- *Monotonicity: The family UC_{random} is monotone (regardless of $\mathcal{V}$ and N).*
- *Spread: For any item $i \in I$, $spread_f(\mathcal{V}, i) = O(t \log(Nk))$ with probability greater than $1 - 1/N$ over the choice of $f \in UC_{random}$.*
- *Load: For any bucket $b \in B$, $load_f(\mathcal{V}, b) = O\left(\left(\frac{|I|}{T} + 1\right) t \log(Nmk)\right)$ with probability greater than $1 - 1/N$ over the choice of $f \in UC_{random}$.*
- *Balance: For any fixed view V and item i, the probability that item i is mapped to bucket b in view V is $O\left(\frac{1}{|V|}\left(\frac{\log(N|V|)}{m} + 1\right)\right) + \frac{1}{N}$.*

The balance claim of theorem 2.2.3 requires clarification. Note that if we choose $m = \Omega(\log(|V|))$, and $N = poly(|V|)$, then the bound simplifies into $O(1/|V|)$ which gives the definition of the balance property.

The proof of theorem 2.2.3 is presented in the remainder of this section as a series of lemmas.

In a number of the proofs, the Chernoff bound (See appendix A) is used where a more direct method could be applied. The reason for this is that in section 2.2.4 the family UC_{random} is modified so that the mapping of points to the circle is not completely random. The modification will be such that the Chernoff bounds will still hold; thus the proofs presented in this section remain valid. Cases in which superfluous use of Chernoff bounds are made are highlighted, and should not make reading the proofs any harder.

Intuitively, the family is monotone since when a new bucket is added, the only items that move are those that are now closest clockwise to points associated with the new bucket. The proof of monotonicity is simple and is given in the following Lemma.

Lemma 2.2.4 *The family UC_{random} is monotone.*

Proof:

Let $V_1 \subseteq V_2 \subseteq B$ be two views of the buckets. Let f be any function in UC_{random}. We need to show that $f_{V_2}(i) \in V_1$ implies $f_{V_1}(i) = f_{V_2}(i)$. Now, $f_{V_2}(i) \in V_1$ implies

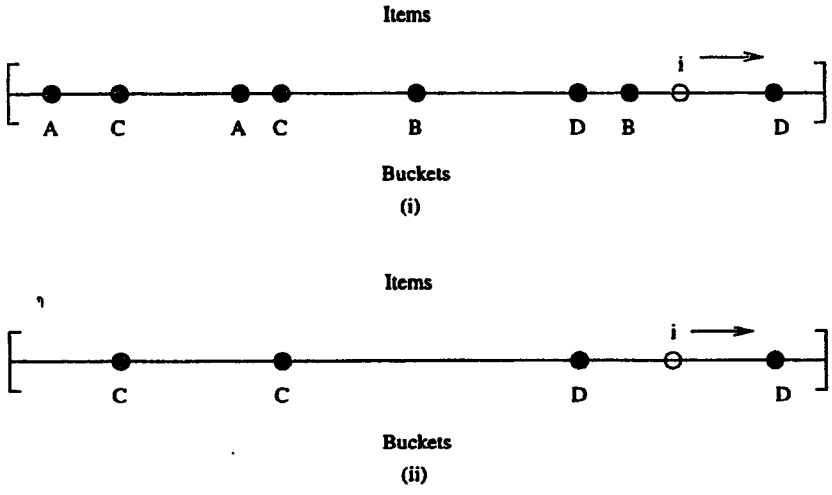

Figure 2-11: Monotonicity for the family UC_{random}. In this figure the unit circle is depicted by an interval of length one, which is obtained by cutting the unit circle at an arbitrary point. (i) The mapping of points to the circle for a view $V_2 = \{A, B, C, D\}$ ($m = 2$ in this example). The closest bucket point clockwise of i's point is one associated with the bucket D. (ii) For any view $V_1 \subseteq V_2$ containing the bucket D (here $V_1 = \{C, D\}$), the point closest to i's point will still be D.

that when adding buckets to V_1 to get V_2, none of the points that we add to the unit circle fall in the arc between i's point and the bucket point that it was previously closest to in V_1. Thus, the item i must be mapped to the same bucket in both V_1 and V_2 (see figure 2-11).

■

Before showing the bound on spread, a technical lemma is derived which is used in both the spread and load bound proofs. The lemma shows that an arc does not have to be very long if we want there to be high probability that at least one bucket point from *every* view falls into the arc.

Lemma 2.2.5 *Any fixed set of measure at least* $\frac{4t\log(2Nk)}{Tm}$ *in the unit circle contains at least one bucket point from every view with probability greater than* $1 - \frac{1}{2N}$

Proof:

Note that the probability is over the choice of the function r_B. Since in the family UC_{random} all functions are equally likely, we can assume in the proof that points for the buckets are distributed uniformly and at random around the unit circle.

Let X_j be a random variable denoting the number of points associated with buckets in view V_j that fall into a set of measure l. There are at least $M = \frac{mT}{t}$ bucket points associated with each view (since every view contains at least T/t buckets). However, we can assume that there are exactly M points in each view since more points would imply that a set with smaller measure would suffice (more precisely, the distribution of X_j when there are more than M points stochastically dominates the distribution when there are M points). Thus, we have $E[X_j] = Ml$. We will choose the value of l so that the probability of X_j being 0 is at most $1/2Nk$.

Following is a use of the Chernoff bound where we could have made a more direct argument. [5]

$$\begin{aligned} \Pr[X_j = 0] &\leq \Pr[|X_j - Ml| \geq Ml] \\ &\leq 2e^{-\frac{Ml}{4}} \end{aligned}$$

So we choose l so that $2e^{-\frac{Ml}{4}} = \frac{1}{2Nk}$. We obtain $l = \frac{4\log(4Nk)}{M} = \frac{4t\log(4Nk)}{Tm}$.

From the union bound we have:

$$\Pr[\ \textit{Some}\ X_j = 0] \leq \sum_{j=1}^{k} \Pr[X_j = 0] \leq k\frac{1}{2Nk} = \frac{1}{2N}$$

Thus, any set of measure $\frac{4t\log(4Nk)}{Tm}$ contains a bucket point from every single view with probability at least $1 - 1/2N$.

■

We now show the bound on spread.

Lemma 2.2.6 *For any item $i \in I$, $spread_f(\mathcal{V}, i) = O(t\log(Nk))$ with probability greater than $1 - 1/N$ over the choice of $f \in UC_{random}$.*

[5] The direct argument is $\Pr[X_j = 0] = (1 - l)^M$ since each point is mapped independently to the circle.

Proof:

The proof of the lemma uses a technique that reoccurs many times in this paper. The basic idea is simple. If we want to show that some event A has low probability, then we can proceed as follows:

1. Find a set of events $\mathcal{B}$ so that the occurrence of event A implies that at least one of the events in $\mathcal{B}$ occurs.

2. Show that there is only a small probability that *any* event in $\mathcal{B}$ occurs.

3. Deduce that the probability of A must also be small.

More formally stated, the technique is to find a set of events $\mathcal{B}$ so that $A \subset \bigcup_{B_i \in \mathcal{B}} B_i$. This implies that:

$$\Pr[A] \leq \Pr\left[\bigcup_{B_i \in \mathcal{B}} B_i\right] \leq \sum_{B_i \in \mathcal{B}} \Pr[B_i]$$

So, if the sum of the probabilities of the events in $\mathcal{B}$ is small then the probability of A must also be small.

The first step is to define the event A and the family of events $\mathcal{B}$. In our case, A is the event that the bound on the spread does *not* hold. The family $\mathcal{B}$ contains two events: B_1 and B_2 which are defined as follows:

1. B_1: Fix an arc α of length $\frac{4t\log(4Nk)}{Tm}$ to the right of the point $r_I(i)$ associated with the item i. Denote by B_1 the event that there is *some* view that has *no* bucket point in the arc α. Note that this is, in some sense, the complement of the event that we considered in the previous lemma (lemma 2.2.5).

2. B_2: Let X be the random variable denoting the total number of bucket points in the arc α. Let B_2 denote the event that X is *more* than $8t\log(4Nk)$; that is, more than $8t\log(4Nk)$ bucket points fall into the arc α.

The next step is to show that the event A implies at least one of the events B_1 and B_2. In this case, it is easier to show the contrapositive, or: $\overline{B_1} \cap \overline{B_2} \Rightarrow \overline{A}$.

If event $\overline{B_1}$ occurs, we know that *every* view has at least one bucket point in the arc α. Now, for each view V, the item i is mapped to the first bucket point from V encountered in a clockwise traversal of the circle. The event $\overline{B_1}$ implies that this bucket point will be found somewhere in the arc α. So, it must be that in every view, i is mapped to some bucket with a point in α. Therefore, the spread of i cannot be larger than the total number of bucket points that fall into the arc α! If in addition to $\overline{B_1}$, event $\overline{B_2}$ occurs, then we know that the number of buckets in α is *less* than $8t\log(4Nk)$, and hence, the spread of i must be less than $8t\log(4Nk) = O(t\log(4Nk))$. This is precisely the event $\overline{A}$! This proves that $\overline{B_1} \cap \overline{B_2} \Rightarrow \overline{A}$.

The last and final step is to bound the probabilities of the events B_1 and B_2.

Lemma 2.2.5 shows that $\Pr[\overline{B_1}] \geq 1 - 1/2N$, and thus $\Pr[B_1] \leq 1/2N$. It remains to bound the probability of event B_2. Note that there are a total of Tm points coming from all the views and thus $E[X] = Tm\frac{4t\log(2Nk)}{Tm} = 4t\log(2Nk)$. The Chernoff bound implies that:

$$\begin{aligned} & \Pr[\ |X - 4t\log(4Nk)| \geq 4t\log(4Nk)] \\ & \leq 2e^{-t\log(4Nk)} = \frac{2}{(4Nk)^t} \leq \frac{1}{2N} \end{aligned}$$

(Assuming that $t \geq 1$.)

Consequently, $\Pr[B_2] \leq 1/2N$. To wrap up the proof we have:

$$\Pr[A] \leq \Pr[B_1 \cup B_2] \leq \Pr[B_1] + \Pr[B_2] \leq 1/2N + 1/2N = 1/N$$

This proves that the probability that the bound on spread does *not* hold is less than $1/N$, and thus, the probability that the bound *does* hold is at least $1 - 1/N$. This concludes the proof of the lemma.

■

The next Lemma shows the load bound.

Lemma 2.2.7 *For any bucket $b \in B$, $load_f(\mathcal{V}, b) = O\left(\left(\frac{|I|t}{T} + 1\right)\log(Nmk)\right)$ with probability greater that $1 - 1/N$ over the choice of $f \in UC_{random}$.*

Proof:

The intuition for the proof is simple: An item that is assigned to a bucket must fall into one of the arcs that the bucket's points are responsible for. Lemma 2.2.5 implies that the length of the arc that a single bucket point is responsible for over all views is not too big. Hence the total fraction of the circle that a bucket is responsible for over all views is relatively small. Thus, only a small fraction of the items are assigned to a bucket even if there are many views. More precisely, lemma 2.2.5 is used to bound the length of the arc that a single bucket point is responsible for over all views. Multiplying this length by m we get a bound on the total measure of the set in which items could fall and be assigned to the bucket in question. We then bound the total number of items that fall into this set, getting an upper bound on the load.

Since in the family UC_{random} all functions are equally likely to be chosen, we can assume in the proof that points for the buckets and items are distributed uniformly and at random in the unit circle. In addition, we note that item points are distributed independently of bucket points.

The bucket b has m points associated with it in the circle. An item is assigned to the bucket b if in some view it is closest clockwise to one of these m points among all other bucket points.

Fix one of the m points associated with the bucket b. We examine an arc starting at this point and going counter-clockwise around the circle that is long enough so that in *every* view, there is another bucket point in the arc with probability at least $1 - \frac{1}{2Nm}$. Invoking lemma 2.2.5 we get that the length of such an arc is given by $\frac{4t\log(8Nmk)}{Tm}$.

Now, by the union bound we have that with probability at least $\frac{1}{2N}$, the length of the arc to the left of *every one* of the m points associated with b is $\frac{4t\log(8Nmk)}{Tm}$. We denote by A this event.

Now we have:

$$\begin{aligned}
\Pr[\ load_f(\mathcal{V}, b) \geq z\ ] &= \Pr[\ load_f(\mathcal{V}, b) \geq z \mid A]\Pr[A] \\
&+ \Pr[\ load_f(\mathcal{V}, b) \geq z \mid \overline{A}]\Pr[\overline{A}] \\
&\leq \Pr[\ load_f(\mathcal{V}, b) \geq z \mid A] + \Pr[\overline{A}] \\
&\leq \Pr[\ load_f(\mathcal{V}, b) \geq z \mid A] + 1/2N
\end{aligned}$$

It remains to bound the load on bucket b given that event A has occurred. If event A occurs then we know that any item assigned to bucket b falls within distance $\frac{4t\log(8Nmk)}{Tm}$ of one of the m points associated with b. Thus, if an item is assigned to bucket b it must fall in a set of total measure at most $m\frac{4t\log(8Nmk)}{Tm} = \frac{4t\log(8Nmk)}{T}$. If there is overlap of arcs then the total measure may be smaller, but this will only lead to a better bound on the load. Therefore, all we need to do is bound the number of item points that fall into a set of measure $\frac{4t\log(8Nmk)}{T}$.

Let X denote the number of item points in the set. The expected number of item points is $E[X] = \frac{4t|I|\log(8Nmk)}{T}$. Since item points are mapped independently of bucket points we can use Chernoff bounds on X. This unfortunately turns out to be a bit technical. There are two cases to be considered according to the value of $\frac{T}{|I|}$:

Case 1: $0 < \frac{T}{|I|} \leq 2e - 1$

In this case we use a Chernoff bound with $\delta = \sqrt{\frac{T}{|I|}}$. This gives us:

$$\Pr\left[X > \left(1 + \sqrt{\frac{T}{|I|}}\right)\frac{4t|I|\log(8Nmk)}{T}\right] \leq \frac{1}{(8Nmk)^t} \leq \frac{1}{2N}$$

(Since $t, k, m \geq 1$.)

Case 2: $2e - 1 \leq \frac{T}{|I|}$

In this case we use a Chernoff bound with $\delta = \frac{T}{|I|}$. This gives us:

$$\Pr\left[X > \left(1+\frac{T}{|I|}\right)\frac{4t|I|\log(8Nmk)}{T}\right] \leq \frac{1}{(8Nmk)^{t\frac{|I|}{T}(1+\frac{T}{|I|})}}$$
$$\leq \frac{1}{(8Nmk)^t}$$
$$\leq \frac{1}{2N}$$

(Since $t, k, m \geq 1$.)

Since both $\left(1+\sqrt{\frac{T}{|I|}}\right)\frac{4t|I|\log(8Nmk)}{T}$ and $\left(1+\frac{T}{|I|}\right)\frac{4t|I|\log(8Nmk)}{T}$ are $O\left(\left(\frac{|I|t}{T}+1\right)\log(Nmk)\right)$ we have shown that for $z = O\left(\left(\frac{|I|t}{T}+1\right)\log(Nmk)\right)$:

$$\Pr[\ load_f(\mathcal{V}, b) \geq z\ ] \leq \Pr[\ load_f(\mathcal{V}, b) \geq z \mid A] + 1/2N$$
$$\leq 1/2N + 1/2N = 1/N$$

This concludes the proof of the lemma.

■

It remains to show the balance property. If we fix a particular view, then the probability that an item is assigned to a particular bucket is exactly the total length of the unit circle that the bucket is "responsible" for. The following lemma bounds the total length of the set that each bucket is responsible for, and thus will be the main step in showing the balance property (which is proved in lemma 2.2.9).

For a bucket b, denote by $length(b)$ the measure of the set of points in the unit circle that b is responsible for.

Lemma 2.2.8 *Let V be a fixed view containing $v = |V|$ buckets. Then with probability at least $1 - 1/N$ for all $b \in V$, $length(b) = O\left(\frac{1}{v}\left(\frac{\log(Nv)}{m}+1\right)\right)$.*

Proof:

The proof of lemma 2.2.8 is based on the same idea as the proof of lemma 2.2.6. If we want to show that some event A has low probability, then we can find a set of

events $\mathcal{B}$ so that the occurrence of event A implies that at least one of the events in $\mathcal{B}$ occurs. Then, if we show that there is only a small probability that *any* event in $\mathcal{B}$ occurs, then the probability of A must also be small.

In our case the event A is that some bucket b has $length(b)$ greater than the value stated in the lemma. Recall that the part of the unit circle that the bucket b is "responsible" for is broken up into m non-overlapping arcs. One (not good) set of events $\mathcal{B}$ is obtained by observing that if b is responsible for a set of measure p, then there is a collection of m arcs of total length p (with right ends at the m points associated with b) in which no other bucket point falls. Thus each event in $\mathcal{B}$ is described by a set of m arcs of total length p, into which no other bucket point falls. The problem is that there are uncountable many ways to divide length p among m arcs. Since we want to use a union bound to bound the probability that any event from $\mathcal{B}$ occurs, the set needs to be finite!

We make $\mathcal{B}$ smaller by discretizing the circle and counting the number of ways to distribute total length p by discrete units. This is of course finite, and the error introduced by the discretization turns out to be small. We now formalize the above argument.

Recall that a bucket point is *responsible* for an arc if the bucket point is on the right end of the arc, there is no bucket point in the interior of the arc, and some other bucket point is on the left end of the arc. A bucket b is responsible for the union of the m arcs that the points associated with b are responsible for.

The main result is that the probability a single bucket b is responsible for more than an $O\left(\frac{1}{v}\left(\frac{\log(Nv)}{m}+1\right)\right)$ fraction of the unit circle is at most $1/Nv$. The union bound then implies that none of the v buckets are responsible for more than an $O\left(\frac{1}{v}\left(\frac{\log(Nv)}{m}+1\right)\right)$ fraction of the circle with probability $1-1/N$.

To show the main result, we begin by fixing a bucket b. The portion of the unit circle for which b is responsible must consist of m non-overlapping arcs in the circle, each bounded on the right by one of b's points. Suppose we shrink all these arcs by moving the left endpoints rightward, until the length of every arc is a multiple of $\Delta = \frac{4\alpha}{mv}$ (α will be set later). Since there are m of these arcs and each shrinks by at

most Δ, the decrease in the total length of all arcs is at most $4\frac{\alpha}{v}$. So if the total length of these arcs after shrinking is $4\frac{\alpha}{v}$, then the total length before shrinking is at most $4\frac{\alpha}{v}+4\frac{\alpha}{v}=8\frac{\alpha}{v}$. This implies that if bucket b is responsible for a $8\frac{\alpha}{v}$ fraction of the unit circle, then b must be responsible for every point in a collection of non-overlapping arcs, each bounded on the right with one of b's points, each a multiple of Δ in length, and with total length $4\frac{\alpha}{v}$.

Now, given a collection of arcs of total length $4\frac{\alpha}{v}$ we will bound the probability that in these arcs there is no point associated with some other bucket. The expected number of the $mv-m$ points falling in this collection of arcs is $4\frac{\alpha m(v-1)}{v}$. So we have from a superfluous use of the Chernoff bounds:

$$\begin{aligned} Pr[X=0] &\leq Pr[(X-4\alpha m\frac{v-1}{v})\geq 4\alpha m] \\ &\leq e^{-\alpha m\frac{v-1}{v}} \end{aligned}$$

The number of collections of m arcs with total length $\frac{4\alpha}{v}$ and with all lengths multiples of Δ is exactly the number of ways to partition $\frac{4\alpha}{v}/\Delta=m$ into m integral parts, which is:

$$\binom{2m-1}{m-1}\leq 2^{2m-1}\leq e^{2m}$$

By the union bound, the probability that any of the above collections of arcs contains no point associated with other buckets is at most:

$$e^{-\alpha m\frac{v-1}{v}}e^{2m}=e^{-(\alpha m\frac{v-1}{v}-2m)}$$

We will choose α so that this probability is at most $1/(Nv)$. which gives:

$$\alpha=O\left(\frac{\log(Nv)}{m}+1\right)$$

This proves that with probability at least $1 - 1/(Nv)$ the total length assigned to a bucket is at most $\frac{8\alpha}{v} = O(\frac{1}{v}\left(\frac{\log(Nv)}{m} + 1\right))$. Now since there are v buckets the same bound holds for all buckets with probability $1 - 1/N$ by the union bound.

■

We now prove the balance bound given in theorem 2.2.3.

Lemma 2.2.9 *For any fixed view V and item i, the probability that item i is mapped to bucket b in view V is $O\left(\frac{1}{|V|}\left(\frac{\log(N|V|)}{m} + 1\right)\right) + \frac{1}{N}$.*

Proof:

Denote by A the event that for every bucket b in the view V we have $length(b) = O\left(\frac{1}{|V|}\left(\frac{\log(N|V|)}{m} + 1\right)\right)$. Lemma 2.2.8 says that the probability of A is at least $1 - 1/N$. Now for any $b \in V$:

$$\begin{aligned}
\Pr[\, f_V(i) = b] &= \Pr[\, f_V(i) = b |\, A] \Pr[A] \\
&+ \Pr[\, f_V(i) = b |\, \overline{A}] \Pr[\overline{A}] \\
&\leq \Pr[\, f_V(i) = b |\, A] + \Pr[\overline{A}] \\
&\leq \Pr[\, f_V(i) = b |\, A] + 1/N \\
&= O\left(\frac{1}{|V|}\left(\frac{\log(N|V|)}{m} + 1\right)\right) + \frac{1}{N}
\end{aligned}$$

Since, given that event A has occurred, the probability that item i is mapped to any particular bucket b is exactly $length(b)$ (the mapping of items is independent of the mapping of buckets) which is $O\left(\frac{1}{|V|}\left(\frac{\log(N|V|)}{m} + 1\right)\right)$.

■

This concludes the proof of theorem 2.2.3. We state and prove one corollary of theorem 2.2.3 that is useful for various applications.

Corollary 2.2.10 *With the same conditions as theorem 2.2.3, the probability that an item i is mapped to a bucket b in at least one of the views is $O\left(\frac{t\log(Nmk)}{T}\right) + \frac{1}{N}$.*

Proof:

We saw in the proof of lemma 2.2.7 that with probability at least $1 - 1/N$, the bucket b is not responsible for more than a $O\left(\frac{t \log(Nmk)}{T}\right)$ fraction of the circle over all views. Using the same argument as in the proof of lemma 2.2.9, the corollary follows.

■

2.2.4 A Practical Consistent Hash Family

In the previous section, we showed that the family UC_{random} has good consistency properties: the family is monotone, spread and load grow logarithmically with the number of views, and the family has the balance property.

However, there is a drawback to the family UC_{random}; it requires manipulation of real numbers. More specifically, storing a function requires infinite space, and furthermore, choosing a function from the family requires an infinite number of random bits.

In this section we remedy these problems by modifying the basic construction in two simple ways. We show that limited independence in the mapping of points to the circle suffices for the family to have the same consistency properties as UC_{random}. Using limited independence reduces the number of random bits required to choose a function from the family, and reduces the space required to store a function from the family. Furthermore, we show how to use limited precision in the real numbers used in the basic construction, thus eliminating the need to manipulate arbitrary real numbers. In section 2.2.4 these two modifications are combined to construct a new, more practical hash family. An implementation of this family, which is remarkably simple and efficient, is presented in section 2.2.5.

Using Limited Independence

We say that a family of functions is k-way independent if any k elements from the domain are mapped independently into the range when we choose a function at random from the family. In other words a family is k-way independent if for any distinct

x_i, $1 \leq i \leq k$ from the domain of the family, and any y_i, $1 \leq i \leq k$ from the range of the family:

$$\Pr[f(x_1) = y_1, f(x_2) = y_2, \ldots, f(x_k) = y_k] = \prod_{i=1}^{k} Pr[f(x_i) = y_i]$$

Where the probability is over a random choice of function f from the family. Another way of looking at this is that the random variables $\{f(i)\}$ where i ranges over the domain of the family are k-way independent.

As an example, consider the linear congruential hash family introduced in example 1. This is a 2-way independent family. We showed that the number of random bits required to choose a function from the family and the space required to store a function from the family are very small compared to a completely random function. These are exactly the reasons that we are interested in using limited independence mappings.

We show that if UC_{random} is modified so that items and bucket copies are mapped to the circle using limited independence, then the consistency properties of the family remain unchanged.

The basic tool used is the following theorem from[17] that shows that Chernoff bounds apply to cases with certain limited amounts of independence.

Theorem 2.2.11 *If X is the sum of k-wise independent random variables each of which is in the range $[0,1]$ with $\mu = E[X]$, then:*

- *For any $\delta \geq 1$ and $k \geq \lceil \delta\mu \rceil$:*

$$\Pr[\mid X - \mu \mid \geq \delta\mu] \leq e^{\frac{-\delta\mu}{3}}$$

- *For any $\delta \leq 1$ and $k \geq \lfloor \delta^2 \mu e^{-1/3} \rfloor$:*

$$\Pr[\mid X - \mu \mid \geq \delta\mu] \leq e^{\lfloor \frac{-\delta^2\mu}{3} \rfloor}$$

The only probabilistic tool used in the proof of theorem 2.2.3, which states the

consistent properties of the family UC_{random}, is the Chernoff bound on sums of indicator variables (other than the union bound which is true regardless of independence of mappings). Theorem 2.2.11 implies that the claims of theorem 2.2.3 are still valid even if item and bucket points are mapped to the circle with only limited independence and not completely randomly as assumed in the proof. In fact, we show that if the bucket and item points are each mapped $\Omega(t\log(NTk))$-way independently then all the bounds of theorem 2.2.3 still hold.

Theorem 2.2.12 *If bucket copies and items are mapped to the unit circle using $\Omega(t\log(NTk))$-way independent families, then, as long as item points are mapped independently of bucket points, theorem 2.2.3 still holds.*

Proof: Monotonicity is not affected by the independence of the mappings.

We need to check that each of the Chernoff bounds used in the proof of spread, load, and balance are still valid with only $\Omega(t\log(NTk))$-way independence. Recall that the proof of theorem 2.2.3 was divided into lemmas 2.2.6, 2.2.7, and 2.2.9.

For each of lemma 2.2.5, lemmas 2.2.6 and 2.2.7, and also lemma 2.2.8 we will show that using a function with $\Omega(t\log(NTk))$-way independence is sufficient.

- Lemma 2.2.5: The proof uses a Chernoff bound with $\delta = 1$ and $\mu = \frac{4t\log(2Nk)}{Tm}$. Invoking the first case of theorem 2.2.11 we see that we need $\left\lceil \frac{4t\log(2Nk)}{Tm} \right\rceil$-way independence. Since this is $O(t\log(NTk))$, lemma 2.2.5 holds.

- Lemma 2.2.6 **(spread):** The proof uses lemma 2.2.5 that we showed above holds. The proof of the spread bound (lemma 2.2.6) contains one more Chernoff bound with $\delta = 1$ and $\mu = 4t\log(2Nk)$. From the first case of theorem 2.2.11 we see that $\Omega(t\log(NTk))$-way independence suffices.

- Lemma 2.2.7 **(load):** The first part of the proof relies on lemma 2.2.5 which we have shown still holds. Item points are mapped independently of bucket points so the remaining Chernoff bound is still valid as long as we have enough independence: There are two cases to consider depending on whether $T/|I|$ is larger or smaller than 1.

Case 1: $T/|I| < 1$

In this case we use a Chernoff bound with parameters $\delta = \sqrt{T/|I|} < 1$ and $\mu = \frac{4t|I|\log(4Nmk)}{T}$. We apply the second case of theorem 2.2.11 and observe that $\Omega(t\log(NTk))$-way independence suffices. This case then parallels the case of $T/|I| < 2e-1$ in the proof of lemma 2.2.7.

Case 2: $T/|I| \geq 1$

In this case we use a Chernoff bound with parameters $\delta = T/|I| \geq 1$ and $\mu = \frac{4t|I|\log(4Nmk)}{T}$. We invoke the first case of theorem 2.2.11. This case then parallels the case of $T/|I| \geq 2e-1$ in the proof of lemma 2.2.7.

- Lemma 2.2.8 **(balance):** The proof relies on a Chernoff bound with $\delta = 1$ and $\mu = O(\log(NT) + m)$. Invoking the first case of theorem 2.2.11 we see that $\Omega(t\log(NTk) + m)$-way independence suffices, however if we choose $q = O(\log(T))$, then $\Omega(t\log(NTk))$-way independence suffices. The proof also relies on the fact the item points are mapped independently of bucket points.

■

Using Limited Precision

In this section we show how to use limited precision for representing real numbers for the unit circle hash function. The next section defines the final practical hash family and proves its properties.

The basic idea is simple. Each function in the family UC_{random} is defined by a total of $|I|+m|B|$ random points on the real unit circle. The important observation is that what really matters about these points is their clockwise *ordering* around the circle. Given just the clockwise ordering of all the points, we can reconstruct the mapping of every item in every view. An item i is mapped to a bucket b with a point closest to the point of i in the clockwise ordering. The following lemma shows that the ordering on a set of $|I| + m|B|$ random points is with high probability, already completely

defined by the $O(\log(|I| + m|B|))$ most significant bits of the binary expansions of the points.

Lemma 2.2.13 *With probability at least $1 - 1/N'$, the clockwise ordering on n random points in the unit circle is determined by the $2\log(N'n)$ most significant bits of the points. ($N' \geq 1$ is an arbitrary confidence factor.)*

Proof:

The probability that any two of the numbers can not be distinguished by their $2\log(N'n)$ most significant bits is $1/(N'n)^2$ (since the probability of an infinite sequence of 0's or 1's is zero). By the union bound, the probability that any of the $\binom{n}{2} \leq n^2$ pairs of points are not distinguished by their $2\log(N'n)$ most significant bits is less than $(1/N')^2 \leq 1/N'$.

Thus, for each point we need no more than $2\log(N'n)$ bits to determine the ordering with probability at least $1 - 1/N'$.

■

The following simple corollary shows that lemma 2.2.13 holds even if the points are distributed only k-way independently for $k \geq 2$.

Corollary 2.2.14 *Lemma 2.2.13 holds if the points are distributed uniformly and k-way independently for any $k \geq 2$.*

Proof: The corollary follows from two observations. The first observation is that if the points are k-way independent, then the bits at a fixed place in the binary expansion of the points are k-way independent. The second observation is that in the proof of Lemma 2.2.13 we only used 2-way independence of these bits. ■

Putting it all Together

This section describes the hash family obtained by modifying UC_{random} to use finite precision, and limited independence mappings. This family is called UC.

Let $c = O(\log(N'(m|B| + |I|)))$ for an arbitrary confidence factor $N' \geq 1$. The family UC is defined in the same way as UC_{random} except in the following aspects: