Method 2: Locality-Sensitive Hashing (LSH)

(Indyk, Motwani '98)

idea - if p, q are close, more likely that $h(p) = h(q)$

Consider special case of Hamming space $\{0,1\}^d$

given $p, q \in \{0,1\}^d$,

define $d(p,q) =$ # bit positions that are diff.

e.g. $p = 1011001$ $\quad d(p,q) = 2.$
$q = 1001011$

Will solve approx decis problem for fixed r, approx factor c.

Preproc Alg'm:

let k be param.

define hash fn $h: \{0,1\}^d \to \{0,1\}^k$

by projecting to k rand dims

eg. $p = 10111110011$

$h(p) = 010$

for each $p \in P$

put p in bucket for $h(p)$. $\leftarrow$ $\tilde{O}(n)$ space

Query(q): if each $p \in P$ in bucket for $h(q)$

if $d(p,q) \le cr$ return p

return no

Prop $\forall p, q \in \{0,1\}^d$,

$$\Pr[h(p) = h(q)] = \left(1 - \frac{d(p,q)}{d}\right)^k$$

$$\Pr[h(p)=h(q)] = \left(1-\frac{d(p,q)}{d}\right)^k$$

e.g. $p = 10\,1\,11\,0\,001$ $\quad h(p) = 010$

$q = 10\,0\,11\,1\,001$ $\quad h(q) = 010$

$d(p,q) = 2.$

$1-x \approx e^{-x}$

Cor (i) if $d(p,q) > cr$,

$$\Pr[h(p)=h(q)] < \left(1-\frac{cr}{d}\right)^k \le e^{-\frac{crk}{d}} = e^{-\ln n} = \frac{1}{n}$$

(ii) if $d(p,q) < r$,

$$\Pr[h(p)=h(q)] > \left(1-\frac{r}{d}\right)^k \gtrsim e^{-\frac{rk}{d}} = e^{-\frac{1}{c}\ln n} = \frac{1}{n^{1/c}}$$

choose $k = \frac{d \ln n}{cr}$

Expected query time

$$= O\Big(d \cdot E\big[\,\big|\underbrace{\{p\in P : h(p)=h(q)\}}_{\text{with } d(p,q)>cr}\big|\,\big]\Big)$$

$$= O\left(d \cdot n \cdot \frac{1}{n}\right) = O(d)$$

Error Analysis:

if $d(p,q) > cr \ \ \forall p \in P$,

return no $\Rightarrow$ always correct

if $d(p,q) < r$ for some $p \in P$.

$\rightarrow$ $\Pr[\text{alg'm correct}] \geq \Pr[h(p)=h(q)]$

$$= \frac{1}{n^{1/c}}$$

wrong most of the time!!

Final idea.

repeat $t = 100 n^{1/c} \ln n$ times

$(1-x \le e^{-x})$

t

repeat $t = 100 n^{1/c} \ln n$ times $(1-x \le e^{-x})$

$$\Rightarrow \text{err prob} \lesssim \left(1 - \frac{1}{n^{1/c}}\right)^t$$

$$\le e^{-\frac{t}{n^{1/c}}} = e^{-100 \ln n}$$

tiny

Space $\boxed{O(d n^{1+1/c})}$

Query time $\boxed{O(d n^{1/c})}$

for approx factor c.

e.g. $c=2$: space $\sim n^{3/2}$, time $\sim \sqrt{n}$

$c=1.1$: space: $n^{1.91}$, time: $n^{0.91}$

$c=10$: space $\sim n^{1.1}$, time $\sim n^{0.1}$

Rmk: extends to L_1 metric in $[U]^d$.

by mapping $[U]^d \to \{0,1\}^{dU}$

e.g. $p = (2,3) \to 1100000\ 111000$

$q = (5,2) \to 1111100\ 110000$

$d(p,q) = 3+1 = 4$ in L_1

(do the mapping implicitly)

Rmk. extends to Euclidean

with some choice of hash fn family ...

(rand proj + rand shift)

(Andoni. Indyk '06: space n^{1+1/c^2}, query n^{1/c^2})

Rmk. improvements

Andoni - Razenshteyn '15: L_1: space $\sim n^{1+\frac{1}{2c-1}}$

time $\sim n^{\frac{1}{2c-1}}$.

space / time trade-offs

Monte Carlo rand. $\rightarrow$ Las Vegas rand.

⋮

GRAPHS

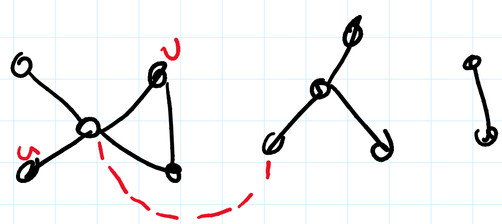
Problem (Dynamic Connectivity)

maintain undirected graph $G = (V, E)$ to support

connectivity queries: given $u, v \in V$, are u & v connected?

insert edge

delete edge



Rmk: insert-only case

$\Rightarrow$ union-find

$O(\alpha(n))$ amort. update/query

fully dynamic?

Without amort:

Frederickson '83: $O(\sqrt{m})$ update time, $O(1)$ query time

Eppstein et al. '92: $O(\sqrt{n})$ "

(assume obliv.

Eppstein et al. '92: $O(\sqrt{n})$ "

Kapron-King-Mountjoy '13: $O(\log^5 n)$ rand. (assume obliv. adversary)

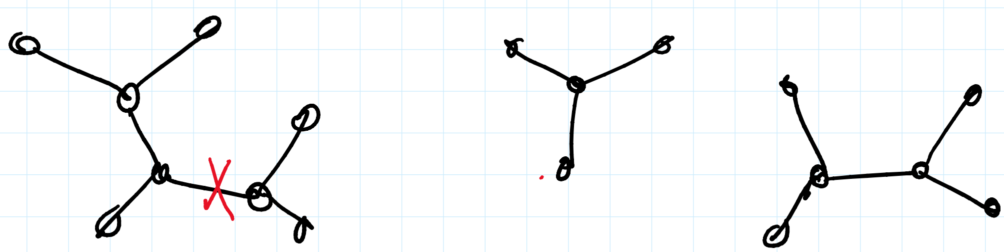
Chuzhoy, Gao, Li, Nanongkai, Peng, Saranurak '20: $O(n^{\varepsilon})$ det. (no assumption)

with amortization:

- Henzinger-King '97: $O(n^{1/3})$ rand. update, $O(1)$ query
- Henzinger-King '95: $O(\log^3 n)$ rand. update, $O\left(\frac{\log n}{\log\log n}\right)$ query
- Henzinger-Thorup '96: $O(\log^2 n)$ rand. update "
- Holm-de Lichtenberg-Thorup '98: $O(\log^2 n)$ det., "
- Thorup '00: $O(\log n (\log\log n)^3)$ rand. $O\left(\frac{\log n}{\log\log\log n}\right)$ query
- — — '17: $O(\log n (\log\log n)^2)$

Warm-Up: Case of Forest

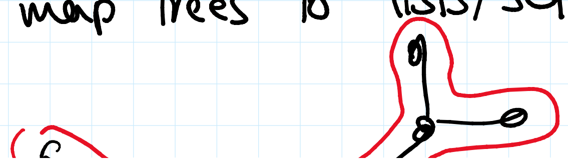
acyclic graphs

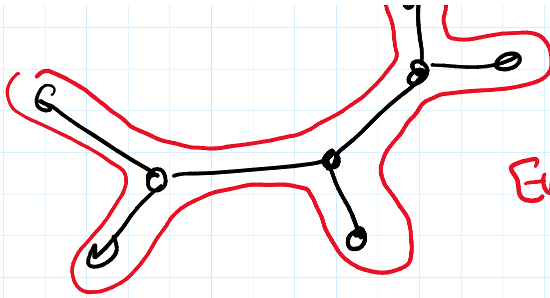


insert: link 2 trees

delete: cut tree into 2

idea - map trees to lists/sequences - -





Euler tour