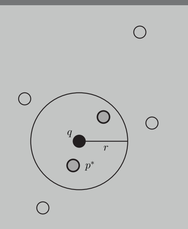
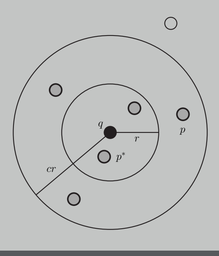
Beyond Locality-Sensitive Hashing

Alexandr Andoni (Microsoft Research), Piotr Indyk (MIT),
Huy L. Nguyen (Princeton) and **Ilya Razenshteyn (MIT)**

(Approximate) Near Neighbor Problem ((A)NN)

- **r**-Near Neighbor Problem (NN)
  - **P** — **n**-point subset of $\mathbb{R}^{\mathbf{d}}$, $\mathbf{r} > \mathbf{0}$
  - Given $\mathbf{q} \in \mathbb{R}^{\mathbf{d}}$ report any $\mathbf{p} \in \mathbf{P}$ s.t. $\|\mathbf{p} - \mathbf{q}\| \leq \mathbf{r}$
  - Hard, if **d** is "large"
- **(c, r)**-Approximate Near Neighbor Problem (ANN)
  - In addition, we are given $\mathbf{c} > \mathbf{1}$
  - Given $\mathbf{q} \in \mathbb{R}^{\mathbf{d}}$ report $\mathbf{p} \in \mathbf{P}$ s.t. $\|\mathbf{q} - \mathbf{p}\| \leq \mathbf{cr}$, **if there exists** $\mathbf{p}^* \in \mathbf{P}$ **s.t.** $\|\mathbf{q} - \mathbf{p}^*\| \leq \mathbf{r}$

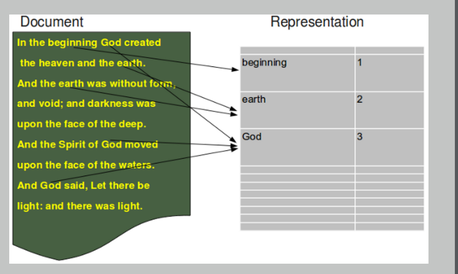




Applications

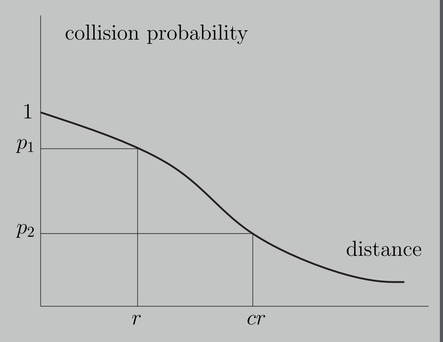
- Pattern recognition, statistical classification, computer vision, computational geometry, databases, recommendation systems, DNA sequencing, spell checking, plagiarism detection, clustering etc
- **Approximate string matching**
  - Text **T**, query **S**: substring of **T** closest to **S**
  - Know $|\mathbf{S}|$ in advance: can reduce to NN wrt Hamming distance
- **ML: nearest neighbor rule**
  - Use the label of the closest labeled example
- **Near-duplicate document retrieval**
  - Bag of words

T = GAGTAACTCAATA
S = AGTA



Locality-Sensitive Hashing (LSH)

- Introduced in (Indyk, Motwani 1998)
- A hash family $\mathcal{H}$ of functions $\mathbf{h}\colon \mathbb{R}^{\mathbf{d}} \to \mathbf{U}$ s.t. for every $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{\mathbf{d}}$
  - if $\|\mathbf{p} - \mathbf{q}\| \leq \mathbf{r}$, then $\Pr_{\mathbf{h} \sim \mathcal{H}}[\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{q})] \geq \mathbf{p_1}$ ("large")
  - if $\|\mathbf{p} - \mathbf{q}\| \geq \mathbf{cr}$, then $\Pr_{\mathbf{h} \sim \mathcal{H}}[\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{q})] \leq \mathbf{p_2}$ ("small")
- Call such a family $(\mathbf{r}, \mathbf{cr}, \mathbf{p_1}, \mathbf{p_2})$-*sensitive*
- Close points collide often



From LSH to ANN

- $\mathcal{H}$ is a "reasonable" $(\mathbf{r}, \mathbf{cr}, \mathbf{p_1}, \mathbf{p_2})$-sensitive family
- "Quality" of $\mathcal{H}$

$$\rho = \frac{\ln(1/\mathbf{p_1})}{\ln(1/\mathbf{p_2})}$$

- Can solve ANN in $\mathbf{O(n^{1+\rho} + d \cdot n)}$ space and $\mathbf{O(d \cdot n^{\rho})}$ query time (Indyk, Motwani 1998)
- Example: $\{\mathbf{0}, \mathbf{1}\}^{\mathbf{d}}$, Hamming distance;
  - $\mathcal{H} = \{\mathbf{h_1}, \ldots, \mathbf{h_d}\}$, $\mathbf{h_i(x) = x_i}$
  - $\rho \leq \mathbf{1/c}$
- The best ρ for $\ell_{\mathbf{p}}$ metric ($\|\mathbf{x} - \mathbf{y}\|_{\mathbf{p}} = (\sum_{\mathbf{i}} |\mathbf{x_i} - \mathbf{y_i}|^{\mathbf{p}})^{\mathbf{1/p}}$) for $\mathbf{1} \leq \mathbf{p} \leq \mathbf{2}$ is $\rho = \mathbf{1/c^p + o(1)}$ (matching lower bounds) (Indyk, Motwani 1998), (Andoni, Indyk 2006), (Motwani, Naor, Panigrahy 2007), (O'Donnell, Wu, Zhou 2011)

The main result

- A data structure for the $(\mathbf{c}, \mathbf{r})$-ANN with space $\mathbf{O(n^{1+\rho} + d \cdot n)}$ and query time $\mathbf{O(d \cdot n^{\rho})}$, where

$$\rho \leq \frac{\mathbf{0.73}}{\mathbf{c^p}} + \mathbf{O}\left(\frac{\mathbf{1}}{\mathbf{c^{3p/2}}}\right) + \mathbf{o(1)}.$$

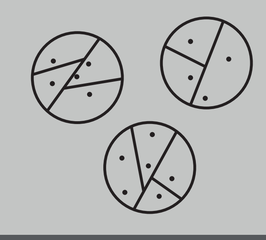
- Go beyond **the best possible** ρ for "vanilla" Locality-Sensitive Hashing
- The first improvement upon (Indyk, Motwani 1998) for ℓ_1 and (Andoni, Indyk 2006) for ℓ_2

The main approach: data-dependent hashing

- LSH is oblivious, try to be data-dependent in spirit of (Fredman, Komlós, Szemerédi 1984)
- **The definition of LSH is too strong!** Enough to satisfy the inequalities for $\mathbf{p} \in \mathbf{P}$ and $\mathbf{q} \in \mathbb{R}^{\mathbf{d}}$.
  - Exploit the geometry of **P**?
  - **That is what we do!**
- Parallels with practice
  - PCA trees (Sproull 1991), (McNames 2001), (Verma, Kpotufe, Dasgupta 2009)
  - Spectral Hashing (Weiss, Torralba, Fergus 2008)
  - Semantic Hashing (Salakhutdinov, Hinton 2009)
  - WTA Hashing (Yagnik, Strelow, Ross, Lin 2011)

The two main steps

- Cluster **P** into low-diameter pieces (of diameter $\mathbf{O(cr)}$)
- Handle each piece using a new "low-diameter" family

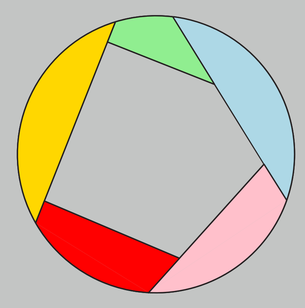


The low-diameter case

- Euclidean case: the previous record is $\rho \leq \mathbf{1/c^2}$
- Points and queries are on a sphere of radius $\mathbf{O(cr)}$
- Can achieve

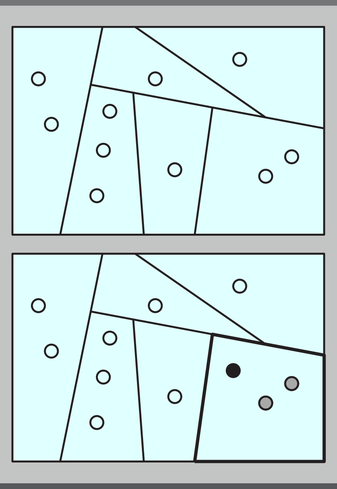
$$\rho = \frac{\ln(1/\mathbf{p_1})}{\ln(1/\mathbf{p_2})} \leq \frac{\mathbf{1 - \Omega(1)}}{\mathbf{c^2}}$$

using "ball carving" (similar to (Karger, Motwani, Sudan 1998))



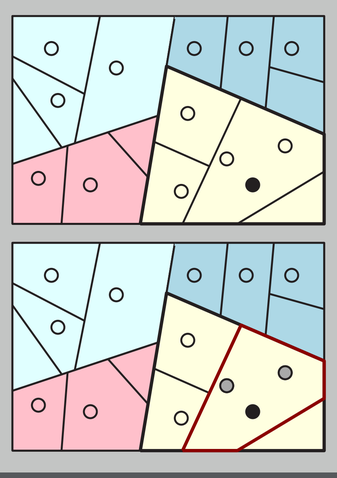
Vanilla LSH

- $\mathcal{H}$ is a $(\mathbf{r}, \mathbf{cr}, \mathbf{p_1}, \mathbf{p_2})$-sensitive family
- Hash **P** using a concatenation of **k** functions from $\mathcal{H}$: $\mathbf{x} \mapsto (\mathbf{h_1(x)}, \mathbf{h_2(x)}, \ldots, \mathbf{h_k(x)})$
- Locate a query **q**, enumerate points from the bucket
- If choose **k** carefully: need $\mathbf{n}^{\rho}$ hash tables
- Overall: $\mathbf{n^{1+\rho}}$ space, $\mathbf{n}^{\rho}$ query time



Two-level hashing

- Coarser partition (smaller **k**)
- Every part has a low diameter (aim at $\mathbf{O(cr)}$)
- The "low-diameter" family to partition space finer
- "Outer" (data-independent, $\rho \leq \mathbf{1/c^2}$) + "inner" (data-dependent, "low-diameter" family) hash tables
- Get $\rho \leq \mathbf{(1 - \Omega(1))/c^2}$: in inner tables better relation between $\mathbf{p_1}$ and $\mathbf{p_2}$!



Conclusions

- Overcome limitations of LSH by using data-dependency
- Open problem: improve ρ
  - For a random instance from (Motwani, Naor, Panigrahy 2007) achieve $\rho \lesssim \mathbf{1/(2c^2)}$
  - Achieve this ρ for the general case?
  - Do even better for that random instance?
- Lower bounds?

http://ilyaraz.org ilyaraz@mit.edu