Coordinated Sampling

Ok, so lets sample..

- Lets start with sampling 1 element..
- We get a stream of items want to stay with 1 uniformly random

Ok, so lets sample..

- Lets start with sampling 1 element..
- We get a steam of items want to stay with 1 uniformly random

1

Sample: 1

Ok, so lets sample..

- Start with 1 element..
- We get a steam of items want to stay with 1 uniformly random

1 2

Sample: 1

Replace with probability 1/2

Ok, so lets sample..

- Start with 1 element..
- We get a steam of items want to stay with 1 uniformly random

Sample: 

Replace with probability 1/3

Ok, so lets sample..

- Start with 1 element..
- We get a steam of items want to stay with 1 uniformly random

Sample: 

Replace with probability 1/3

Reservoir sampling (Vitter 85)

- Start with 1 element..
- We get a steam of items want to stay with 1 uniformly random

1 2 3 4

Sample: 3

Replace with probability 1/4

If an item appears more than once ?

Min hash

Choose at random a hash function h from some family H

Your sample is the item x with min h(x)

Need h to be min-wise independent..

“Permanent random numbers”

We shall assume that $h(x) \in U[0,1]$

$h(x_i)$ are independent

Ok, so lets **sample k items**

- **k-mins:** Repeat the experiment k times with $h_1, h_2, \ldots, h_k$

- **Bottom-k:** Take the k smallest

Ok, so lets **sample k items**

- **k-mins:** Repeat the experiment k times with $h_1, h_2, \ldots, h_k$ **Sampling with replacement**

- **Bottom-k:** Take the k smallest **Sampling without replacement**

Develop estimators

Develop estimators

- For the size **of the set = # of distinct items ?**

- For the size of selected subsets (say, the number of **green** items) ?

Estimators (# greens: |G|)

- Let **n = #distinct items** **(assume it is known)**
- Estimate the total # of green items to be **g·(n/k)**

Adjusted weights (HT 52)

- Let p **(=k/n)** be the probability that item x is sampled
- Let the adjusted weight of x be

$$a(x) = \begin{cases} 1/p = n/k & \textit{if } x \textit{ is sampled} \\ 0 & \textit{otherwise} \end{cases} \qquad E[a(x)] = 1$$

- Estimate |G| by

$$g\frac{n}{k} = \sum_{x \in S \cap G} \frac{1}{p} = \sum_{x \in G} a(x)$$

Adjusted weights (HT 52)

- Let p **(=k/n)** be the probability that item x is sampled
- Let the adjusted weight of x be

$$a(x) = \begin{cases} 1/p = n/k & \text{if } x \text{ is sampled} \\ 0 & \text{otherwise} \end{cases} \qquad E[a(x)] = 1$$

- Unbiased:

$$E\left[g\frac{n}{k}\right] = E\left[\sum_{x \in S \cap G} \frac{1}{p}\right] = E\left[\sum_{x \in G} a(x)\right] = \sum_{x \in G} E[a(x)] = |G|$$

Caveats

- We do not really know n = #distinct items
- So we do not know the adjusted weight (n/k)
- In fact, often we want to estimate n itself ?

Terminology

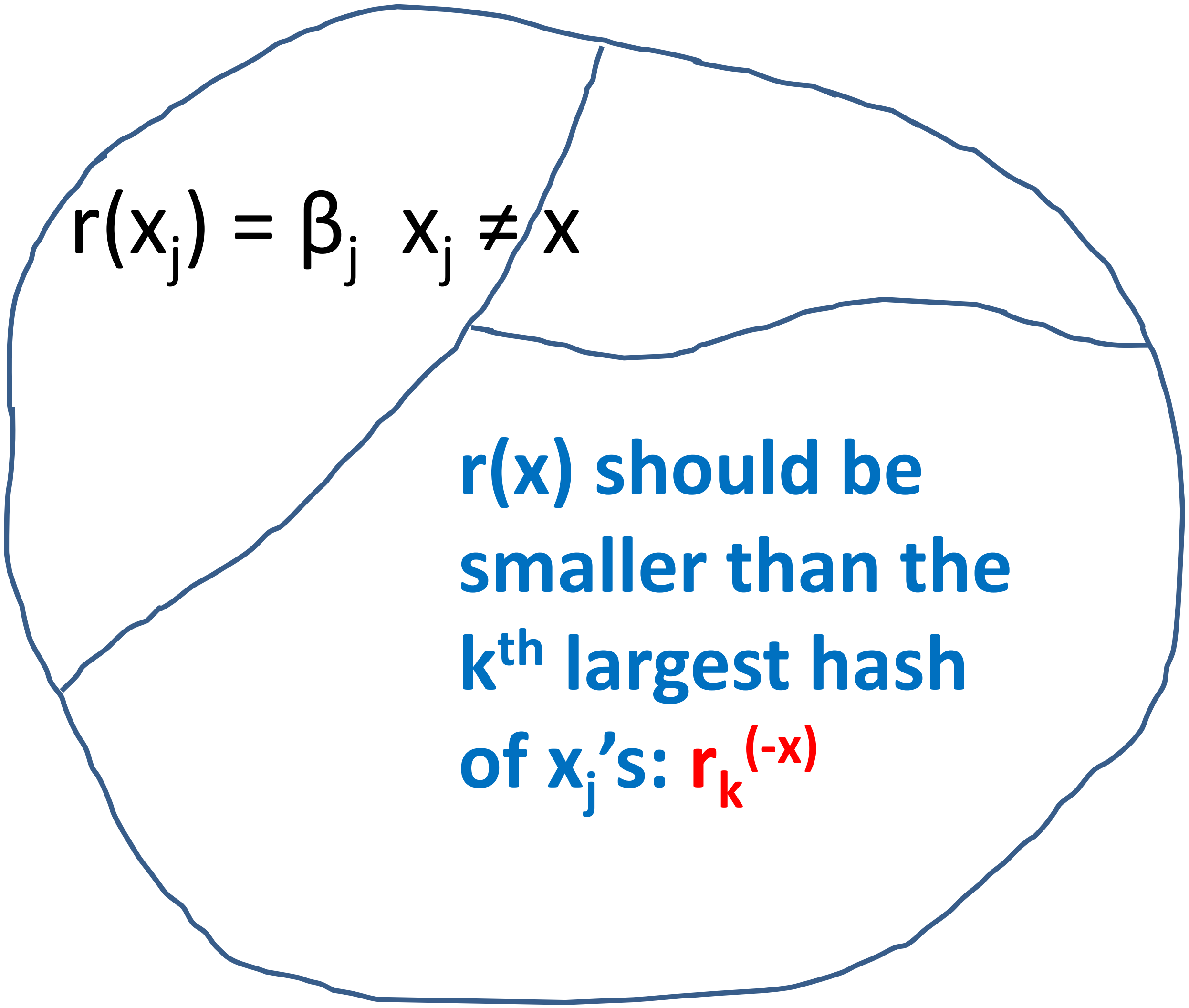
- **When items are unweighted I refer to $h(x)=r(x)$ as the rank of x**

Partition the sample space

Partition the sample space

- Recall: a ``point'' in our sample space is an assignment of ranks $\in U[0,1]$ to the elements
- Lets condition on the ranks assigned to all elements but **x**
- **What is the probability that we sample x ?**

Partition the sample space



Conditioned adjusted weights

- Let p be the probability that item x is sampled conditioned on $r(x_j) = \beta_j$, $x_j \neq x$
- **$p = r_k^{(-x)}$**

$$a(x) = \begin{cases} 1/p = 1/r_k^{(-x)} & \text{if } x \text{ is sampled} \\ 0 & \text{otherwise} \end{cases} \qquad E[a(x)] = 1$$

- Estimate $|G|$ by

$$g \frac{1}{r_k^{(-x)}} = \sum_{x \in S \cap G} \frac{1}{p} = \sum_{x \in G} a(x)$$

Partition the sample space

- How do we know $r_k^{(-x)}$?
- Lets keep with the sample the k+1 smallest hash value, r_{k+1}
- $r_k^{(-x)} = r_{k+1}$ **if x is in the sample**

Conditioned adjusted weights

- Let p be the probability that item x is sampled conditioned on $r(x_j) = \beta_j$, $x_j \neq x$
- **p = r_{k+1}**

$$a(x) = \begin{cases} 1/p = 1/r_{k+1} & \text{if } x \text{ is sampled} \\ 0 & \text{otherwise} \end{cases} \qquad E[a(x)] = 1$$

  - Estimate $|G|$ by

$$g\frac{1}{r_k} = \sum_{x \in S \cap G} \frac{1}{p} = \sum_{x \in G} a(x)$$

Coordinated sampling

Coordinated sampling

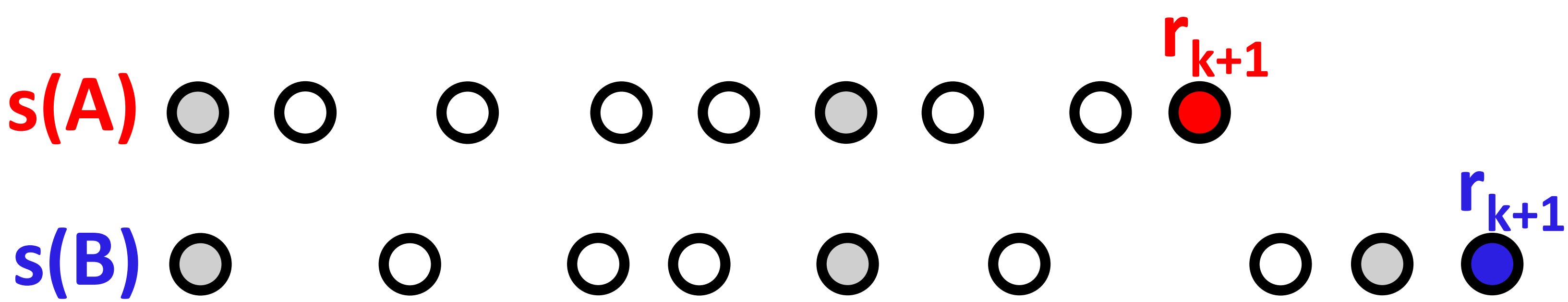
- Say we have two sets **A** and **B**
- We have a bottom-k sketches **s(A)** of **A** and **s(B)** of **B**, **prepared with the same hash function !**
- Want to estimate $|A \cap B|$?

Multiple set scenario

item	Router 1	Router 2
132.169.1.1	✓	
132.66.235.47		✓
157.166.238.17	✓	✓
128.112.132.86	✓	
170.149.172.130		
72.52.4.119	✓	✓
192.115.76.44		
128.139.199.7		✓
107.23.224.136	✓	✓

Coordinated sampling

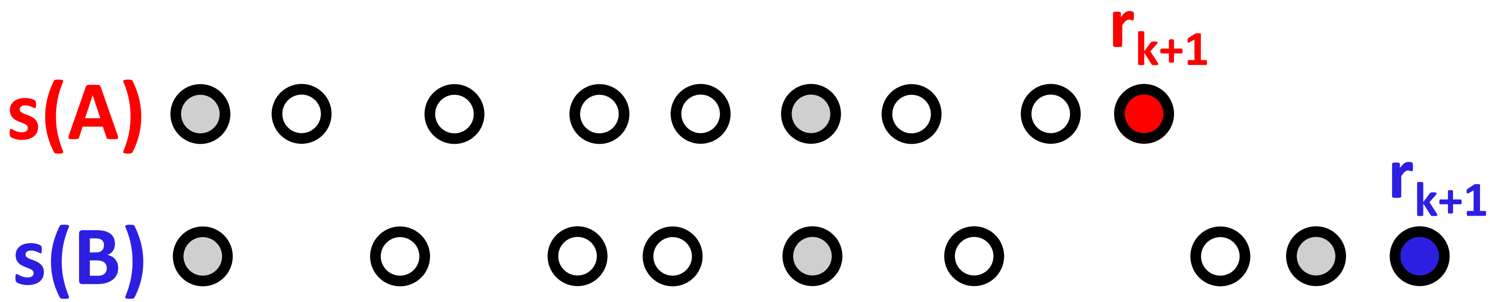
Give adjusted weights to the elements in A∩B:



$$a(x)=\begin{cases}\dfrac{1}{p} & if\ x\in s(A)\cap s(B)\\ 0 & otherwise\end{cases}$$

p is the probability that $x \in s(A) \cap s(B)$

Estimating A∩B with coordination



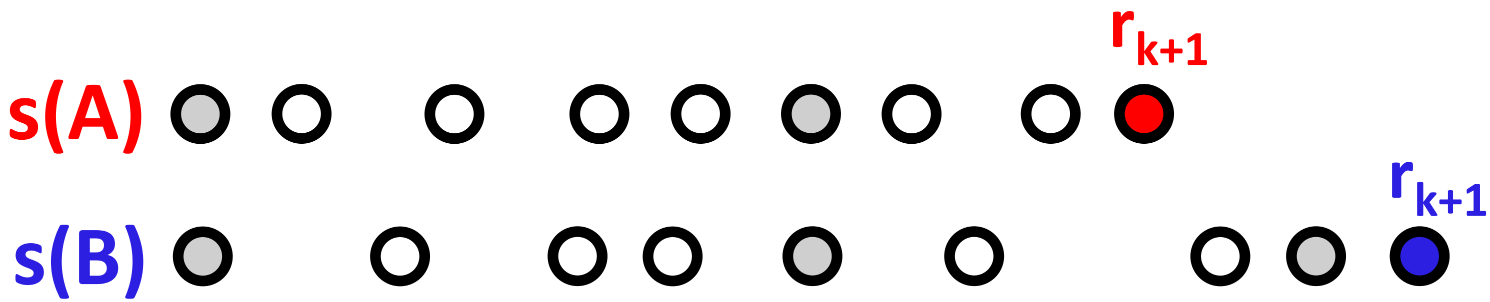
$$a(x) = \begin{cases} \dfrac{1}{p} & \text{if } x \in s(A) \cap s(B) \\ 0 & \text{otherwise} \end{cases}$$

p is the probability that x ∈S(A)∩S(B)

conditioned on all ranks of y≠x

What is p ??

Estimating A∩B with coordination



$$a(x) = \begin{cases} \dfrac{1}{p} & if\ x \in s(A) \cap s(B) \\ 0 & otherwise \end{cases}$$

p = min{r_{k+1},r_{k+1}}

rather than p = $r_{k+1} \cdot r_{k+1}$ without coordination !

Homework

- **Develop an estimate for |B-A|** ?

Weighted items

- Each item has a weight $w(x)$
- Want to estimate total weight of a selected subset: $w(G)$=sum of the weights of the greens

Multiple set scenario

item	Router 1	Router 2
132.169.1.1	30	
132.66.235.47		40
157.166.238.17	100	100
128.112.132.86	1000	1000
170.149.172.130		
72.52.4.119	440	440
192.115.76.44		
128.139.199.7		515
107.23.224.136	w	**w**

Weighted items

- Assign adjusted weights:

$$a(x) = \begin{cases} \dfrac{w(x)}{p} = \dfrac{w(x)}{r_{k+1}} & \textit{if } x \textit{ is sampled} \\ 0 & \textit{otherwise} \end{cases} \qquad E[a(x)] = w(x)$$

- Estimate w(G)=weight of greens by

$$\sum_{x \in S \cap G} \frac{w(x)}{r_{k+1}} = \sum_{x \in S \cap G} \frac{1}{p} = \sum_{x \in G} a(x)$$

Weighted items

- Assign adjusted weights:

$$a(x) = \begin{cases} \frac{w(x)}{p} = \frac{w(x)}{r_{k+1}} & \textit{if } x \textit{ is sampled} \\ 0 & \textit{otherwise} \end{cases} \qquad E[a(x)] = w(x)$$

- This is an unbiased estimator

$$E\left[\sum_{x \in S \cap G} \frac{w(x)}{r_{k+1}}\right] = E\left[\sum_{x \in S \cap G} \frac{1}{p}\right] = E\left[\sum_{x \in G} a(x)\right] = \sum_{x \in G} E[a(x)] = \sum_{x \in G} w(x)$$

HT Variance

$$Var[a(x)] = p\left(\frac{w(x)}{p}\right)^2 - w^2(x) = w^2(x)\left(\frac{1}{p} - 1\right)$$

- Want p to be large
- In particular if w(x) is large

Weighted items

➔ **Draw the random rank of x proportionally to w(x)**

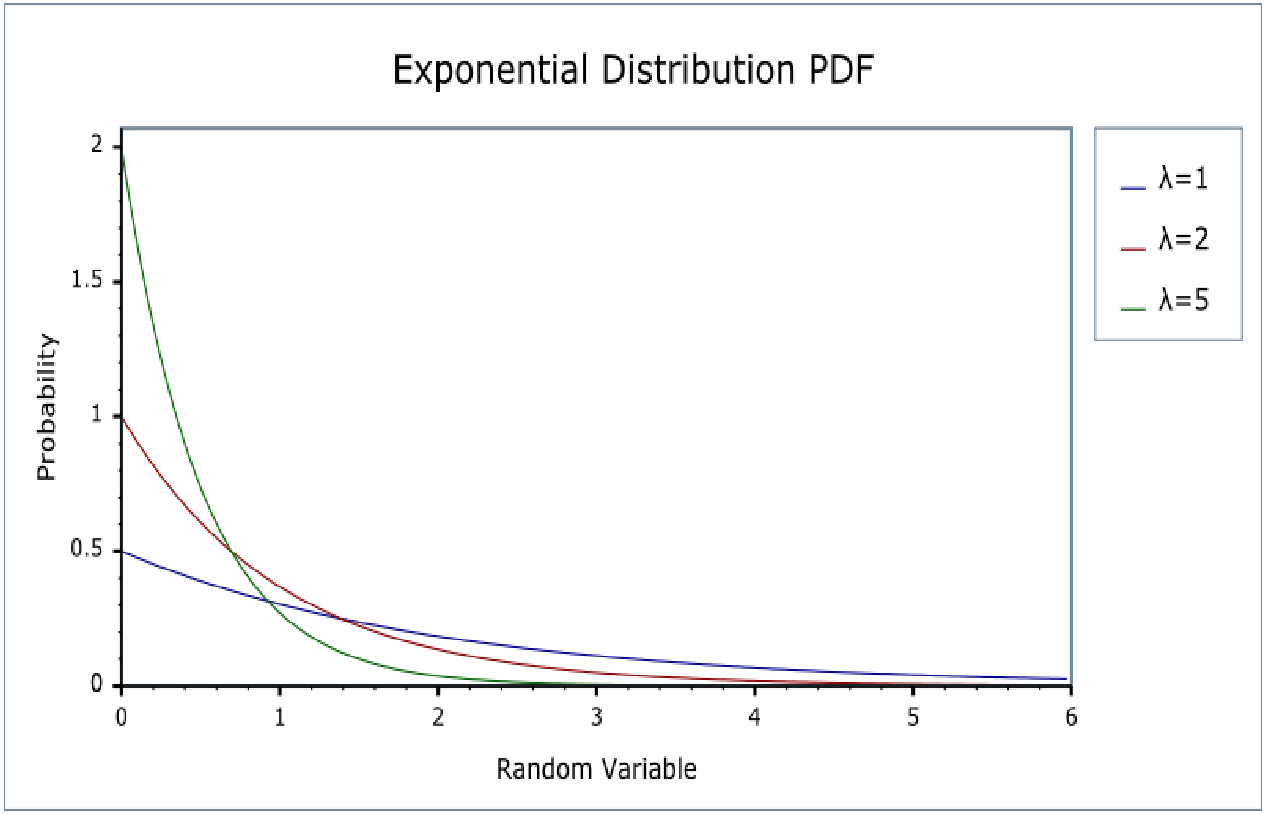
There are different ways to do it

Priority ranks

- Set **$r(x) = h(x)/w(x)$**, $h(x) \in U[0,1]$

Draw from: Exp(w)

- PDF: $we^{-wx}, x \geq 0$
- CDF: $1 - e^{-wx}$
- $\mu = \sigma = \frac{1}{w}$



$\mathrm{Exp}(w)$ ranks

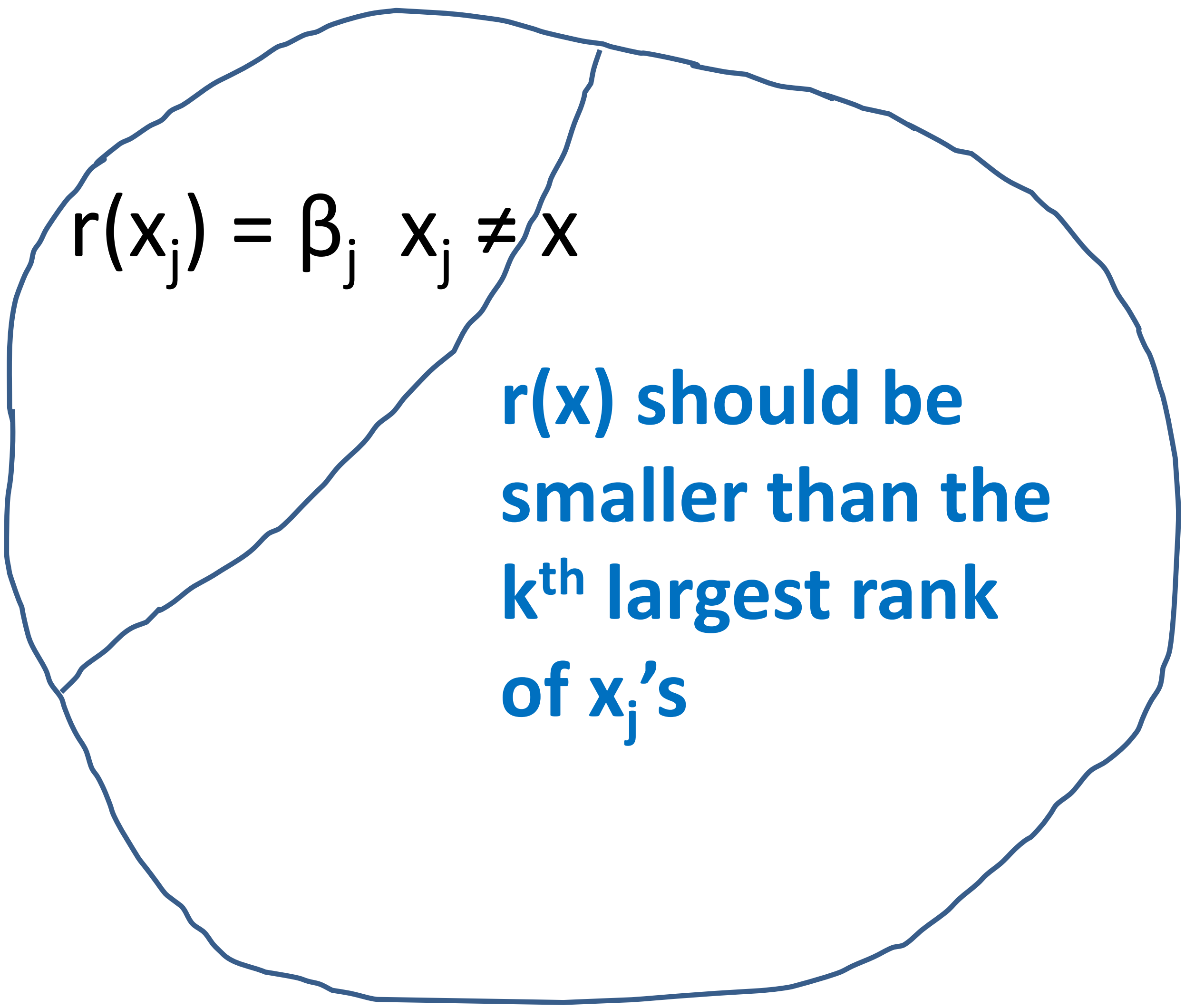
- Larger weight ➔ smaller rank
- Equivalent to weighted sampling with replacement
- Can draw using the following:

$$\mathrm{h(x)} \in \mathrm{U}[0,1], \quad \mathrm{r(x)} = -\frac{\ln(1-\mathrm{h(x)})}{\mathrm{w}} \sim \mathrm{Exp(w)}$$

Partition the sample space

- A ``point'' in our sample space is an assignment of ranks $\in$ U[0,1]/w(x) to each element x
- Lets condition on the ranks assigned to all elements but one, say **x**
- **What is the probability that we sample x ?**

Partition the sample space



Weighted items

- Adjusted weights for weighted sampling

$$a(x) = \begin{cases} \dfrac{w(x)}{p(h(x) < r_{k+1} w(x))} & \textit{if } x \textit{ is sampled} \\ 0 & \textit{otherwise} \end{cases}$$

- Estimate w(G)=weight of greens by

$$\sum_{x \in S \cap G} \frac{w(x)}{p(h(x) < r_{k+1} w(x))} = \sum_{x \in S \cap G} \frac{1}{p} = \sum_{x \in G} a(x)$$

Multiple weights

Multiple set scenario

item	Router 1	Router 2	Router 3
132.169.1.1	30		200
132.66.235.47		40	
157.166.238.17	100	200	90
128.112.132.86	1000	500	
170.149.172.13	102		9999
72.52.4.119		440	450
192.115.76.44			330
128.139.199.7		515	111
107.23.224.136	w_1	w_2	w_3

Weighted sampling and Coordination

- We want the samples to be weighted
- We also want coordination, that is if we sampled x at router 1 then x is more likely to be sampled also at router 2
- How could we achieve this ?

Coordinated priority ranks

$$\boldsymbol{h(x) \sim U[0,1]}$$

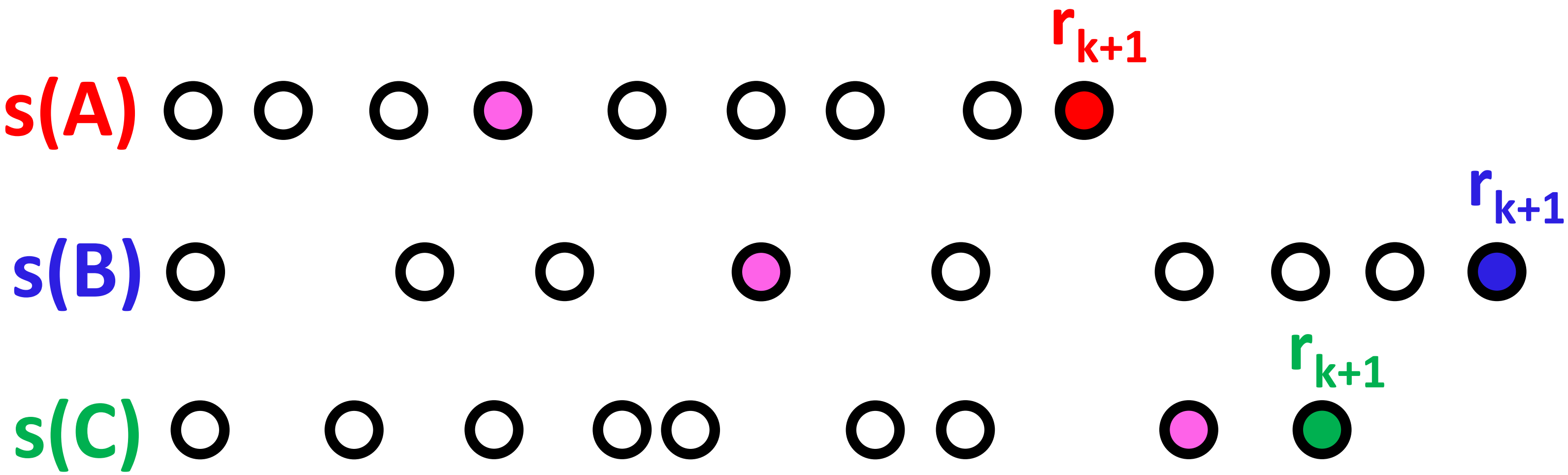
$r_1 = \mathbf{h(x)}/w_1$

$r_2 = \mathbf{h(x)}/w_2$

$r_3 = \mathbf{h(x)}/w_3$

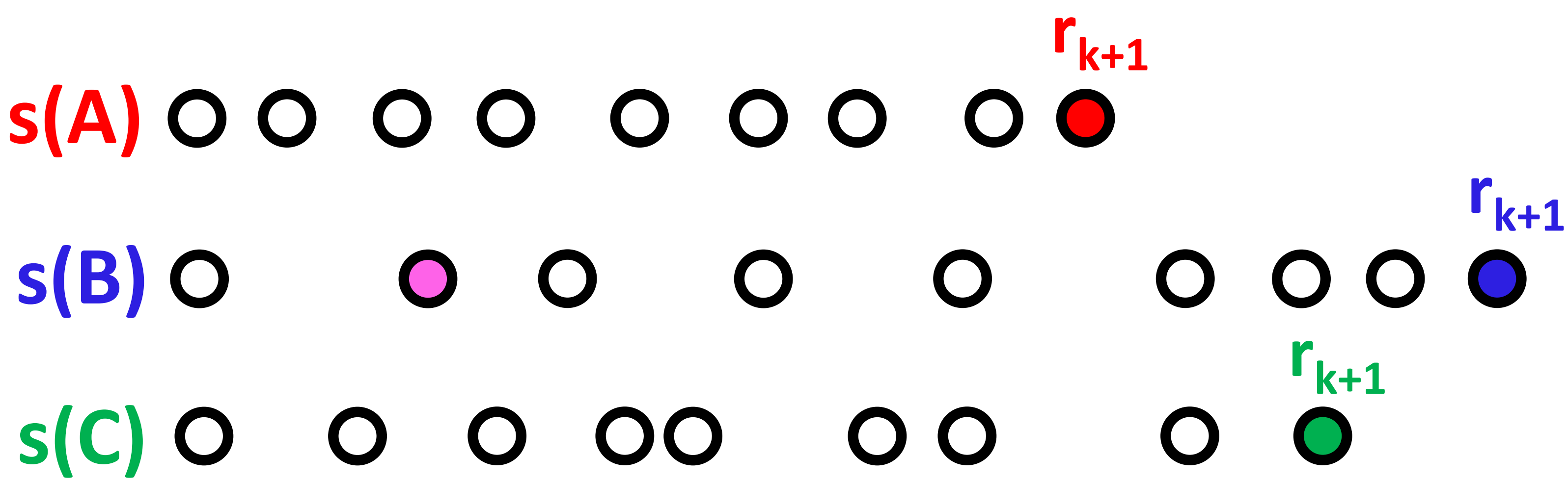
Coordinated weighted samples

- Estimate $\sum_x \max\{w_1(x), w_2(x), w_3(x)\}$?



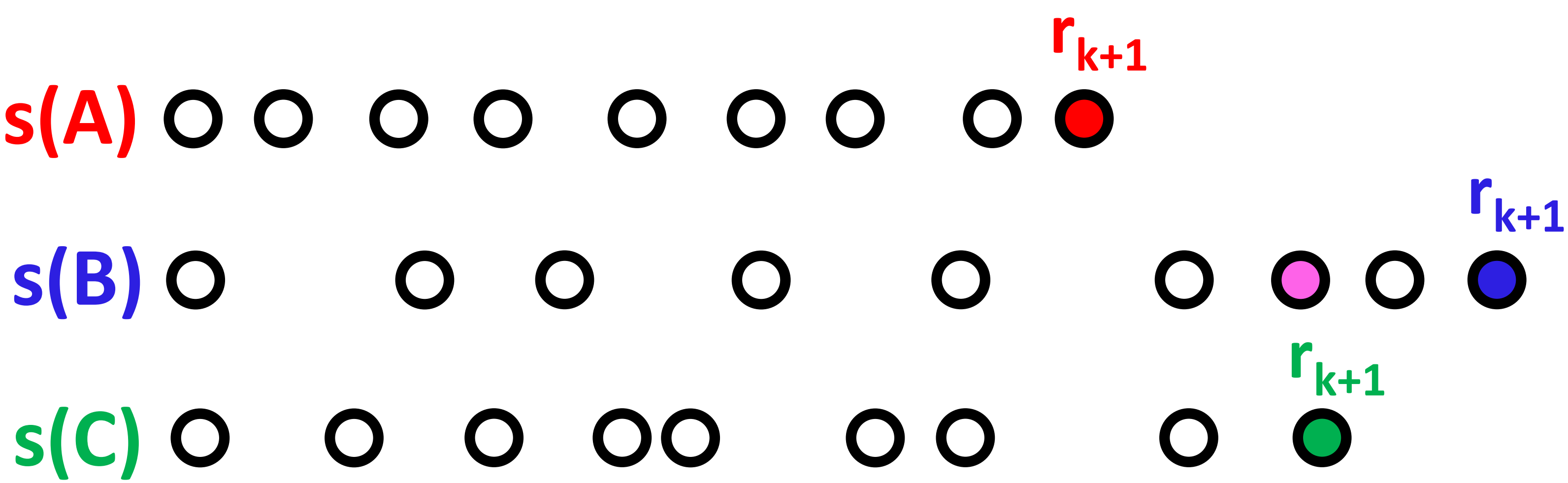
Coordinated weighted samples

- Estimate $\sum_x \max\{w_1(x), w_2(x), w_3(x)\}$?



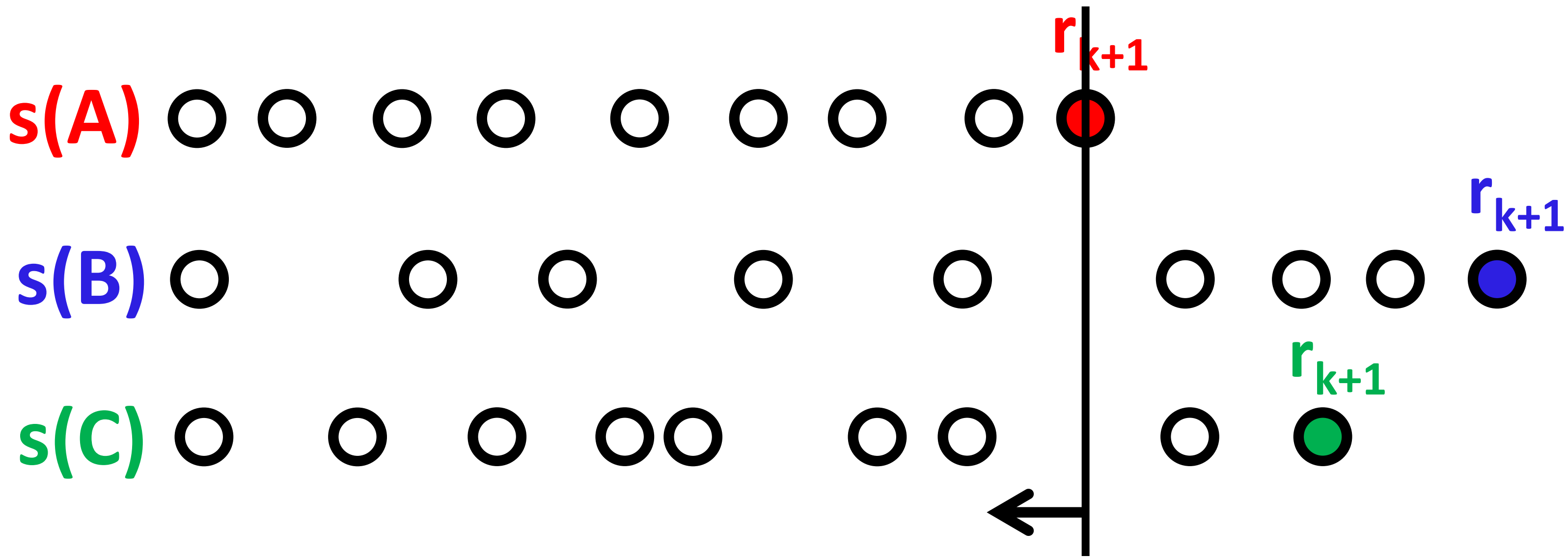
Coordinated weighted samples

- Estimate $\sum_x \max\{w_1(x), w_2(x), w_3(x)\}$?



Coordinated weighted samples

- Estimate $\sum_x \max\{w_1(x), w_2(x), w_3(x)\}$?



- We know $\max\{w_1(x), w_2(x), w_3(x)\}$ for all x such that $\exists \mathbf{i}\ \ r_i(x) < r_{k+1}$

Coordinated sampling

$$a(x) = \begin{cases} \dfrac{\max\{w_i(x)\}}{p} & \min\{r_1(x), r_2(x), r_3(x)\} < \\ & \qquad \min\{r_{k+1}, r_{k+1}, r_{k+1}\} \\ 0 & otherwise \end{cases}$$

p =P(min{r_1(x),r_2(x),r_3(x)} < min{$\mathbf{r_{k+1}}$,$\mathbf{r_{k+1}}$,$\mathbf{r_{k+1}}$})

p =P(h(x) < min{$\mathbf{r_{k+1}}$,$\mathbf{r_{k+1}}$,$\mathbf{r_{k+1}}$}max{$\mathbf{w_1(x)}$,$\mathbf{w_2(x)}$,$\mathbf{w_3(x)}$})

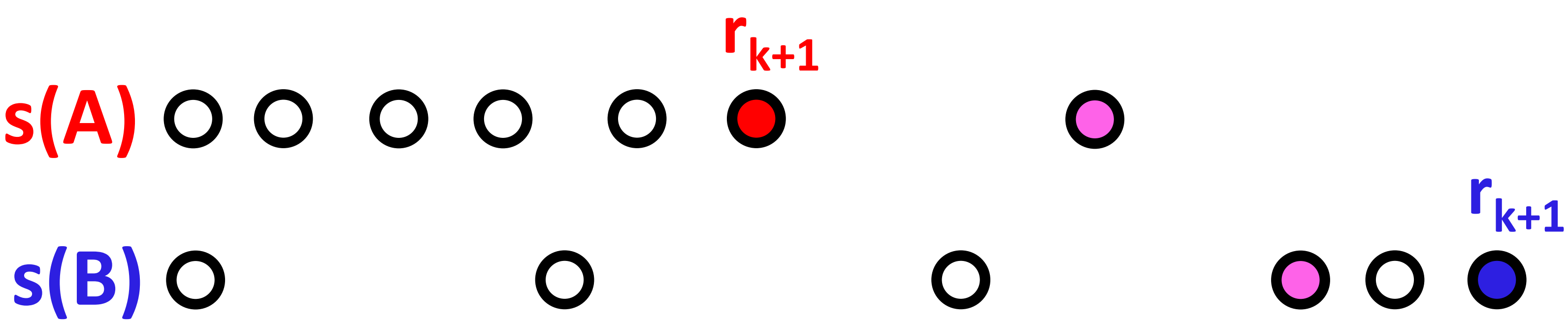
Homework

- Develop an estimator for $\sum_x \text{median}\{\mathbf{w_1(x)}, \mathbf{w_2(x)}, \mathbf{w_3(x)}\}$?

Using partial information

Partial information

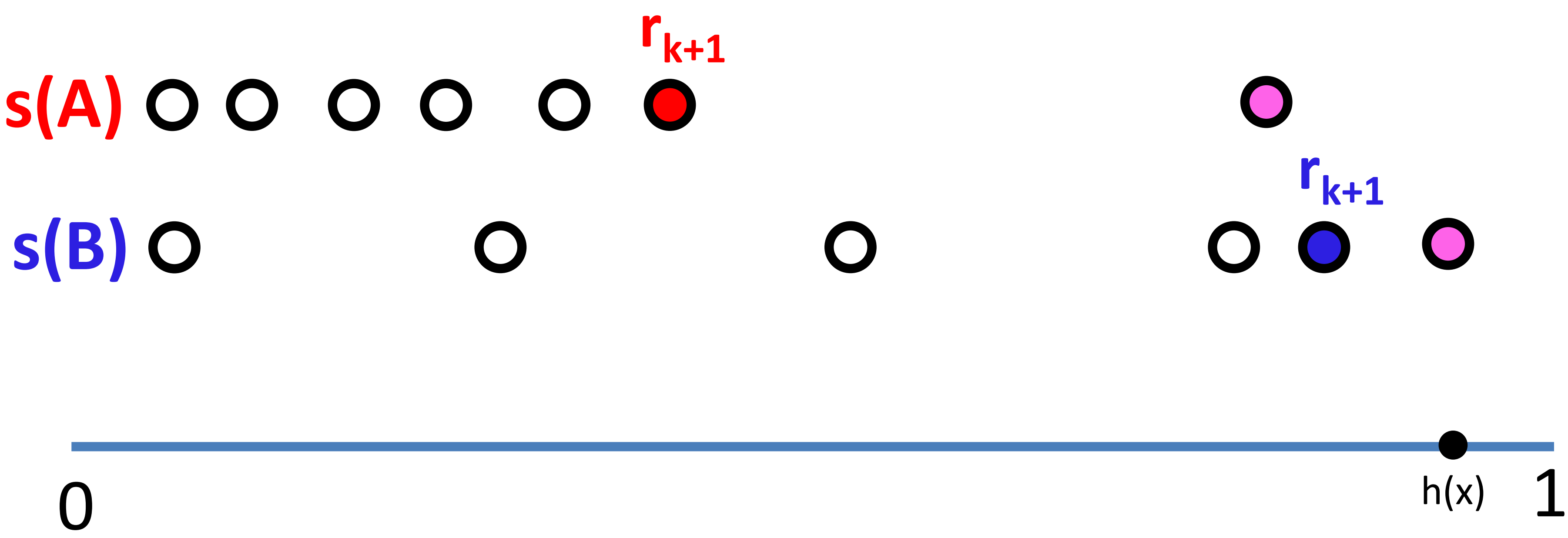
- Estimate $\sum_{\mathbf{x}} \max\{\mathbf{w_1(x)}, \mathbf{w_2(x)}\}$?



- We may not know $\max\{\mathbf{w_1(x)}, \mathbf{w_2(x)}\}$ but we do know $\mathbf{w_2(x)}$ **which is a lower bound**

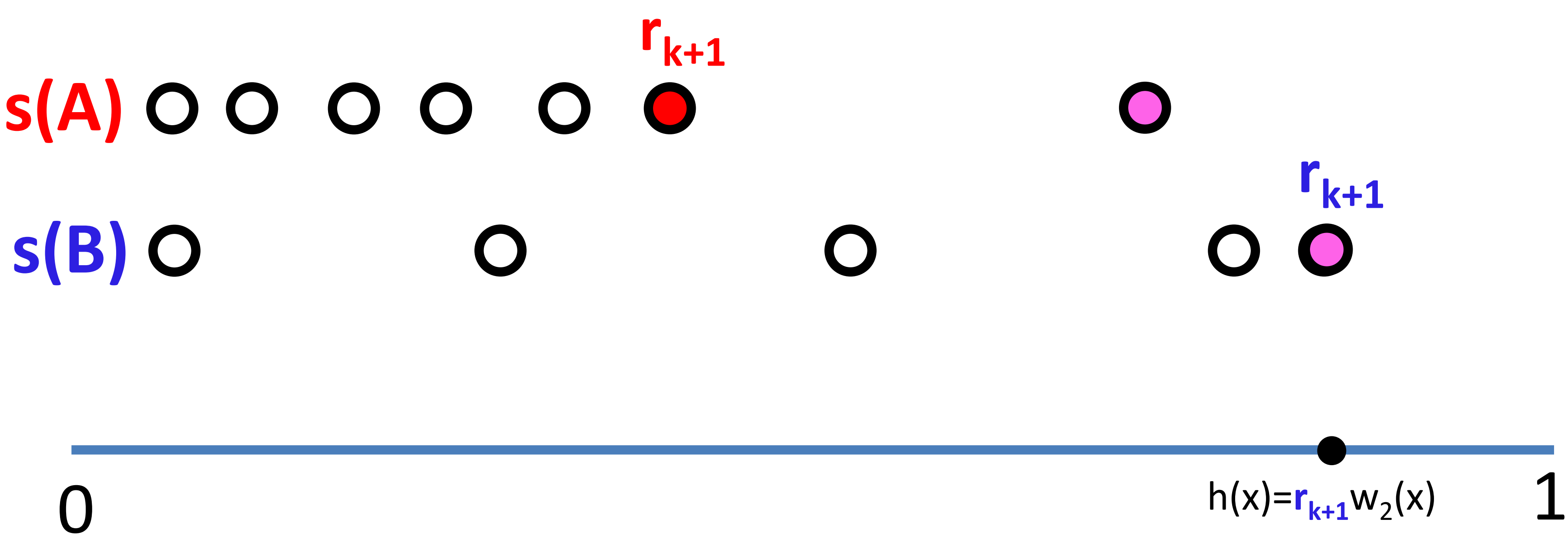
Partial information

- Estimate $\sum_x \max\{w_1(x), w_2(x)\}$?



Partial information

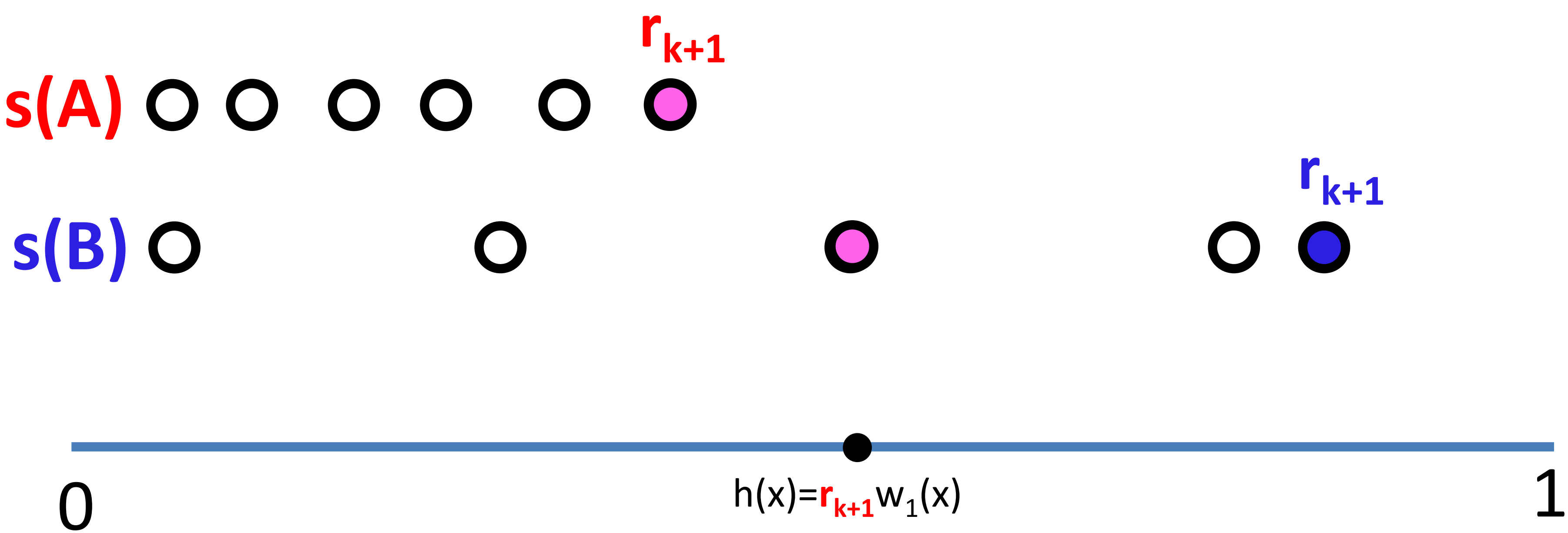
- Estimate $\sum_{\mathbf{x}} \max\{\mathbf{w_1(x)}, \mathbf{w_2(x)}\}$?



Estimate: $w_2(x)/(r_{k+1}w_2(x))=1/r_{k+1}$

Partial information

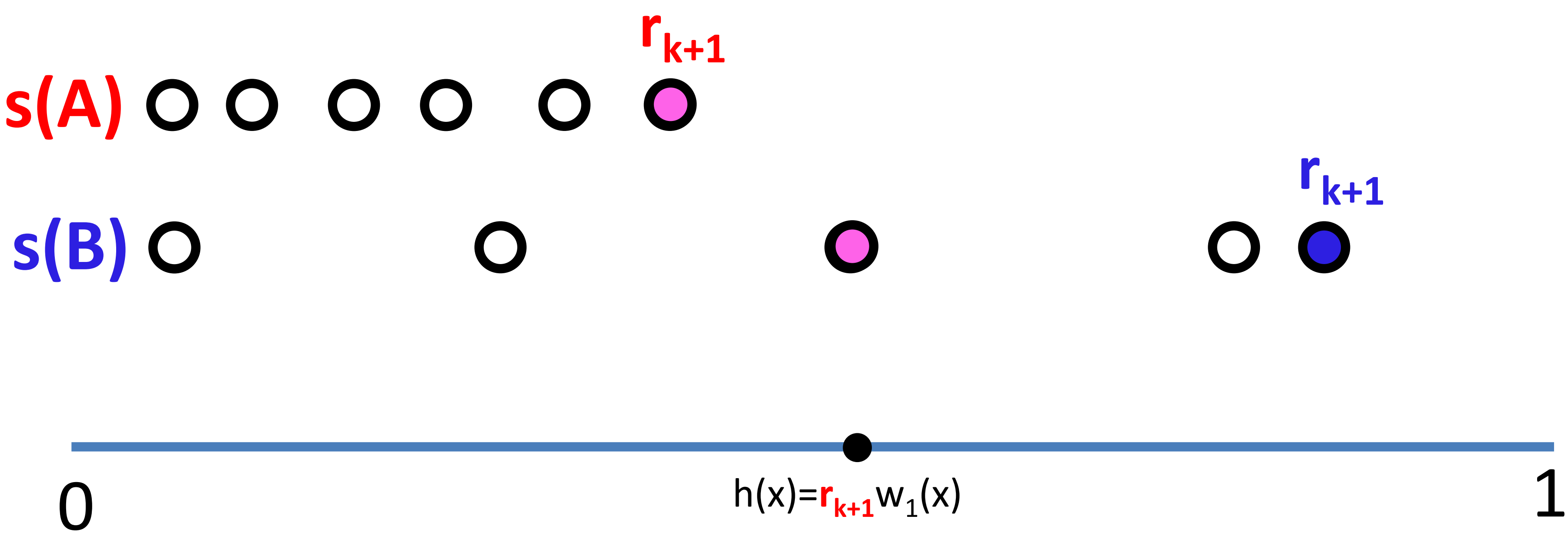
- Estimate $\sum_{x}\max\{w_1(x),w_2(x)\}$?



Estimate: ??

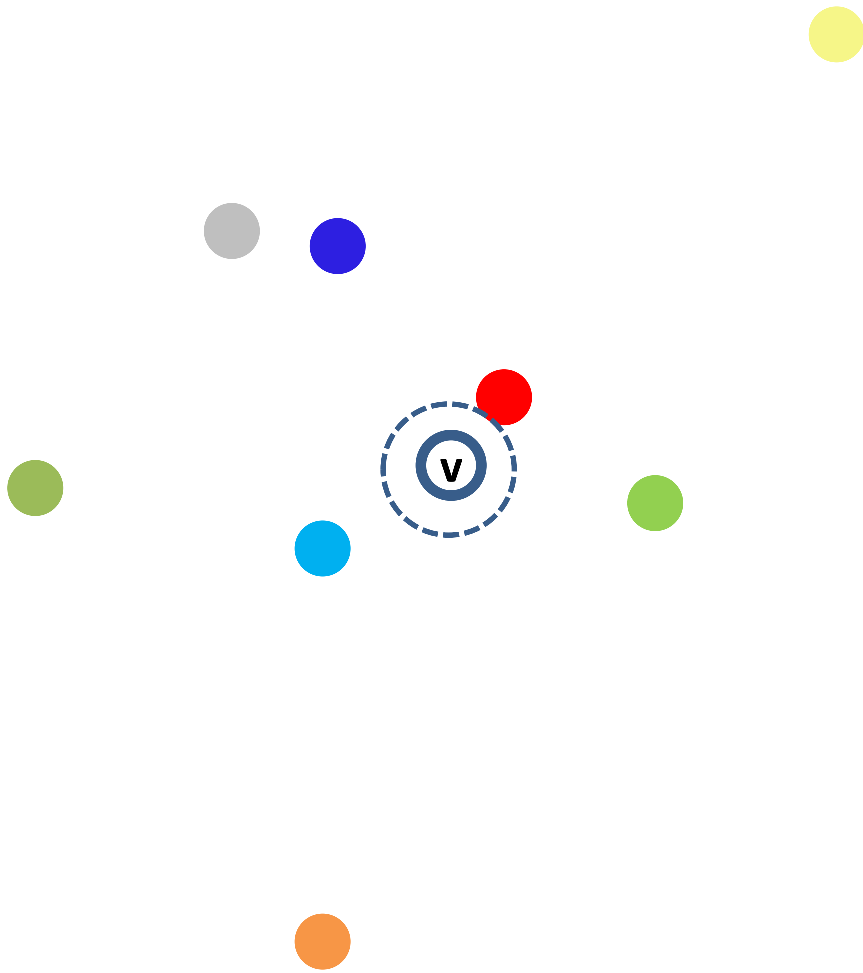
Partial information

- Estimate $\sum_{\mathbf{x}} \max\{\mathbf{w_1(x)}, \mathbf{w_2(x)}\}$?

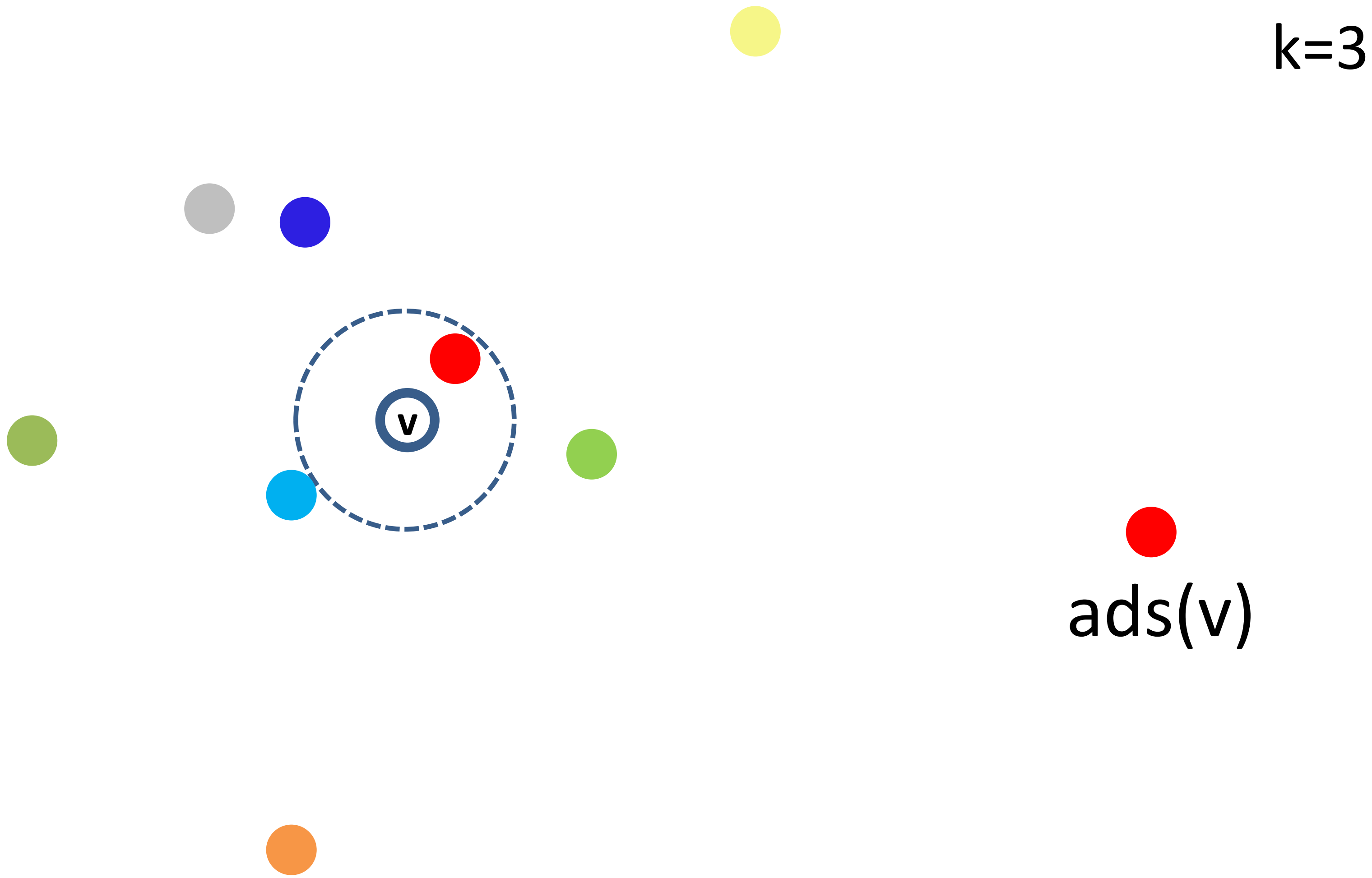


$$(1/\mathbf{r_{k+1}})(\mathbf{r_{k+1}}w_2(x)-\mathbf{r_{k+1}}w_1(x))+\mathbf{E}\times\mathbf{r_{k+1}}w_1(x)=w_1(x)$$

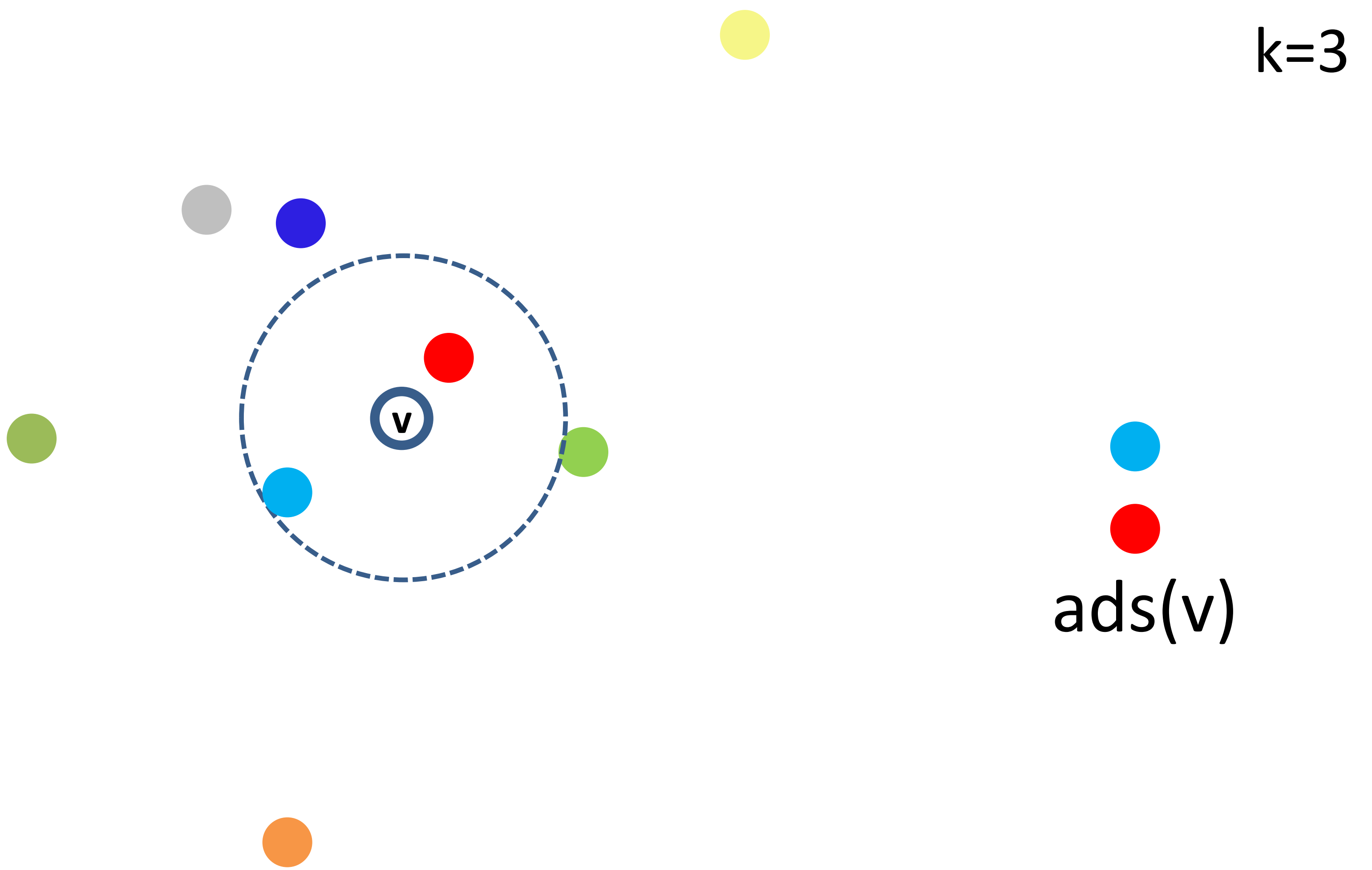
All Distances (coordinated) Sketches (ADS)



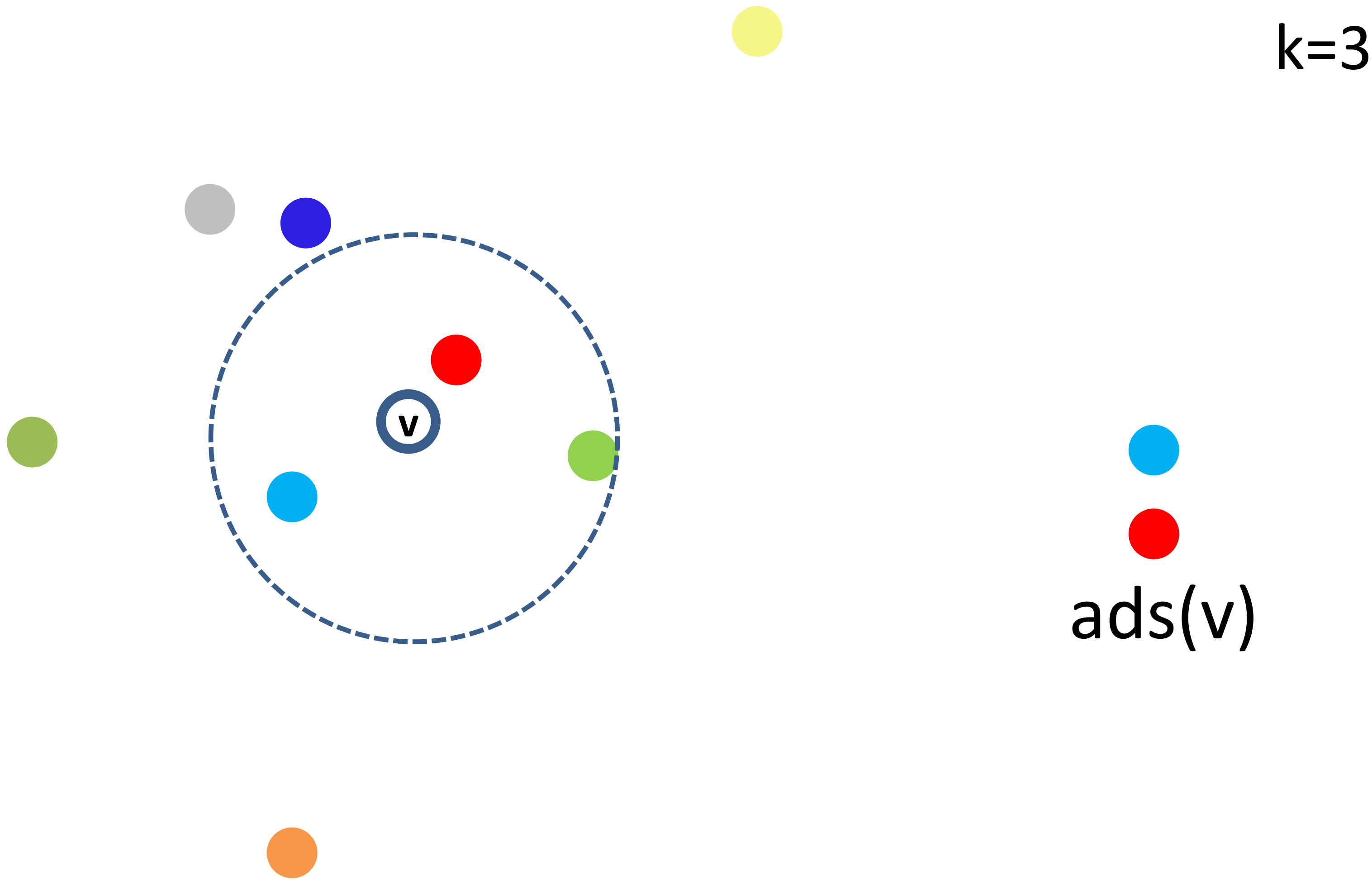
k=3

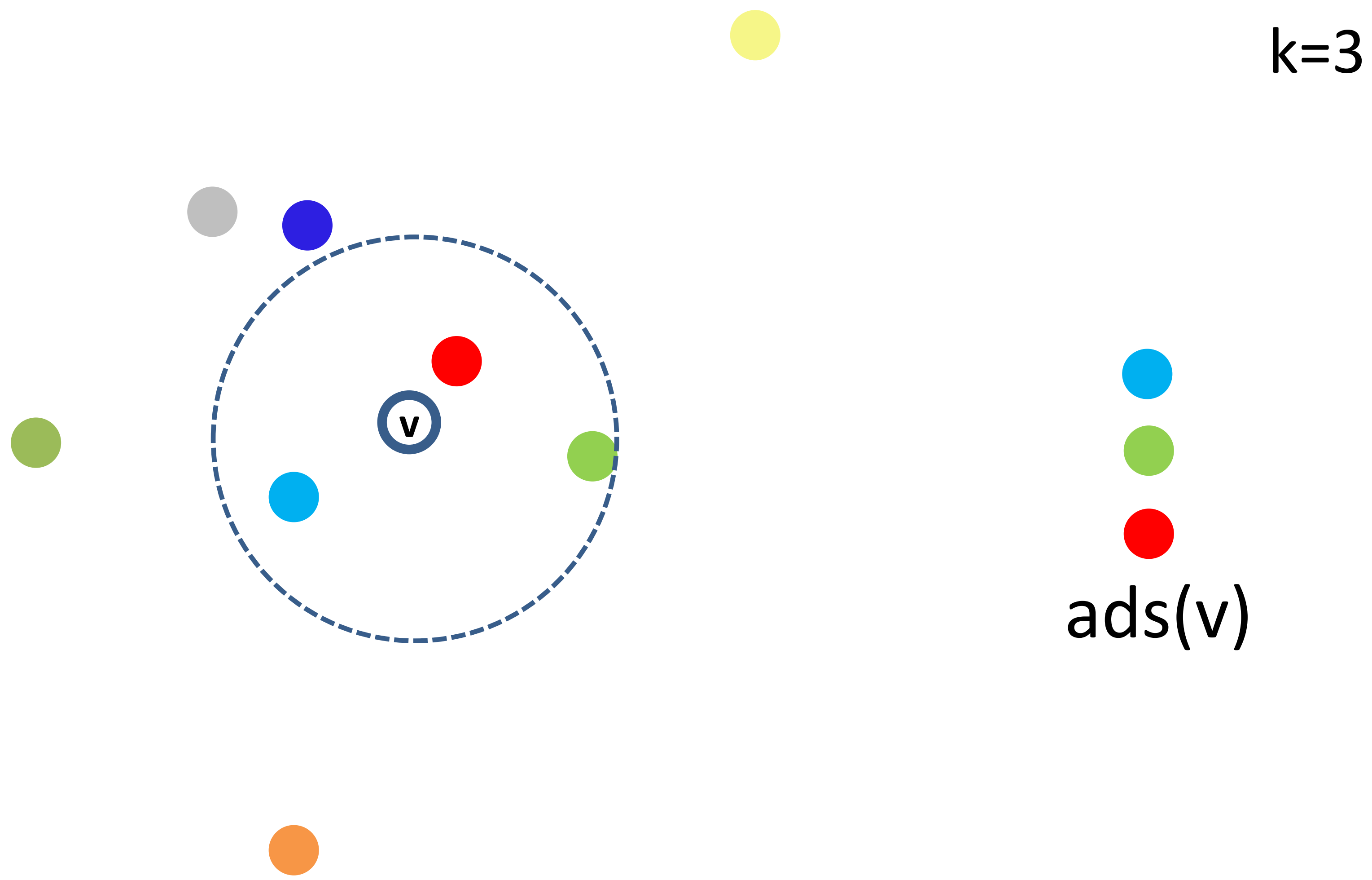


k=3
v
ads(v)

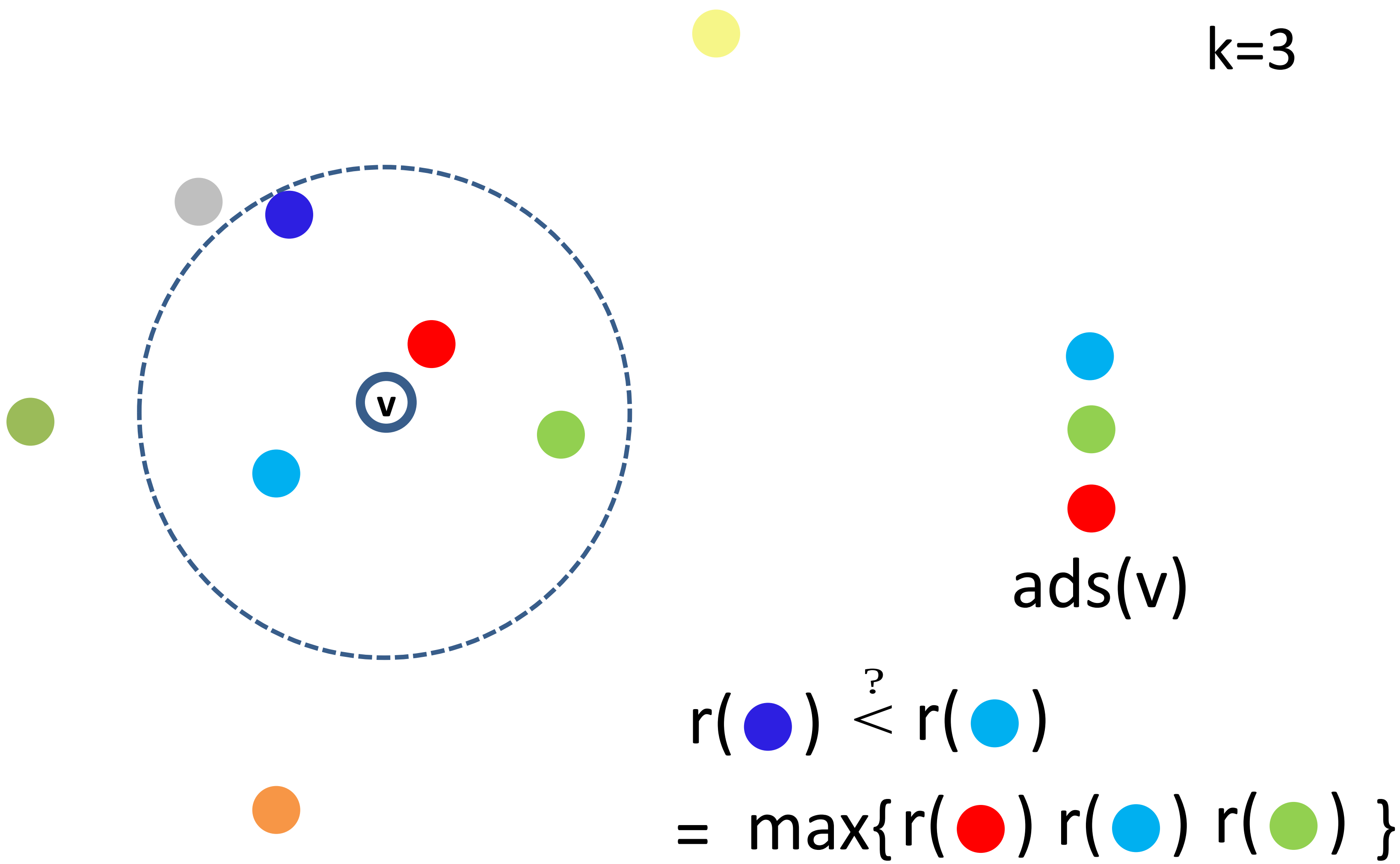


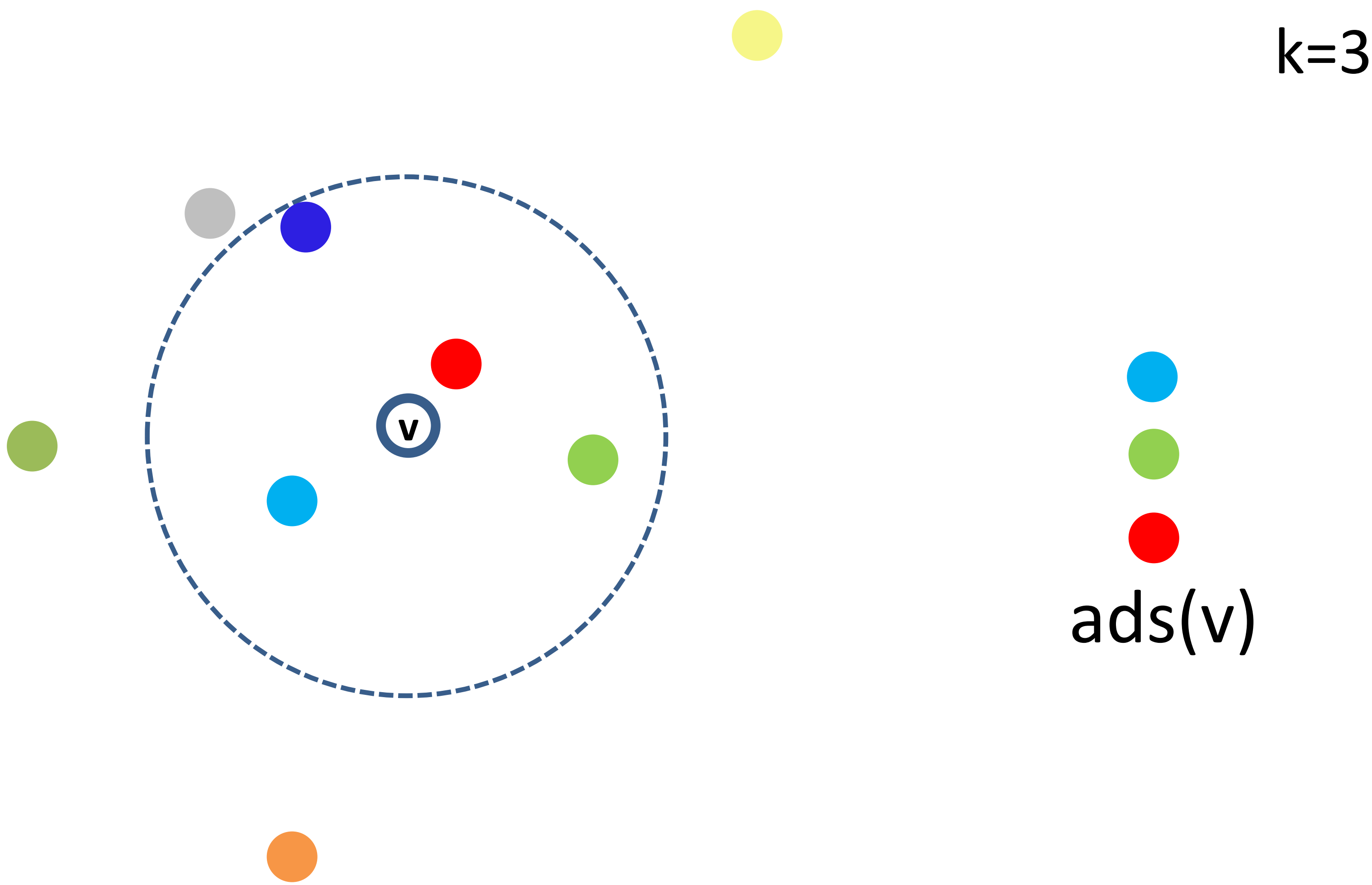
k=3
v
ads(v)

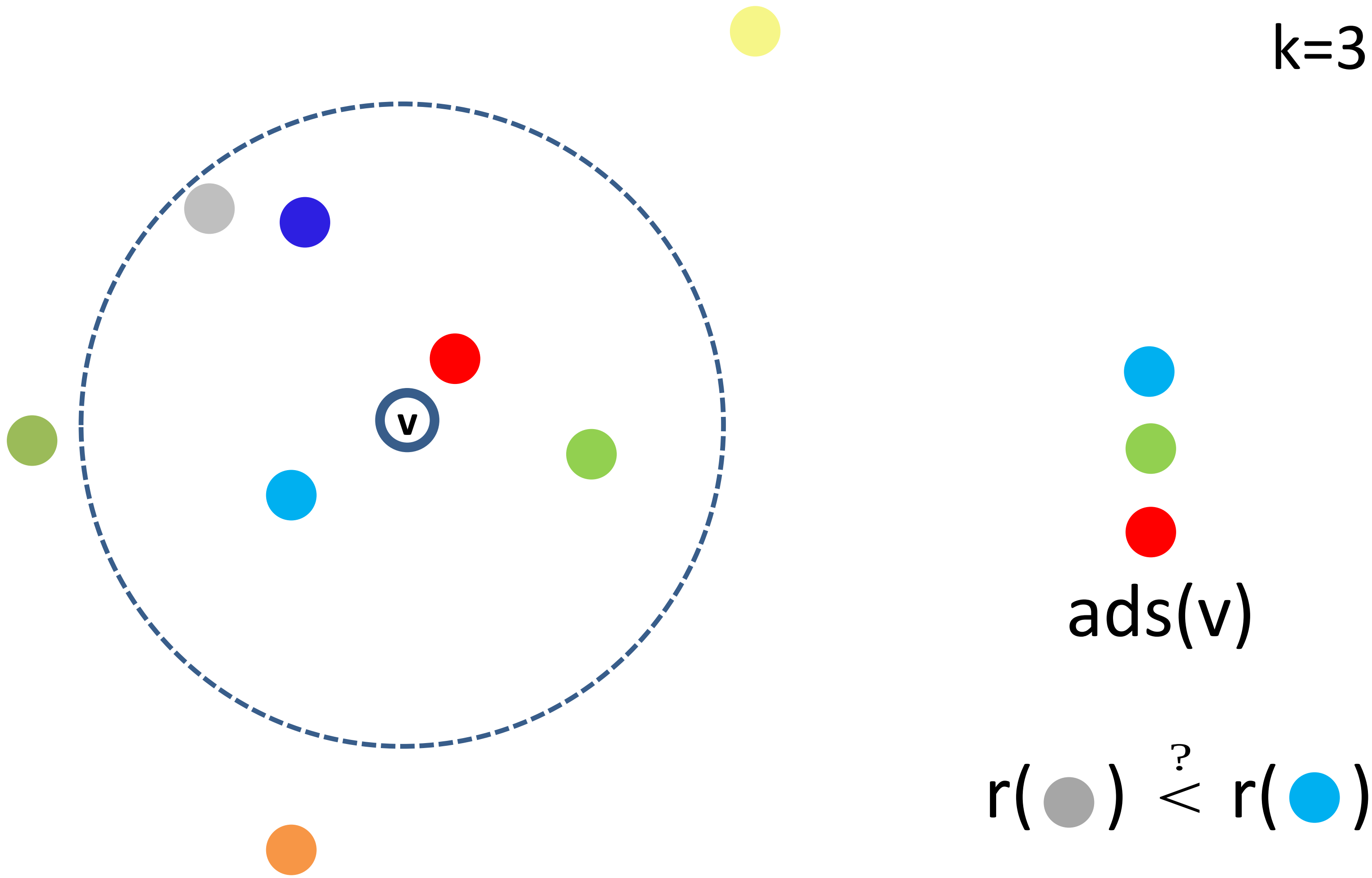

k=3
v
ads(v)

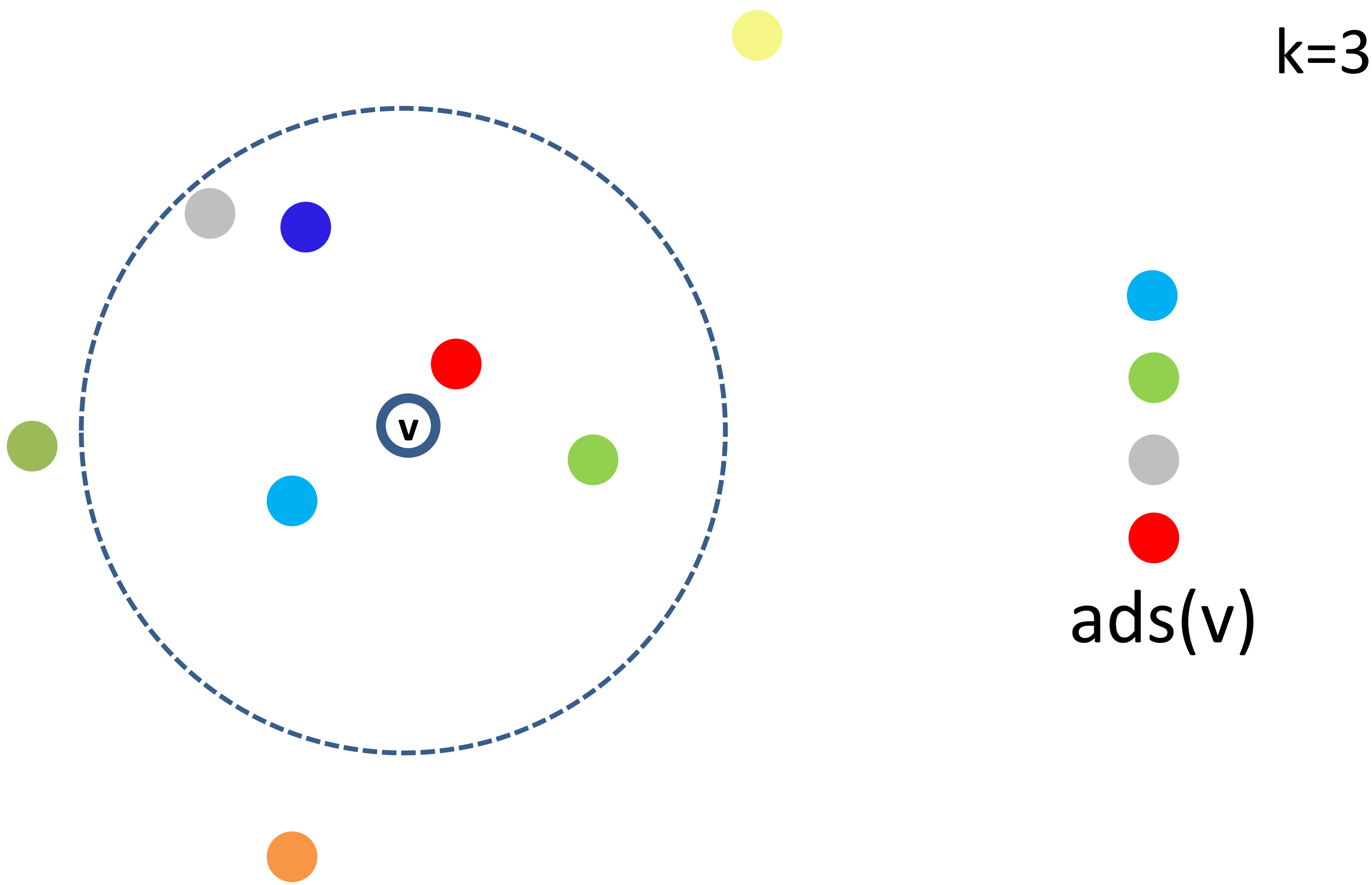


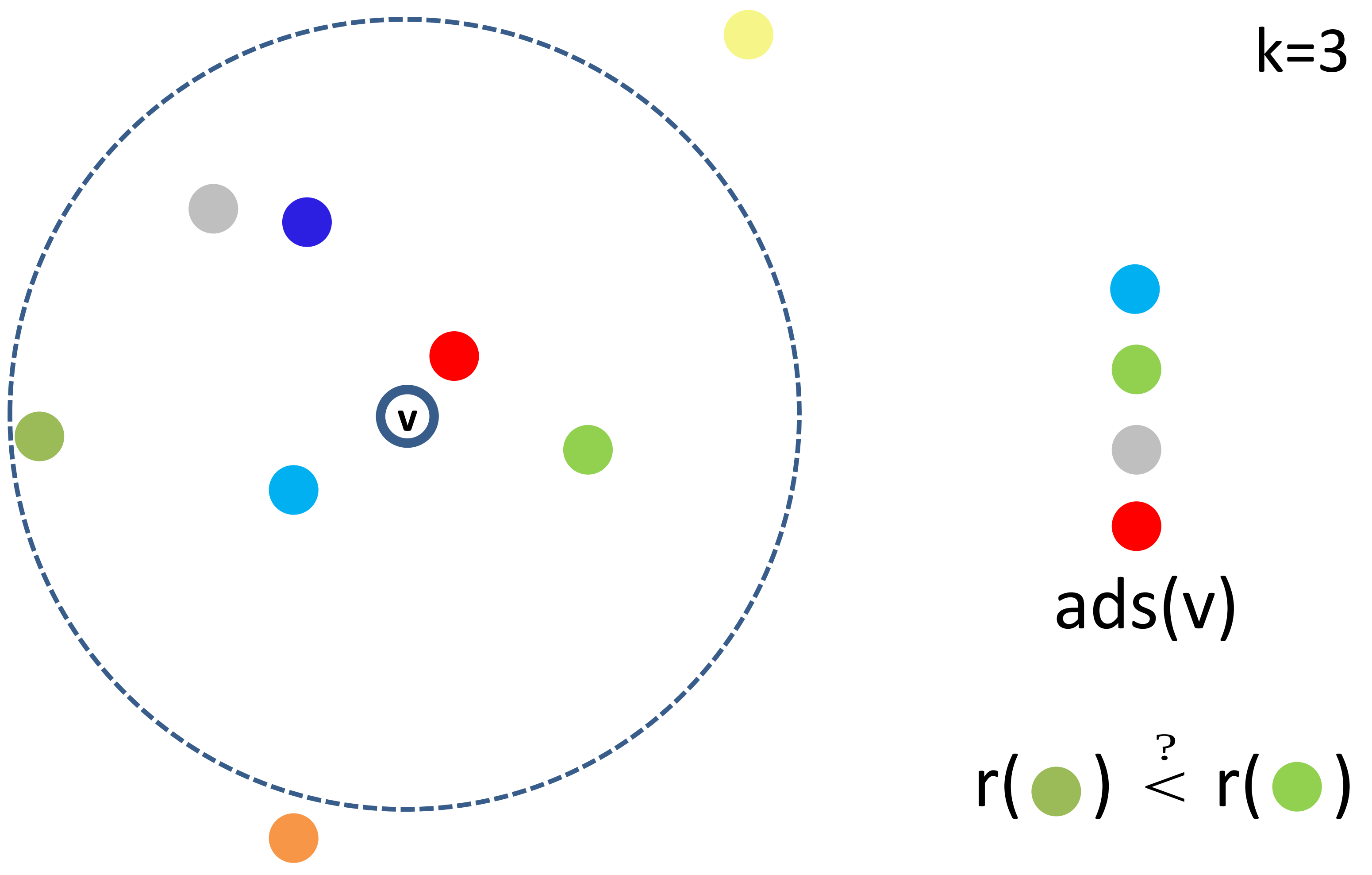
k=3
v
ads(v)


k=3
v
ads(v)
r( ) <? r( )
= max{r( ) r( ) r( ) }

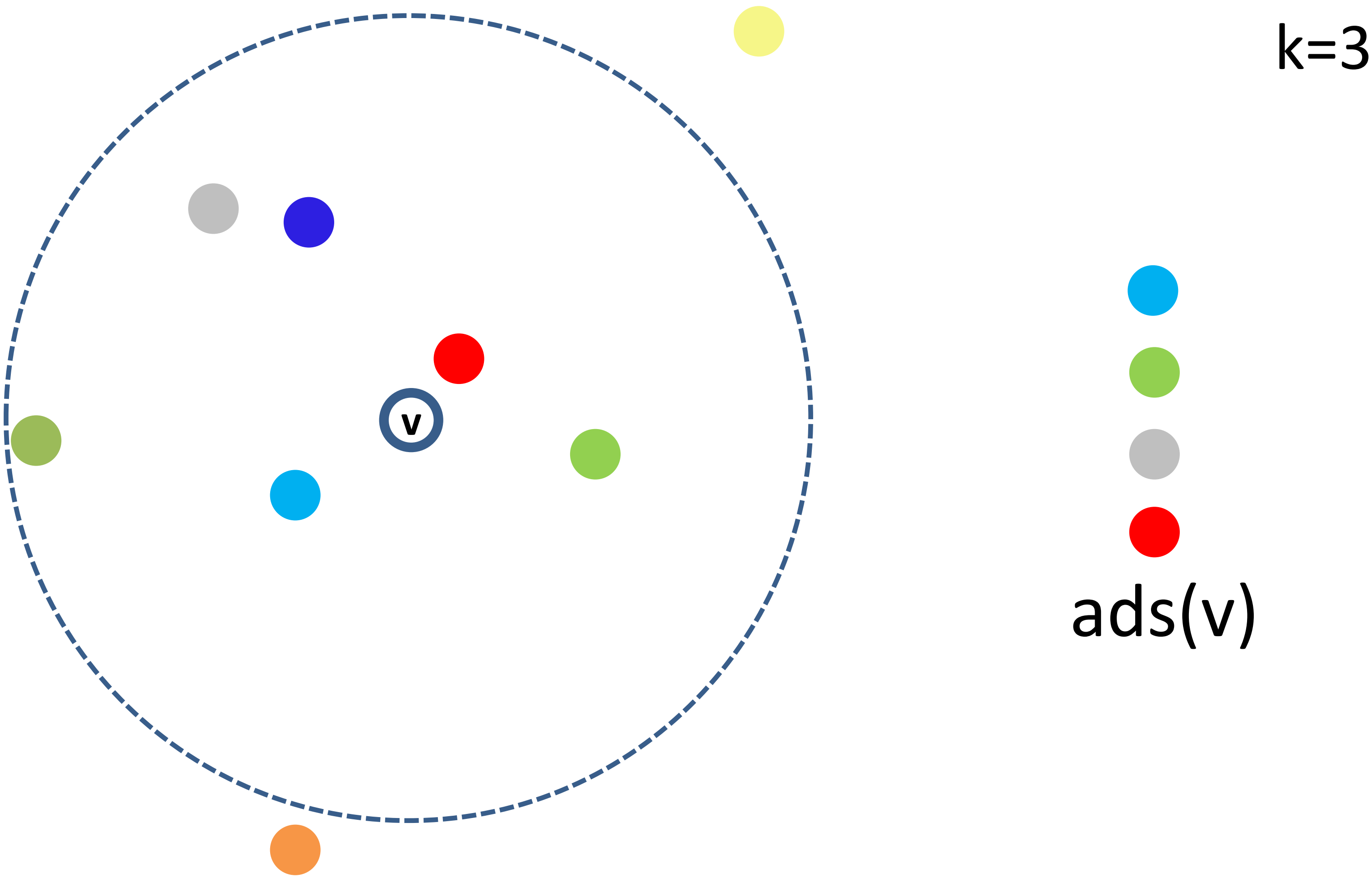

k=3
v
ads(v)


k=3
v
ads(v)
r( ) $\overset{?}{<}$ r( )

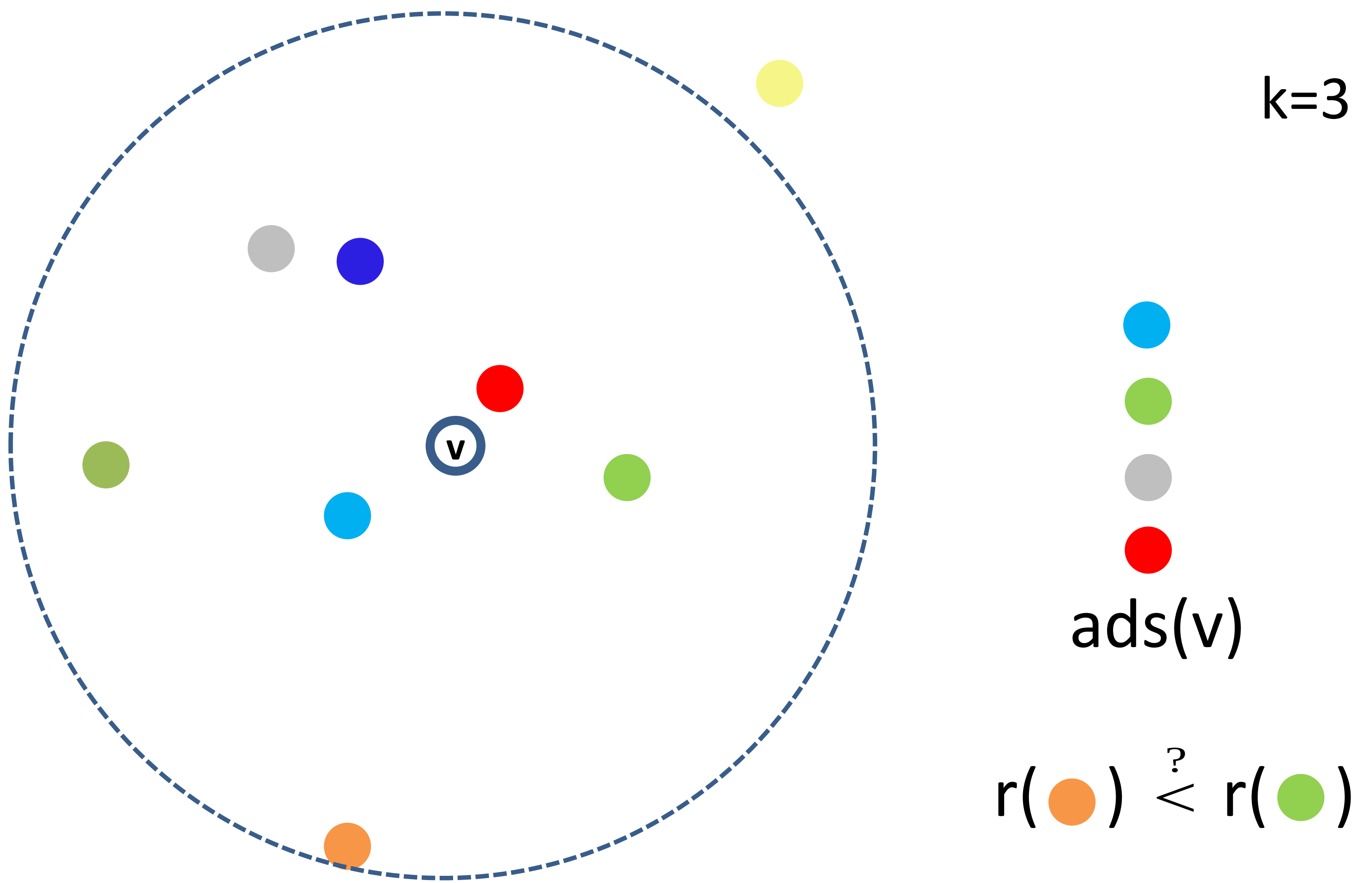

k=3
v
ads(v)

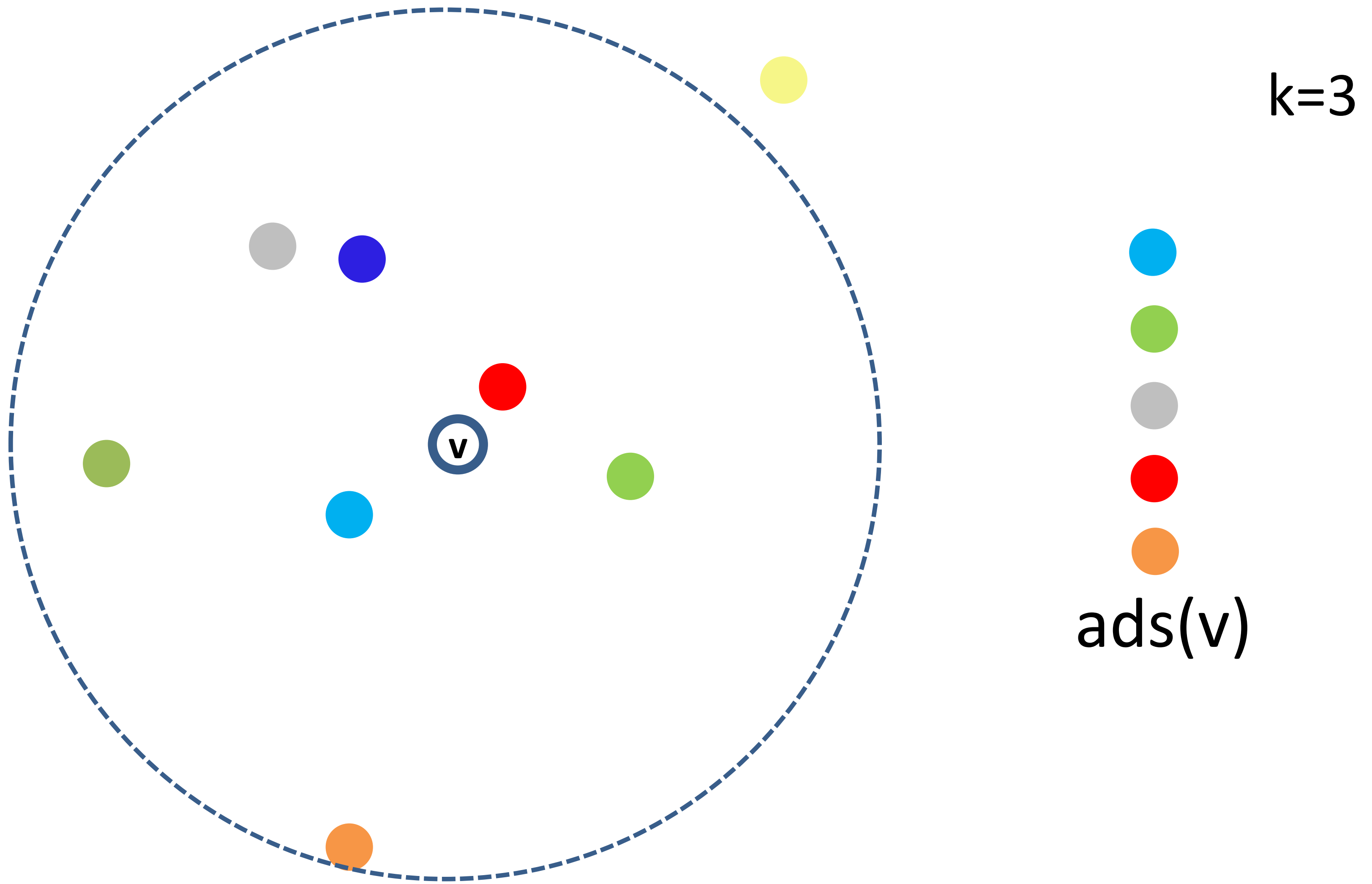


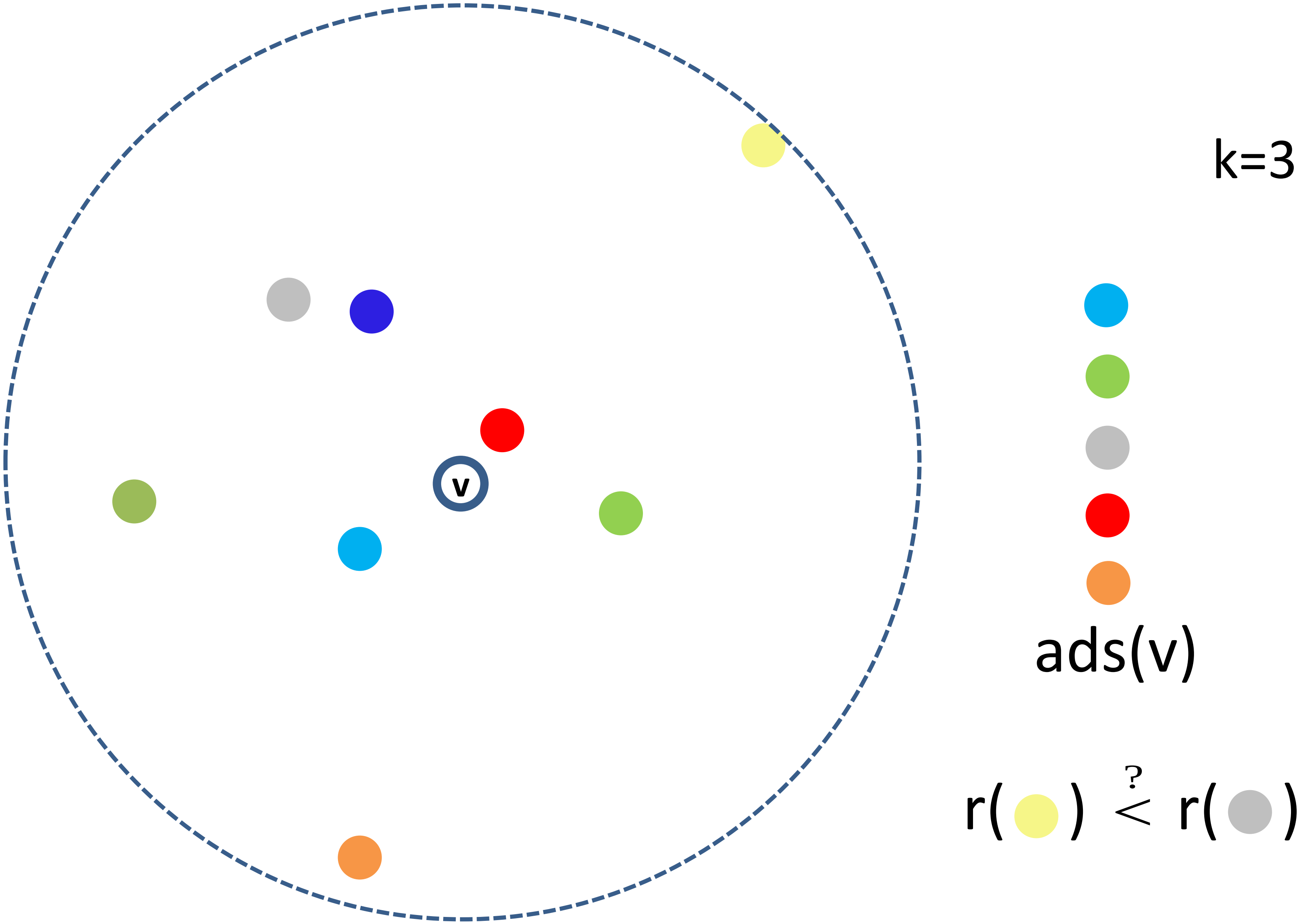
v
k=3
ads(v)
r( ) ≤? r( )

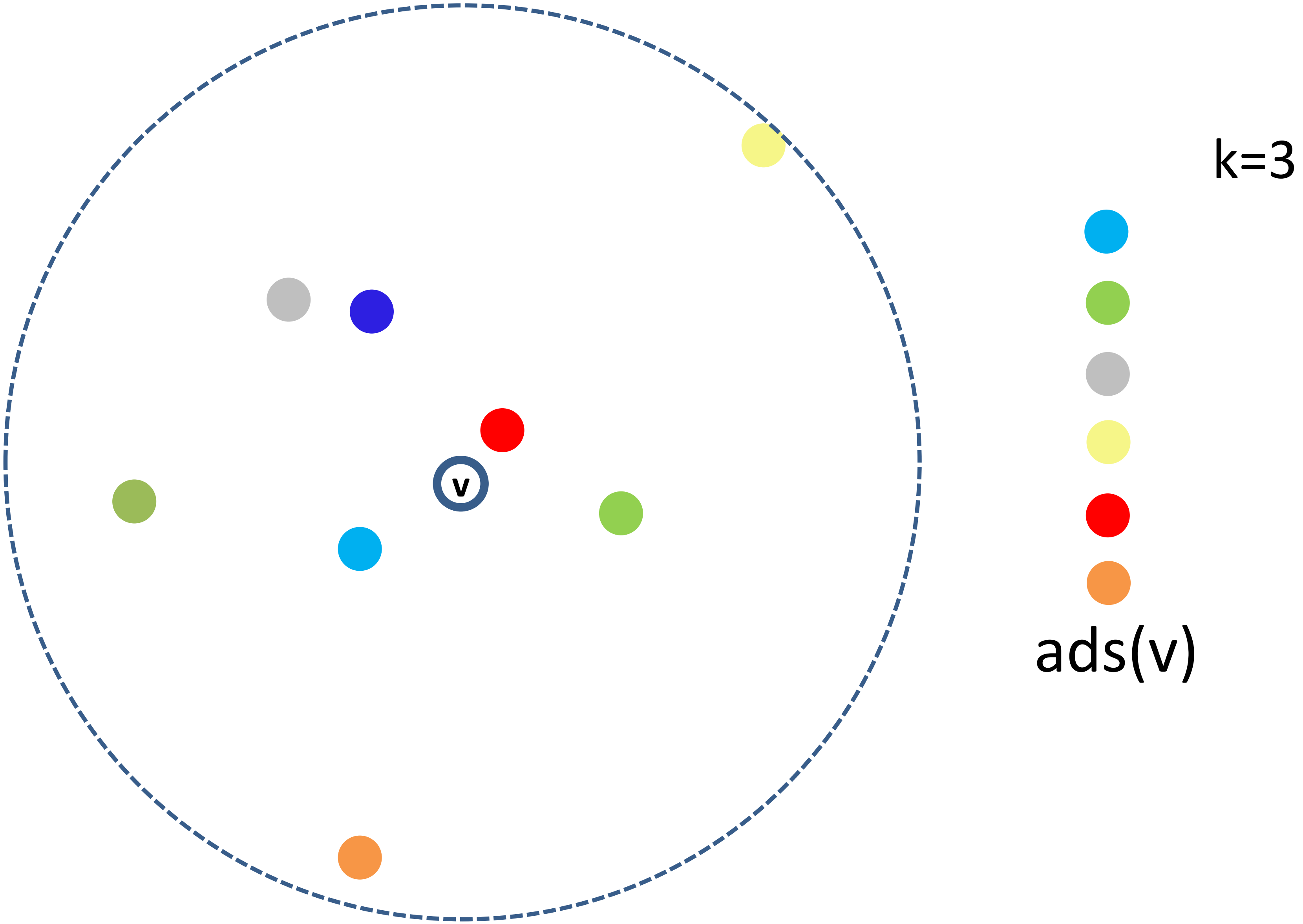


k=3
v
ads(v)


v
k=3
ads(v)
r( ) <? r( )

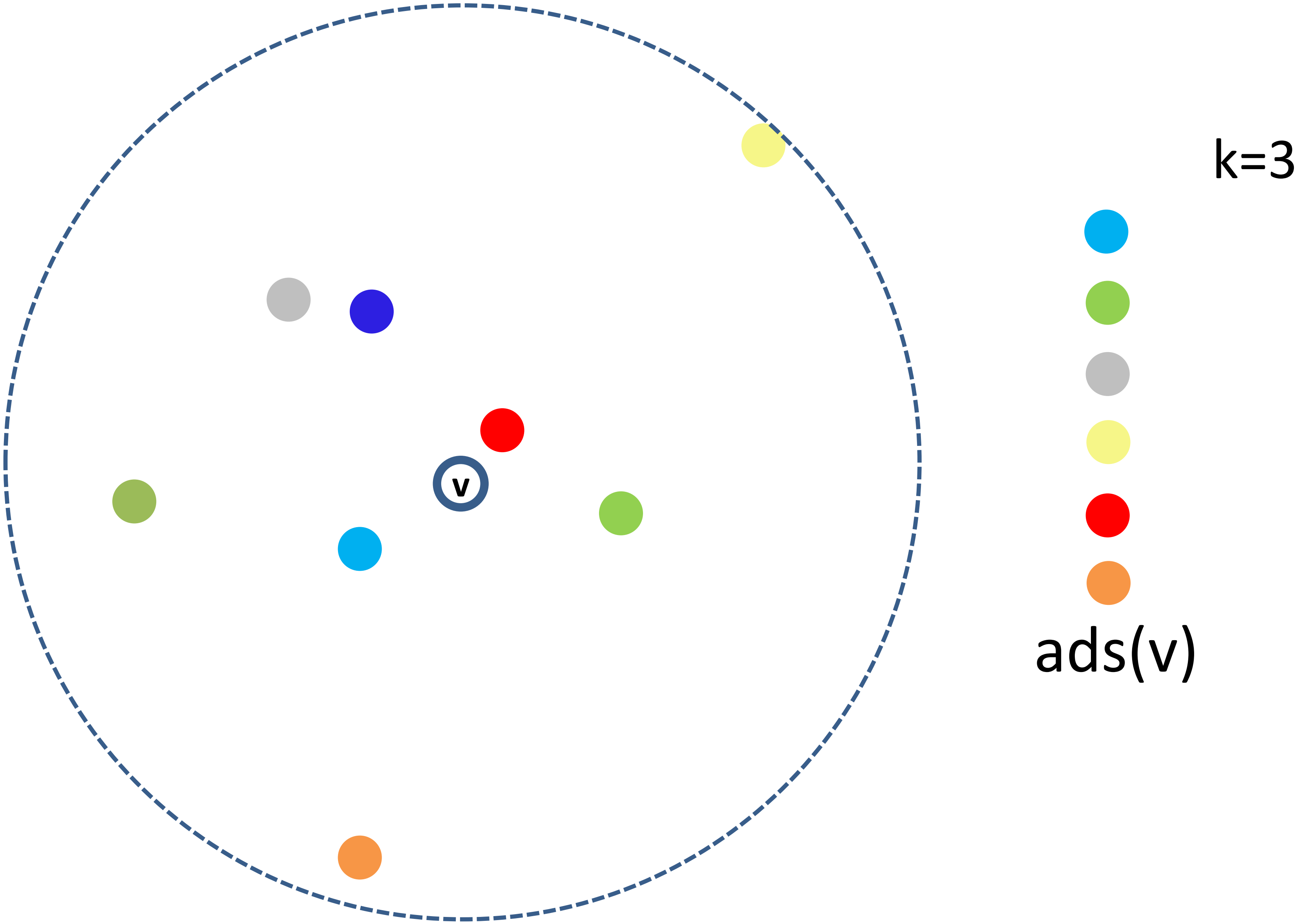

k=3
v
ads(v)


k=3
v
ads(v)
r( ) $\overset{?}{<}$ r( )



v
k=3
ads(v)

Properties of the ADS

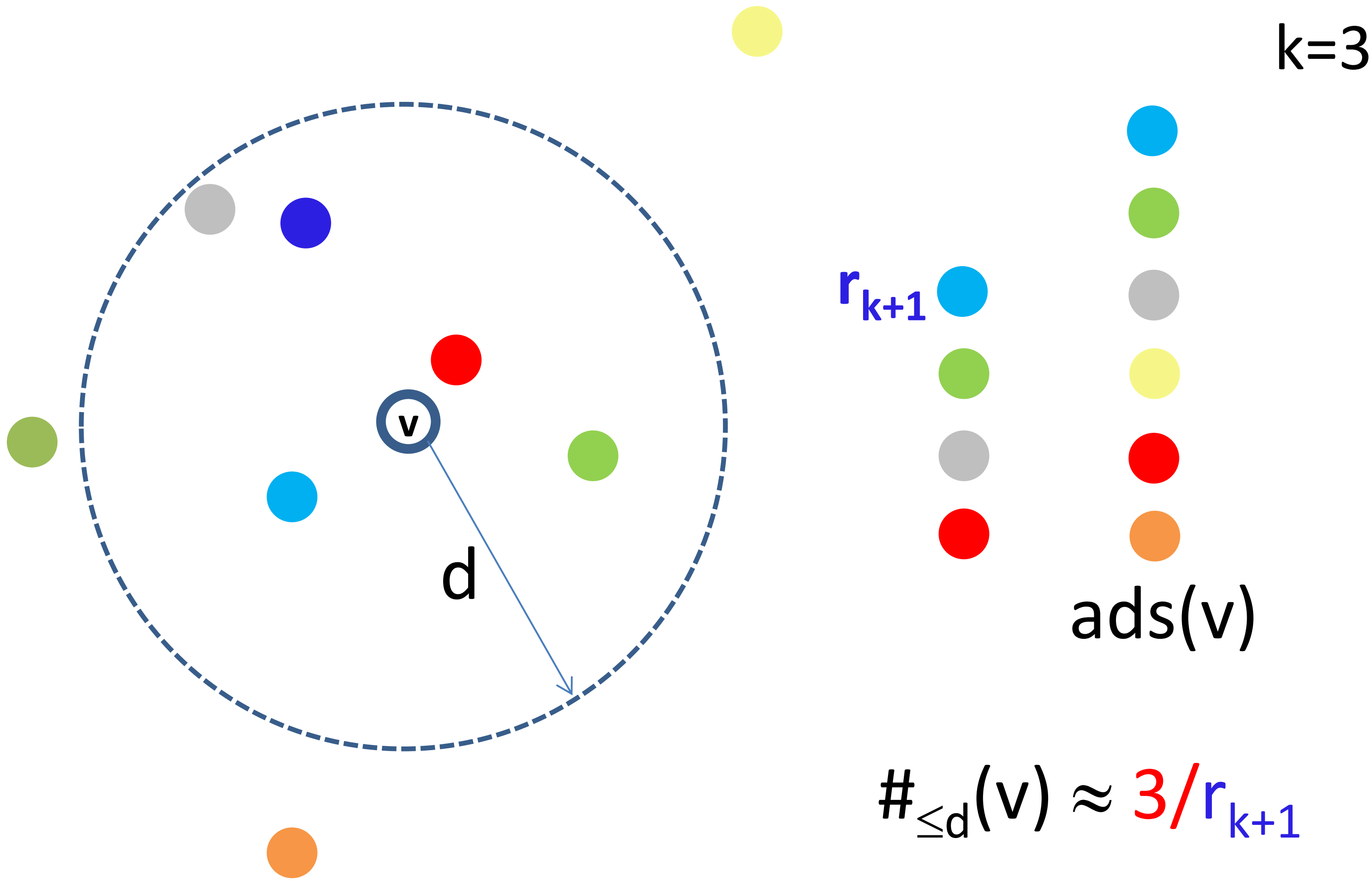


Properties of the ADS

- Includes a bottom-k sketch for every ball around v

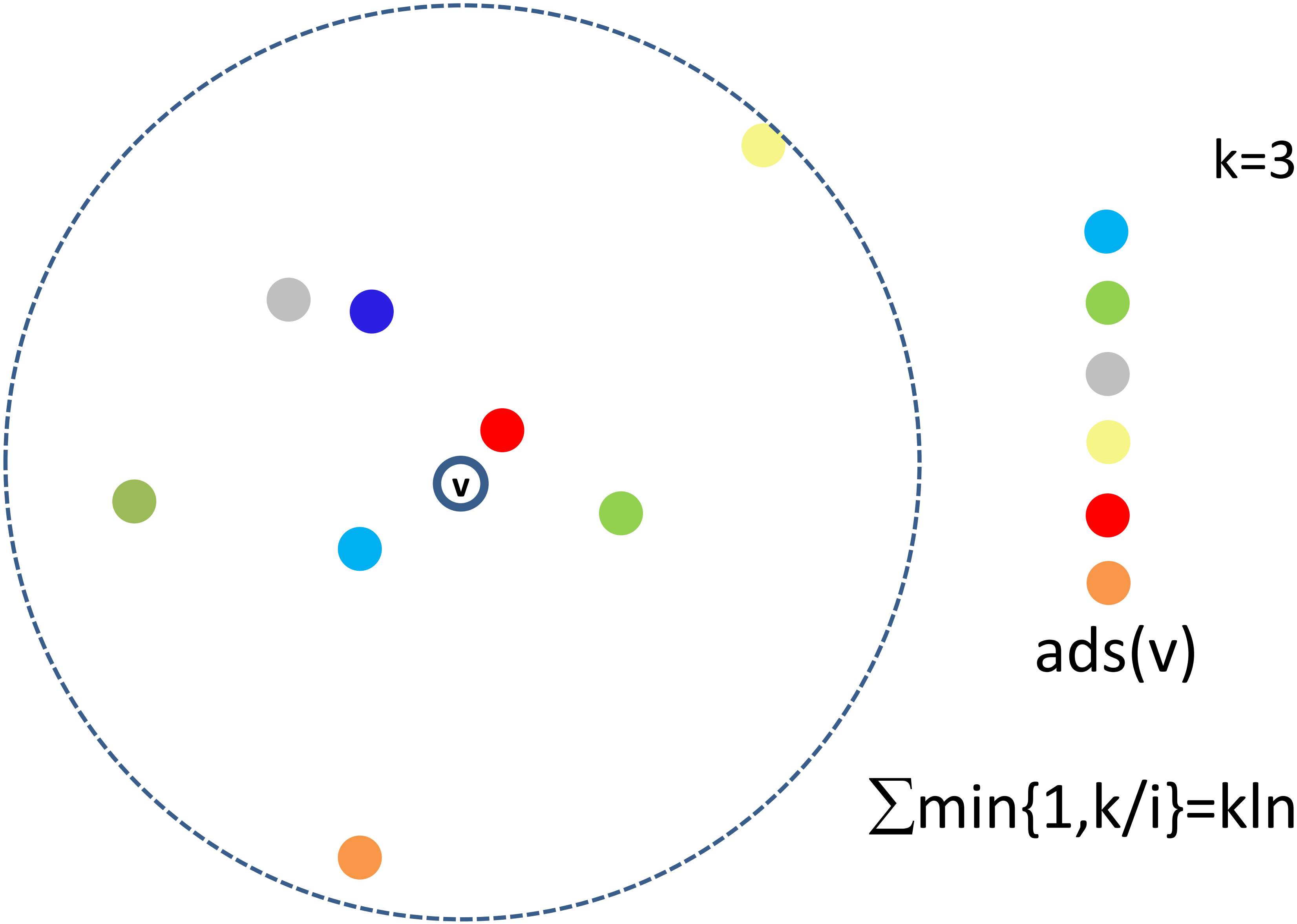
➔Estimate how many vertices are there in every ball around v

Estimating using the ADS



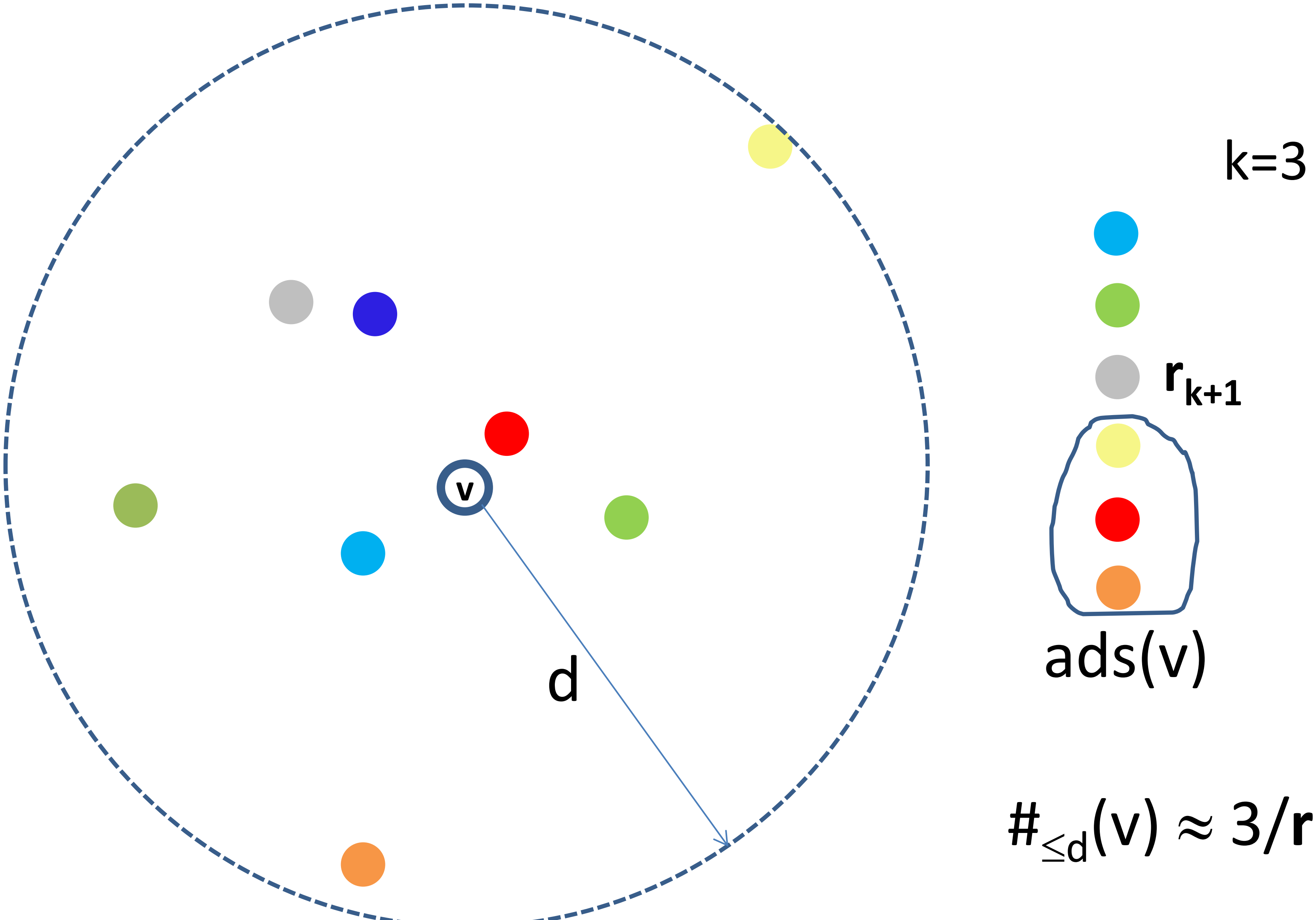
$$\#_{\le d}(v) \approx 3/r_{k+1}$$

Expected size of the ADS



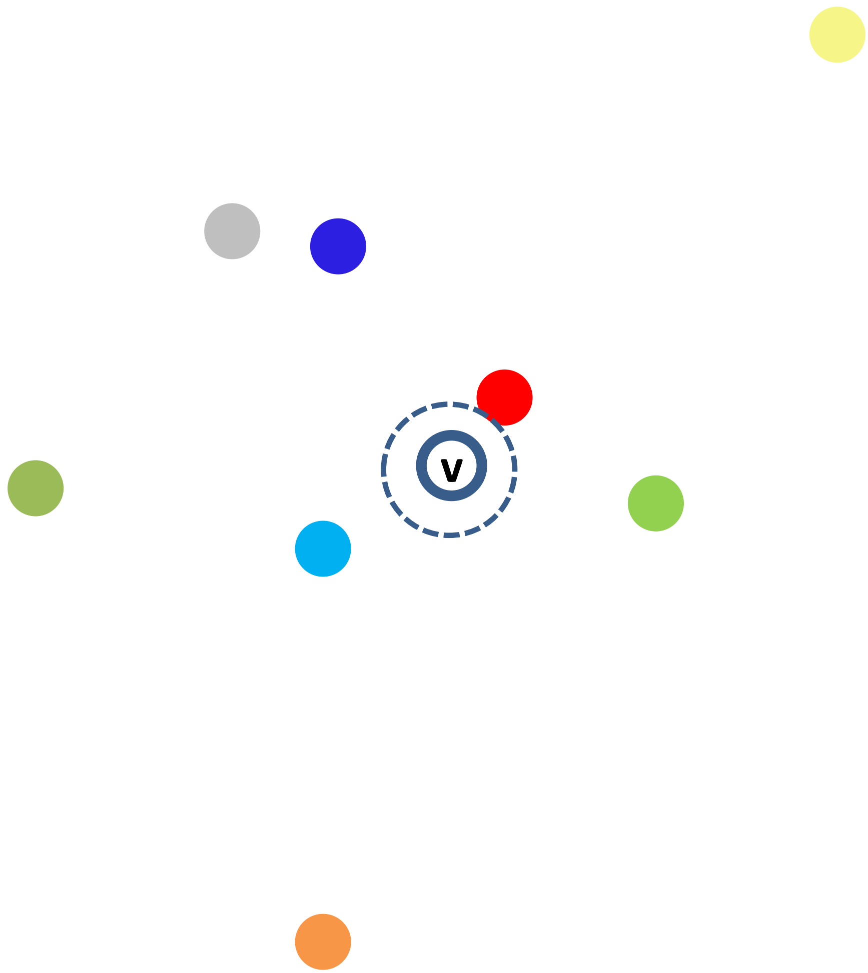
$\sum \min\{1, k/i\} = k \ln(n)$

Do we use the most out of the ADS ?

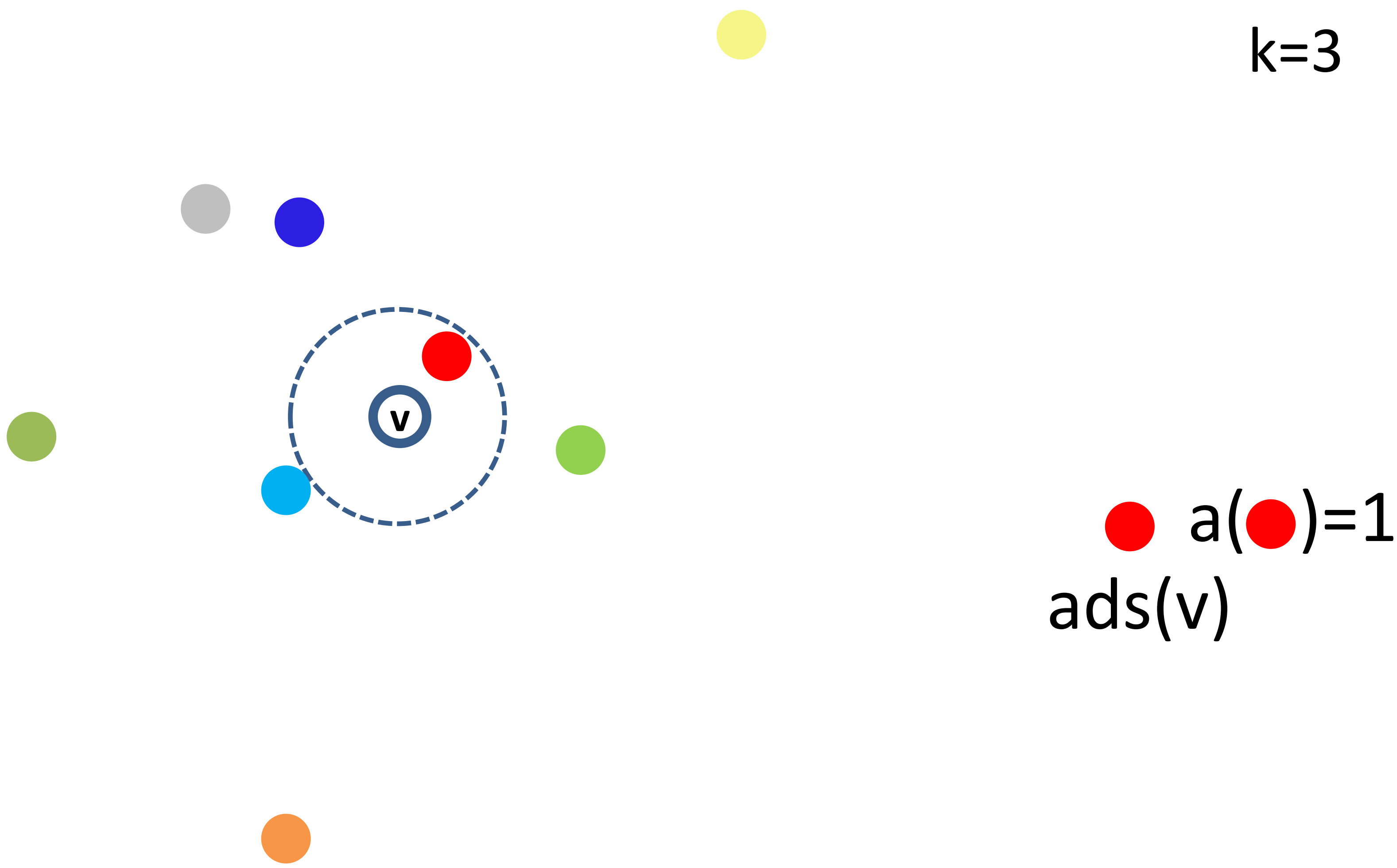


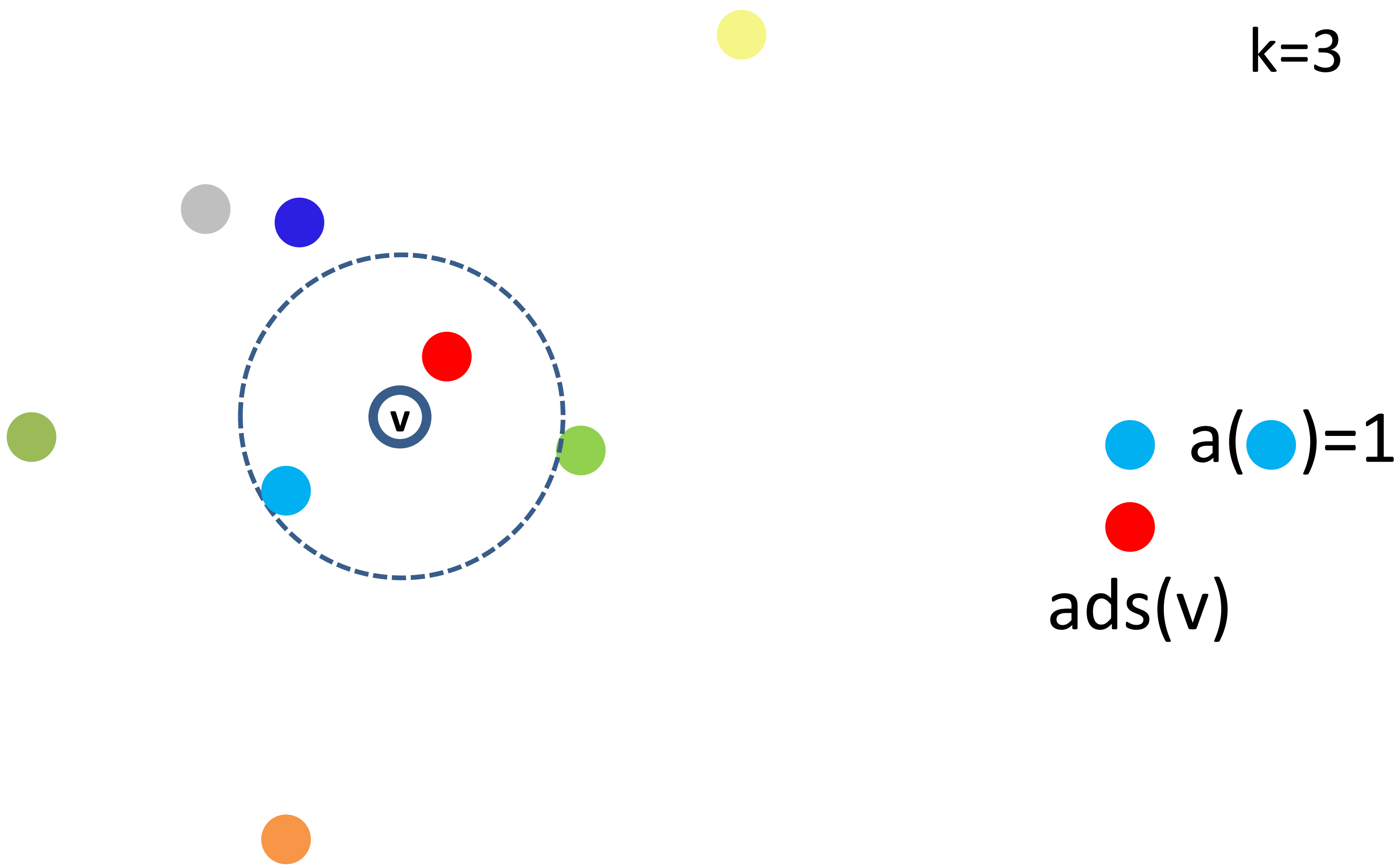
$\#_{\leq d}(v) \approx 3/\mathbf{r}_{k+1}$

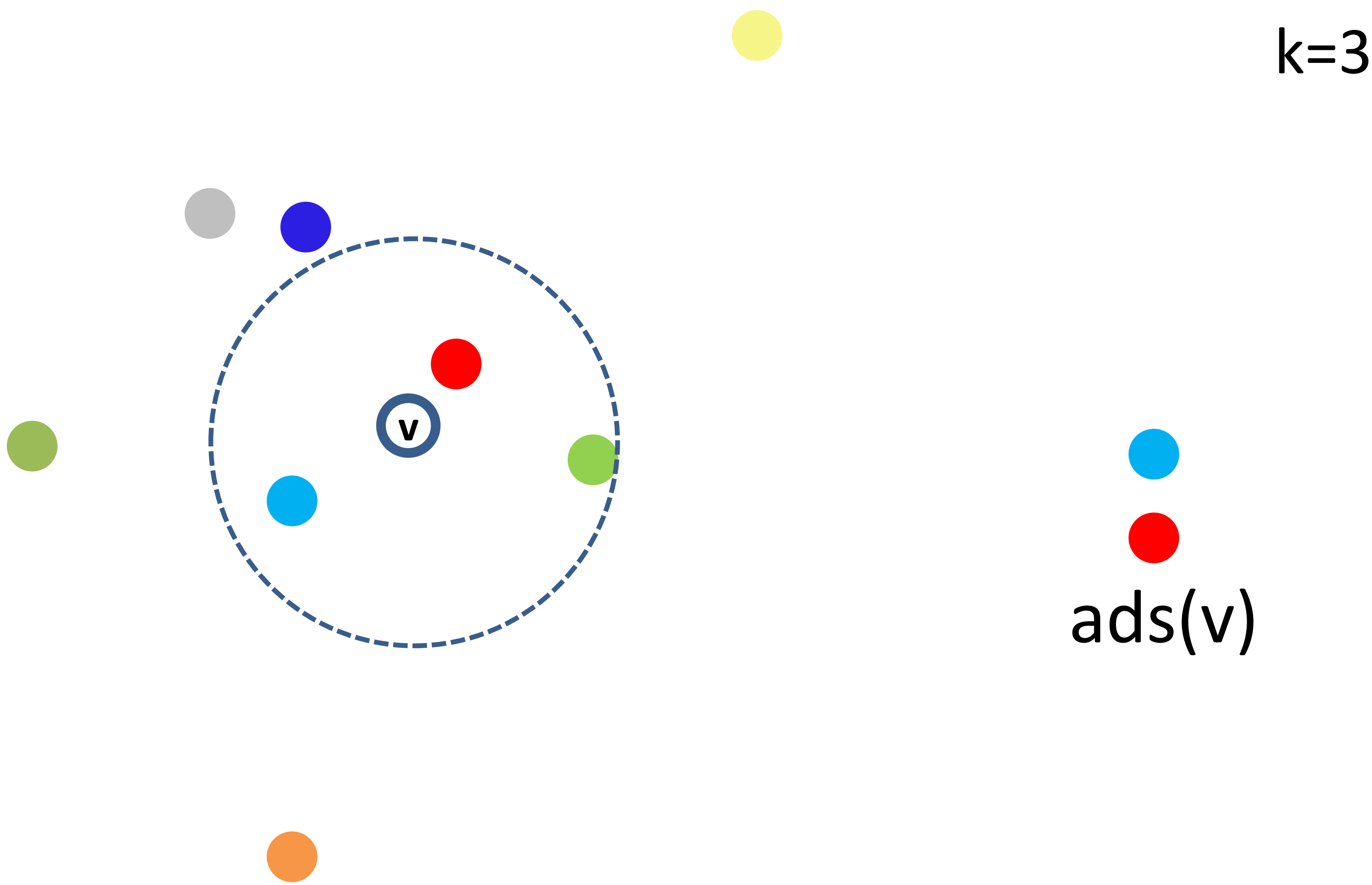
Give adjusted weights to all elements

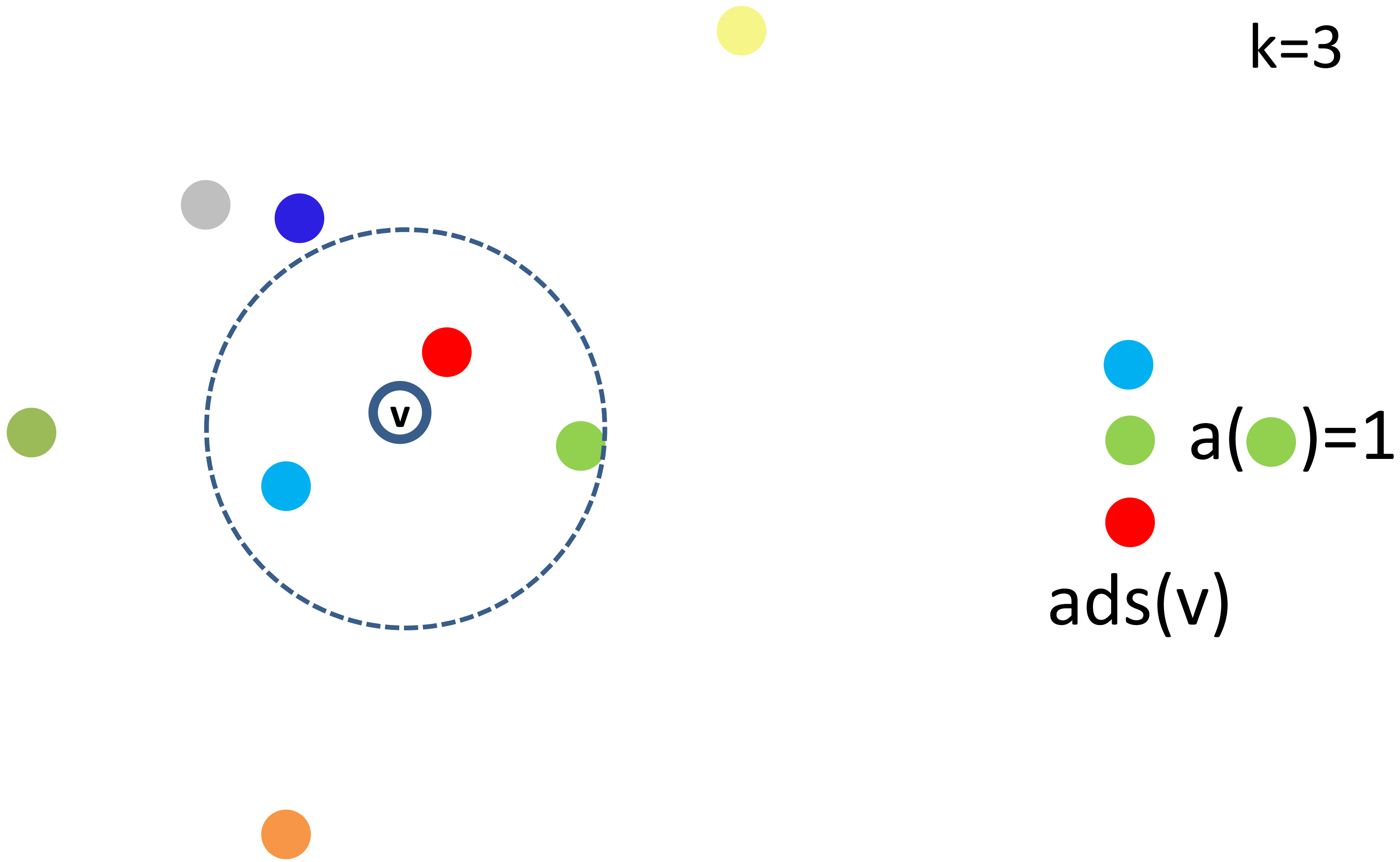


k=3

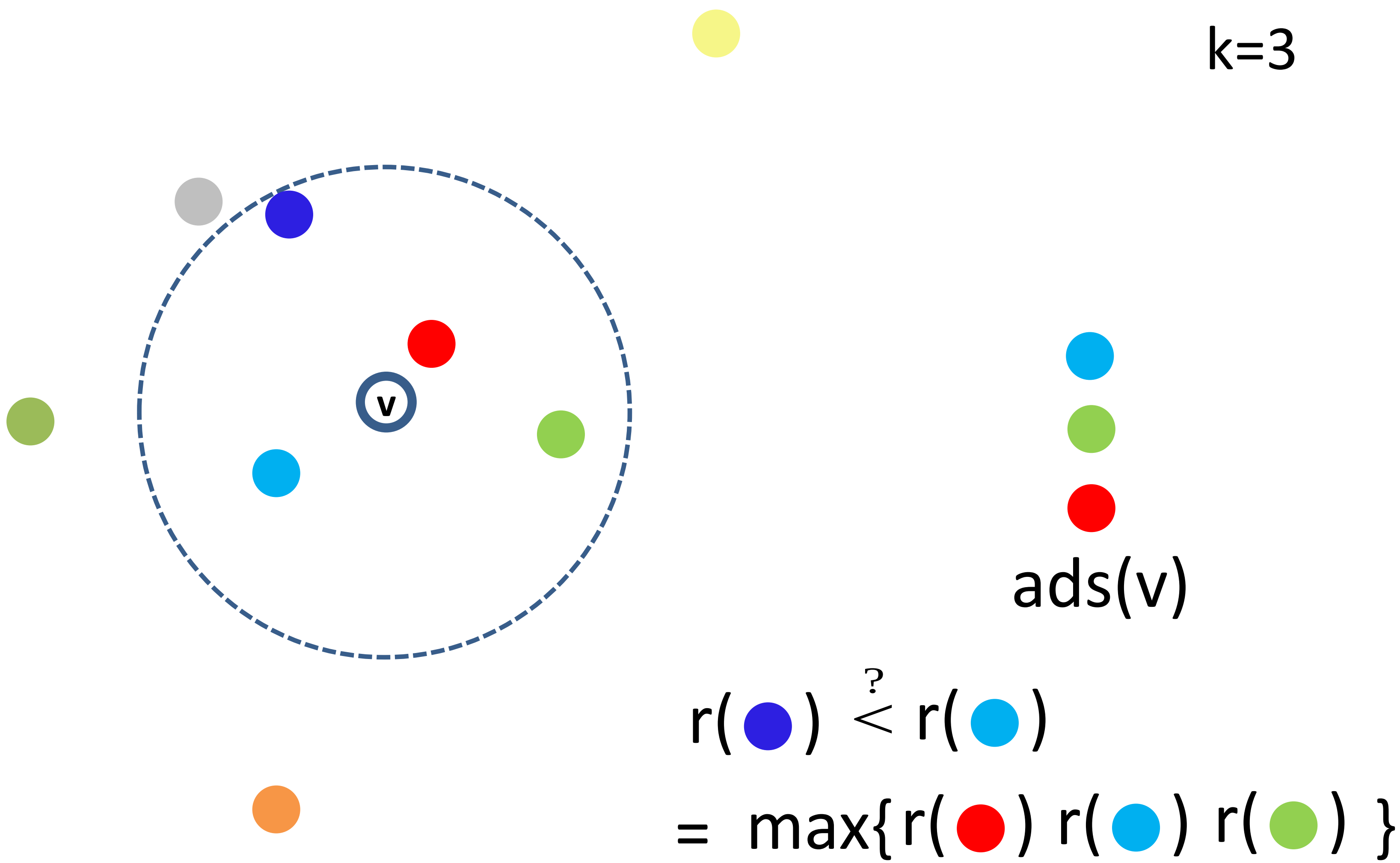

k=3
v
a( )=1
ads(v)

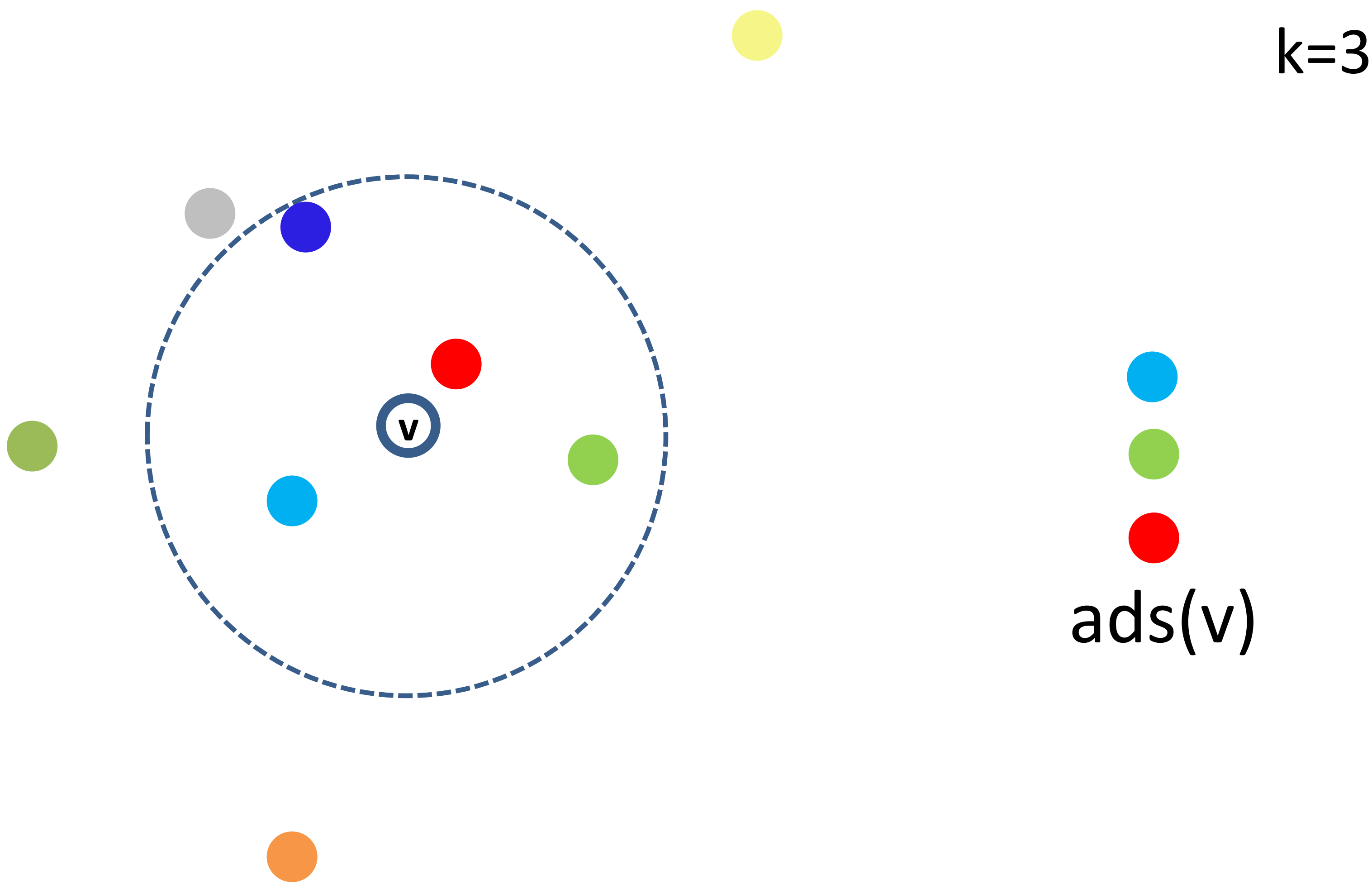

k=3
v
a( )=1
ads(v)

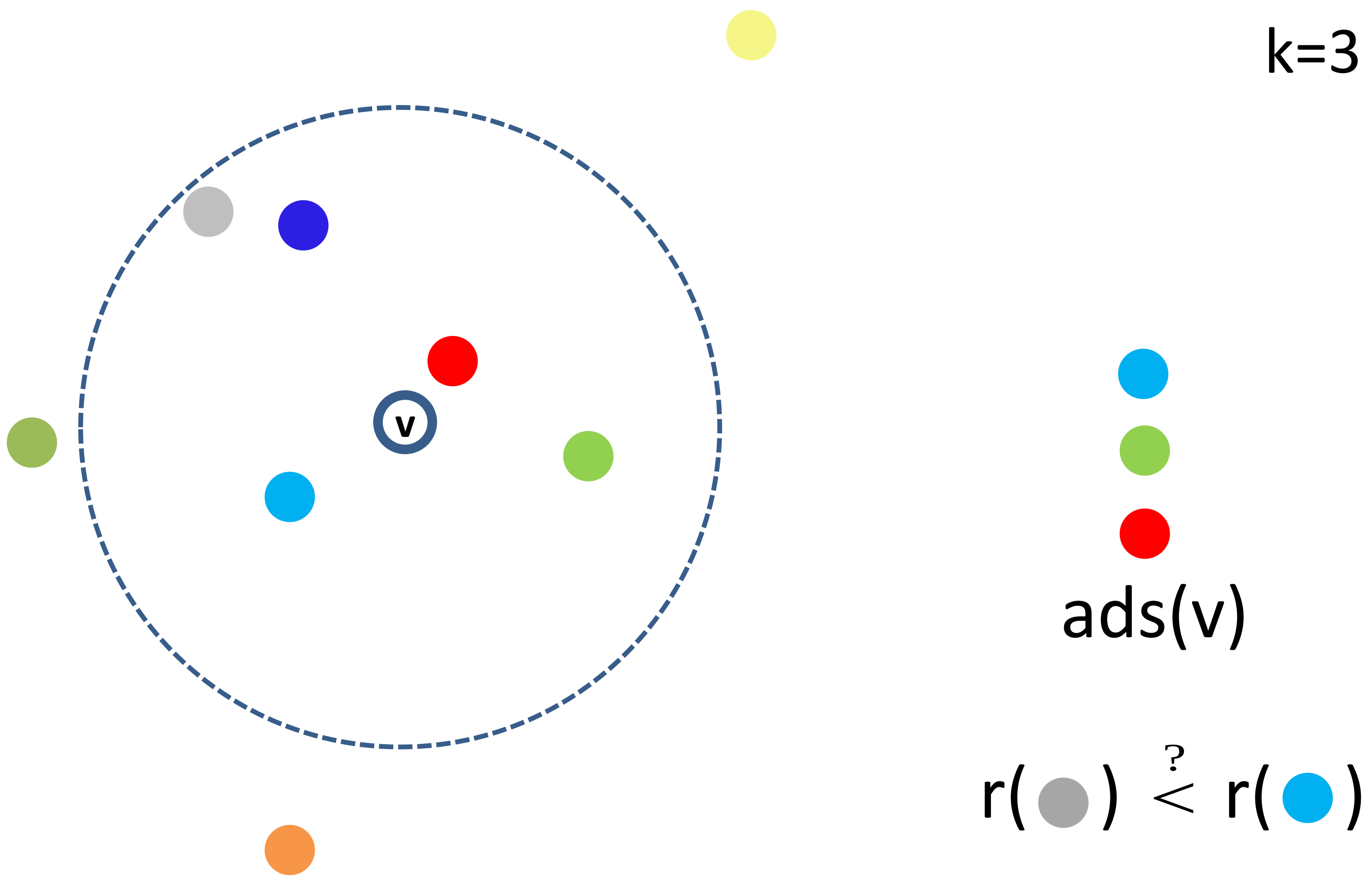

k=3
v
ads(v)

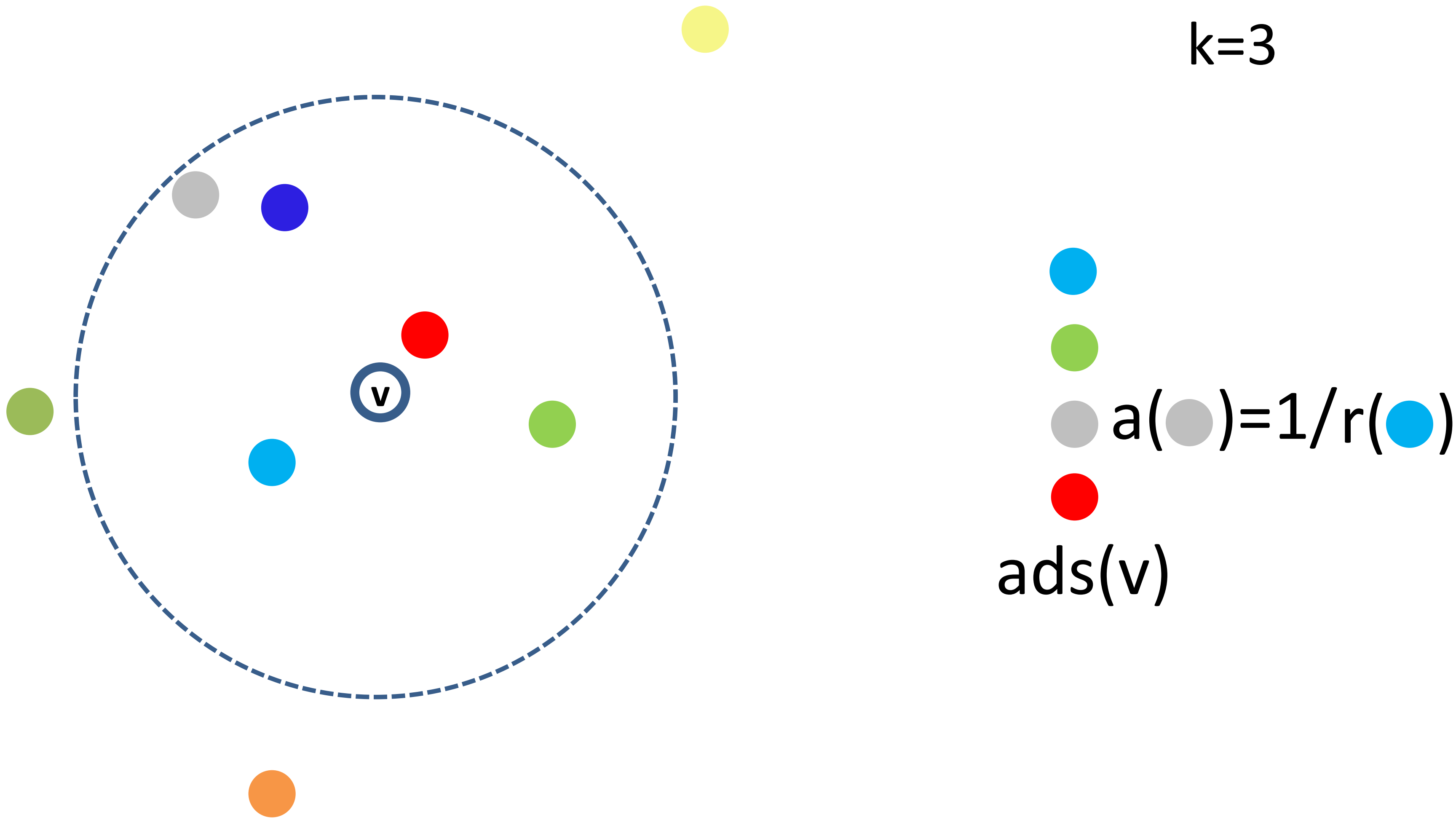


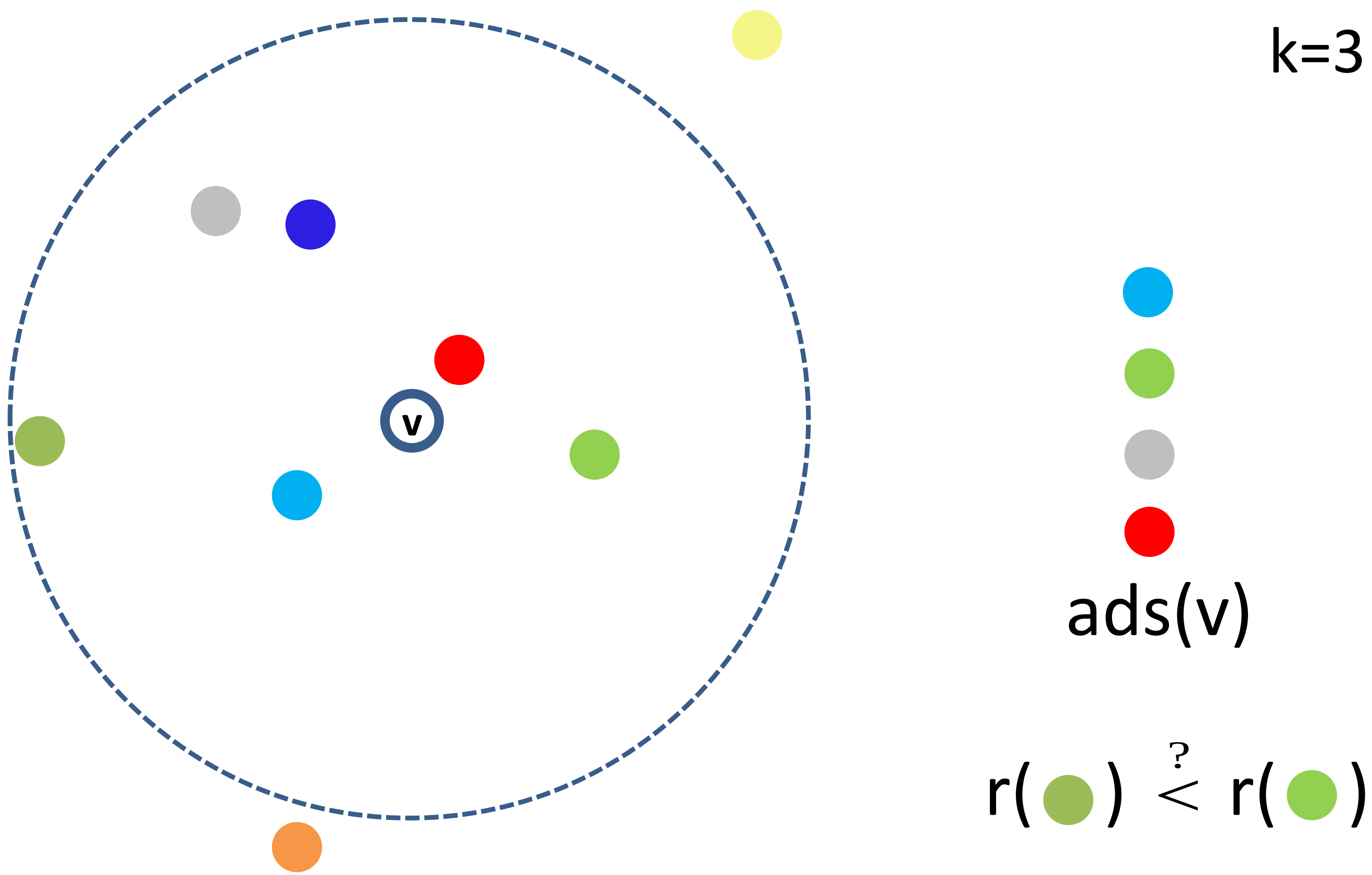
k=3
v
a( )=1
ads(v)


k=3
v
ads(v)
r( ) <? r( )
= max{r( ) r( ) r( ) }

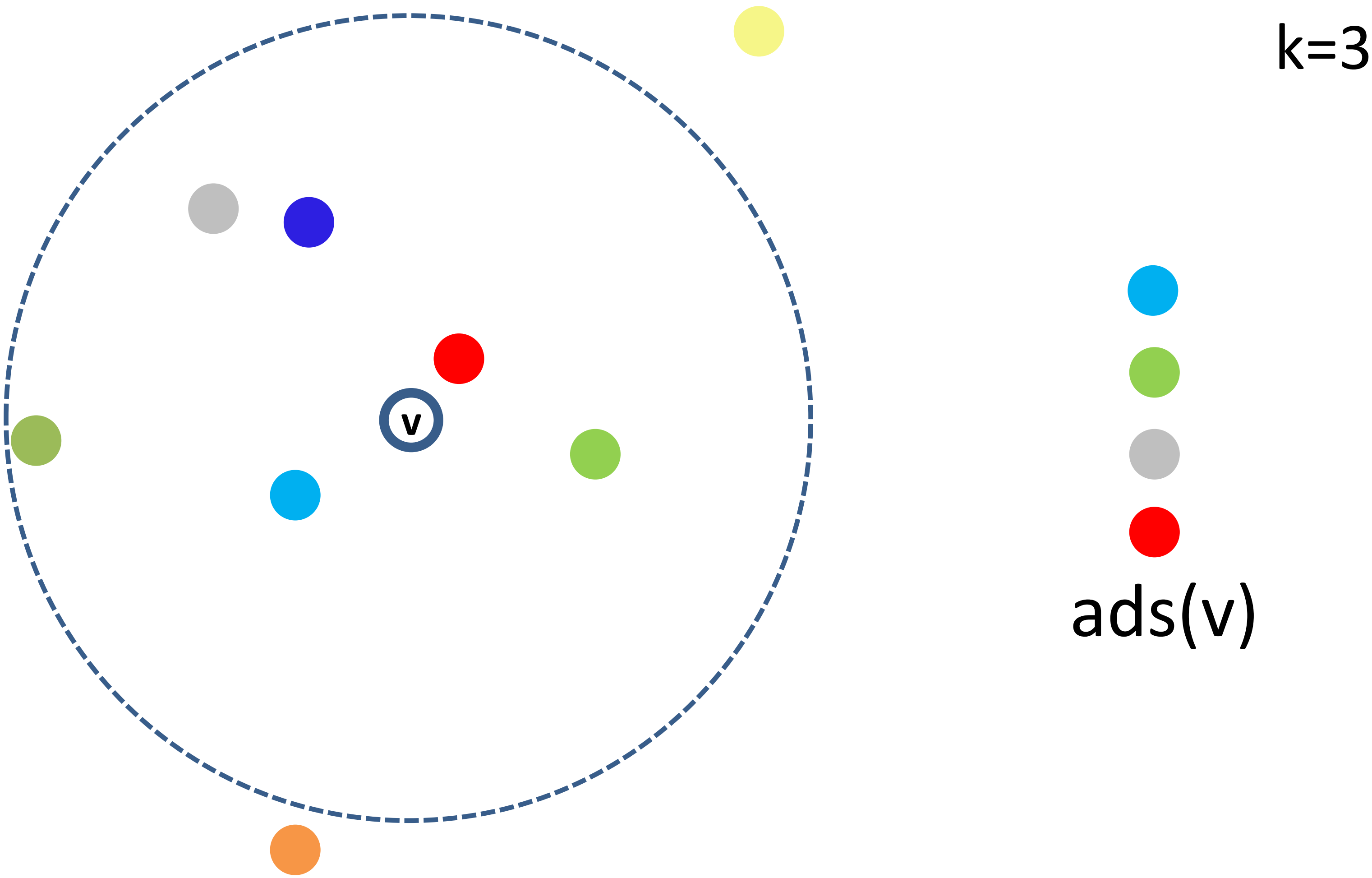

k=3
v
ads(v)

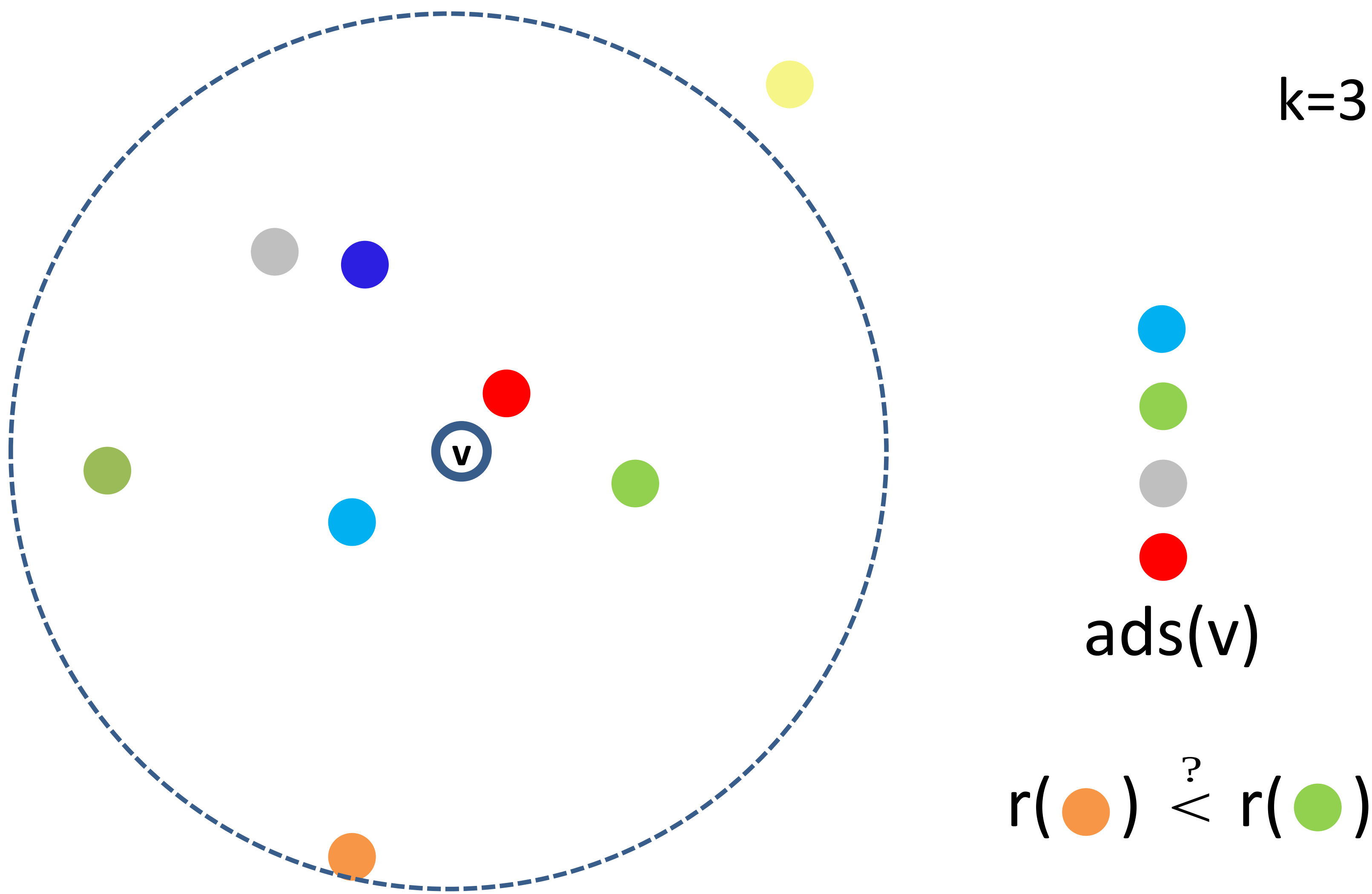

k=3
v
ads(v)
r( ) <? r( )

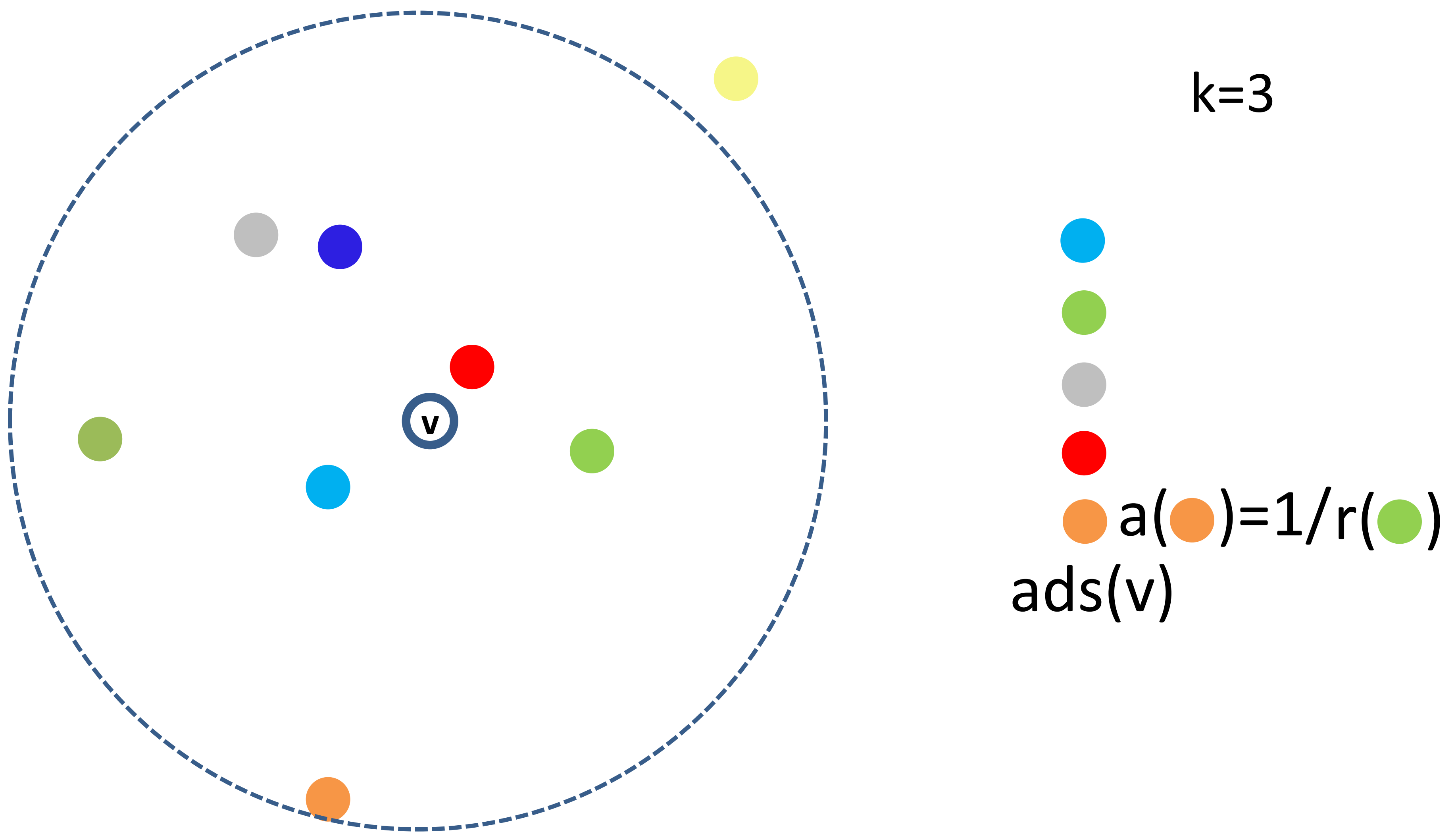

k=3
v
a( )=1/r( )
ads(v)

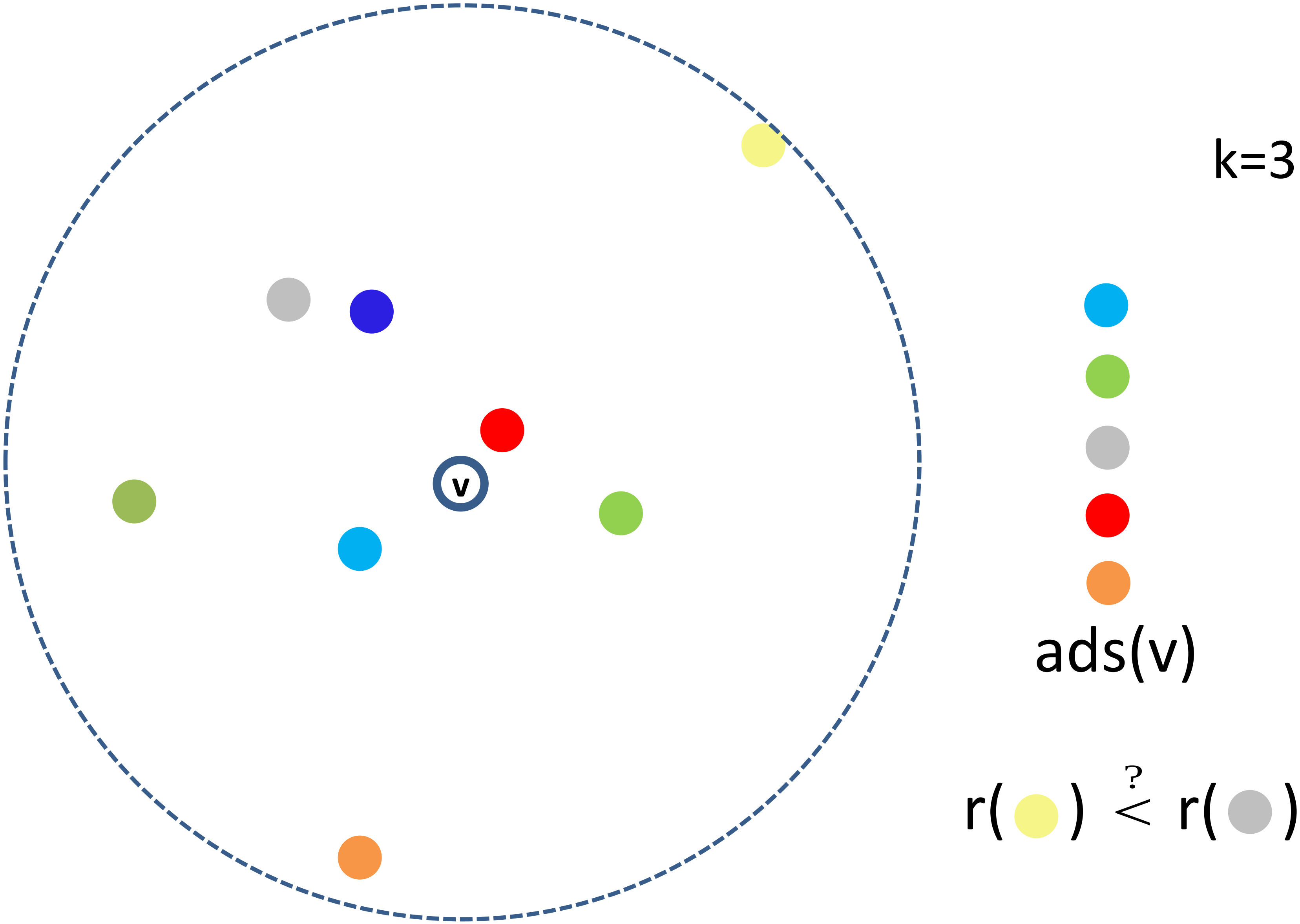


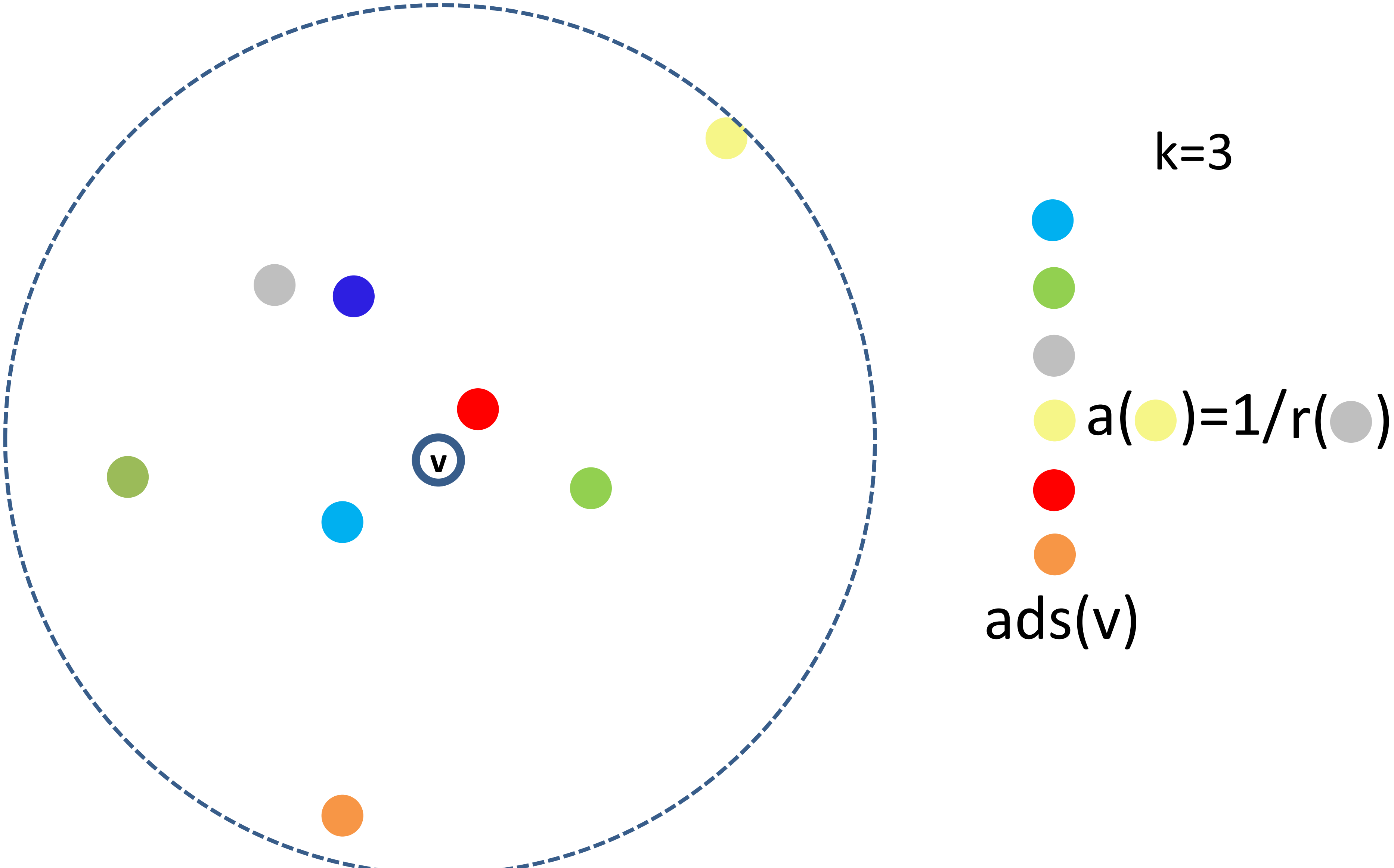
k=3
v
ads(v)
r( ) $\overset{?}{<}$ r( )


v
k=3
ads(v)

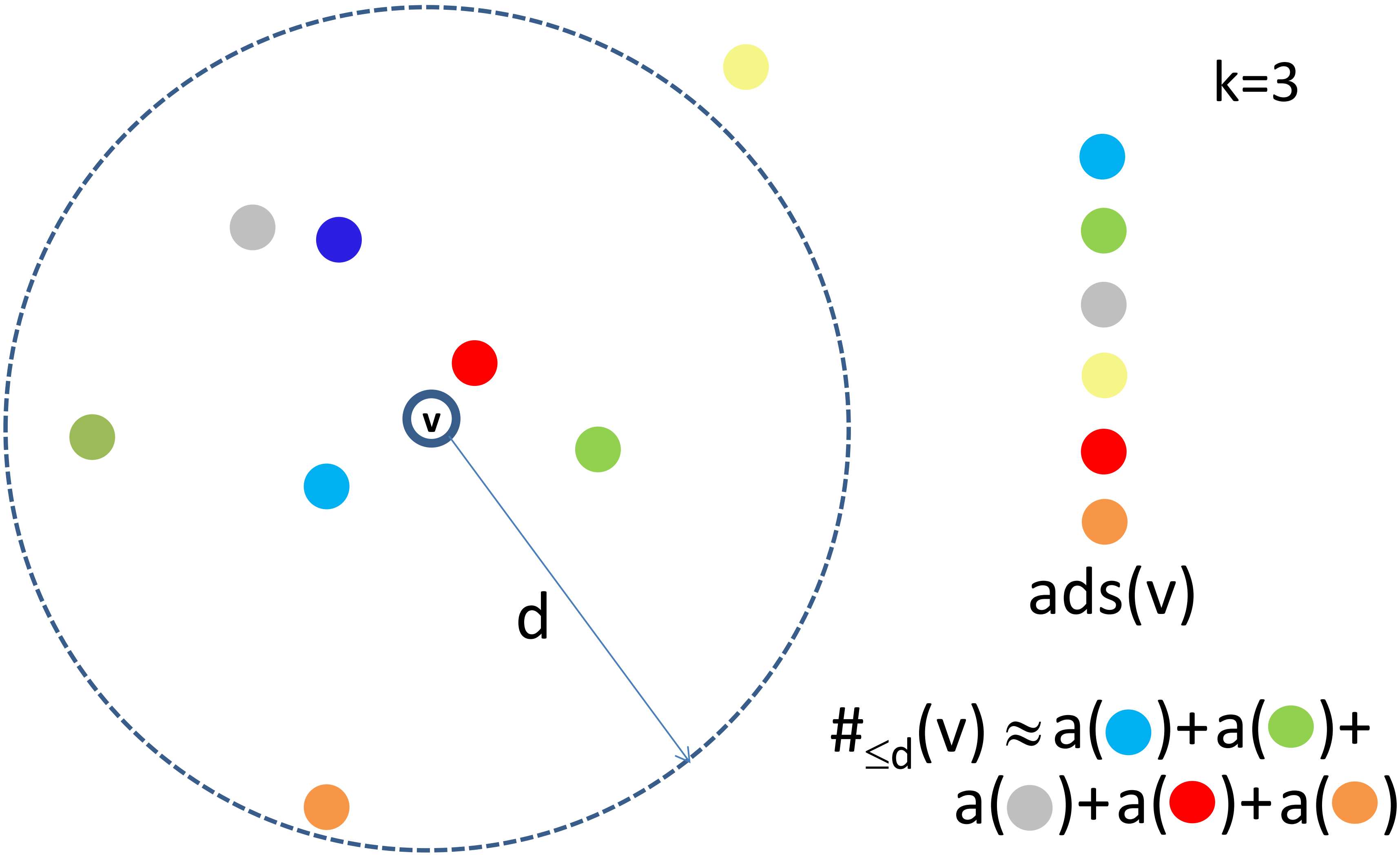

k=3
v
ads(v)
r( ) $\overset{?}{<}$ r( )


v
k=3
a( )=1/r( )
ads(v)


v
k=3
ads(v)
r( ) $\overset{?}{<}$ r( )


v
k=3
a( )=1/r( )
ads(v)

Estimating



More material

- Edith Cohen: **All-distances sketches, revisited: HIP estimators for massive graphs analysis.** PODS 2014
- **Estimation for Monotone Sampling: Competitiveness and Customization. PODC 2014**
- Edith Cohen: **Distance Queries from Sampled Data: Accurate and Efficient. KDD 2014**
- Edith Cohen, Haim Kaplan: **What You Can Do with Coordinated Samples.** APPROX-RANDOM 2013
- Edith Cohen, Haim Kaplan: **Leveraging discarded samples for tighter estimation of multiple-set aggregates.** SIGMETRICS/Performance 2009

- **And earlier refs that you can find in the above**