{i=1}^{L-1}\log(1-P_i^l).$$

To estimate the right hand side we use Lemma 3.3. It is clear that if the algorithm runs entirely perfectly, then the term $p_i|A_i(v)|$ in the lower bound of P_i^l (in Lemma 3.3) is the dominating one. Since $i \le L$, $i\theta^2 = o(1)$ (recall that $\theta = \log^{-2} q$). Thus $P_i^l \ge \frac{1}{2}i\theta^2$ by Remark 3.4, and we have

$$\log(1-P_i^l) \le \log(1-\frac{1}{2}i\theta^2) \le -\frac{1}{4}i\theta^2.$$

This implies

$$\sum_{i=1}^{L-1}\log(1-P_i^l) \le -\frac{1}{4}\sum_{i=1}^{L-1} i\theta^2 \le -\frac{1}{10}L^2\theta^2 \le -2\log q,$$

a contradiction. Therefore, $N \le L = 10\theta^{-1}\log^{1/2} q$.

Using a similar argument together with the estimate on P_i^u, we can show that $N \ge \frac{1}{10}\theta^{-1}\log^{1/2} q$ and this finishes the proof. □

Remark 3.6 *The lower bound on N is not so crucial as we do not need a lower bound on the size of the arc in Theorem 1.1. On the other hand, this bound helps us to derive Corollary 1.3.*

It follows directly from the previous lemma and the bound on $|M_i|$ (Properties **(1)** in both phases) that

Corollary 3.7 *If the algorithm runs entirely perfectly , then* $|A_N| = \Theta(q^{1/2}\log^{1/2} q)$.

Now we are ready to deduce the key fact that $P_i^u - P_i^l$ is sufficiently small for all i.

Lemma 3.8 *If the algorithm runs entirely perfectly, then for every i*

$$P_i^u - P_i^l = O(\theta^2 \log q).$$

Proof. Lemmas 3.2 and 3.3 imply:

$$\begin{aligned} P_i^u - P_i^l &\leq p_i(1 + \max_v |A_i(v)| - \min_v |A_i(v)|) + p_i^2 \max_v |T_i(v)| \\ &+ 2p_i^2 \max_v |A_i(v)|^2 + p_i^3 \max_v |A_i(v)| \max_v |T_i(v)|. \end{aligned} \tag{2}$$

Recall that $p_i = \theta(b_i q^{3/2})^{-1}$. Using the estimate of $|T_i(v)|$ in the main lemma, $|T_i(v)| \sim \frac{1}{2} b_i^2 q^3$, we have that $p_i^2 \max_v |T_i(v)| = O(\theta^2)$. Next, using $|A_i(v)| \sim a_i b_i q^{3/2}$, we have $p_i^3 \max_v |A_i(v)| \max_v |T_i(v)| = O(a_i \theta^3) = o(\theta^2 \log q)$, as $a_i = O(\log^{1/2} q)$ by Corollary 3.7 (recall that $a_i = |A_i|/q^{1/2} \leq |A_N|/q^{1/2}$). Moreover, $p_i^2 \max_v |A_i(v)|^2 \sim a_i^2 \theta^2 = O(\theta^2 \log q)$. Thus, it remains to show that $p_i(\max_v |A_i(v)| - \min_v |A_i(v)|) = O(\theta^2 \log q)$.

Due to the estimates in the main lemma, the quantity $p_i(\max_v |A_i(v)| - \min_v |A_i(v)|)$ is larger when we are in the second phase (in this phase the error term concerning $|A_i(v)|$ is larger than in the first phase). In the second phase, due to property **(3)**,

$$(\max_v |A_i(v)| - \min_v |A_i(v)|) = O(i\theta^2 a_i b_i q^{3/2}).$$

As $p_i = \theta(b_i q^{3/2})^{-1}$, we have

$$p_i(\max_v |A_i(v)| - \min_v |A_i(v)|) = O(i a_i \theta^3). \tag{3}$$

On the other hand, $a_i \leq a_N = \Theta(\log^{1/2} q)$ and $i \leq N = O(\theta^{-1}) \log^{1/2} q$. These, together with (3), imply the lemma. □

3.3 Proof of Theorem 1.2 via the Main Lemma

We shows that if the algorithm runs entirely perfectly, then the output satisfies the following three estimates: $|A_N| = \Theta(q^{1/2} \log^{1/2} q)$, $|\Omega_N| = O(q^{1/2} \log^c q)$, and $|\cup_{i=1}^N B_i \backslash M_i| = O(q^{1/2} \log q)$. As noted in subsection 3.1, the set of uncovered points at the end of the algorithm is a subset of $\Omega_N \cup \left(\cup_{i=1}^N B_i \backslash M_i \right)$, so these estimates imply the theorem.

The first estimate has already been derived in Corollary 3.7. To prove the second estimate, note that by the definition of the stopping time and Property **(4)** of the second phase, $|S_N| \leq q^{1/2} \log^c q$. Thus, it is sufficient to show that $|\Omega_N|/|S_N| = O(1)$. In fact, more will be true, namely, $|\Omega_N|/|S_N| = 1 + o(1)$. To see this, first observe that by Property **(14)** of the first phase and Property **(5)** of the second phase,

$$|\Omega_N| \leq (1 + \log^{-10} q)^N b'_N q^2 = (1 + o(1)) b'_N q^2 = (1 + o(1)) \prod_{i=0}^{N-1} (1 - P_i^l) q^2.$$

On the other hand, $|S_N| = (1 + o(1)) b_N q^2 = (1 + o(1)) \prod_{i=0}^{N-1} (1 - P_i^u) q^2$. Therefore

$$|\Omega_N|/|S_N| = (1+o(1))\Big(\frac{\prod_{i=0}^{N-1}(1-P_i^l)}{b_N}\Big) = (1+o(1))\Big(\prod_{i=0}^{N-1}\frac{1-P_i^l}{1-P_i^u}\Big). \tag{4}$$

On the other hand, due to Lemma 3.8

$$\log\frac{1-P_i^l}{1-P_i^u} = O(P_i^u - P_i^l) = O(\theta^2\log q).$$

To conclude, recall $\theta = \log^{-2} q$ and by Lemma 3.5 $N = \Theta(\theta^{-1}\log^{1/2} q)$, so

$$\log\frac{|\Omega_N|}{|S_N|} = O(N\theta^2\log q) = o(1),$$

proving our claim.

The last estimate is a trivial consequence of the upper bound on N and the upper bound on $|B_i|$ provided in the second half of Property **(1)** in both phases (see subsection 3.1). This completes our proof. □

Remark 3.9 *We will use the following (under the hypothesis that the algorithm runs perfectly up to the i^{th} step) frequently in later proofs*

$$\begin{aligned} a_i\theta &\sim i\theta^2 \le N\theta^2 = O(\log^{1/2} q\theta) = o(1) \\ a_i &\sim i\theta = O(\log^{1/2} q) = o(\log q). \end{aligned}$$

Notice that if the algorithm runs perfectly up to the i^{th} step then $P_i^u = o(1)$. On the other hand, $b_{i+1} = b_i(1-P_i^u)$. So under the above assumption $b_{i+1} \sim b_i \ge 0.9b_i$.

Remark 3.10 *We have seen that the quantities $|A_i(v)|, |T_i(v)|, |M_i|, |S_i|$ and $|\Omega_i|$ are directly involved in the proof of Theorem 1.2 above. This explains the necessity of controlling these quantities in Main Lemma. In order to control $|A_i(v)|, |T_i(v)|$ and $|S_i(v)|$ in the first phase, it is sufficient to control $|S_i(l)|$ for all l (Property* **(2)***). The later is possible since in this phase $|S_i(l)|$ is sufficiently large ($|S_i(l)| \ge \log^{c_1} q$) which enables us to show that $|S_i(L)|$'s (as random variables) are strongly concentrated. Properties* **(4)** *and* **(5)** *of Phase 1 were introduced in order to help us to control $|S_i(l)|$ (see the first paragraph of subsection 5.1). In the second phase, $|S_i(l)|$ can be very small and strong concentration no longer holds. Therefore, we need to consider $|A_i(v)|, |T_i(v)|$ and $|S_i(v)|$ directly. In order to control $|A_i(v)|$ in this phase, we have to prepare in the first phase by introducing Property* **(6)***. Properties* **(7)** *and* **(8)** *were introduced to help us to prove* **(6)***. Finally, similar to the situation with $|S_i|$, in order to control $|\Omega_i|$ in the first phase, it is sufficient to control $|\Omega_i(l)|$ (Property* **(3)***) and Properties* **(9)** *and* **(10)** *were introduced to make this possible).*

4 CONCENTRATION

The heart of many proofs using the probabilistic method is to show that certain random variables are strongly concentrated around their means. To carry out such a task, one frequently uses a concentration (i.e., large deviation) result from probability theory, such as Csernoff's, Azuma's or Talagrand's inequalities (see [3] for many examples).

While concentration is also the main issue in our proof, none of the existed tools seems to be sufficiently strong to prove the properties in the main lemma. The following two paragraphs are intended to give the reader some idea about the main obstacle in our situation. Many known results have been developed for *smooth* functions, i.e., functions where each individual atom random variable has relatively small effect (or in other words, the Lipschitz coefficient is small). To be more concrete, let Y be a function depending on n variables $t_1, \ldots, t_n$, where the t_i's are independent binary random variables. The (discrete) Lipschitz coefficient of Y is the smallest number r such that whenever two n-dimensional binary vectors t and t' differ at only one coordinate, $|Y(t) - Y(t')| \leq r$. If r is sufficiently small, compared to n and the mean of Y, then Y is strongly concentrated with variance of order at most $r^2 n$. A typical example is Theorems 7.2.1 of [3], which is a variant of Azuma's inequality. This theorem is, perhaps, one of the most commonly used concentration results in probabilistic combinatorics.

Unfortunately, small Lipschitz coefficient is something we cannot afford in our situation. To illustrate this, let us consider the quantity $|M_1|$. (See the first property in Phase 1, subsection 3.1.) To each point $x \in S_0$, let $t_x = 1$ if x is chosen in B_i and 0 otherwise. Thus, $|M_1|$ is a random variable depending on the t_x's, $x \in S_0$. We show that the Lipschitz coefficient of M_1 can be as large as $\Omega(q)$ in the worst case. To see this, assume our plane is the Galois plane $PG(2, q)$ and let $\mathcal{C}$ be the conic $xy = z^2$ in it. It is well-known (and easy to see) that $\mathcal{C}$ is an arc. For a point $v \in S_i$ let $l_1, \ldots, l_{q+1}$ be the lines through it; about half of these lines intersect $\mathcal{C}$ in exactly two points. We denote these pair of points by $(x_1, y_1), \ldots, (x_K, y_K)$ where $K \approx q/2$. Imagine that in the "Choose" operation we have considered all points but v and among the considered points, we have chosen all x_i, y_i $(i = 1, \ldots, K)$ and nothing else. Then the choice of v has a huge effect on $|M_1|$. If v is chosen, then it spoils the whole configuration and thus $|M_1| = 0$; if v is not chosen then $M_1 = \{x_1, y_1, \ldots, x_K, y_K\}$ and $|M_1| \approx q$.

Roughly speaking, this kind of obstacle may be overcome when the sum of squares of Lipschitz coefficients are not too large. Kahn [33] showed that if all Lipschitz coefficients are not too large, then a strong concentration hold with variance essentially at most the sum instead of $r^2 n$ (the square of the maximum Lipschitz coefficient multiplied by the number of basic random variables). Alon, Kim and Spencer [5] considered similar cases in slightly different point view and found a so-called dynamic version of Kahn's result, which may also be regarded as a handy version. They applied it to find an almost optimal matching in a simple hypergraph, in particular to improve Brouwer's lower bound [12] for the size of a largest packing in a Steiner triple system. Grable [24] found a nice handy version too and applied it to find an almost optimal matching for a hypergraph with certain conditions. It was Kahn and the first author (see [34]) who considered the case that the maximum coefficient is too large but all coefficients are still small enough if one excludes a bad event, usually, of small probability. They showed that the over all maximum in Azuma's bound can be replaced by the maximum over only the complement of the bad event, which is called an essential maximum, as long as we add the probability of the bad event in the final concentration inequality. (See Lemma 4.1 for more details.)

In this paper, two new concentration inequalities are presented. The first one will cover both of the two cases mentioned above. It roughly says that upper bounds for the maximum and the sum excluding a bad event may replace the over all maximum multiplied by the number of basic random variables. Notice that the maximum Lipschitz coefficient in

Azuma's inequality may be regarded as the maximum over the supremum, or l_∞-, norms of all first order derivatives. The second inequality basically states that if all higher order derivatives are included, then their expectations, or l_1-norms, give enough information to obtain a concentration inequality which is almost as strong as Azuma's inequality. We believe that the method developed based on these concentration results is very systematic and robust. It has turned out later that this method can be applied to analyze the nibble process in several other problems in a convenient way, leading to notable improvements (see [61, 65]). For instance, [61] contains an extension of above mentioned Alon-Kim-Spencer's and Grable's results on matching of hypergraphs. Several variants of these concentration results are proved in consequent papers [60, 63] and applications have been found in diverse areas, ranging from additive number theory [62] to random graphs [37, 38, 66]. The reader who is interested can find a comprehensive account about these developments in a recent survey [60].

In the first two subsections of this section, we present our concentration results (Lemma 4.1 and Lemma 4.2). The proofs of these lemmas appeared in a separate paper [37], which was originally intended as an appendix to this paper.

4.1 Martingale

In this subsection we consider a probability space generated by n independent binary random variables $t_1, ..., t_n$, equipped with the product measure. To this end, p_i denotes the expectation of t_i. The asymptotic notation is used under the assumption that $n \to \infty$.

Let Y be a function depending on $t_1, ..., t_n$. For any vector $v = (t_1, \ldots, t_n)$ and any $1 \leq i \leq n$, define $C_i(v)$ as follows. First let $v^{(1)}$ and $v^{(0)}$ be the vector obtained from v by setting its i^{th} coordinate to 1 and 0, respectively, and

$$C_i(v) = \Big|\mathbb{E}\Big(Y(v^{(1)}) - Y(v^{(0)})|t_1, \ldots, t_{i-1}\Big)\Big|.$$

We call $C_i(v)$ the *(conditioned) average effect* of the random variable t_i when $t_1, .., t_{i-1}$ are given. By definition, C_i depends on $t_1, .., t_{i-1}$ and $p'_j s, j > i$. The corresponding *(conditioned) variance bound* $p_i C_i^2$ also plays an important role. The variance bound is exactly the variance of the random variable X which attains two values 0 and C_i with $Pr[X = C_i] = p_i$. As mentioned earlier, we cannot apply a classical Azuma type concentration inequality. To overcome this obstacle, a bad event of small probability must be excluded. As we want that all average effect $C_i(v)$ and the sum of the variance bounds are small enough, define a bad event with respect $\mathbf{C}$ and $\mathbf{V}$

$$\mathbb{B}_0 = \mathbb{B}_0(\mathbf{C}, \mathbf{V}) = \{v | \max_i C_i(v) \geq \mathbf{C} \text{ or } \sum_i p_i C_i^2(v) \geq \mathbf{V}\}. \tag{5}$$

It is sometimes convenient to consider

$$\mathbb{B}_1 = \mathbb{B}_1(\mathbf{C}, \mathbf{V}) = \{v | \max_i C_i(v) \geq \mathbf{C} \text{ or } \sum_i p_i C_i(v) \geq \mathbf{V}/\mathbf{C}\}.$$

For $\sum_i p_i C_i^2(v) \leq \max_i C(v) \sum_i p_i C_i(v)$, we have $\mathbb{B}_0 \subseteq \mathbb{B}_1$.

Lemma 4.1 *For any positive numbers $\lambda, \mathbf{C}$ and $\mathbf{V}$ satisfying $0 \leq \lambda \leq \mathbf{V}/\mathbf{C}^2$,*

$$Pr\Big(|Y - \mathbb{E}(Y)| \geq (\lambda \mathbf{V})^{1/2}\Big) \leq 2e^{-\lambda/4} + Pr(\mathbb{B}_0).$$

In particular,

$$Pr\Big(|Y - \mathbb{E}(Y)| \geq (\lambda \mathbf{V})^{1/2}\Big) \leq 2e^{-\lambda/4} + Pr(\mathbb{B}_1).$$

In order to apply this lemma, in subsection 4.3 and 4.4 we shall show how to define the decisive parameters $\mathbf{C}$ and $\mathbf{V}$ for our situation. For an application of Lemma 4.1 in the theorem of random graphs, we refer to [66]. Variants of this lemma can be found in [60, 38] along with more comprehensive discussion.

4.2 Concentration of polynomials

Let H be a hypergraph with the vertex set $\mathcal{V}(H) = \{1, 2, ..., n\}$. We allow H to have an empty edge. To this end $\mathcal{E}(H)$ denotes the edge set of H. Each edge e has some at most k vertices. Furthermore, we assign to each e a positive weight $w(e)$. Suppose t_i, $i = 1, , 2..., n$ are independent random variables, where t_i is either a binary random variable with expected value p_i or $t_i = p_i$ with probability 1. We consider the following function

$$Y_H = \sum_{e \in \mathcal{E}(H)} w(e) \prod_{s \in e} t_s$$

We call H the *supporting hypergraph* of $Y = Y_H$. Notice that Y is a polynomial of degree at most k. If e is the empty set, then we set $\prod_{s \in e} t_s = 1$.
Example. If $V(H) = \{1, 2, 3\}$ and $\mathcal{E}(H) = \{\{1, 2\}, \{3\}, \emptyset\}$ with weights $2, 0.2, 1$, respectively then:

$$Y_H = 2t_1t_2 + 0.2t_3 + 1$$

Truncated subhypergraphs. For each (non-empty) subset A of $V(H)$, H_A (the A-truncated subhypergraph of H) is defined as follows:

$\mathcal{V}(H_A) = \mathcal{V}(H) \backslash A$.
$\mathcal{E}(H_A) = \{B \subset \mathcal{V}(H_A), B \cup A \in \mathcal{E}(H)$.
If $B \in \mathcal{E}(H_A)$ then $w(B) = w(B \cup A)$.

Now let $\mathbb{E}_i(Y) = \max_{A \subset V(H), |A| = i} \mathbb{E}(Y_{H_A})$; by definition $\mathbb{E}_0(Y)$ is the expectation $\mathbb{E}(Y)$ of Y. Intuitively, $\mathbb{E}_i(Y)$ can be interpreted as the expected effect of a group of i random variables. Furthermore, set $E = \max_{i \geq 0} \mathbb{E}_i(Y)$ and $E' = \max_{i \geq 1} \mathbb{E}_i(Y)$.

Lemma 4.2 *There exist positive numbers c_k, d_k depending only on k so that for any positive number λ*

$$Pr\Big(|Y - \mathbb{E}(Y)| \geq c_k (EE')^{1/2} \lambda^k\Big) \leq d_k \exp(-\lambda + k \log n)$$

Corollary 4.3 *Assume that $k \leq 5$. Under the assumptions of Lemma 4.2, we have*

$$Pr\Big(|Y - \mathbb{E}(Y)| \geq (EE')^{1/2} \log^{k+1} n\Big) = \exp(-\omega(\log n)),$$

meaning

$$-\frac{\log Pr\Big(|Y-\mathbb{E}(Y)| \geq (EE')^{1/2}\log^{k+1} n\Big)}{\log n} \to \infty,$$

as $n \to \infty$.

Lemma 4.2 implies that if $\mathbb{E}_0(Y)$ is much larger than $\max_{i>0}\mathbb{E}_i(Y)$, then Y concentrates very strongly around its expected value. For instance, assume that k is a constant and $n \to \infty$ and $\mathbb{E}_0(Y) = E \geq E'\log^{2k+1} n$. In this case, we can choose $\lambda = \log^{1+1/3k} n = \omega(\log n)$ so that the tail $c_k(EE')^{1/2}\lambda^k = o(E) = o(\mathbb{E}(Y))$ and the bound $d_k \exp(-\lambda + k\log n) = \exp(-\omega(\log n))$.

The strength of Lemma 4.2 relies on the fact that we need to consider only the expected effects $\mathbb{E}_i(Y)$ instead of the worst-case effect, which is usually much larger. Thus, Lemma 4.2 can be used in several situation where classical tools such as Azuma's inequality fails. For a more comprehensive discussion about Lemma 4.2, we refer to [37] or [60].

It is clear from Corollary 4.3 that if $k \leq 5$ and $\mathbb{E}_i$'s are bounded by a constant for all $i \geq 0$, then with very high probability Y is $O(\log^{k+1} n)$.

We apply Lemma 4.2 and its corollary in the following way. Given a function Y, we first approximate it with a low degree polynomial Y' and next apply our results to show that Y' is concentrated. If the approximation is sufficiently fine, this would imply that Y itself is strongly concentrated. In the case we need only show that with high probability Y is upper (lower) bounded by some number, then it is sufficient to find a polynomial Y' which bounds Y from above (below). This method, which we call the polynomial method, will be used throughout Section 5.

4.3 Bounding the effects

In the following, we consider a generic step i, where the inputs are Ω_i, S_i and A_i. Set $n = |S_i|$ and index the points in Ω_i by $1, 2, \ldots, n$.

There are two sources of randomness in a step. One is the "Choose" operation, the other is the "Compensation" operation. To this end, t_j is the indicator of the event that the point j is chosen in the "Choose" operation, and u_j is the indicator of the event that j is deleted by "Compensation". Together, we have $2n$ independent binary random variables; the t_j's are i.i.d., but the u_j's are independent but not necessarily identically distributed. The order of random variables is $t_1, ..., t_n$, $t_{n+1} = u_1, ..., t_{2n} = u_n$. As an order does not particularly plays an important role here, we just choose a convenient one. (The order is sometimes important in other applications, see e.g., [34].)

For a set $L \subset S_i$ with $|L| \geq \log^{100} q$, let L' be the set of its remaining vertices after the step i, namely, $L' = L \cap S_{i+1}$. Typically, we want to show that

$$Pr(||L'| - \mathbb{E}(|L'|)| \geq T) \leq \exp(-\omega(\log q)),$$

for some appropriate error term T. As already discussed, the t_j's can have huge effect on L'. On the other hand, u_j is of effect at most 1 and no effect if $j \notin L$. Thus in the sum of variance bounds in $\mathbb{B}_0(\mathbf{C}, \mathbf{V})$ (see (5)) the effects of u_j's contribute at most $|L|$. We will take $\mathbf{C} \geq 1, \mathbf{V} \geq 2|L|$ so that

$$\mathbb{B}_0(\mathbf{C}, \mathbf{V}) \subseteq \mathbb{B} = \mathbb{B}(\mathbf{C}, \mathbf{V}) := \Big\{v | \max_{j=1,\ldots,n} C_j(v) \geq \mathbf{C} \text{ or } \sum_{j=1}^{n} \mathbb{E}(t_j)C_j(v) \geq \mathbf{V}/2\mathbf{C}\Big\}. \quad (6)$$

We now want to find the parameters $\mathbf{C} \geq 1$ and $\mathbf{V} \geq 2|L|$ such that

$$Pr(\mathbb{B}) = \exp(-\omega(\log q)).$$

As the sub-index i is fixed, we do not explicitly write it in the rest of this subsection. In other words, A, S, Ω stand for A_i, S_i, Ω_i, respectively. We also assume that the algorithm runs perfectly up to step $i-1$. The argument below will hold for any fixed u_j's.

For any point j, let

$$A(L, j) = \{t \in L | (tj) \cap A \neq \emptyset\}.$$

and $a(L) = \max\Big\{\max_{j\in\Omega} |A(L,j)|, |L| \log^{-100} q\Big\}$.

Lemma 4.4 *With very high probability, the effect $C_k(v)$ is $o(\log q)a(L)$ for every k.*

Lemma 4.5 *With very high probability*

$$\sum_{k=1}^{n} pC_k(v) = o(\log q)|L|.$$

The above two lemmas particularly imply that we can set $\mathbf{C} = a(L)\log q \geq 1$ and $\mathbf{V} = a(L)|L|\log^4 q \gg |L|$. (See Lemma 4.7.) Before presenting the proofs of these lemmas, let us give some intuition why the quantity $a(L)$ is relevant. Suppose that by switching t_k from 1 to 0, j is dropped out of the arc. Then all the point $g \in L$, which are deleted by a secant through j and a point from the current arc A now have a chance to survive. The number of these points is at most $a(L)$.

The technicality here is that beside the above mentioned situation, many other situations can occur. We consider these situations below. By adding a $\log q$ factor, we can handle all these situations in a relatively simple manner.

Proof of Lemma 4.4. Suppose that we switch t_k from 1 to 0. Let H_k denote the influence of this switch on $|L'|$, where all other t_j's are fixed. Trivially, $H_k \leq \alpha + \beta$, where α is the number of new points L' receives, and β is the number of new deleted points from L caused by the switch. It is important to keep in mind that C_k is a random variable depending on $t_1, \ldots, t_{k-1}$.

Denote by $\sigma(k)$ the sigma-algebra generated by $t_1, .., t_k$. We have

$$C_k = \Big|\mathbb{E}\Big(H_k | \sigma(k-1)\Big)\Big|. \tag{7}$$

On the other hand,

$$H_k \leq \sum_{g\in L} H_k(g), \tag{8}$$

where $H_k(g) = 1$ if by switching t_k the inclusion relation between g and L' is changed and 0 otherwise. Observe that the only reason L' is changed is that the new arc A' (recall that A stands for A_i and A' stands for A_{i+1}) is modified due to the t_k switch. Due to this switch, A' may gain a few more points; moreover, the only point it can possibly lose is k itself. Let us consider the effect of these events on L'.

(a) The arc A' gets a few new points, so more points will be deleted from L'. Suppose $g \in L$ is such a point (g was not deleted when $t_k = 1$, but becomes deleted when t_k switches to 0). To make this possible one of the following situations must occur.

(I) There are a point j and $a, b \in A$ such that $t_j = 1$, $[gaj]$ and $[jkb]$ hold. (In this case the following can happen: By switching t_k, j can be included in A' and this deletes g as g, j and a are co-linear. The reasoning for (II-VI) is similar)

(II) There are $a \in A$ and j, j' such that $t_j = t_{j'} = 1$ and $[ajg]$ and $[jj'k]$ hold.

(III) There are j, j' such that $t_j = t_{j'} = 1$ and $[g, j, j']$ hold and the line (jk) intersects A.

(IV) There are j, j', j'' such that $t_j = t_j = t_{j''} = 1$ and $[gjj']$ and $[jj''k]$ hold.

(b) By changing t_k the new arc A' loses a point. This occurs if and only if the point k itself was in A'. In this case, L' could get few additional points. If g is such a point, then g was deleted when $t_k = 1$, and g survives when $t_k = 0$. This could only happen if one of the following situations take place.

(V) There is a point $j \in A_i$ such that $[gjk]$ holds.

(VI) There is a point j such that $t_j = 1$ and $[gjk]$ holds.

We next split $\sum_g H_k(g)$ into the sum of six terms $H_k(\mathrm{I})$ through $H_k(\mathrm{VI})$ corresponding to the six situations. (Since we only need to bound H_k from above, we can ignore the overlaps between the cases.) First let us bound $H_k(\mathrm{I})$. Observe that

$$\begin{aligned} H_k(\mathrm{I}) &\leq \sum_{j \neq k} t_j \mathbf{1}_{\{j \in A(k)\}} \sum_{g \in L} \mathbf{1}_{\{j \in A(g)\}} \\ &\leq \sum_{j \in A(k)} t_j |A(L, j)|. \end{aligned}$$

Setting $C_k(I)$ in a similar way (by splitting C_k into six terms). We have

$$C_k(\mathrm{I}) \leq \mathbb{E}\Big(H_k(I) | \sigma(k-1)\Big) \leq \sum_{j \in A(k), j<k} t_j |A(L, j)| + \sum_{j \in A(k), j>k} p |A(L, j)|. \tag{9}$$

The second term on the right hand side is $O(1)$ since

$$\sum_{j \in A(k), j>k} p |A(L, j)| \leq a(L) p |A(k)|,$$

where $A(k) = A_i(k)$ is defined as in subsection 2.3. As we assume that the algorithm runs perfectly up to the current stage, we have $|A(k)| = |A_i(k)| \sim a_i b_i q^{3/2}$ and $p = p_i = \theta(b_i q^{3/2})^{-1}$ (see Remark 3.4). Thus

$$a(L) p |A(k)| \leq 2 a_i \theta a(L) = o(a(L)).$$

Now consider the first term in the right hand side (9), which is a sum of independent random variables. Since the expectation of $\sum_{j \in A(k), j>k} t_j$ is at most $a_i \theta = o(1)$, with very high probability this sum is $o(\log q)$. Consequently, with very high probability:

$$C_k(\mathrm{I}) = o(\log q) a(L). \tag{10}$$

The remaining cases (II), (III), (IV) can be handled similarly and we omit the details.

For situation (V), it is clear that the number of such a points g is at most $\max_j A(L,j) \leq a(L)$.

To complete the proof, let us now consider situation (VI). By a reasoning similar to the one used for (I), we have

$$C_k(\text{VI}) \leq \sum_{t \in L} \sum_{j,[jkt]} t_j = Y$$

Suppose that we are in the first phase of the algorithm; since we assume that the algorithm runs perfectly up to the current stage, every line through j or k has less than $2b_i q$ points. Thus $\mathbb{E}(Y) \leq |L|(2b_i q)p_i \leq |L|q^{-1/2} \leq |L|\log^{-102} q$, given that q is sufficiently large. Since Y is a sum of independent random variables, it is easy to show (using Csernoff's bound or Lemma 4.2) that with very probability $Y \leq |L|\log^{-100} q \leq a(L)$.

Now suppose that we are in the second phase. In this case, we need to use the property that every line has at most $2\log^{c_1} q$ points (this is a corollary of Property **(2)** in the first phase and the definition of the second phase). Given this, $\mathbb{E}(Y) \leq 2|L|p_i \log^{c_1} q$. On the other hand, $p_i = \theta(b_1 q^{3/2})^{-1} \leq \log^{-c} q$ by the description of the algorithm. Therefore, $\mathbb{E}(Y) \leq |L|\log^{c_1 - c} q \leq |L|\log^{-102} q$ (remember that in subsection 3.1 we set $c = 300$ and $c_1 = 100$). Again by Csernoff's bound or Lemma 4.2, we can conclude that with very high probability, $Y \leq |L|\log^{-100} q \leq a(L)$. □

Proof of Lemma 4.5 Again split C_k into the sum of six terms $C_k(\text{I}) - C_k(\text{V})$. Consider

$$p\sum_{k=1}^{n} C_k(\text{I}) \leq p\sum_{k=1}^{n} \Big(\sum_{j \in A(k), j<k} t_j |A(L,j)| + \sum_{j \in A(k), j>k} p|A(L,j)| \Big).$$

Now let us split the right hand side into a constant and a sum of random variables. The constant is:

$$\begin{aligned}
\sum_{k=1}^{n} \sum_{j \in A(k), j>k} p|A(L,j)| &\leq p^2 \sum_{j=1}^{n} |A(L,j)||\{k : j \in A(k)\}| \\
&= p^2 \sum_{j=1}^{n} |A(L,j)||A(j)| \\
&\leq p^2 \max_j |A(j)| \sum_{j=1}^{n} |A(L,j)| \\
&= p^2 \max_j |A(j)| \sum_{t \in L} |A(t)| \\
&\leq |L|p^2 (\max_j |A(j)|)^2.
\end{aligned}$$

Recall that at step i, $p\max_j |A(j)| \sim a_i \theta = o(1)$. Thus the last formula is $o(|L|)$. We now show that the other term which is the sum of many random variables could be bound (with very high probability) by $o(\log q)|L|$. The sum in question is

$$p\sum_{k=1}^{n} \sum_{j \in A(k), j<k} t_j |A(L,j)|, \tag{11}$$

which can be upper bounded by,

$$p \max_j |A(j)| \sum_{j=1}^{n} t_j |A(L,j)|.$$

The expectation of $\sum_{j=1}^{n} t_j |A(L,j)|$ is $p \sum_{j=1}^{n} |A(L,j)| = p \sum_{j \in L} |A(j)|$. As already shown, the last quantity is at most $p|L| \max_{t \in L} |A(t)| = o(|L|)$. Given this, it is easy to prove that with very high probability $\sum_{j=1}^{n} t_j |A(L,j)|$ is $O(\log q / a_i \theta)|L|$ since $|A(L,j)| \leq |L|$. Therefore, with very high probability (11) is at most $p \max_j |A(j)| O(\log q)|L| = o(\log q)|L|$, since $p \max_j |A(j)| = o(1)$. The proofs regarding the remaining cases (II-VI) are similar and omitted. □

Remark 4.6 • *In this and the previous subsection, we assume, for the sake of simplicity, that $L \subset S$. On the other hand, all statements also hold for $L \subset \Omega$. For the analysis of the previous subsection, the points in $L \backslash S$ could only help, as the "Compensation" operation does not act on them. In this subsection, we never use the fact that $L \subset S$.*

• *The proofs of Lemmas 4.4 and 4.5 also hold for a multi-set L, where some points $t \in L$ might have multiplicity larger than 1. To see this, note that in all summations over $t \in L$, the fact that t are different is not essential. Of course, when we use these Lemmas for multi-set, $A(L,j)$ should also be defined with multiplicities.*

4.4 Consequences

Lemma 4.7 *Fix a (multi-)set L in Ω. Then with very high probability*

$$|L' - E(L')| \leq a(L)^{1/2} |L|^{1/2} \log^5 q.$$

Proof. Set $\mathbf{C} = a(L) \log q$, $\mathbf{V} = a(L)|L| \log^4 q$ and $\lambda = \log^{3/2} q$. By Lemma 4.1,

$$Pr\Big(|L' - E(L')| \geq (\lambda \mathbf{V})^{1/2}\Big) \leq \exp(-\omega(\log q)) + Pr(\mathbb{B}),$$

with $\mathbb{B}$ defined with respect to $\mathbf{C}$ and $\mathbf{V}$ as in subsection 4.1. Note that $(\lambda V)^{1/2} = o(a(L)^{1/2} |L|^{1/2} \log^5 q)$ with room to spare. On the other hand, by Lemmas 4.4 and 4.5,

$$Pr(\mathbb{B}) = \exp(-\omega(\log q)),$$

completing the proof. □

The following corollary is immediate.

Corollary 4.8 *Let K be a fixed positive constant. For any set L satisfying $a(L) \leq \frac{|L|}{\log^{2(K+5)} q}$, we have with very high probability that $|L' - E(L')| \leq |L| \log^{-K} q$.*

In the proof of the main lemma, we frequently need to prove a statement of the following form "with very high probability $|L - E(L)| \leq |L| \log^{-d} q$", where d is properly chosen constant (notice that all the relative error terms in the properties of the main lemma is of the

form $\log^{-d} q$ for some d). Consider the quantity $a(L) = \max\Big\{ \max_{j\in\Omega_i} |A(L,j)|, |L|\log^{-100} q\Big\}$. It is clear that if $a(L) = |L|\log^{-100} q$, then Lemma 4.7 or Corollary 4.8 immediately prove the statement described above (in all applications, d will be much smaller than 100, so the constant 100 in the exponent provides plenty of room). Therefore, we have to focus only on the case $a(L) = \max_{j\in\Omega_i} |A(L,j)|$ and we will assume that this is the case in all applications of Lemma 4.7 and Corollary 4.8.

5 PROOF OF THE MAIN LEMMA

In this proof, we consider a generic step i, assuming that the algorithm runs perfectly in the first $i-1$ steps (for $i = 1$, this assumption holds trivially); we sometime refer to this assumption as the induction hypothesis. We shall show that each of the properties holds for $i+1$ with very high probability. Since the number of properties (taking into account all possible choices for l, u, v, w, z) is $O(q^{10})$, it follows that all properties hold simultaneously with very high probability.

5.1 Phase one

The ten properties of Phase one split into four groups: **(1)**, **(2)(4)(5)**, **(3)(9)(10)**, and **(6)(7)(8)**. The strategy for the last three groups is similar. First, we prove the highest indexed properties (**(5)**, **(10)**, **(8)**, respectively), using the polynomial method discussed in subsection 4.2. Next, we prove the middle properties (**(4)**, **(9)**, **(7)**, respectively), using Lemma 4.7 and again the polynomial method. The reason we need to prove **(5)**, **(10)**, **(8)** first is that the quantities considered in these properties play the role of $a(L)$ (see Lemma 4.7) with respect to the quantities considered in **(4)**, **(9)**, **(7)**, respectively. Similarly, the quantities in **(4)**, **(9)**, **(7)** play the role of $a(L)$ with respect to the quantities in **(2)**, **(3)**, **(6)**, respectively. Thus, we can again use Lemma 4.7 to prove **(2)**, **(3)** and **(6)**. The proof of **(1)** is based entirely on the polynomial method.

Property **(6)** is not primary, but it plays an essential role in the analysis of the second phase.

5.1.1 Proof of (1)

Set $U_i = B_i \backslash M_i$. It suffices to show that with very high probability $|B_i| \le (1+o(1))\theta q^{1/2}$ and $|U_i| = o(1)\theta q^{1/2}$. The first inequality is easy by Csernoff bound, as B_i is the sum of i.i.d random variables and $\mathbb{E}(B_i) = \theta q^{1/2} \ge \log^2 q$ (one can also use Lemma 4.2). To prove the second inequality, we apply the polynomial method, discussed in the last paragraph of subsection 4.2. To apply this method, we first bound $|U_i|$ by a low degree polynomial as follows. Notice for any point j in U_i, j should be chosen (that is, $t_j = 1$) but j is not in M_i. The later has two possible reasons. The first is that there is a point $j' \in B_i$ such that (jj') intersect A_i, i.e, $\sum_{j\in A_i(j)} t_{j'} \ge 0$. The second is that there are two points j' and j'' in B_i such that j, j' and j'' are co-linear. This occurs if $\sum_{j',j''|[jj'j'']} t_{j'}t_{j''} \ge 0$. Together, we have

$$|U_i| \le \sum_{j\in S_i} t_j 1_{j\notin M_i} \le \sum_{j\in S_i} t_j \Big(\sum_{j'\in A_i(j)} t_{j'} + \sum_{j',j''|[jj'j'']} t_{j'}t_{j''}\Big). \tag{12}$$

Consider the first sum in the last formula

$$\sum_{j\in S_i} t_j \sum_{j'\in A_i(j)} t_{j'} = \sum_{j\in S_i, j'\in A_i(j)} t_j t_{j'} = Y.$$

To bound Y, we apply Lemma 4.2 as Y is a polynomial of degree 2. Furthermore, by the induction hypothesis, $|S_i| = b_i q^2$, $a_i \sim i\theta$ and $|A_i(j)| \sim a_i q^{1/2}(b_i q) \leq 2a_i b_i q^{3/2}$. So

$$\mathbb{E}_0(Y) = \mathbb{E}(Y) \leq 2(b_i q^2)(a_i b_i q^{3/2}) p_i^2 = 2a_i \theta^2 q^{1/2}.$$

To bound $\mathbb{E}_1(Y)$, observe that for any fixed j, there are at most $\max_j |A_i(j)|$ points j' such that the product $t_j t_{j'}$ appears in Y. As mentioned above, $|A_i(j)| \leq 2a_i b_i q^{3/2}$. Thus, we have

$$\mathbb{E}_1(Y) \leq 2(a_i b_i q^{3/2}) p_i = 2a_i \theta \leq 4i\theta^2 = o(1).$$

In the last inequality, we use the fact that $i = O(\theta^{-1}\log^{1/2} q)$ and $\theta = \log^{-2} q$. Finally, notice that each product $t_j t_{j'}$ could appear at most twice so $\mathbb{E}_2(Y) \leq 2$.

Now Lemma 4.2 yields that with very high probability,

$$Y \leq 3a_i \theta^2 q^{1/2} = o(\theta q^{1/2}).$$

Using a similar argument, we can also prove the same statement for the other term in (12), namely, with very high probability

$$\sum_{j\in S_i} t_j \Big(\sum_{j',j''|[jj'j'']} t_{j'} t_{j''} \Big) = o(\theta q^{1/2}).$$

This completes the proof. □

Remark 5.1 *A detailed calculation shows that we can bound, with very high probability, the second term* $\sum_{j\in S_i} t_j (\sum_{j',j''|[jj'j'']} t_{j'} t_{j''})$ *by* $10\theta^3 q^{1/2}$. *This yields that with very high probability*

$$a_{i+1} - a_i \geq \theta - (3a_i\theta^2 + 10\theta^3).$$

Moreover, we can also prove that with very high probability $|B_i| \leq (\theta + \theta^3) q^{1/2}$. *So (with very high probability)* $a_{i+1} - a_i \leq \theta + \theta^3$.

5.1.2 Proof of (5)

We proved that $|S_{i+1}(l,u,v)| \leq (i+1)\log^4 q$. Consider a line l and denote by $B_i(l,v)$ the set of vertices $x \in l$ such that the line (vx) intersects B_i. Moreover, set $B_i(l,u,v) = B_i(l,u) \cap B_i(l,v)$. Observe that any point $x \in S_{i+1}(l,u,v) \backslash S_i(l,u,v)$ must belong to either $B_i(l,u,v), S_i(l,u) \cap B_i(l,v)$, or $S_i(l,v) \cap B_i(l,u)$. It follows that

$$|S_{i+1}(l,u,v)| \leq |S_i(l,u,v)| + |B_i(l,u,v)| + |S_i(l,u) \cap B_i(l,v)| + |S_i(l,v) \cap B_i(l,u)|.$$

By the induction hypothesis, we have $|S_i(l,u,v)| \leq i\log^4 q$. So it suffices to show that the last three terms on the right hand side are, with very high probability, at most $\frac{1}{3}\log^4 q$.

Let us consider the first term $|B_i(l,u,v)|$. Again we apply the polynomial method in a way similar to the proof of **(1)**. Notice that a point $x \in l$ is in $B_i(l,u,v)$ then there exists

$j \in (xu)$ and $j' \in (xv)$ where both j and j' are chosen in B_i (in other words, $t_j = t_{j'} = 1$). Therefore,

$$B_i(l,u,v) \leq \sum_{x \in l} (\sum_{j \in (xu)} t_j)(\sum_{j' \in (xv)} t_{j'}) = \sum_{jj' \in \mathcal{E}(H)} t_j t_{j'} = Y.$$

Here H is a graph on S_i, the set $(\mathcal{E}(H))$ of edges of H is the set of all pairs jj' obtained by expressing the product of the two sums. Next, we show that with very high probability $Y = O(\log^3 q) \ll \frac{1}{3}\log^4 q$. This is a straightforward consequence of Lemma 4.2. The only thing we need is to verify is that $\mathbb{E}_i(Y) = O(1)$ for $i = 0,1$ and 2 (see the paragraph following Corollary 4.3). It is trivial that $\mathbb{E}_2(Y) \leq 2$ as each pair (j,j') can occur at most twice. Moreover,

$$\mathbb{E}_0(Y) \leq p_i^2 (\max_{l'} |S_i(l')|)^2 \leq 4p_i^2 (b_i q)^2 = 4\theta^2 q^{-1} \leq 1,$$

where the maximum is taken over the set of all lines l'. Here we use the fact $|S_i(l')| \leq 2b_i q$ for every line l' (as we assume the algorithm runs perfectly up to step $i-1$), and the definition of p_i, $p_i = \theta(b_i q^{3/2})^{-1}$.

Similarly,

$$\mathbb{E}_1(Y) \leq p_i \max_{l'} |S_i(l')| \leq 2\theta q^{-3/2} \leq 1.$$

The proofs for the other two terms are similar and omitted. □

5.1.3 Proof of (4)

We show inductively that the following holds for all plausible j with very high probability

$$|S_j(l,v)| \leq 8ja_j b_j q^{1/2} + j\log^{40} q. \tag{13}$$

To apply the results in subsections 4.4. and 4.5, we set $L = S_i(l,v)$. Provided that $|S_i(l,v)|$ satisfies (13), we prove that with very high probability

$$|S_{i+1}(l,v)| \leq 8(i+1)a_{i+1}b_{i+1}q^{1/2} + (i+1)\log^{40} q \tag{14}$$

As in subsection 4.5, let L' be the set of surviving points in L; it is clear that any point $x \in S_{i+1}(l,v) \backslash L$ should be in $B_i(l,v)$. Therefore,

$$|S_{i+1}(l,v)| \leq |L'| + |B_i(l,v)|. \tag{15}$$

We bound the terms in the right hand side separately. First observe that

$$\mathbb{E}(|L'|) = |L|(1 - P_i^u) \leq 8ia_i b_{i+1} q^{1/2} + i\log^{40} q, \tag{16}$$

since $b_{i+1} = b_i(1 - P_i^u)$ by definition. Next, we apply Lemma 4.7. In order to apply this lemma, observe that by the proof of Property (5),

$$a(L) \leq \max_{l,u,v} |S_i(l,u,v)| \leq i\log^4 q \leq \log^7 q,$$

as $i \leq \log^3 q$. By Lemma 4.7, with very high probability

$$|L'| \leq \mathbb{E}(L') + \mathbb{E}(L')^{1/2}\log^9 q. \tag{17}$$

Now let us bound $|B_i(l,v)|$ using the polynomial method. It follows from the description of the algorithm that if a point x is in $B_i(l,v)$, then some point j on the line (xv) must be chosen, i.e, $t_j = 1$. Therefore

$$|B_i(l,v)| \leq \sum_{x\in l} \sum_{j\in(xv)} t_j = Y.$$

Next, we apply Lemma 4.2 (in fact Csernoff's bound also applies as Y is a sum of independent variables). Again using the induction hypothesis $|S_i(l')| \sim b_i q_i \leq 2b_i q_i$ for any line l', we have

$$\mathbb{E}_0(Y) \leq p_i \max_{l'} |S_i(l')|^2 \leq p_i (2b_i q_i)^2 = 4\theta b_i q^{1/2}. \tag{18}$$

Moreover, as each point j appear at most once $\mathbb{E}_1(Y) = 1$. By Lemma 4.2, we have, with very high probability

$$|B_i(l,v)| \leq \mathbb{E}(Y) + (\max(\mathbb{E}(Y),1))^{1/2} \log^2 q. \tag{19}$$

By (14–19), we obtain that with very high probability

$$\begin{aligned} |S_{i+1}(l,v)| &\leq 8ia_i b_{i+1} q^{1/2} + i\log^{40} q + \left(ia_i b_{i+1} q^{1/2} + i\log^7 q\right)^{1/2} \log^9 q + \\ &+ 4\theta b_i q^{1/2} + \max(4\theta b_i q^{1/2}, 1)^{1/2} \log^2 q. \end{aligned} \tag{20}$$

Recall $a_i \sim i\theta$, $1.1b_i \geq b_{i+1} \geq 0.9b_i$, $i \leq \log^3 q$ (b_{i+1} can be estimated based on the induction hypothesis). Moreover, **(1)** shows $a_{i+1} \sim a_i + \theta$. Using these estimates, we next verify that the right hand side of (20) is at most

$$8(i+1)a_{i+1}b_{i+1}q^{1/2} + (i+1)\log^{40} q,$$

with lots of room to spare in the exponent of the logarithm.

To start, notice that

$$8(i+1)a_{i+1}b_{i+1}q^{1/2} - 8ia_i b_{i+1} q^{1/2} \geq 8a_i b_{i+1} q^{1/2}.$$

So we need only show,

$$8a_i b_{i+1} q^{1/2} + \log^{40} q \geq \left(ia_i b_{i+1} q^{1/2} + i\log^7 q\right)^{1/2} \log^9 q + 4\theta b_i q^{1/2} + \max(4\theta b_i q^{1/2}, 1)^{1/2} \log^2 q. \tag{21}$$

If $b_i q^{1/2} > \log^8 q$, then

$$\max(4\theta b_i q^{1/2}, 1)^{1/2} \log^2 q \leq 3\theta b_i q^{1/2}$$

and

$$\left(ia_i b_{i+1} q^{1/2} + i\log^7 q\right)^{1/2} \log^9 q \leq a_i b_{i+1} q^{1/2}.$$

Thus, the right hand side of (21) is at most

$$4\theta b_i q^{1/2} + 3\theta b_i q^{1/2} + a_i b_{i+1} q^{1/2} \leq 8a_i b_{i+1} q^{1/2}.$$

Now assume that $b_i q^{1/2} \leq \log^8 q$. In this case, it is easy to check that the term $\log^{40} q$ on the left hand side of (21) swallows everything. This completes the proof $\square$

5.1.4 Proof of (2)

Now we are ready to prove the crucial property **(2)**. Set $L = S_i(l)$ and $K = 14$ (in order to apply Corollary 4.8). Notice that by Property **(4)**

$$a(L) \leq \max_{l',v} |S_i(l',v)| = O(ia_ib_iq^{1/2} + \log^{40} q).$$

Moreover, by the induction hypothesis $|L| \sim b_iq \geq \frac{1}{2}b_iq$. As we are in the first phase of the algorithm, $b_iq \geq \log^{100} q$, so $b_iq \geq a(L)\log^{2(K+5)} q$. By Corollary 4.8, with very high probability

$$\Big||L'| - \mathbb{E}(|L'|)\Big| \leq |L|\log^{-14} q \leq 2b_iq\log^{-14} q,$$

where (as usual) $L' = S_{i+1}(l)$.

Observe that if $|L| = b_iq(1+\alpha)$, then $\mathbb{E}(|L'|) = b_{i+1}q(1+\alpha)$. Therefore, by the induction hypothesis, $\mathbb{E}(|L'|)$ satisfies $|\mathbb{E}(|L'|) - b_{i+1}q| \leq ib_{i+1}q\log^{-13} q$. The triangle inequality yields

$$\Big||L'| - b_{i+1}q\Big| \leq \Big(i\log^{-13} q + 2\log^{-14} q\Big)b_{i+1}q \leq (i+1)b_{i+1}q\log^{-13} q,$$

completing the proof. □

5.1.5 Proofs of (3),(9), (10)

The proofs of the properties in this group are more or less identical to those of **(2)**, **(4)** and **(5)**. The only formal differences are that we use b'_i instead of b_i and we need only the upper bound in **(3)**. □

5.1.6 Proof of (8)

This proof is similar to that of **(5)**. However, we provide some details in order to illustrate our polynomial method.

Let $X = \{u, v, w, z\}$. For any non-empty subset X' of X, set $A_i(X') = \cap_{a\in X} A_i(a)$ and $B_i(X') = \cap_{a\in X} B_i(a)$. Assuming $|A_i(X)| \leq i\log^6 q$, we prove that with very high probability $|A_{i+1}(X)| \leq (i+1)\log^6 q$. To start, observe

$$A_{i+1}(X) \subset A_i(X) \cup B_i(X) \cup \sum_{X', 1\leq|X'|\leq 3} A_i(X') \cap B_i(X\backslash X').$$

Therefore, it suffices to show that with very high probability $|B_i(X)| \leq \log^5 q$ and $|A_i(X') \cap B_i(X\backslash X')| \leq \log^5 q$ for all X'.

To handle $|B_i(X)|$, notice that for any $x \in B_i(X)$, there should be $j, j', j'', j''' \in B_i$ such that $[jxu], [j'xv], [j''xw]$ and $[j'''xz]$ hold. This implies

$$\begin{aligned} B_i(X) &\leq \sum_{x\in S_i} \Big(\sum_{j,[jxu]} t_j\Big)\Big(\sum_{j',[j'xv]} t_{j'}\Big)\Big(\sum_{j'',[j''xw]} t_{j''}\Big)\sum_{j''',[j'''xz]} t_{j'''}\Big) \\ &= \sum_{(j,j',j'',j''')\in\mathcal{I}} t_jt_{j'}t_{j''}t_{j'''} = Y \end{aligned}$$

where $\mathcal{I}$ is the index set consisting of all possible tuples (j, j', j'', j''').

As Y is a polynomial of degree 4, we can use Corollary 4.3 to conclude the proof. It is sufficient to show that $\mathbb{E}_i(Y) = O(1)$ for all $0 \le i \le 4$. Trivially, $\mathbb{E}_4(Y) = 1$. Moreover,

$$\begin{aligned}
\mathbb{E}_0(Y) &\le p_i^4|S_i|(\max_l |S_(l)|)^4 \le 2p_i^4 b_i q^2 (b_i q)^4 = 2\theta^4 b_i = O(1) \\
\mathbb{E}_1(Y) &\le p_i^3(\max_l |S_i(l)|)^4 \le 2p_i^3 (b_i q)^4 = 2\theta^3 b_i q^{-1/2} = O(1) \\
\mathbb{E}_2(Y) &\le p^2(\max_l S_i(l)^2 \le 2p_i^2 (b_i q)^2 = 2\theta^2 q^{-1} = O(1) \\
\mathbb{E}_3(Y) &\le p(\max_l S_i(l)) \le 2p(b_i q) = 2\theta q^{-1/2} = O(1).
\end{aligned}$$

The proof concerning $|A_i(X') \cap B_i(X \backslash X')|$ is quite similar. In this case, the degree of the resulting polynomial is $|X \backslash X'|$. Thus, we can use $\log^4 q$ instead of $\log^5 q$ as a upper bound but this makes no essential difference. □

5.1.7 Proofs of (7) and (6)

First we consider **(7)**. Let $X = \{u, v, w\}$, where u, v, w are three arbitrary point in Ω_i. Set $L = A_i(u, v, w)$ and $L' = L \cap S_{i+1}$. Proceed as in the previous proof, we have

$$|A_{i+1}(X)| \le |L'| + |B_i(X)| + \sum_{X' \subset X} |A_i(X) \cap B_i(X \backslash X')|, \tag{22}$$

with the last sum taken over all proper subsets of X. Assuming $A_i(X) \le i b_i q^{1/2} + i \log^{10} q$, we will show that $A_{i+1}(X) \le (i+1) b_{i+1} q^{1/2} + (i+1) \log^{10} q$.

To start, observe that since each point in S_i survives with probability $1 - P_i^u$,

$$\mathbb{E}(|L'|) \le \Big(i b_i q^{1/2} + i \log^{10} q\Big)(1 - P_i^u) \le i b_{i+1} q^{1/2} + i \log^{10} q.$$

On the other hand, by **(8)** (see also the remark following Lemma 4.7 and the remark at the end of this subsection),

$$a(L) \le \max_{z \in \Omega_i} |A_i(u, v, w, z)| \le \log^6 q. \tag{23}$$

So Lemma 4.7 implies that with high probability

$$|L'| \le \mathbb{E}(L') + (|L| a(L))^{1/2} \log^5 q \le i b_{i+1} q^{1/2} + i \log^{10} q + |L|^{1/2} \log^8 q. \tag{24}$$

We next show that with very high probability

$$|B_i(X)| \le o(b_{i+1} q^{1/2}) + \log^9 q, \tag{25}$$

$$|A_i(X') \cap B_i(X \backslash X')| \le o(b_{i+1} q^{1/2}) + \log^9 q, \tag{26}$$

for all X'. As the sum of the right hand sides in (24)-(26) is upper bounded by $(i+1) b_{i+1} q^{1/2} + (i+1) \log^{10} q$, our proof is complete given (25) and (26).

To verify (25), notice that by a similar argument as in the previous subsection, we have

$$|B_i(X)| = |B_i(u, v, w)| \le \sum_{x \in S_i} \sum_{j.[jux]} t_j \sum_{j', [j'vx]} t_{j'} \sum_{j'', [j''wx]} t_{j''} = Y.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}_0(Y) &\leq p_i^3|S_i|(\max_l|S_i(l)|)^3 \leq 2p_i^3(b_iq^2)(b_iq)^3q_i^{1/2} = 2\theta^3 b_i q^{1/2} = \alpha \\
\mathbb{E}_1(Y) &\leq p_i^2(\max_l|S_i(l)|)^3 \leq 2p_i^2(b_iq)^3 = 2\theta^2 b_i = O(1) \\
\mathbb{E}_2(Y) &\leq p_i \max_l(|S_i(l)|) \leq 2p_i(b_iq) = 2\theta q^{-1/2} = O(1) \\
E_3(Y) &= 1.
\end{aligned}$$

Here we used the hypothesis that $|S_i(l)| \sim b_iq \leq 2b_iq$ for every l. So Lemma 4.2 yields

$$|B_i(X)| \leq \alpha + \alpha^{1/2}\log^4 q.$$

Next, by Cauchy's inequality

$$\alpha + \alpha^{1/2}\log^4 q \leq 2\alpha + \log^8 q.$$

Furthermore, $2\alpha = 4\theta^3 b_i q^{1/2} = o(b_{i+1}q^{1/2})$ since $\theta = o(1)$ and $b_{i+1} \geq b_i/2$. This concludes the proof of (25). The proof for (26) is similar and omitted.

The proof of **(6)** is similar to that of **(7)**. □

Remark 5.2 *If we use $a(L) \leq |L|\log^{-100} q$ in (23), the right most formula in (24) becomes $ib_{i+1}q^{1/2} + i\log^{10} q + |L|\log^{-45} q$. As $|L|\log^{-45} q = o(b_{i+1}q^{1/2})$, this does not influence the rest of the proof.*

5.2 Phase two

The proof of Property **(1)** is the same as before. The proofs of Properties **(4)** and **(5)** are based essentially on Lemma 4.7 and more or less identical. The difficult part of this phase is Properties **(2)** and **(3)**, whose analysis is fairly technical.

5.2.1 Proofs of (4) and (5)

Let us consider **(5)**. As usual, we set $L = \Omega_i$ and define $L' = \Omega_{i+1}$.

Observe that Property **(12)** of Phase 1 and that $b_i'/b_i \leq 2$ (which is a consequence of the induction hypothesis) imply that $|\Omega_i(l)| \leq K\log^{c_1} q$ for some constant K at every step in Phase 2. Therefore, $a(L) \leq 2K\log^{c_1} q a_i q^{1/2}$. On the other hand, since $|L| \sim b_iq^2 \geq \frac{1}{2}b_iq^2 \geq \frac{1}{2}(\log^{c_0} q)q^{1/2}$, we have $|L| \geq a(L)\log^{100} q$. Thus, Lemma 4.7 yields that with very high probability.

$$\Big||L'| - \mathbb{E}(|L'|)\Big| \leq (a(L)|L|)^{1/2}\log^5 q.$$

Recall that $\mathbb{E}(|L'|) \leq L(1-P_i^l)$, we have, with very high probability, that

$$\begin{aligned}
|\Omega_{i+1}| = |L'| &\leq (1-P_i^l)|L| + (a(L)|L|)^{1/2}\log^5 q \\
&= |L|(1 - P_i^l + (\frac{a(L)}{|L|})^{1/2})\log^5 q.
\end{aligned}$$

As $(\frac{a(L)}{|L|})^{1/2})\log^5 q \leq \log^{-20} q$, we have

$$|L|\Big(1-P_i^l+(\frac{a(L)}{|L|})^{1/2})\log^5 q\Big) \le |L|(1-P_i^l)(1+\log^{-10} q),$$

concluding the proof of **(5)**. The proof of **(4)** is similar and omitted. □

5.2.2 Proof of (2)

To prove **(2)**, we first need to estimate the expectation of $|T_{i+1}(v)|$. This involves the expectation of the event that a pair (x,y) survives in S_{i+1}. The obstacle here is that if x and y are two points in S_i, then the events $x \in S_{i+1}$ and $y \in S_{i+1}$ are not independent. The following lemma helps us to overcome this problem by saying that these two events are almost independent in a certain sense, and therefore one can compute the desired expectation with appropriate accuracy. For the purpose of this subsection, set $\delta = \log^{-13} q$.

Lemma 5.3 *For every $x,y \in S_i$,*

$$|Pr(x,y \in S_{i+1}) - Pr(x \in S_{i+1})Pr(y \in S_{i+1})| = o(\delta).$$

The proof of this lemma is complicated and we defer it to the end of this subsection. Let us now complete the proof of **(2)**, provided Lemma 5.3.

Since $\delta = \log^{-13} q$, it suffices to prove that with very high probability

$$\Big||T_{i+1}(v)| - \frac{1}{2}b_{i+1}^2 q^3\Big| \le 3(i+1)\delta b_{i+1}^2 q^3.$$

First let us estimate $\mathbb{E}(|T_{i+1}(v)|)$; by definition

$$\mathbb{E}(|T_{i+1}(v)|) = \sum_{x,y \in S_i, [xyv]} Pr(x,y \in S_{i+1}).$$

By Lemma 5.3

$$|T_i(v)|Pr(x \in S_{i+1})^2(1-\delta) \le \mathbb{E}(|T_{i+1}(v)|) \le |T_i(v)|Pr(x \in S_{i+1})^2(1+\delta).$$

Recall that $b_i Pr(x \in S_{i+1}) = b_{i+1}$, by the induction hypothesis we have

$$\frac{1}{2}b_{i+1}^2 q^3(1-3i\delta)(1-o(\delta)) \le \mathbb{E}(|T_{i+1}(v)|) \le \frac{1}{2}b_{i+1}^2 q^3(1+3i\delta)(1+o(\delta)). \tag{27}$$

It now remains to show that $|T_{i+1}(v)|$ strongly concentrates around its expected value and we can again use Lemma 4.7. For convenience, instead of $T_i(v)$, we will consider the multi-set

$$T_i'(v) = \{x^{m(x)} | x \in S_i \backslash v\},$$

where $m(x) = ((xv)-2)$ is the multiplicity of x. It is clear that $|T_i'(v)| = 2|T_i(v)|$. Next, we use Lemma 4.7 to show that $|T_{i+1}'(v)|$ is strongly concentrated. As usual, set $L = T_i'(v)$ and $L' = T_{i+1}(v)$. To bound $a(L)$, notice that for any u and v

$$|A_i(T_i'(v),u)| \le |A_i(u)| \max_x m(x) \le |A_i(u)| 2\log^{c_1} q \le 4a_i b_i q^{3/2}\log^{c_1} q = \alpha.$$

Since $|L| = |T_i'(v)| \sim \frac{1}{2}b_i^2 q^3 = \beta$, by Lemma 4.7, we have with very high probability that

$$\Big||L'| - \mathbb{E}(|L'|)\Big| \leq (\alpha\beta)^{1/2}\log^5 q. \tag{28}$$

As $c_1 = 100$ and $a_i \leq \log q$ (see Remark 3.9), we have that $(\alpha\beta)^{1/2}\log^5 q \leq (b_i q^{3/2})^{3/2}\log^{60} q$. On the other hand, $b_i q^{3/2} \geq \log^c q$, where $c = 300$, and $b_{i+1} \geq .9b_i$ (see Reamrk 3.9). It thus follows that $(b_i q^{3/2})^{3/2}\log^{60} q = o(\delta b_{i+1}^2 q^3)$. This implies **(2)** via (27) and (28), with lots of room to spare (for instance we can replace 3 in the error term by any constant larger than 1). □

Proof of Lemma 5.3

In this proof, we ignore unnecessary sub-indices. Since

$$Pr(A\cap B) - Pr(A)Pr(B) = Pr(\bar{A}\cap\bar{B}) - Pr(\bar{A})Pr(\bar{B}),$$

we consider $Pr(u \notin S \text{ and } v \notin S)$ instead of $Pr(u \in S \text{ and } v \in S)$. First of all, the compensations are independent of all other. It is enough to show that the two events $\cup_{x\in A(u)\cup T(u)}\{x \in M\}$ and $\cup_{y\in A(v)\cup T(v)}\{y \in M\}$ are almost independent, where $x \in A(v)\cup T(v)$ means that x is a point in $A(v)$ or (unordered) pair in $T(v)$ and a pair $x = (x_1, x_2) \in M$ means that both of x_1, x_2 are in M. Our proof is based on inclusion-exclusion. It is easy to see that

$$\begin{aligned} & Pr\Big(\bigcup_{x\in A(u)\cup T(u)}\{x \in M\}\Big)Pr\Big(\bigcup_{y\in A(v)\cup T(v)}\{y \in M\}\Big) \\ = & \sum_l (-1)^{l-1}\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr\Big(x_1,...,x_l \in M\Big)Pr\Big(y_1,...,y_l \in M\Big), \end{aligned}$$

where ordered pairs (x, y) are chosen from $(A(u)\cup T(u))\times(A(v)\cup T(v))$. One way to see the above equation is the following. Consider two identical independent experiments which create M and M^*. Then

$$\begin{aligned} & Pr\Big(\bigcup_{x\in A(u)\cup T(u)}\{x \in M\}\Big)Pr\Big(\bigcup_{y\in A(v)\cup T(v)}\{y \in M\}\Big) \\ = & Pr\Big(\bigcup_{x\in A(u)\cup T(u)}\{x \in M\}\Big)Pr\Big(\bigcup_{y\in A(v)\cup T(v)}\{y \in M^*\}\Big) \\ = & Pr\Big(\bigcup_{x\in A(u)\cup T(u)}\{x \in M\}\cap\bigcup_{y\in A(v)\cup T(v)}\{y \in M^*\}\Big) \\ = & Pr\Big(\bigcup_{(x,y)}\{x \in M\}\cap\{y \in M^*\}\Big). \end{aligned}$$

On the other hand,

$$\begin{aligned} & Pr\Big(\bigcup_{x\in A(u)\cup T(u)}\{x \in M\}\cap\bigcup_{y\in A(v)\cup T(v)}\{y \in M\}\Big) \\ = & Pr\Big(\bigcup_{(x,y)}\{x \in M\}\cap\{y \in M\}\Big) \\ = & \sum_l (-1)^{l-1}\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr\Big(x_1,...,x_l \in M \ \ and \ \ y_1,...,y_l \in M\Big). \end{aligned}$$

We claim that for $l = 1, \cdots, 100$

$$\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr(x_1,...,x_l \in M \quad and \quad y_1,...,y_l \in M)$$
$$= \sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr(x_1,...,x_l \in M)Pr(y_1,...,y_l \in M) + O(\log^{-20} q) \ ,$$

and for $l = 101, 102$ both of

$$\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr(x_1,...,x_l \in M \quad and \quad y_1,...,y_l \in M)$$

and

$$\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr(x_1,...,x_l \in M)Pr(y_1,...,y_l \in M)$$

are $O(\log^{-15} q)$. It is easy to see that this claim implies the lemma.

Since x_i, y_j are not all distinct, it is convenient to rearrange them, say $\alpha_1, \cdots, \alpha_r$, $\beta_1, \cdots, \beta_s, \gamma_1, \cdots, \gamma_t$ so that α_i's are distinct points in $A(u) \cup A(v)$, and all β_i, γ_i's are in $T(u) \cup T(v)$ and distinct. Moreover, each β_i does not use any previously used point (in $\alpha_1, \cdots, \alpha_r, \beta_1, \cdots, \beta_{i-1}$) and each γ_i contains exactly one previously used point. The rest must not contain any unused point and will not be rearranged. Since l is bounded there are only $O(1)$ number of $\{x_i, y_i\}_{i=0,\cdots,l}$ which yield given $\{\alpha, \beta, \gamma\}$. Clearly,

$$Pr(x_1, \ldots, x_l, y_1, \cdots, y_l \in M) \leq Pr(x_1, \cdots, x_l, y_1, \cdots, y_l \in B) = p^{r+2s+t} \ ,$$

and

$$Pr(x_1, \cdots, x_l \in M)Pr(y_1, \cdots, y_l \in M) \leq Pr(x_1, \cdots, x_l \in B)Pr(y_1, \cdots, y_l \in B) \leq p^{r+2s+t} \ .$$

Each α_i is in $A(u) \cup A(v)$ and there are $O(abq^{3/2})$ possible choices of α_i. Similarly, there are $O(b^2q^3)$ choices for each β_i. For each γ_i, there are only O(1) choices for the used point and once the used point z is chosen the unused point must be in lines (uz) or (vz). Since each line contains $O(\log^{c_1} q)$ surviving points (**(3)** of Phase 1), there are $O(\log^{c_1} q)$ choices for γ_i. All together, we have $O((abq^{3/2})^r (b^2q^3)^s (\log^{c_1} q)^t)$ choices. If one requires $\alpha_i \in A(u) \cap A(v)$ or $\beta_i \in T(u) \cap T(v)$, then the number of choices reduces to $O(\log^{c_1+3} q)$ (**(6)** of Phase 1) or $O(\log^{2c_1} q)$ (**(3)** of Phase 1) respectively. Notice that $pabq^{3/2} = O(\log^{-3/2} q)$, $p^2b^2q^3 = O(\log^{-2} q)$, and $p \log^{c_1} q = O(\log^{-30} q)$. Therefore, the sum of all cases other than $t = 0$, $\alpha_i \notin A(u) \cap A(v)$ for all α_i, and $\beta_i \notin T(u) \cap T(v)$ for all β_i, would be negligible, say $O(\log^{-30} q)$. Furthermore, if $l \geq 100$, then there are at least 10 distinct x_i or y_i. This implies that $t \neq 0$ or $r + 2s \geq 10$ and so both of

$$\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr(x_1,...,x_l \in M \quad and \quad y_1,...,y_l \in M)$$

and

$$\sum_{(x_1,y_1),\cdots,(x_l,y_l)} Pr(x_1,...,x_l \in M)Pr(y_1,...,y_l \in M)$$

are $O(\log^{-15} q)$.

Suppose $t = 0$, $\alpha_i \notin A(u) \cap A(v)$ for all α_i, and $\beta_i \notin T(u) \cap T(v)$ for all β_i. Then $\{x_i\}$ and $\{y_i\}$ do not share any single point, which yields

$$Pr\Big(x_1, \cdots, x_l, y_1, \cdots, y_l \in B\Big) = Pr\Big(x_1, \cdots, x_l \in B\Big) Pr\Big(y_1, \cdots, y_l \in B\Big) \ . \tag{29}$$

(It is still possible $x_i = x_j$.) Now consider

$$Pr(x_1, ..., x_l \in M \ and \ y_1, ..., y_l \in M | x_1, ..., x_l \in B \ and \ y_1, ..., y_l \in B) \ .$$

If we impose any condition

$$w \in A(z), \ or \quad (w, w') \in T(z), \tag{30}$$

for some points z, w, w' consisting of x and y, a similar argument as above would yield that the corresponding sum is negligible. We exclude these cases too. Let F_w be the event $\{A(w) \cap B = \emptyset$ and $T(w) \cap B = \emptyset\}$. Then clearly

$$\begin{aligned} &Pr(x_1, ..., x_l \in M \ and \ y_1, ..., y_l \in M | x_1, ..., x_l \in B \ and \ y_1, ..., y_l \in B) \\ &= Pr(\textstyle\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j} | x_1, ..., x_l \in B \ and \ y_1, ..., y_l \in B) \ . \end{aligned}$$

Once $z \in B$, the conditions $x_i \in B$ and $y_i \in B$ make the event $z \in M$ less likely. (One may apply FKG inequality though a direct coupling argument would be easier.) Thus

$$Pr\Big(\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j} | x_1, ..., x_l \in B \ and \ y_1, ..., y_l \in B\Big) \le Pr\Big(\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j}\Big) \ .$$

On the other hand, if all surviving points in lines containing z and some point consisting of x_i or y_i were not in B, say the set of such points is $R(z)$, the conditions are irrelevant to the event $z \in M$ unless (30) holds for w, w' consisting x_i and y_i, which we have excluded. Thus for $R = \cup_z R(z)$ where the union is taken all points z consisting x_i and y_i,

$$\begin{aligned} &Pr\Big(\textstyle\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j} | x_1, ..., x_l \in B \ and \ y_1, ..., y_l \in B\Big) \\ &= Pr\Big(R \cap B = \emptyset\Big) Pr\Big(\textstyle\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j} | R \cap B = \emptyset\Big) \\ &\ge Pr\Big(R \cap B = \emptyset\Big) Pr(\textstyle\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j}) \ . \end{aligned}$$

(The inequality again uses an FKG type argument.) Consequently, since $|R| \le (4l)^2 \log^{c_1} q$ and $p|R| = O(\log^{-30} q)$,

$$\begin{aligned} &Pr\Big(\textstyle\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j} | x_1, ..., x_l \in B \ and \ y_1, ..., y_l \in B\Big) \\ &\ge (1 - O(\log^{-20})) Pr(\textstyle\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j}) \ . \end{aligned}$$

Finally, it is sufficient to show that

$$Pr(\bigcap_i F_{x_i} \cap \bigcap_j F_{y_j}) = Pr\Big(\bigcap_i F_{x_i}\Big) Pr\Big(\bigcap_j F_{y_j}\Big) \ + O(\log^{-20} q) \tag{31}$$

and

$$Pr\Big(\bigcap_i F_{x_i} | x_1, \cdots, x_l \in B\Big) = (1+O(\log^{-20}))Pr\Big(\bigcap_i F_{x_i}\Big)$$

except for negligible cases. However, we may use exactly the same argument as above. For example, consider

$$Pr\Big(\overline{\bigcap_i F_{x_i}} \cap \overline{\bigcap_j F_{y_j}}\Big) = Pr\Big(\bigcup_{i,j} \bar{F}_{x_i} \cap \bigcup \bar{F}_{y_j}\Big)$$

and let

$$K = \bigcup_i (A(x_i) \cup T(x_i)) \ , \ and \ L = \bigcup_j (A(y_i) \cup T(y_i)) \ .$$

Then

$$Pr\Big(\bigcup_{i,j} \bar{F}_{x_i} \cap \bar{F}_{y_j}\Big) = Pr\Big(\bigcup_{(w,z)\in K\times L} \{w \in B\} \ and \ \{z \in B\}\Big) \ .$$

Now the same argument used for (29) would yield (31).

5.2.3 Proof of (3)

By the induction hypothesis, we can assume that

$$a_i b_i q^{3/2}(1-K(i-1)\theta^2) \le |A_i(v)| \le a_i b_i (1+K(i-1)\theta^2),$$

where K is a constant larger than 16.

Let $U_i = B_i \backslash M_i$. Denote by $U_i(v)$ ($B_i(v)$, $M_i(v)$) the set of points x in S_i such that there is $u \neq x \in U_i$ (B_i, M_i respectively) satisfying that x, u, v are co-linear. Next, let $B_i'(v)$ and $M_i'(v)$ be the intersection of $B_i(v)$ and $M_i(v)$ with S_{i+1}.

As usual, we set $L = A_i(v)$, and $L' = L \cap S_{i+1}$. Since A_i and A_{i+1} are arcs, we have for any $v \in \Omega_{i+1}$

$$|A_{i+1}(v)| = |L'| + |M_i'(v)|.$$

By the usual argument, it is easy to show that $|L'|$ sufficiently concentrates around its expected value. The hard part of the proof is to estimate $M_i'(v)$. For the purpose of this section, set $\delta = \log^{-10} q$. We shall use the following observation to bound $M_i'(v)$.

$$|B_i'(v)| - |U_i(v)| \le |M_i'(v)| \le |B_i'(v)| \le |B_i(v)|.$$

The next three claims give bounds on $|B_i(v)|$, $B_i'(v)|$ and $|U_i(v)|$, respectively.

Claim 5.4 *With very high probability,*

$$|B_i(v)| \le \theta b_i q^{3/2}(1+o(\delta)).$$

Proof. By definition, $x \in B_i(v)$ if there is $j \in (xv)$ such that $j \in B_i$, namely, $t_j = 1$. Therefore,

$$|B_i(v)| \le \sum_{x\in S_i} (\sum_{j,[jxv]} t_j). \tag{32}$$

Every t_j appears in the double sum exactly m_j times, where m_j is the number of points on the line (vj) (excluding v and j). Therefore, the right hand side in (32) can be rewritten as $\sum_{x\in S_i} t_x m(x) = Y$.
Now we apply Corollary 4.3 to bound Y. Notice that $\sum_{x\in S_i} m(x)$ is exactly the quantity $T_i'(v) = 2T_i(v)$ defined in the previous subsection. So we have

$$\mathbb{E}_0(Y) = 2p_iT_i(v) \sim p_ib_i^2q^3(1+o(\delta)) \sim \theta b_iq^{3/2} \geq \log^{150} q, \tag{33}$$

as $b_iq^{3/2} \geq \log^{300} q$. Moreover when we start Phase two, each line has roughly $\log^{c_1} q = \log^{100} q$ points, therefore

$$\mathbb{E}_1(Y) \leq \max_x m(x) \leq \max_l |S_i(l)| \leq 2\log^{c_1} q.$$

Thus, Corollary 4.3 implies that with very high probability

$$|B_i(v)| \leq |\mathbb{E}_0(Y)|(1+o(\delta)).$$

Using the estimate of $|T_i(v)|$ in Property **(2)** together with (33), we have $\mathbb{E}_0(Y) \leq \theta b_iq^{3/2}(1+o(\delta))$, concluding the proof. □

Claim 5.5 *With very high probability*

$$|B_i'(v)| \geq \theta b_iq^{3/2}(1-o(\delta)) - 8a_i\theta^2 b_iq^{3/2}.$$

Proof. Note that $x \in B_i'(v)$ if and only if $x \in S_{i+1}$ and there is at least one point on (xv) (different from x and v) belonging to B_i. Let us denote by $B_i''(v)$ the set of $x \in S_{i+1}$ so that there is *exactly* one point on (xv) with this property. It is clear that $|B_i'(v)| \geq |B_i''(v)|$. We shall prove that even $|B_i''(v)|$ satisfy the claim. The trick here is that while the restriction makes $B_i''(v)$ easy to handle, we do not lose too much as the probability that B_i intersect any line in more than one point is negligible. We can bound $|B_i''(v)|$ by a polynomial as follows

$$|B_i''(v)| \geq \sum_{x\in S_i, |(xv)|\geq 2} \mathbf{1}_{\{x\in S_{i+1}\}} \sum_{j,[jxv]} t_j(1 - \sum_{j'\neq j,[j'xv]} t_{j'}), \tag{34}$$

where $|(xv)| \geq 2$ means that the line (xv) (including x and v) has at least three points. Moreover, taking into account the reasons that make a point deleted, we have

$$\mathbf{1}_{\{x\in S_{i+1}\}} \geq 1 - (\sum_{g\in A_i(x)} t_g + \sum_{g,g',[gg'x]} t_gt_{g'}).$$

So

$$|B_i''(v)| \geq \sum_{x\in S_i, |(xv)|\geq 2} \Big(1 - (\sum_{g\in A_i(x)} t_g + \sum_{g,g',[gg'x]} t_gt_{g'})\Big)\Big(\sum_{j,[jxv]} t_j(1 - \sum_{j'\neq j,[j'xv]} t_{j'})\Big).$$

Next, we expose the product and then split the right hand side as a sum of two terms, where main term is $\sum_{x,|(xv)|\geq 2}\sum_{j,[jxv]} t_j$ and the error term contains everything else. Similar to the proof of Claim 5.4, we can show that with very high probability the main term is at least $\theta b_iq^{3/2}(1-o(\delta))$. Using Lemma 4.2, we can also prove that the error term is at least $-8a_i\theta^2 b_iq^{3/2}$. We omit the technical but rather straightforward calculation. □

Claim 5.6 *With very high probability,*

$$|U_i(v)| \le 6\theta^2 a_i b_i q^{3/2}.$$

Proof. Note that for any $x \in U_i(v)$ there should be a point $j \in (xv)$ different from both x and v such that $j \in B_i$ but $j \notin M_i$. Therefore

$$|U_i(v)| \le \sum_{x,|(xv)|\ge 2} \sum_{j,[jxv]} t_j \mathbf{1}_{\{j\notin M_i\}}.$$

By the description of the algorithm, there are two possible reasons that exclude j from M_i: either there are j' and j'' in B_i such that j, j', j'' are co-linear, or there is $j' \in B_i$ such that (jj') intersects A_i. So

$$\sum_{x,|(xv)|\ge 2} \sum_{j,[jxv]} t_j \mathbf{1}_{\{j\notin M_i\}} \le \sum_{x,|(xv)|\ge 2} \sum_{j,[jxv]} t_j \Big(\sum_{j',j'',[jj'j'']} t_{j'} t_{j''} + \sum_{j'\in A_i(j)} t_{j'} \Big).$$

Split the right hand side into two terms α and β, where

$$\alpha = \sum_{x,|(xv)|\ge 2} \sum_{j,[jxv]} t_j \Big(\sum_{j',j''.,[jj'j'']} t_{j'} t_{j''} \Big).$$

$$\beta = \sum_{x,|(xv)|\ge 2} \sum_{j,[jxv]} t_j \Big(\sum_{j'\in A_i(j)} t_{j'} \Big).$$

Using Lemma 4.2 (or Corollary 4.3), it is relatively simple to show that with very high probability $\alpha \le 5\theta^3 b_i q^{3/2} = o(\theta^2 a_i b_i q^{3/2})$ and $\beta \le 5\theta^2 a_i b_i q^{3/2}$ (again the details are omitted). The claim follows instantly. □

Claim 5.7 *If the alogithm runs perfectly up to step $i-1$, then $b_i/b_{i+1} \le 1+2\theta a_i$.*

Proof. By definition

$$\frac{b_i}{b_{i+1}} = \frac{1}{1-P_i^u} = \frac{1}{1-p_i \max_v |A_i(v)|(1+o(1))},$$

due to Lemma 3.2 (note that in this lemma $p_i \max_v |A_i(v)|$ is the dominating term in the upper bound). Next, due to the induction hypothesis $|A_i(v)| \sim a_i b_i q^{3/2} \le \frac{3}{2} a_i b_i q^{3/2}$,

$$p_i \max_v |A_i(v)| \le \theta (b_i q^{3/2})^{-1} \frac{3}{2} a_i b_i q^{3/2} \le \frac{3}{2}\theta a_i.$$

Therefore

$$\frac{1}{1-p_i \max_v |A_i(v)|(1+o(1))} \le 1+2\theta a_i,$$

since $\theta a_i = o(1)$. This completes the proof. □

Now we are ready to bound $|M_i'(v)|$. By Claims 5.4 and 5.7, we have, with very high probability, that

$$\begin{aligned} |M_i'(v)| &\leq |B_i(V)| \leq \theta b_i q^{3/2}(1+o(\delta)) \\ &\leq \theta b_{i+1} q^{3/2}(1+2\theta a_i)(1+o(\delta)) \\ &\leq \theta b_{i+1} q^{3/2}(1+3\theta a_i). \end{aligned} \tag{35}$$

On the other hand, by the second and third claims

$$\begin{aligned} |M_i'(v)| &\geq |B_{i+1}'(v)| - |U_i(v)| \geq \theta b_i q^{3/2}(1-o(\delta)) - 8\theta^2 a_i b_i q^{3/2} - 6\theta^2 a_i b_i q^{3/2} \\ &\geq \theta b_{i+1} q^{3/2}\Big(1-15\theta a_i\Big). \end{aligned} \tag{36}$$

In (35) and (36), we use the fact that $\delta = \log^{-10} q \ll \theta a_i$. This way, the contribution of $o(\delta)$ is swallowed by an additional θa_i.

Finally, let us estimate $|L'|$. We use Corollary 4.8; in order to apply this corollary, we first need to estimate $a(L)$ and here Property **(6)** of the first phase becomes crucial. Recall that $L = A_i(v)$ so $a(L) = \max_{u,v} |A_i(u,v)|$. On the other hand, due to Property **(6)** of Phase one, for any u and v

$$|A_i(u,v)| \leq i b_i q + i \log^{40} q.$$

Since $i \leq N \leq \log^3 q$ and $b_i q \leq \log^{c_1} q = \log^{100} q$, we conclude that $a(L) \leq \log^{103} q$. As

$$\mathbb{E}(|L'|) \sim a_i b_{i+1} q^{3/2} \geq a_i \log^c q = a_i \log^{300} q \geq \log^{298} q,$$

Corollary 4.8 applies and implies that with very high probability

$$a_i b_{i+1} q^{3/2}(1-K(i-1)\theta^2)(1-o(\delta)) \leq |L'| \leq a_i b_{i+1}(1+K(i-1)\theta^2)(1+o(\delta)). \tag{37}$$

The expectation of $|L'|$ is bounded between $b_{i+1} q^{3/2}(1 \pm K(i-1)\theta^2)$; the error term $o(\delta)$ results from the deviation tail. By (37) and (36), we now have with very high probability

$$\begin{aligned} |A_{i+1}(v)| = |L'| + |M_i'(v)| &\geq a_i b_{i+1} q^{3/2}(1-K(i-1)\theta^2)(1-o(\delta)) + \theta b_{i+1} q^{3/2}(1-15\theta a_i) \\ &\geq b_{i+1} q^{3/2}(a_i + \theta - K a_i i \theta^2 - 15 a_i \theta^2) \\ &\geq b_{i+1} q^{3/2}(a_{i+1} - \theta^3 - K a_i i \theta^2 - 15 a_i \theta^2). \end{aligned}$$

In the second inequality, the difference between $K a_i i$ and $K a_i (i-1)$ swallows the contribution of $o(\delta)$. For the third inequality, we use the fact that $a_{i+1} \leq a_i + \theta + \theta^3$ (see Remark 5.1). As $a_{i+1} \geq a_i$ and $K \geq 16$, it follows that

$$|A_{i+1}(v)| \geq a_{i+1} b_{i+1} q^{3/2}(1-K(i+1)\theta^2).$$

For the upper bound, it follows from (37) and (35) that

$$\begin{aligned} |A_{i+1}(v)| &\leq a_i b_{i+1} q^{3/2}(1+K(i-1)\theta^2)(1+o(\delta)) + \theta b_{i+1} q^{3/2}(1+3\theta a_i) \\ &\leq b_{i+1} q^{3/2}(a_i + \theta + K a_i i \theta^2 + 3 a_i \theta^2). \end{aligned}$$

Again notice that the difference between i and $i-1$ swallows the contribution of $o(\delta)$. Due to Remark 5.1

$$a_{i+1} \geq a_i + \theta - (3a_i\theta^2 + 10\theta^3) \geq a_i + \theta - 4a_i\theta^2,$$

which implies

$$\begin{aligned} |A_{i+1}(v)| &\leq a_{i+1}b_{i+1}q^{3/2}(1+Ki\theta^2+7\theta^2) \\ &\leq a_{i+1}b_{i+1}q^{3/2}(1+K(i+1)\theta^2), \end{aligned}$$

completing the proof. □

6 REMARKS AND OPEN QUESTIONS

The constant c used in the proof is fairly large ($c = 300$). With a tighter analysis, we can reduce it significantly ($c = 10$ is possible). On the other hand, it is not clear that we can obtain $c = 1/2$. Achieving $c = 1/2$ would be best possible with respect to our method and would imply that $n(\mathcal{P}) = O(q^{1/2}\log^{1/2} q)$.

While the best lower bound for $n(\mathcal{P})$ is still linear in $q^{1/2}$, we feel that the truth might be $\omega(q^{1/2})$ (perhaps, even $\Theta(q^{1/2}\log^{1/2} q)$. (Here the asymptotic notation is used assuming $q \to \infty$). The first step toward showing $n(\mathcal{P}) = \omega(q^{1/2})$ would be to prove that there is a point which is covered by $\omega(1)$ secants. Even this does not seem to be known and we make the following conjecture.

Conjecture 1. *Given a plane $n(\mathcal{P})$ of order q and a complete arc A in it. Then there is a point in $\mathcal{P}$ which is covered by $\omega(1)$ secants of A, where $\omega(1)$ tends to infinity with q.*

The proof of Theorem 1.2 gives us a way to generalize a complete arc. Although this arc is not completely random, the algorithm suggests that it is close to be one. If it was the case then it would imply that Fisher's conjecture on the average size of a complete arc is true, up to a polylog factor.

Question 2. *Is it true that there is a constant c such that for any plane $\mathcal{P}$ a random complete arc A (chosen uniformly from the set of all complete arcs in $\mathcal{P}$) has, with probability close to 1, at most $q^{1/2}\log^c q$ points, where q is the order of $\mathcal{P}$?*

Corollary 1.3 asserts that there is always a complete arc of size between $\frac{1}{c}q^{1/2}\log^{1/2} q$ and $q^{1/2}\log^c q$, for some constant c. Somewhat surprisingly, our method cannot be used to produce a larger complete arc. For instance, it is still not clear whether there is a complete arc of size between $q^{1/2+\varepsilon}$ and $q^{1/2+\varepsilon}\log^c q$, for any small constant ε and any constant c. On the other hand, results of Hadnagy and Szőnyi shown that for the Galois plane $PG(2,q)$, the possible sizes of a complete arc is dense in the interval $[q^{3/4}, q]$ (see [53] and the references therein). It is interesting to prove a similar result for the interval $[q^{1/2}, q^{3/4}]$.

Acknowledgement. The authors are grateful to Professor J. Kahn for communicating this problem.

References

[1] V. Abatangelo, *A class of complete ((q+8)/3)-arcs of $PG(2,q)$, with $q = 2^h$ and $h \geq 6$ even* , Ars Combin., 16 (1983), 103-111.

[2] M. Ajtai and J. Komlós and E. Szemerédi", *A dense infinite Sidon sequence*, Europ. J. Combi. 2 (1981), 1-11.

[3] N. Alon and J. Spencer, "The Probabilistic Methods," Wiley, NewYork, 1992.

[4] N. Alon, B. Bollobás, J.H. Kim and V. H. Vu, *Economical covers and geometric applications,* submitted.

[5] N. Alon, J. H. Kim and J. Spencer, *Nearly perfect matchings in regular simple hypergraphs*, IJM 100 (1997), 171–187.

[6] S. Ball, *On small complete arcs in a finite plane*, Discrete Math, **174** (1997), 29–34.

[7] U. Bartocci, *k-insiemi densi nei piani di Galois,* Boll. Un. Mat. Ital. D (6), **2** (1983), 71–77.

[8] A.A. Bruen and J.C. Fisher, *Blocking sets and complete arcs*, Pacific J. Math. 53 (1974), 73-84.

[9] A. Blokhuis, *Polynomials in finite geometry and combinatorics*, Survey in Combinatorics, Cambridge Univ. Press, Cambridge (1993), 35-52.

[10] A. Blokhuis, *Extremal problems in finite geometries*, Extremal problems for finite sets, Bolyai Soc. Mathematical studies 3, (1994) 111-135.

[11] A. Blokhuis, *Blocking sets in desarguesian planes*, Erdős is eighty, Bolyai society mathematical studies, 2, eds D. Miklós, V.T. Sós, T. Szőnyi (1996) 133-155.

[12] A.E. Brouwer, *On the size of a maximum transversal in a Steiner Triple System*, Canadian J. of Math. 33 (1981), 1202-1204.

[13] A.A. Bruen and M.J. de Resmini, *Blocking sets, k-arcs and nets of order ten*, Combinatorics 81, North-Holland Math. Stud. 78, North-Holland, Amsterdam-NewYork (1983) 169-175.

[14] A.A. Bruen and M.J. de Resmini, *Blocking sets in affine planes*, Annals of Discrete Math. 18 (1983) 169-176.

[15] A.A. Bruen and R. Silverman, *Arcs and Blocking sets II*, Euro. J. Combi. 8 (1987), 351-356.

[16] P. Dembowski, Finite geometries, Springer 1968.

[17] C. Di Comite, *Su k-archi deducibili da cubiche piane* , Atti dell' Accad. Naz Lincei Rend. (8) 33 (1962), 429-435.

[18] N. G. De Brujin and P. Erdős, *On a combinatorial problem*, Indag. Math, 10 (1948) 421-423.

[19] C. Di Comite, *Su k-archi deducibili da cubiche piane* , Atti dell' Accad. Naz Lincei Rend. (8) 35 (1963), 274-278.

[20] C. Di Comite, *Intorno a certi* $(q+9)/2$*-archi de* $S_{q,2}$, Atti dell'Accad. Naz. Lincei Rend. (8) 47 (1967) 240-244.

[21] J.C. Fisher, *Random k-arcs*, Preliminary report, 1989.

[22] J.C. Fisher, J.W.P. Hirschfeld and J.A. Thas, *Complete arcs in planes of square order*. Annal of Discrete Math 30 (1986) 243-250.

[23] P. Frankl and V. Rödl, *Near perfect coverings in graphs and hypergraphs*, Europ. J. Combinatorics 6 (1985), 317-326.

[24] D. A. Grable, *More-than-nearly perfect packings and partial designs*, Combinatorica 19(1999), 221-239.

[25] R. Graham, M. Grötchel ad L. Lovász eds, Handbook of Combinatorics, Chapter 13, North Holland, 1995.

[26] H. R. Halder, *Zur Existenz von k-Kurven in endlichen Ebenen,* J. Geom., **14** (1980), 71–74.

[27] J.W.P. Hirschfeld, Projective geometries over finite fields, Claredon Press, Oxford (1971).

[28] J.W. P. Hirschfeld, *Algebraic curves and arcs over finite fields* , Quad. del Dip. di Mat. Lecce, Q-6 (1987).

[29] J.W.P. Hirschfeld, *Maximum sets in finite projective spaces*, Survey in combinatorics, LMS Lecture Note Series, 82, Cambridge Univ. Press (1983) 55-76.

[30] A. Johansson, *An improved upper bound on the choice number for triangle free graphs*, submitted.

[31] A. Johansson, *The choice number of sparse graphs*, preprint.

[32] L. Kadison and M. T. Kronmann, "Projective Geometry and Modern Algebra," Birkhäuser, Boston, 1996.

[33] J. Kahn, *Asymptotically good list colorings*, J. of Combi. Th. A. **73** (1996), 1-59.

[34] J. H. Kim, *On Brooks' theorem for sparse graphs,* Combinatorics, Probability and Computing **4** (1995), 97-132.

[35] J.H. Kim, *The Ramsey number $R(3,t)$ is $t^2/\log t$*, Random structures and algorithms, 7 (1995), 173-207.

[36] J.H. Kim, *Steiner partial systems,* preprint.

[37] J. H. Kim and V. H. Vu, *Concentration of multivariate polynomials and its applications*, submitted to this Journal.

[38] J. H. Kim and V. H. Vu, *On the number of triangles in $G(n,p)$,* in preparation.

[39] G. Korchmáros, *New example of complete k-arcs in $PG(2,q)$* , Europ. J. Comb. 4 (1983), 329-334.

[40] S. J. Kovács, *Small saturated sets in finite projective planes,* Rend. Mat. Appl. (7), **12** (1992), 157–164.

[41] L. Lunelli and M. Sce, *Considerazioni arithmetiche e risultati sperimentali sui* $\{K;n\}_q$ *archi*, Ist. Lombardo Accad. Sci. Rend. A 98 (1964), 3-52.

[42] L. Lombardo-Radice, *Sui problema dei k-archi completi di* $S_{2,q}$, Boll. Un. Mat. Ital. 11 (1956), 178-181.

[43] M. Molloy and B. Reed, *A bound on the total chromatic number*, Combinatorica **18** (1998), 241-280.

[44] N. Pippenger and J. Spencer, "Asymptotic behavior of the chromatic index for hypergraphs", J. of Combi. Th. A. 51 (1989), 24-42.

[45] B. Reed and B. Sudakov, *Asymptotically the list constants are 1*, submitted.

[46] V. Rödl, *On a packing and covering problem*, Europ. J. Combi. 5 (1985), 69-78.

[47] R. Schoof, *Non-singular plane cubic curves over finite fields*, J. Combin. Theory Ser. A, 46 (1987), 183-211.

[48] B. Segre, *Le geometrie di Galois,* Ann. Mat. Pura Appl. 48 (1959), 1-97.

[49] B. Segre, *Introduction to Galois geometry,* ed. J.W.P. Hirschfeld, Mem. Accad. Naz. Lincei, (8), 1967, 133-263.

[50] B. Segre, *Ovali e curve* σ *nei piani de caratteristica due*, Atti dell'Accad. Naz. Lincei Rend. (8) 32 (1962), 785-790.

[51] T. Szőnyi, *Small complete arcs in Galois planes*, Geom. Dedicata 18 (1985), 161-172.

[52] T. Szőnyi *Arcs, caps, codes and 3-independent subsets*, Giornate di Geometrie Combinatorie, (edits G. Faina, G. Tallini) Uni. Perugia (1993), 57-80,

[53] T. Szőnyi, *Some applications of algebraic curves in finite geometry and combinatorics*, Surveys in Combinatorics 1997, 197-236, LMS Lecture Notes **241**, Cambridge Univ. Press, 1997.

[54] T. Szőnyi, *Complete arcs in* $PG(2,q)$, a survey in Quad. del Sem. Geom. Comb. Univ. di Roma ("La Sapienza"), 94 (1989).

[55] T. Szőnyi, *Arcs in cubic curves and 3-independent subsets of abelian groups*, in Combinatorics, Eger Colloquia Mathematica Societatis János Bolyai, 52, North-Holland, Amsterdam (1987), 499-508.

[56] T. Szőnyi, *Small complete arcs in André planes of square order*, Graphs and Comb., **8** (1992), 81–89.

[57] E. Ughi, *Saturated configurations of points in projective Galois spaces,* Eur. J. Comb., **8** (1987), 325–334.

[58] J.F. Voloch, *On the completeness of certain plane arcs II*, Europ. J. Comb., 11 (1990) 491-496.

[59] J.F. Voloch, *Complete arcs in Galois planes of non-square order*, Advance in Finite Geometries and Designs (eds J.W.P. Hirschfeld, D.R. Huges, J.A. Thas) Oxford Univ. Press, Oxford (1991) 401-406.

[60] V. H. Vu, *Concentration of non-Lipschitz functions and applications,* submitted.

[61] V. H. Vu, *New bounds on nearly perfect matchings in hypergraphs: higher codegrees do help,* Random Structures and Algorithms, **17** (2000), 29-63.

[62] V. H. Vu, *On a refinement of Waring's problem,* Duke Math. Journal, **105** (2000), 107-134.

[63] V. H. Vu, *On the concentration of multivariate polynomials with small expectation,* Random Structures and Algorithms, **16** (2000), 344-363.

[64] V. H. Vu, *On some degree conditions which guarantee the upper bound of chromatic (choice) number of random graphs,* Journal of Graph Theory, **31** (1999), 201-226.

[65] V. H. Vu, *A general upper bound on the list chromatic number of locally sparse graphs,* to appear in Combinatorics, Probability and Computing.

[66] V. H. Vu, *A large deviation result on the number of small subgraphs of a random graph,* Combinatorics, Probability and Computing, **10** (2001), 79-94.

[67] F. Zirilli, *Su una classe di k-archi di un piano di Galois*, Rend. Accad. Naz. Lincei, 54 (1973) 393-397.