# Corrigendum to "Efficient similarity search and classification via rank aggregation" by Ronald Fagin, Ravi Kumar and D. Sivakumar, Proc. SIGMOD'03.


Alexandr Andoni
MIT
32 Vassar St.
Cambridge, MA 02139

Ronald Fagin
IBM Almaden
650 Harry Road
San Jose, CA 95120

Ravi Kumar
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089

Mihai Pătraşcu
MIT
32 Vassar St.
Cambridge, MA 02139

D. Sivakumar
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043


In this corrigendum, we correct an error in the paper [FKS03]. The error was discovered by Alexandr Andoni, and the corrected theorem is due to the three authors of [FKS03], along with Alexandr Andoni and Mihai Pătraşcu.

Theorem 4 of [FKS03] states:

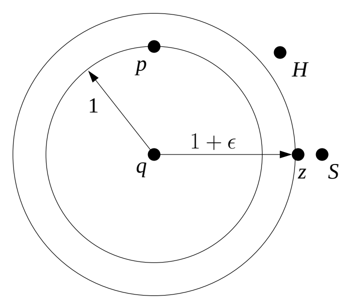
> *Let $D$ be a collection of $n$ points in $\mathbb{R}^d$. Let $r_1, \ldots r_m$ be random unit vectors in $\mathbb{R}^d$, where $m = \alpha\epsilon^{-2}\log n$ with $a$ suitably chosen. Let $q \in \mathbb{R}^d$ be an arbitrary point, and define, for each $i$ with $1 \leq i \leq m$, the ranked list $L_i$ of the $n$ points in $D$ by sorting them in increasing order of their distances to the projection of $q$ along $r_i$. For each element $x$ of $D$, let $medrank(x) = median(L_1(x), \ldots, L_m(x))$. Let $z$ be a member of $D$ such that $medrank(x)$ is minimized. Then with probability at least $1 - 1/n$, we have $d(z,q) \leq (1+\epsilon)d(x,q)$ for all $x \in D$.*

As stated, the above theorem does not hold, but a slight modification of it holds. Below, we first give a counter-example to the original theorem, and then present a modified theorem.

## 1. A COUNTER-EXAMPLE

In our counter-example, we give a specific set of $n$ points in 2-dimensional space. Consider the following point set for very small $\epsilon$, illustrated in Fig. 1:

- point $q = (0,0)$, the query;
- point $p = (0,1)$, the nearest neighbor;
- point $z = (1+\epsilon, 0)$, the false nearest neighbor;
- set $H$ of 10 points $h$ all at distance $(1+\epsilon)^2$ from $q$, specifically $h = (1+\epsilon)^2 \cdot (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$;
- set $S$ of the rest $n - 12$ points, all situated at $s = ((1+\epsilon)^2, 0)$.




1

Let $r$ be random unit vector in $\mathbb{R}^2$, $L$ is the list of the pointset $D$ sorted by increasing distance from $q$; and $rank(x)$ is the rank of point $x$ in $L$. Then we have the following two claims.

Claim 1.1. $\Pr_r[rank(z) \leq 2] > 1/2 + \Omega(\epsilon)$.

The claim follows immediately from Lemma 3 of [FKS03].

Claim 1.2. $\Pr_r[rank(p) \geq |H|] \geq 1/2 + 1/16$.

It is sufficient to consider $r$'s with non-negative $x$ coordinate, and identify $r$'s by their angle with the $Ox$ axis. First, $rank(p) \leq rank(z)$ if $r \in [\alpha, \beta]$, where $\alpha$ is angle formed by the perpendicular to the line connecting $q$ to midpoint of $pz$, and $\beta$ is the angle formed by the perpendicular to $pz$. We can estimate $\alpha$ and $\beta$:

$$\alpha = \arctan\frac{p_y + z_y}{p_x + z_x} - \pi/2 = \arctan(1+\epsilon) - \pi/2 = -\pi/4 - \Theta(\epsilon)$$

$$\beta = \arctan\frac{z_x - p_x}{p_y - z_y} = \arctan(1+\epsilon) = \pi/4 + \Theta(\epsilon).$$



*SIGMOD'08*, June 9–12, 2008, Vancouver, BC, Canada.
ACM 978-1-60558-102-6/08/06.

Thus, if $r \in [\alpha, \beta]$, $rank(p) \geq rank(z)$, and then $rank(p) \geq |S| + 1$.

Moreover, as we will see, if the angle of $r$ is around $-\pi/4$, then $rank(p) > rank(h)$. Indeed, consider any angle $\gamma \in [-\pi/4, -\pi/4 + \pi/16]$. Then, $|\langle p, r\rangle| = |\sin\gamma| \geq 0.5$ and $|\langle h, r\rangle| = |(1+\epsilon)^2 \frac{1}{\sqrt{2}} \cdot (\sin\gamma + \cos\gamma)| \leq 0.2(1+\epsilon)^2$.

Thus, if the angle of $r$ is in the range $(-\pi/2, -\pi/4 + \pi/16)$ or $(\beta, \pi/2)$, $rank(p) \geq |H|$, and this happens with probability at least $\frac{\pi/4+\pi/16-\Theta(\epsilon)}{\pi/2} \geq 1/2 + 1/16$.

Standard high concentration bounds will yield that $medrank(z) < medrank(p)$ with high probability. For completeness, we include one such lemma, due to Indyk:

Lemma 1.3 (cf. [Ind00], Lemma 2). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}$ and $F$ be its cumulative distribution function. Then, for $\epsilon, \delta > 0$ and $k = O(\frac{\log 1/\delta}{\epsilon^2})$, if $X_1 \dots X_k$ are iid from $\mathcal{D}$, then $X = median\{X_1, \dots X_k\}$ satisfies $\Pr[F(x) \in (1/2 - \epsilon, 1/2 + \epsilon)] \geq 1 - \delta$.*

## 2. A NEW ALGORITHM

To correct the theorem, we propose to use the following new function $medrank(x)$:

$$medrank(x) = median_i(|xr_i - qr_i|).$$

The resulting algorithm is presented in Fig. 2. Next, we show that this algorithm gives a $1+\epsilon$ nearest neighbor data structure.

We note that, for Step 2 of the query algorithm, we can use also other aggregation functions instead of the median function. In particular, if we use $\ell_2$ norm of the *score* vector instead of the median, then the same lemma as above holds, implied by Johnson-Lindernstrauss lemma [JL84]. Furthermore, if use $\ell_1$ norm of the *score* vector, then again the same lemma as above holds, and is implied by the $\ell_2$ to $\ell_1$ embedding of [JS82].

## 3. REFERENCES


Ronald Fagin, Ravi Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 301–312, New York, NY, USA, 2003. ACM.

P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. *Proceedings of the Symposium on Foundations of Computer Science*, 2000.

W.B. Johnson and J. Lindenstrauss. Extensions of lipshitz mapping into hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

W.B. Johnson and G. Schechtman. Embedding $l_p^m$ into $l_1^n$. *Acta Mathematica*, 149:71–85, 1982.


**Preprocessing.** Input: a set of points $P \subset \mathbb{R}^d$, $|P| = n$, and $\epsilon > 0$.

1. Choose $k = O(\frac{\log n}{\epsilon^2})$ vectors $r_i \in \mathbb{R}^d$, $i = 1 \dots k$, where each coordinate of $r_i$ is drawn from a Gaussian $N(0,1)$ distribution. Vectors $r_i$ represent random projections.
2. Construct $k$ lists, where the $i^{th}$ list contains all the points from $p \in P$ sorted according to the value $p \cdot r_i$.

**Query.** Input: a query point $q \in \mathbb{R}^d$.

1. For fixed $i$ and $p \in P$, define $score_i(p) = q \cdot r_i - p \cdot r_i$.
2. Find the point $p^* \in P$ that minimizes $median_{i \in [k]}\{|score_i(p^*)|\}$.
3. Return $p^*$.



Lemma 2.1. *The algorithm from Figure 2 returns a $1+\epsilon$ nearest neighbor of $q$ with probability at least $1 - 1/n$.*

Proof. Fix some $p$ and let $\Delta = \|p - q\|_2$. For each $i \in [k]$, $score_i(p)$ is distributed as $N(0, \Delta^2)$, normal distribution with standard deviation $\Delta$. We will once again use Lemma 1.3 for estimating the median of iid samples.

Let $M_p = median_{i\in[k]}\{|score_i(p)|\}$. Applying Lemma 1.3 to the distribution $N(0, \Delta^2)$, we conclude that $F(M_p) \in (1/2 - \epsilon, 1/2 + \epsilon)$ with probability at least $1 - 1/n^2$, where $F$ is the cumulative of $N(0, \Delta^2)$. Since the value $x$ that satisfies $F(x) = 1/2$ is $x = c\Delta$ where $c$ is an absolute constant, and $F$ has derivate $\Theta(1/\Delta)$ around this $x$, we conclude that $M_p \in (x - O(\epsilon\Delta), x + O(\epsilon\Delta))$. Thus, $M_p$ is a $1+\epsilon$ approximation to $\|q - p\|$ with probability at least $1 - 1/n^2$.

We conclude that $M_p$ is a $1+\epsilon$ approximation to $\|q-p\|$ for all $p$, with probability at least $1 - 1/n$. Thus the algorithm returns a $1+\epsilon$ approximate nearest neighbor with probability at least $1 - 1/n$. $\square$