A Theoretical Justification of Link Prediction Heuristics

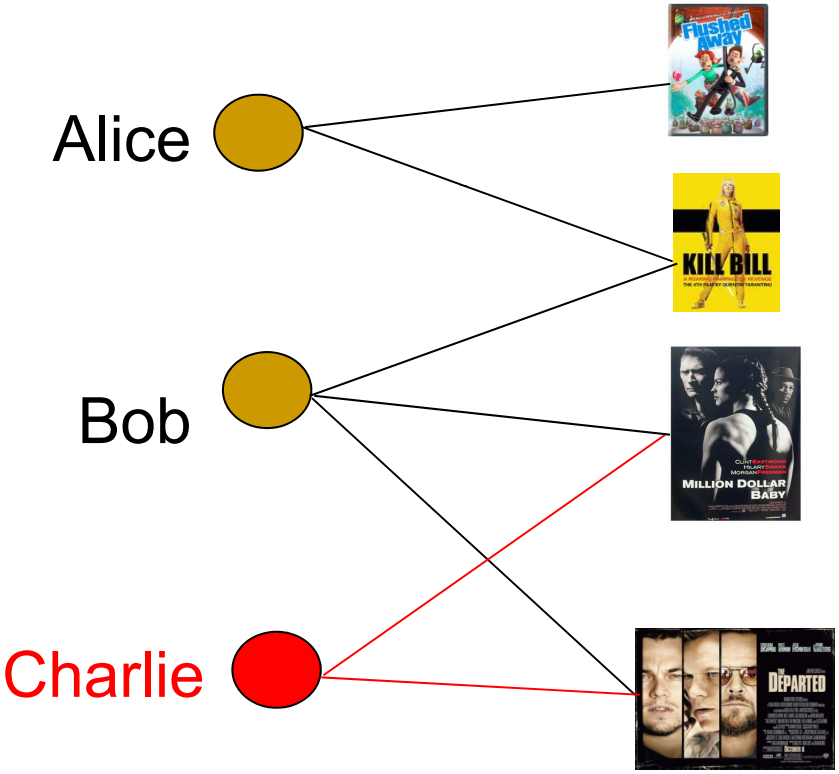
Deepayan Chakrabarti (deepay@cs.cmu.edu)

Purnamrita Sarkar

Andrew Moore

Link Prediction

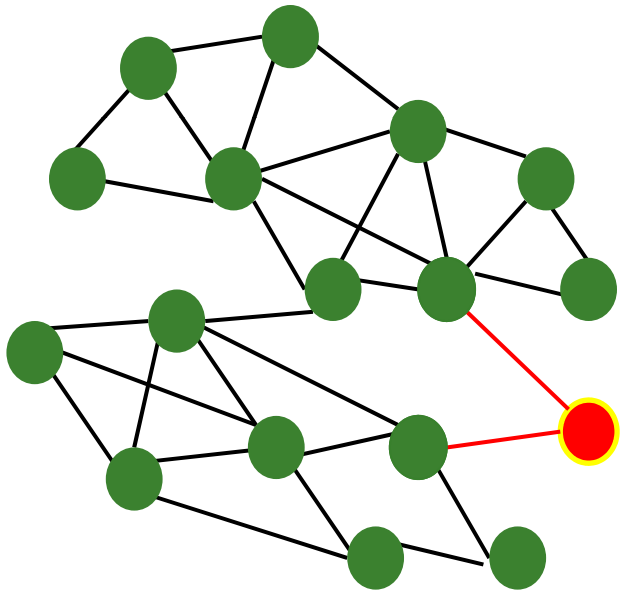
- Which pair of nodes {i,j} should be connected?



Goal: Recommend a movie

Link Prediction

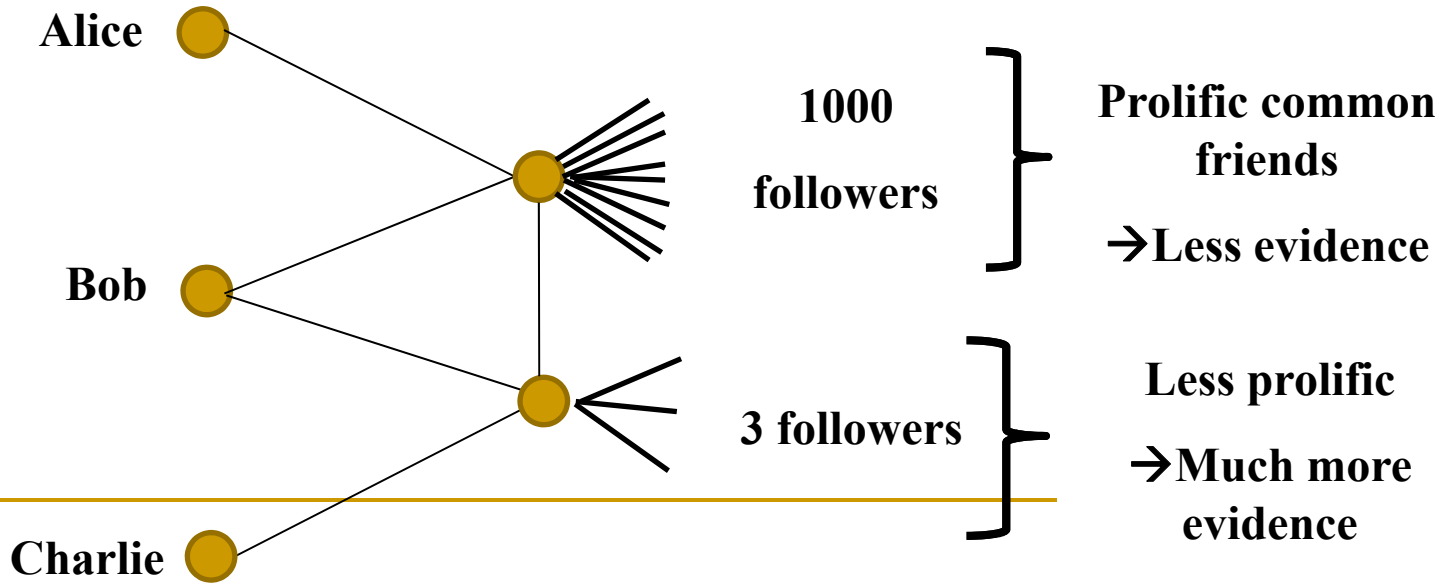
- Which pair of nodes {i,j} should be connected?



Goal: Suggest friends

Link Prediction Heuristics

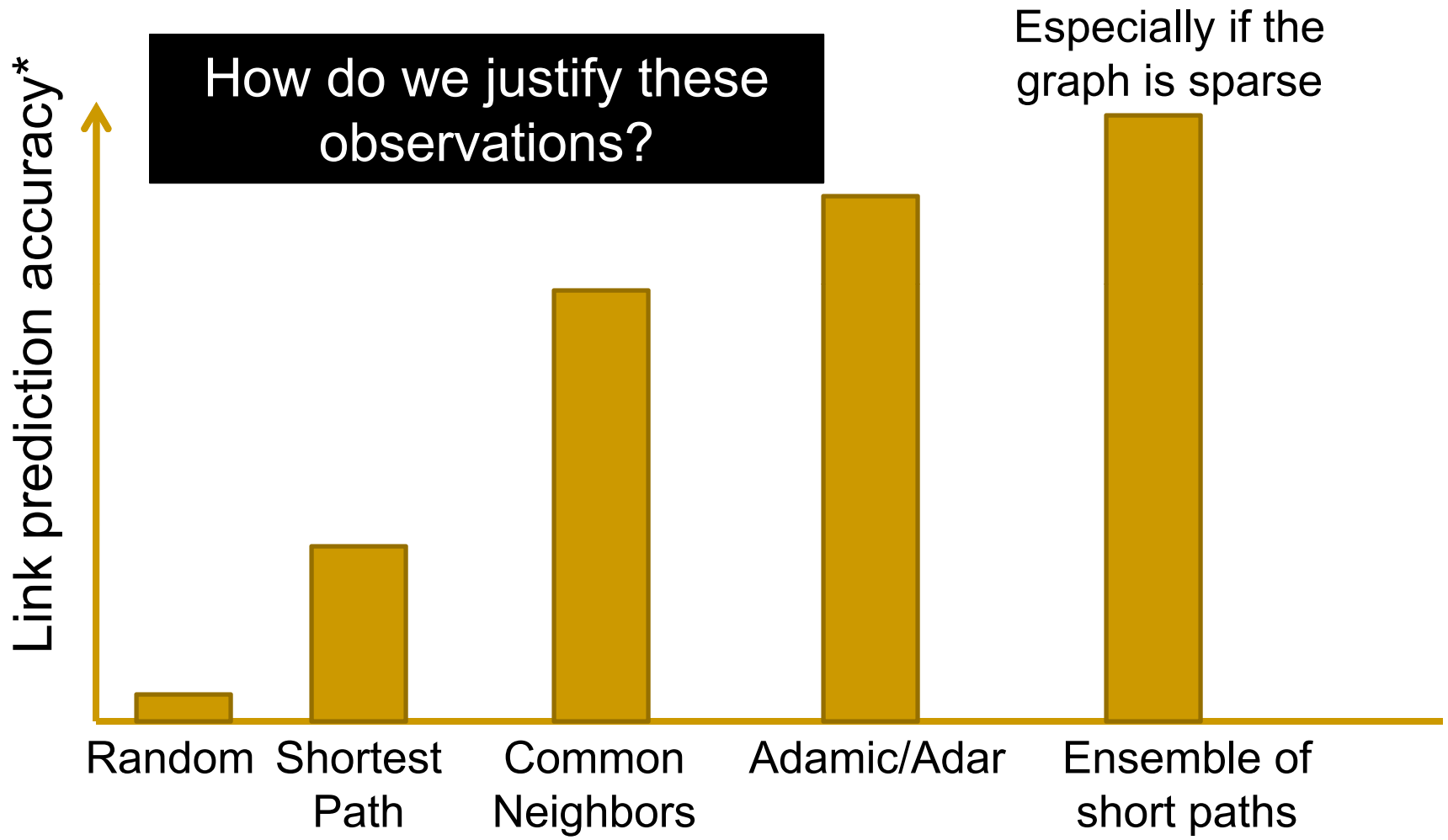
- Predict link between nodes
  - Connected by the shortest path
  - With the most common neighbors (length 2 paths)
  - More weight to low-degree common nbrs (Adamic/Adar)



Link Prediction Heuristics

- Predict link between nodes
  - Connected by the shortest path
  - With the most common neighbors (length 2 paths)
  - More weight to low-degree common nbrs (Adamic/Adar)
  - With more *short* paths (e.g. length 3 paths)
    - exponentially decaying weights to longer paths (Katz measure)
  - …

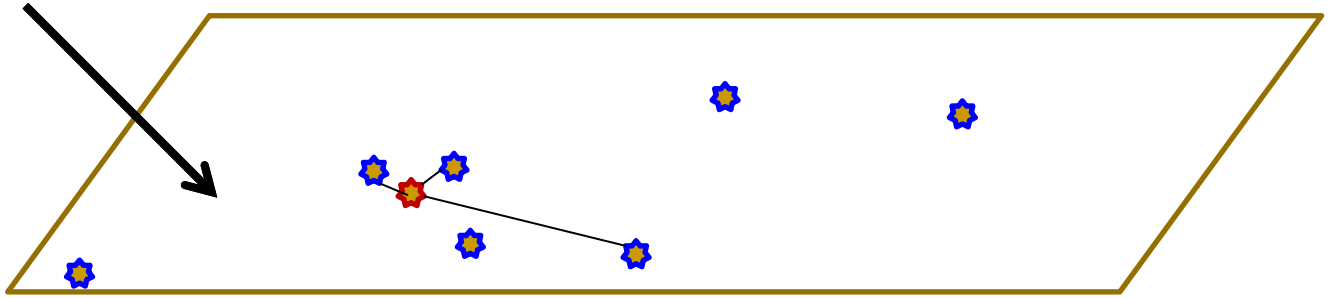
Previous Empirical Studies*



***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

Link Prediction – Generative Model

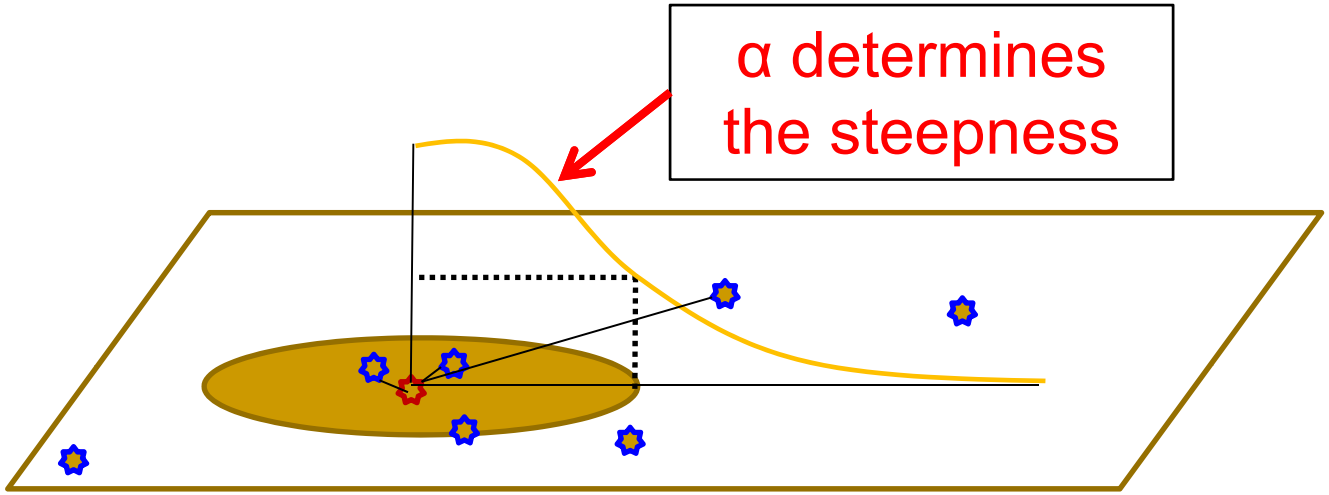
Unit volume universe



Model:

1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph
   - Logistic distance function (Raftery+/2002)

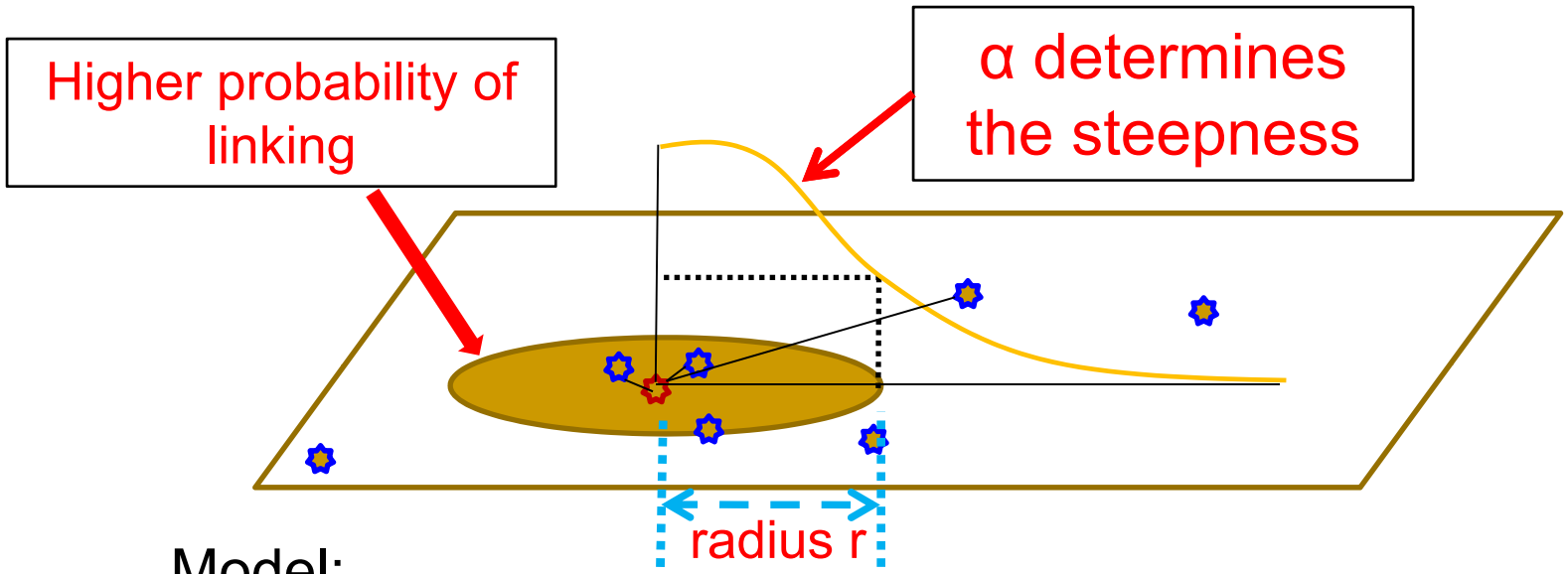
Link Prediction – Generative Model



Model:

1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph

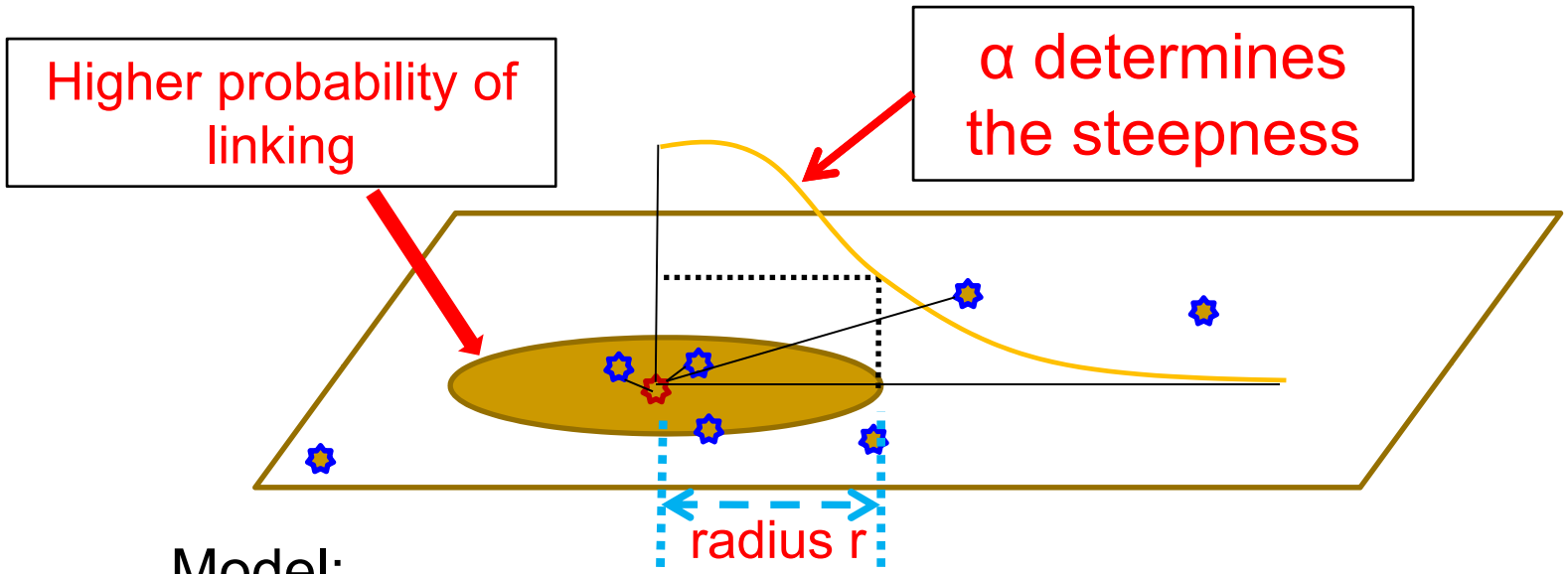
Link Prediction – Generative Model



Model:

1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph

Link Prediction – Generative Model



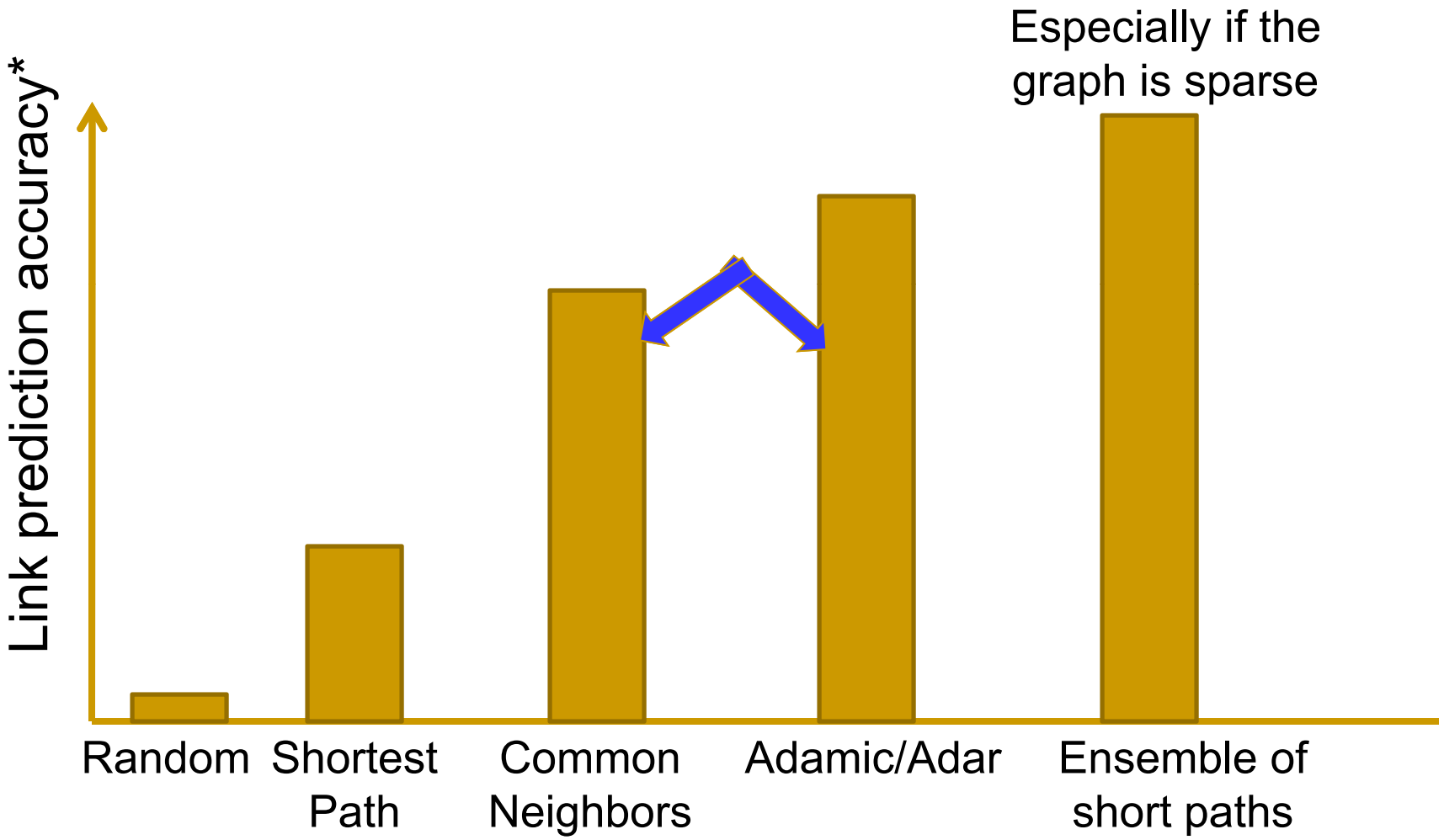
Model:

1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph

Link prediction ≈ find nearest neighbor who is not currently linked to the node.

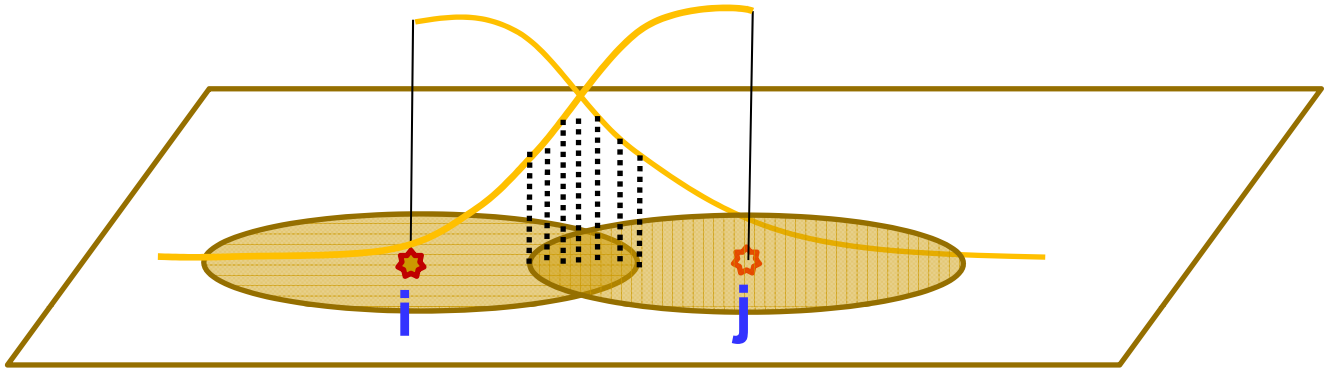
❖ Equivalent to inferring distances in the latent space

Previous Empirical Studies*



***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

Common Neighbors



- $Pr_2(i,j) = Pr(\text{common neighbor}|d_{ij})$

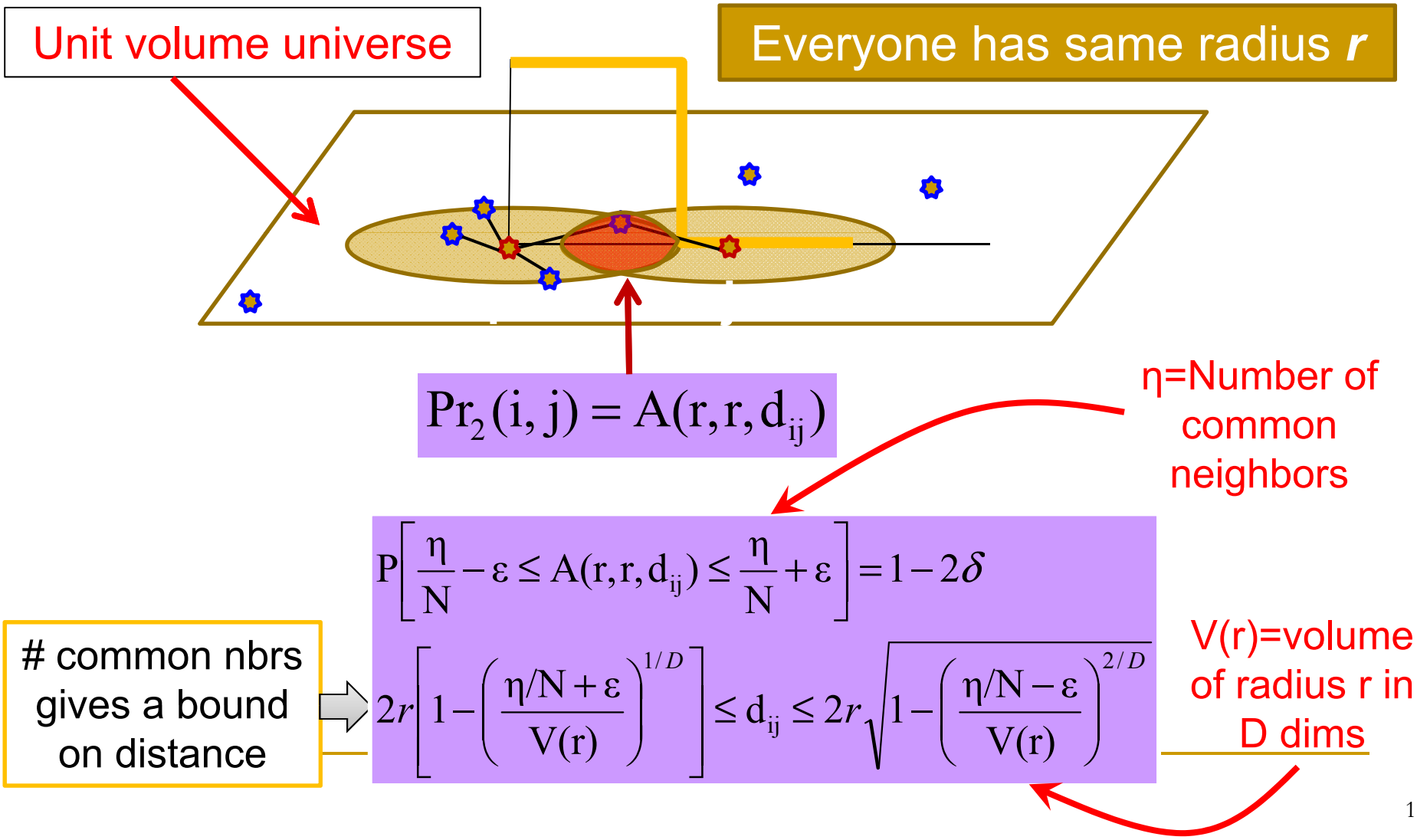
$$\Pr_2(\mathrm{i},\mathrm{j}) = \int \Pr(i \sim k \mid \mathrm{d}_{\mathrm{ik}}) \Pr(j \sim k \mid \mathrm{d}_{\mathrm{jk}}) P(\mathrm{d}_{\mathrm{ik}}, \mathrm{d}_{\mathrm{jk}} \mid \mathrm{d}_{\mathrm{ij}}) \partial \mathrm{d}_{\mathrm{ik}} \partial \mathrm{d}_{\mathrm{jk}}$$

Product of two logistic probabilities, integrated over a volume determined by d_{ij}

As $\alpha \rightarrow \infty$ Logistic → Step function

Much easier to analyze!

Common Neighbors



Common Neighbors

- OPT = node closest to i
- MAX = node with max common neighbors with i

- Theorem:

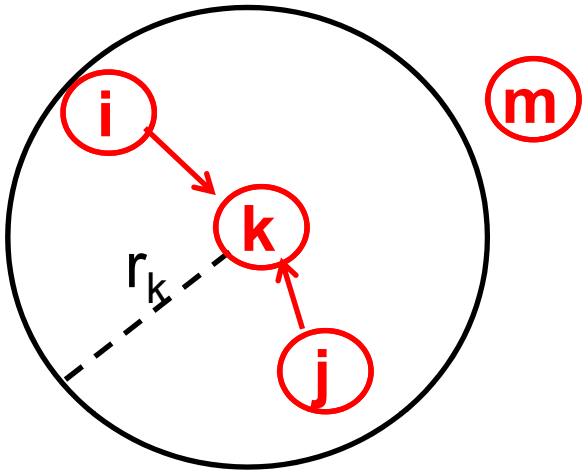
$$d_{OPT} \leq d_{MAX} \overset{w.h.p}{\leq} d_{OPT} + 2[\varepsilon/V(1)]^{1/D}$$

Link prediction by common neighbors is asymptotically optimal

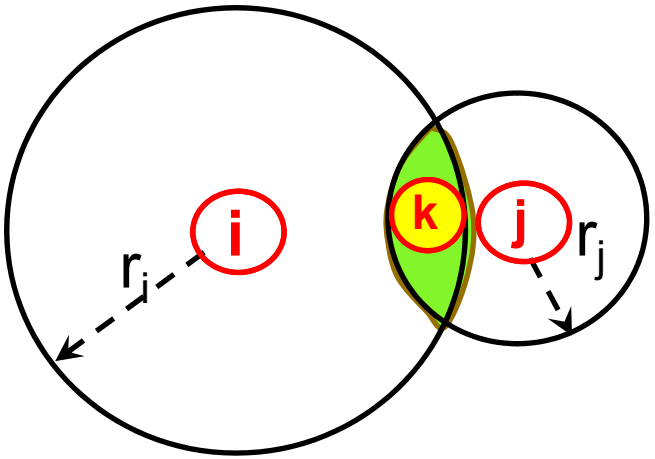
Common Neighbors: Distinct Radii

- Node *k* has radius r_k .

- i→k if $d_{ik} \le r_k$ (Directed graph)
  - r_k captures popularity of node k

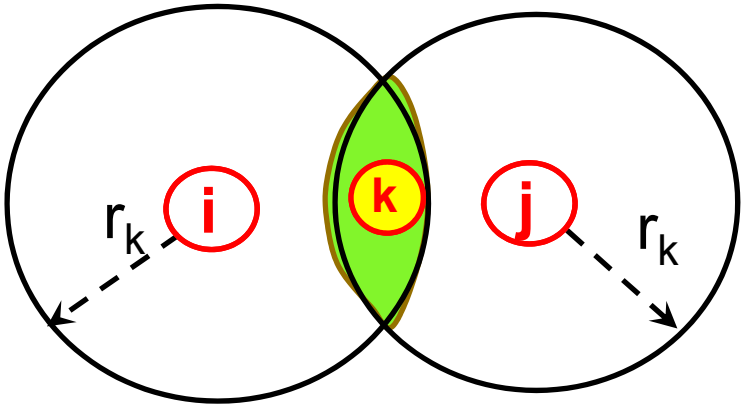


Type 1: i← k → j



$A(r_i , r_j , d_{ij})$

Type 2: i → k ← j



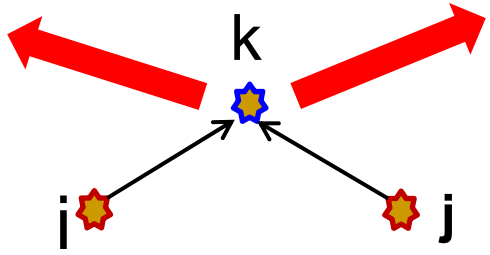
$A(r_k , r_k , d_{ij})$

Type 2 common neighbors

Example graph:

- N_1 nodes of radius r_1 and N_2 nodes of radius r_2
- $r_1 << r_2$

$\eta_1 \sim \text{Bin}[N_1 , A(r_1, r_1, d_{ij})]$ $\eta_2 \sim \text{Bin}[N_2 , A(r_2, r_2, d_{ij})]$



Pick d^* to maximize $\Pr[\eta_1 , \eta_2 \mid d_{ij}]$

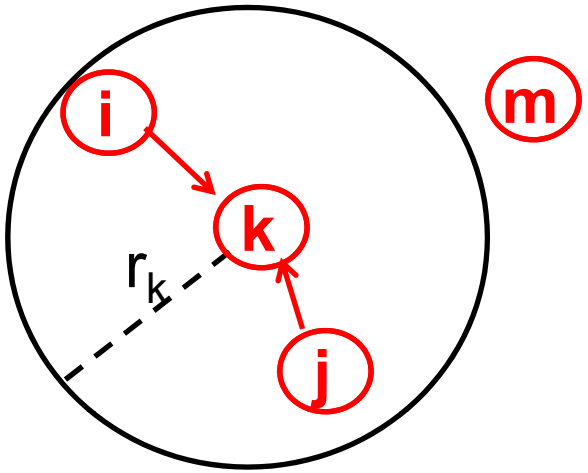
➔ $w(r_1)\, E[\eta_1 \mid d^*] + w(r_2)\, E[\eta_2 \mid d^*] = w(r_1)\eta_1 + w(r_2)\, \eta_2$

Inversely related to d^*

Weighted common neighbors

Common Neighbors: Distinct Radii

- Node *k* has radius r_k .



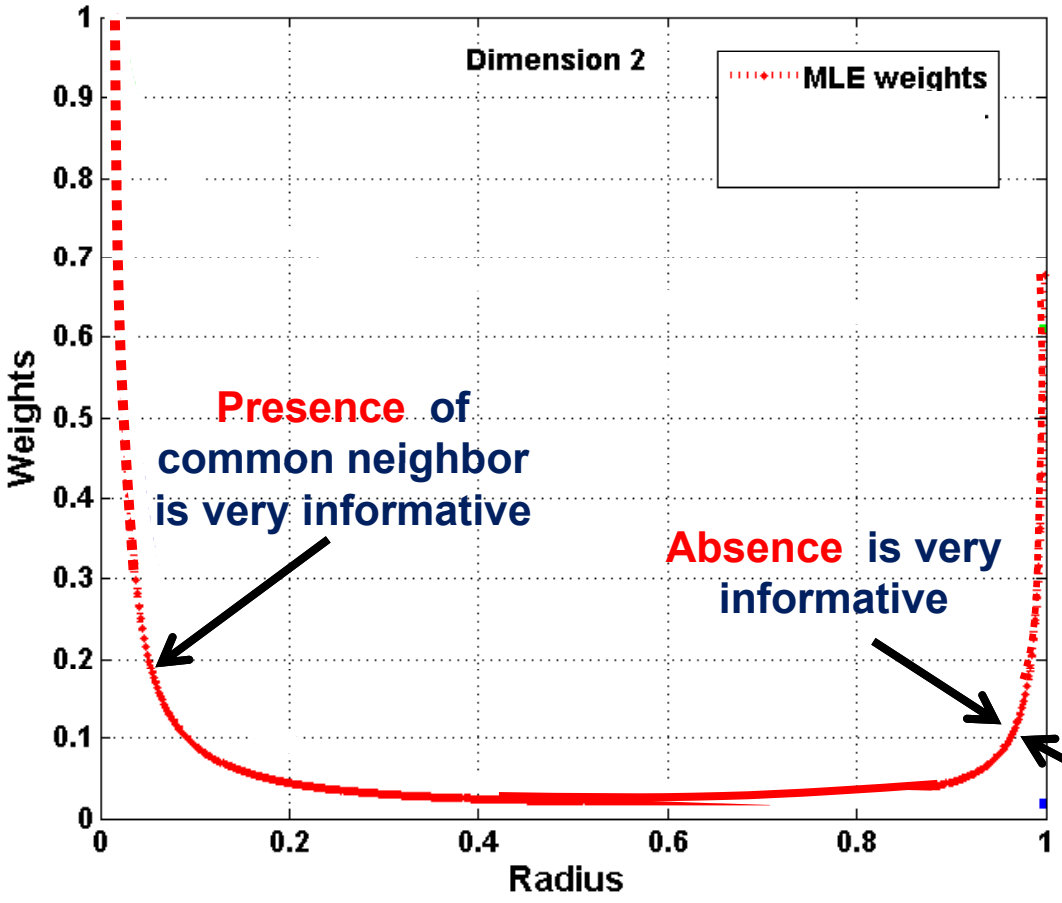
- i→k if $d_{ik} \leq r_k$ (Directed graph)
  - r_k captures popularity of node k

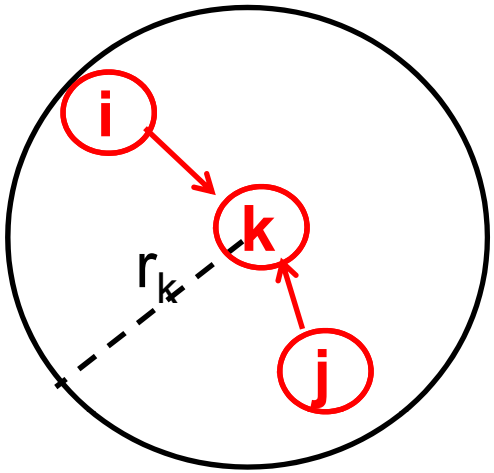
- "Weighted" common neighbors:
  - Predict (i,j) pairs with highest $\sum w(r)\eta(r)$

$\eta(r)$: # common neighbors of radius r

$w(r)$: Weight for nodes of radius r

Type 2 common neighbors



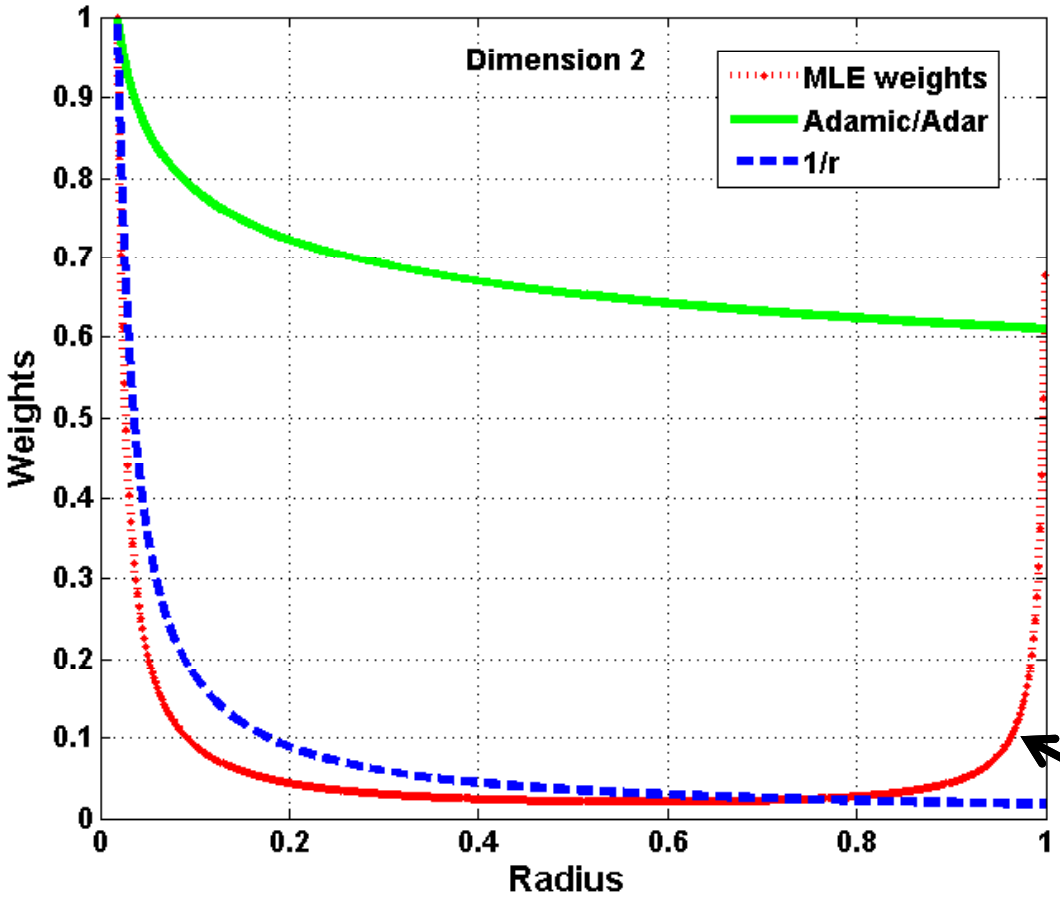


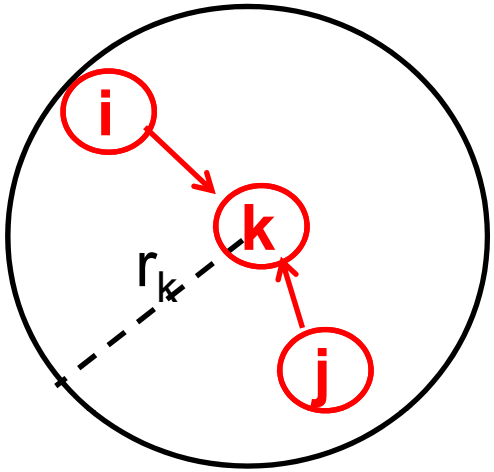
$$w(r) \approx \frac{const}{r} \approx \frac{const}{deg^{1/D}}$$

r is close to max radius

Real world graphs generally fall in this range

Type 2 common neighbors



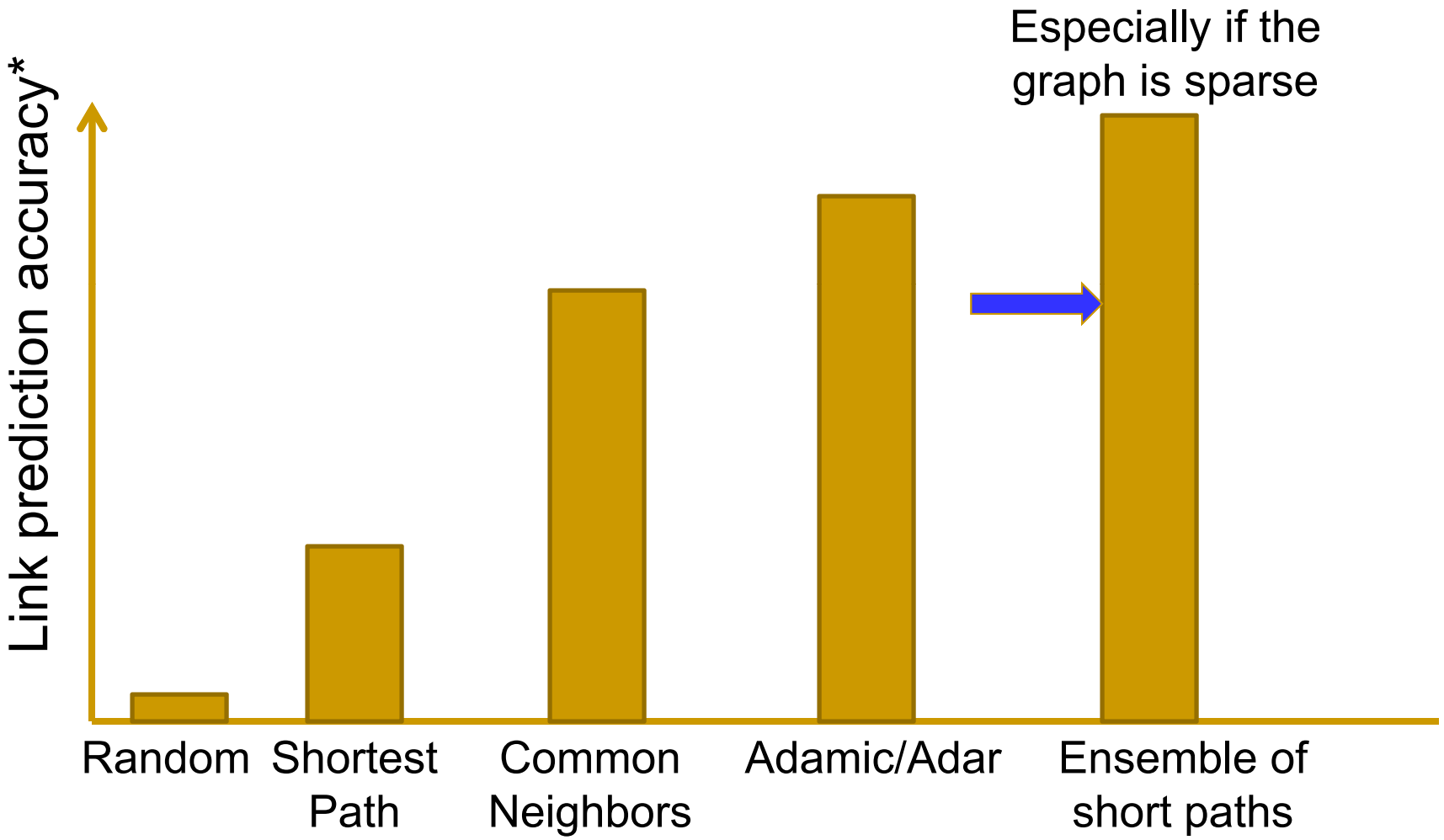


$$w(r) \approx \frac{const}{r} \approx \frac{const}{deg^{1/D}}$$

r is close to max radius

Real world graphs generally fall in this range

Previous Empirical Studies*

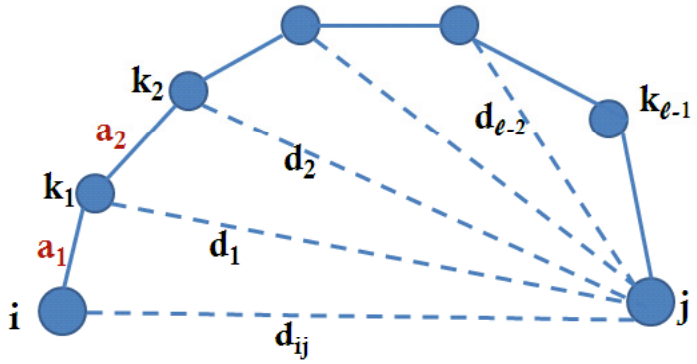


***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

ℓ hop Paths

- Common neighbors = 2 hop paths

- Analysis of longer paths: two components
  1. Bounding $E(\eta_\ell \mid d_{ij})$. [η_ℓ = # ℓ hop paths]
     - Bounds $Pr_\ell(i,j)$ by using triangle inequality on a series of common neighbor probabilities.
  2. $\eta_\ell \approx E(\eta_\ell \mid d_{ij})$



Triangulation

ℓ hop Paths

- Common neighbors = 2 hop paths

- Analysis of longer paths: two components
  1. Bounding $E(\eta_\ell \mid d_{ij})$. [η_ℓ = # ℓ hop paths]
     - Bounds $Pr_\ell(i,j)$ by using triangle inequality on a series of common neighbor probabilities.
  2. $\eta_\ell \approx E(\eta_\ell \mid d_{ij})$
     - Bounded dependence of η_ℓ on position of each node

     ➔ Can use McDiarmid's inequality to bound

     $|\eta_\ell - E(\eta_\ell \mid d_{ij})|$

ℓ-hop Paths

- Common neighbors = 2 hop paths
- For longer paths:

$$d_{ij} \leq r + (\ell - 1)r[1 - g(\eta_\ell, N, \delta)]$$

- Bounds are weaker
- For $\ell' \geq \ell$ we need $\eta_{\ell'} >> \eta_\ell$ to obtain similar bounds
  - ➔ justifies the exponentially decaying weight given to longer paths by the Katz measure

Summary

- Three key ingredients

1. Closer points are likelier to be linked.

   Small World Model- Watts, Strogatz, 1998, Kleinberg 2001

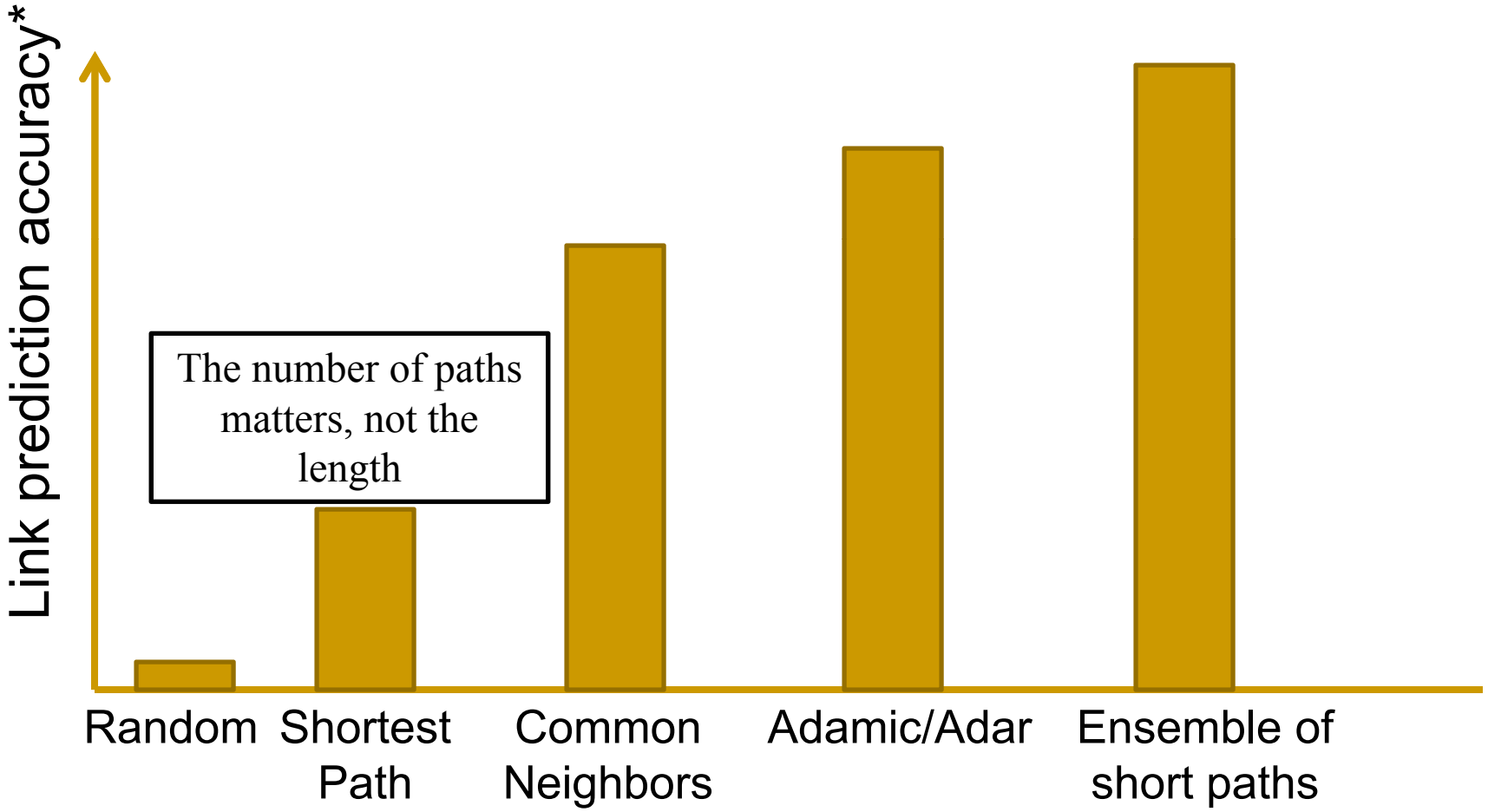
2. Triangle inequality holds

   ➔necessary to extend to ℓ-hop paths

3. Points are spread uniformly at random

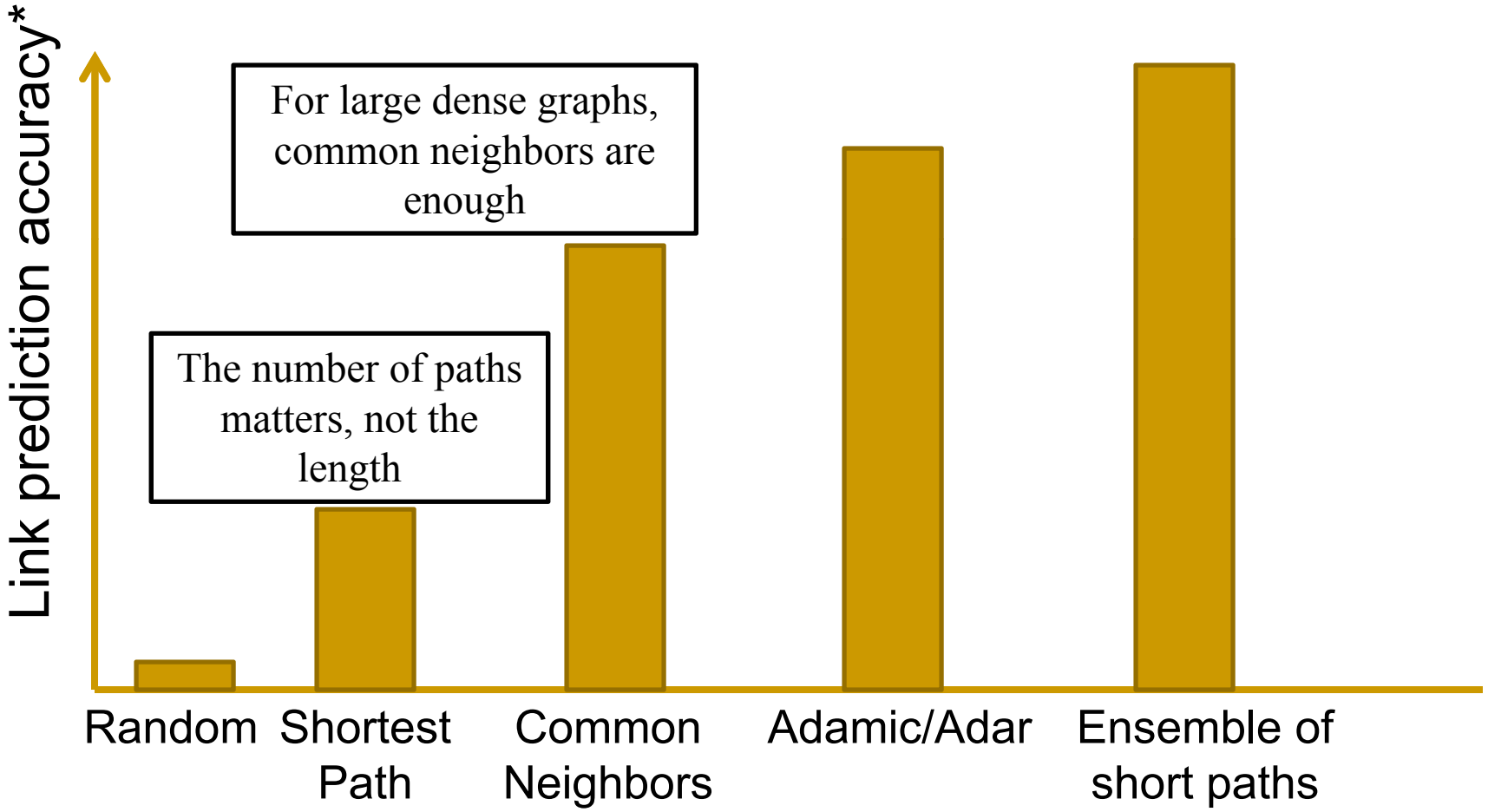
   ➔ Otherwise properties will depend on location as well as distance

Summary



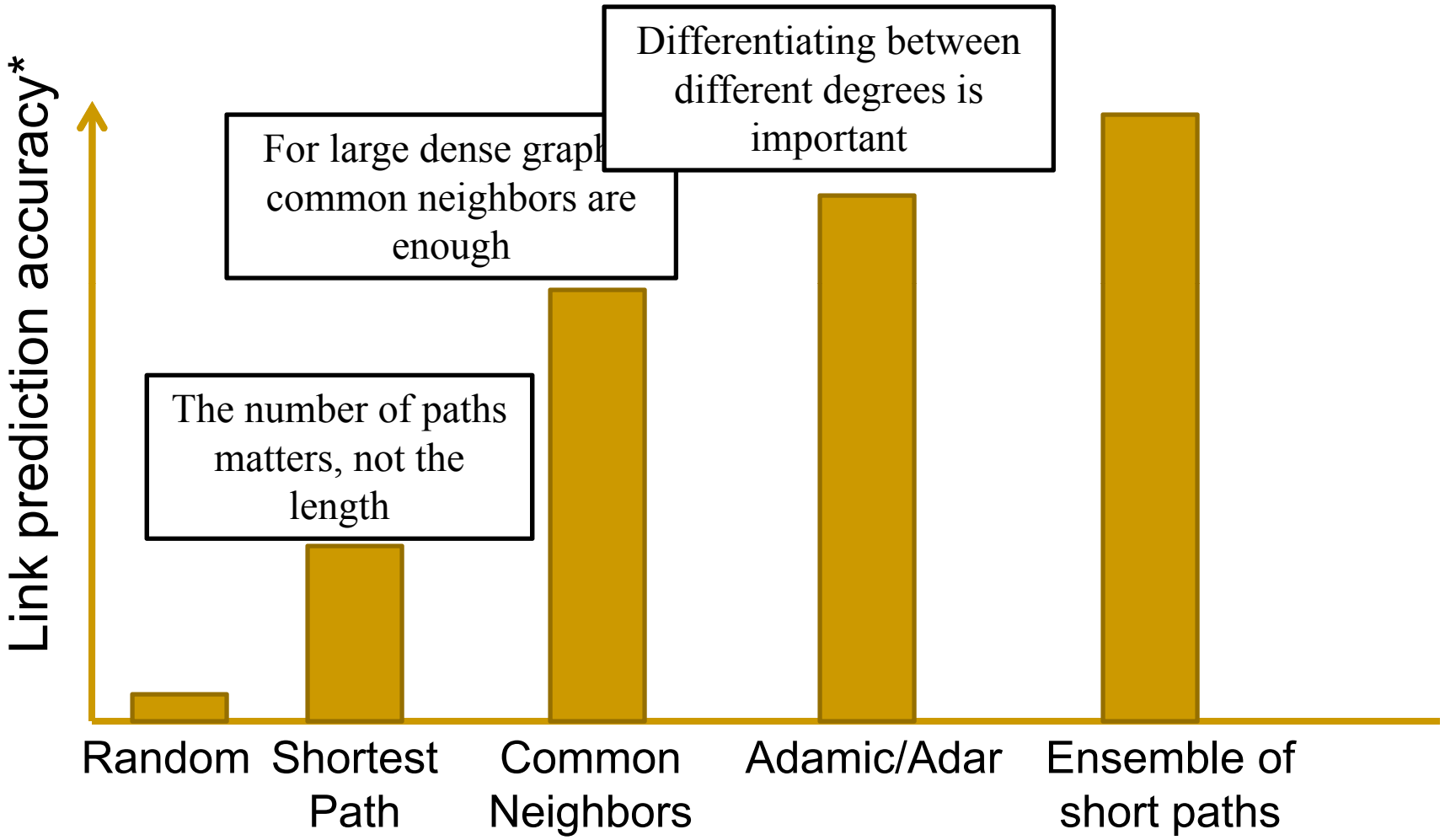
***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

Summary



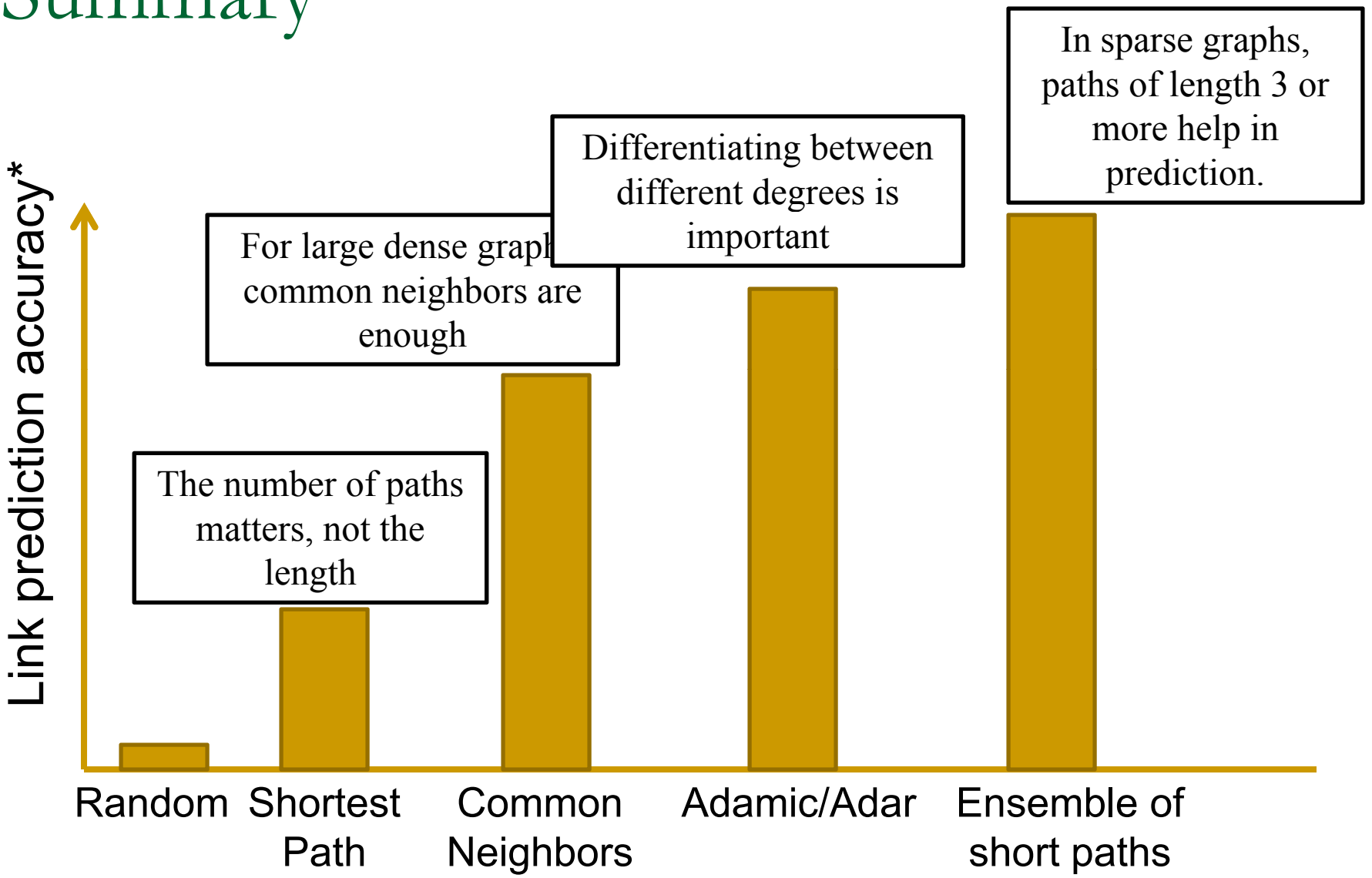
***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

Summary



***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

Summary



***Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007**

Sweep Estimators