All-Distances Sketches, Revisited:
HIP Estimators for Massive Graphs Analysis

Edith Cohen

Microsoft Research

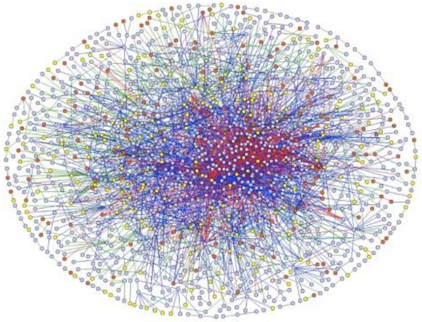
Presented by:

Thomas Pajor

Microsoft Research

Microsoft Research

Very Large Graphs



- Model many types of relations and interactions
  - Call detail data, email exchanges
  - Web crawls
  - Social Networks: Twitter, Facebook, linkedIn
  - Web searches, Commercial transactions,…
- Need for scalable analytics:
  - **Centralities/Influence** (power/importance/coverage of a node or a set of nodes): Viral marketing,…
  - **Similarities/Communities** (how tightly related are 2 or more nodes): Recommendations, Advertising, Marketing

All-Distances Sketches (ADS) [C '94]

- **Summary structures: For each node $i \in [n]$:** $\mathbf{ADS}(i)$ **"samples" the distance relations of i to all other nodes.**

  Useful for queries involving a single node: Neighborhood cardinality and statistics

- **Sketches of different nodes are *coordinated*:** **related in a way that is useful for queries that involve multiple nodes (similarities, influence, distance)**

All-Distances Sketches (ADS) [C '94]
Basic properties

- m edges, n nodes, parameter $k \geq 1$ which controls trade-off between sketch size and information

- ❑ **ADSs work for directed or undirected graphs**
- ❑ **Compact size:** $\mathbf{E}[|\mathbf{ADS}(i)|] \leq k \ln n$
- ❑ **Scalable Computation**: $km \ln n$ edge traversals to compute $\mathbf{ADS}(i)$ for *all* nodes i
- ❑ Many applications

All-Distances Sketches: Definition

$\mathbf{ADS}(\boldsymbol{v})$ is a list of pairs of the form $(\boldsymbol{i}, \boldsymbol{d}_{\boldsymbol{vi}})$

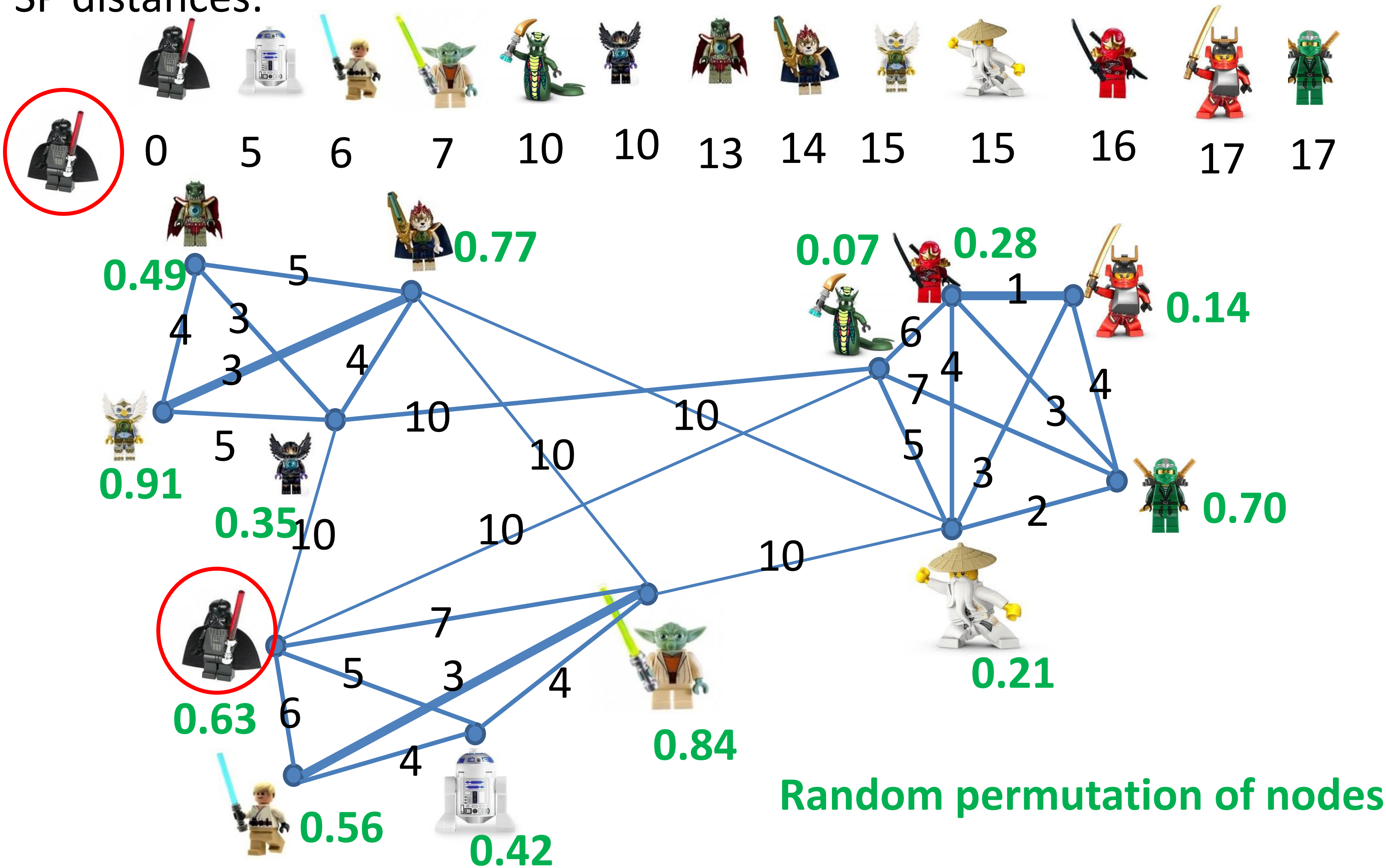
- Draw a random permutation of the nodes: $\boldsymbol{r} : [\boldsymbol{n}] \rightarrow [\boldsymbol{n}]$
- $i \in \mathbf{ADS}(\boldsymbol{v}) \Leftrightarrow \boldsymbol{r}(\boldsymbol{i}) < \boldsymbol{k}^{\text{th}}$ smallest rank amongst nodes that are closer to $\boldsymbol{v}$ than $\boldsymbol{i}$

This is a ***bottom-k ADS***, it is the union of bottom-k MinHash sketches (k smallest rank) of all "neighborhoods."

There are other ADS "flavors", vary by the rank distribution $\boldsymbol{r}$ (e.g. can use $r(i) \sim U[0,1]$) or sketch structure.

ADS example

SP distances:



ADS example $k = 1$

All nodes sorted by SP distance from 

0.63 0.42 0.56 0.84 0.07 0.35 0.49 0.77 0.91 0.21 0.28 0.14 0.70

$k = 1$:

0.63 0.42 0.07

ADS example $k = 2$

Sorted by SP distances from 

0.63 0.42 0.56 0.84 0.07 0.35 0.49 0.77 0.91 0.21 0.28 0.14 0.70

$k = 2$:

0.63 0.42 0.56 0.07 0.35 0.21 0.14

“Basic” use of ADSs (90’s– 2013)

Extract MinHash sketch of the d neighborhood of v, $N_d(v)$, from ADS(v) :

bottom-$k\{i \in ADS(v) | d_{vi} < d\}$

From MinHash sketches, we can estimate:

- Cardinality $|N_d(v)|$
  - Estimate has CV $\frac{\sigma}{\mu} \leq \frac{1}{\sqrt{k-2}}$ (*optimally* uses the information in the MinHash sketch)
- Jaccard similarity of $N_d(v)$ and $N_d(u)$,
- Other relations of $N_d(v)$ and $N_d(u)$,

Historic Inverse Probability (HIP) inclusion probability & estimator

- For each node i, we estimate the "**presence" of $\boldsymbol{i}$ with respect to $\boldsymbol{v}$**: $I_{v \rightsquigarrow i}$ (=1 if $v \rightsquigarrow i$, 0 otherwise)
- Estimate is $\boldsymbol{a_{vi} = 0}$ if $i \notin \mathrm{ADS}(v)$.
- If $i \in \mathrm{ADS}(v)$, we compute the probability p that it is included, *conditioned* on fixed rank values of all nodes that are closer to v than i. We then use the *inverse-probability* estimate $\boldsymbol{a_{vi} = \frac{1}{p}}$. [HT52]
- This is unbiased (when $p > 0$):

$$\mathrm{E}[a_{vi}] = p\frac{1}{p} + (1-p)0 = 1$$

Bottom-k HIP

- For bottom-k *and* $r \sim U[0,1]$

  $p = \text{k}^{\text{th}}\{r(u) | u \in \text{ADS}(v) \wedge d_{vu} < d_{vi}\}$

⇨ HIP can be used with all flavors of MinHash sketches. Over distance (ADS) or time (Streams)

Example: HIP estimates

Bottom-2 ADS of 

	0.63	0.42	0.56	0.07	0.35	0.21	0.14
p:	1	1	0.63	0.56	0.42	0.35	0.21
$a = \frac{1}{p}$:	1	1	1.59	1.79	2.38	2.86	4.76

p:2nd smallest r value among closer nodes

HIP cardinality estimate

Bottom-2 ADS of 

distance:	0	5	6	10	10	15	17
a:	1	1	1.59	1.79	2.38	2.86	4.76

Query: $n_6(v) = |\{ i \mid d_{vi} \le 6 \}| = \sum_i I_{d_{vi} \le 6}$

$$\widehat{n_6(v)} = \sum_{(i,d_{vi}) \in ADS(v) \mid d_{vi} \le 6} a_{vi} = 1 + 1 + 1.59 = 3.59$$

Quality of HIP cardinality Estimate

Lemma: The HIP neighborhood cardinality estimator

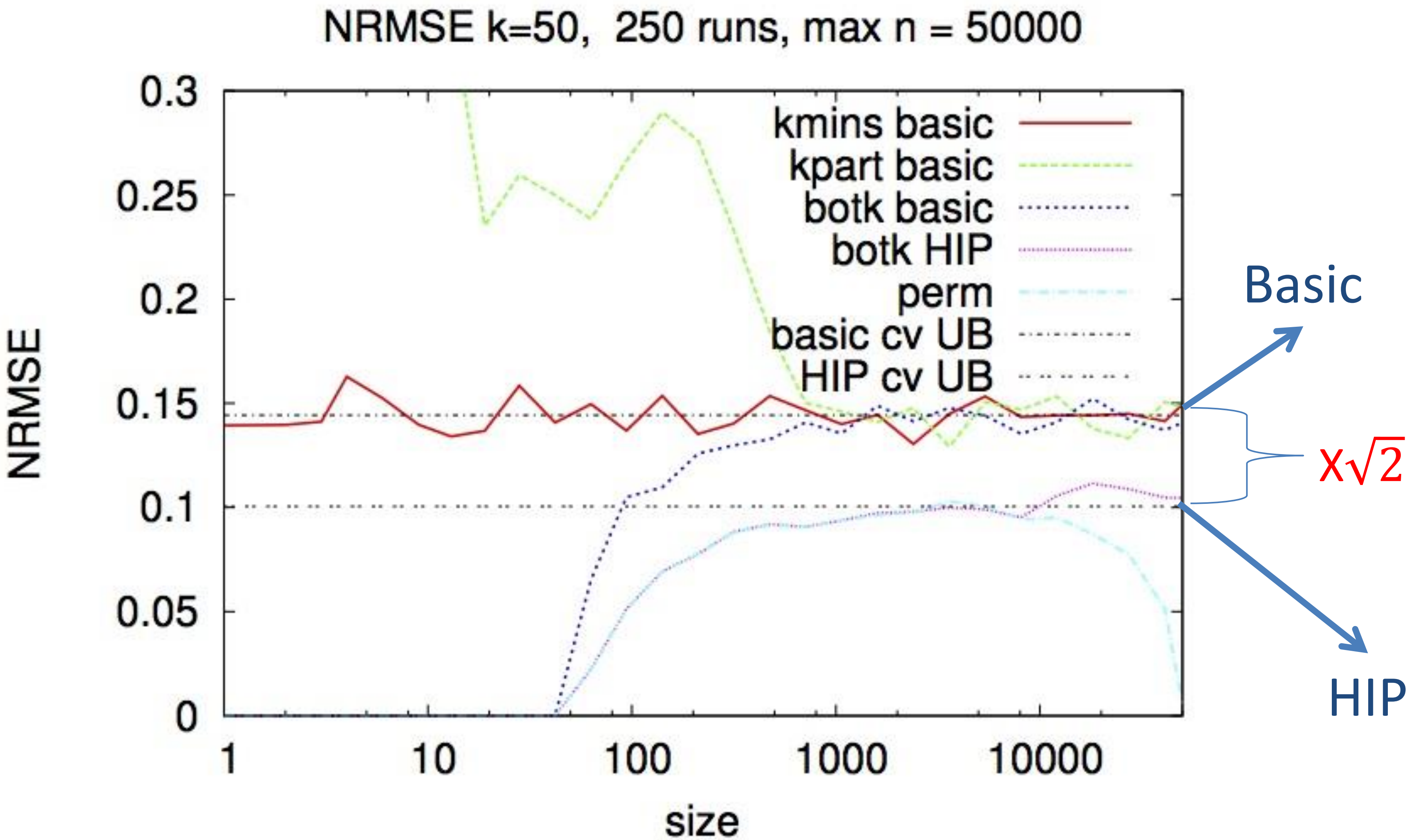
$$\widehat{n_d(v)} = \sum_{(i,d_{vi}) \in ADS(v) \mid d_{vi} \le d} a_{vi}$$

has CV $\frac{\sigma}{\mu} \le \frac{1}{\sqrt{2k-2}}$

This is $\sqrt{2}$ improvement over “basic” estimators, which have CV $\frac{\sigma}{\mu} \le \frac{1}{\sqrt{k-2}}$

See paper for the proof

HIP versus Basic estimators



HIP: applications

Querying ADSs:

- Cardinality estimation: $\sqrt{2}$ gain in relative error over “basic” (MinHash based) estimates
- More complex queries: closeness centrality with topic awareness (gain can be polynomial)
- Estimating relations (similarities, coverage) of pairs (sets) of nodes .

Streaming:

- Approximate distinct counting on streams.

Topic-aware Distance-decay Closeness Centrality

$$C_v = \sum_i \alpha(d_{vi})\beta(i)$$

- α non increasing; β some filter

Centrality with respect to a filter $\beta(i)$:

- Topic, interests, education level, age, community, geography, language, product type
- Applications for filter: attribute completion, targeted advertisements

....Closeness Centrality

$$\boldsymbol{C_v = \sum_i \alpha(d_{vi})\beta(i)}$$

- $\boldsymbol{\alpha}$ non increasing; $\boldsymbol{\beta}$ some filter

- Polynomial (Harmonic) decay: $\alpha(x) = \frac{1}{x}$
- Exponential decay $\alpha(x) = e^{-x}$
- Threshold ($\in N_d(v)$): $\alpha(x) = 1 \Leftrightarrow x \leq d$

HIP estimates of Centrality

$$\boldsymbol{C_v = \sum_i \alpha(d_{vi})\beta(i)}$$

- $\boldsymbol{\alpha}$ non increasing; $\boldsymbol{\beta}$ some filter

$$\boldsymbol{\widehat{C_v} = \sum_{i \in ADS(v)} a_{vi}\, \alpha(d_{vi})\, \beta(i)}$$

HIP estimates: closeness to good/evil

Bottom-2 ADS of 

							
distance:	0	5	6	10	10	15	17
a:	1	1	1.59	1.79	2.38	2.86	4.76
$\boldsymbol{\beta}$:	**0**	**1**	**1**	**0.2**	**0.1**	**1**	**0.9**

Filter: $\beta \in [0,1]$ measures "goodness"

Distance-decay: $e^{-d_{vi}}$

$$\widehat{C_v} = \sum_{i \in ADS(v)} a_{vi}\beta(i)e^{-d_{vi}} = e^{-5} + e^{-6} + 0.3e^{-10} + \cdots$$

Counting Distinct Elements on Data Stream

32, 12, 14, 32, 7, 12, 32, 7, 6, 12, 4,

Elements occur multiple times, we want to count the number of *distinct* elements *approximately* with "small" storage, about $O(\log \log n)$

- Best practical and theoretical algorithms maintain a MinHash sketch. Cardinality is estimated by applying an estimator to sketch [Flajolet Martin 85],…
- Best (in practice) is the HyperLogLog (HLL) algorithm and variations. [Flajolet + FGM 2007],…

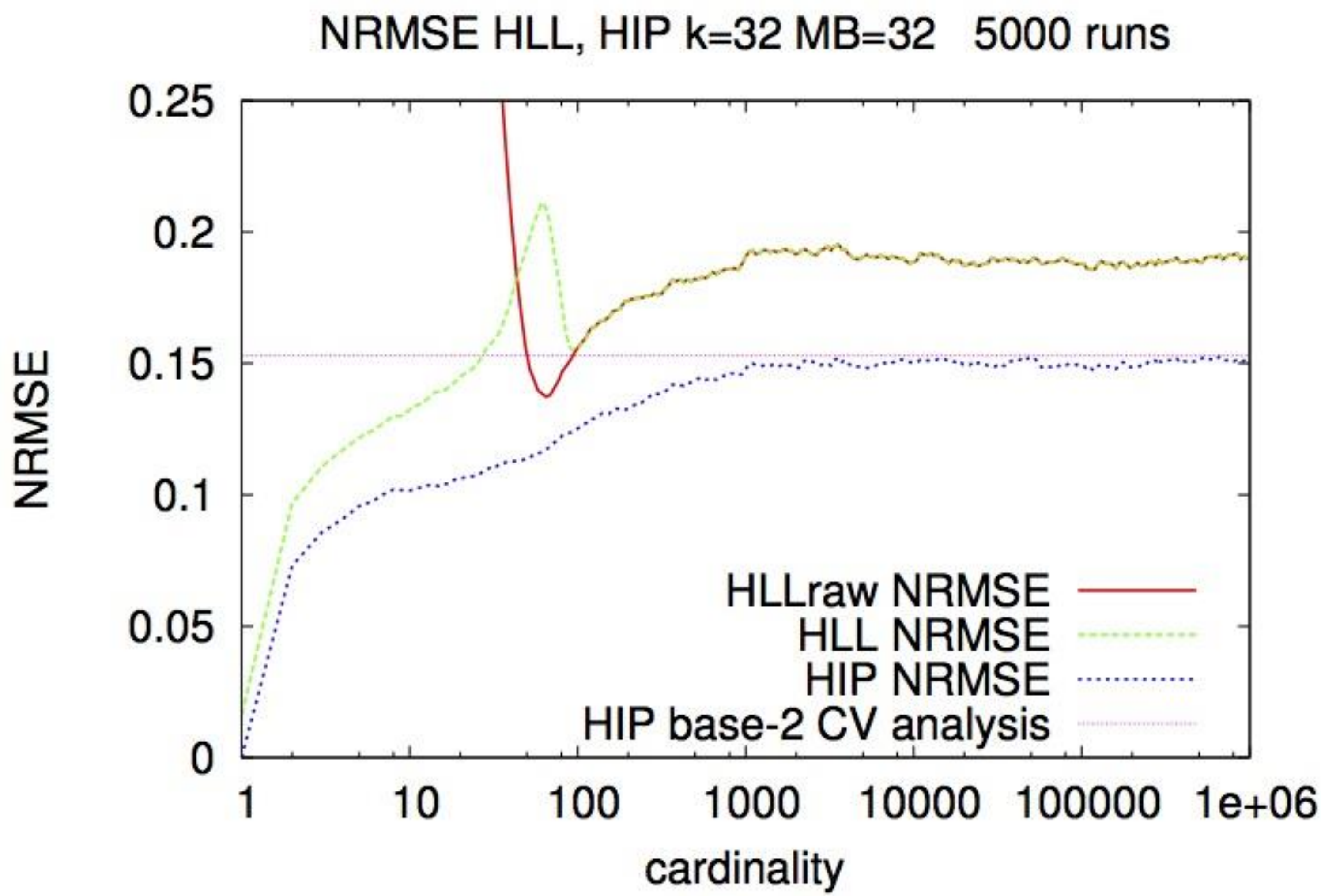
Counting Distinct Elements with HIP

We maintain a MinHash sketch **and** an approximate counter -- variation on [Morris77]. The counter explicitly maintains an *approximate distinct count*.

- Each time the sketch is updated ($\mathrm{E} \leq k \ln n$ times), we increase the counter (add the HIP estimate for the inserted new distinct element)
- The approximate counter can be represented with few bits (e.g., can be a relative correction to sketch-based estimate or share its “exponent”)

This works with any MinHash sketch. In experiments, for comparison, we use the same sketch as HyperLogLog (HLL).

HLL vs. HIP (on HLL sketch)



Conclusion

- **ADS**: old but a very versatile and powerful tool for (scalable approximate) analytics on very large graphs: distance/similarity oracles, distance distribution, closeness, coverage, influence, tightness of communities
- **HIP**: simple and practical technique, applicable with ADSs and streams

Further ADS+HIP applications:

- closeness similarity (using ADS+HIP) [CDFGGW COSN 2013]
- … Timed-influence oracle

Thank you!!

Legends of Chima

Cragger

Laval

Eris

Rascal

Ninjago

Acidicus

Nya
“Samurai X”

Lloyd
“The Green Ninja”

Kai
“Ninja of Fire”

Sensei Wu

Star Wars

Darth Vader

Luke
Skywalker

R2-D2

Yoda