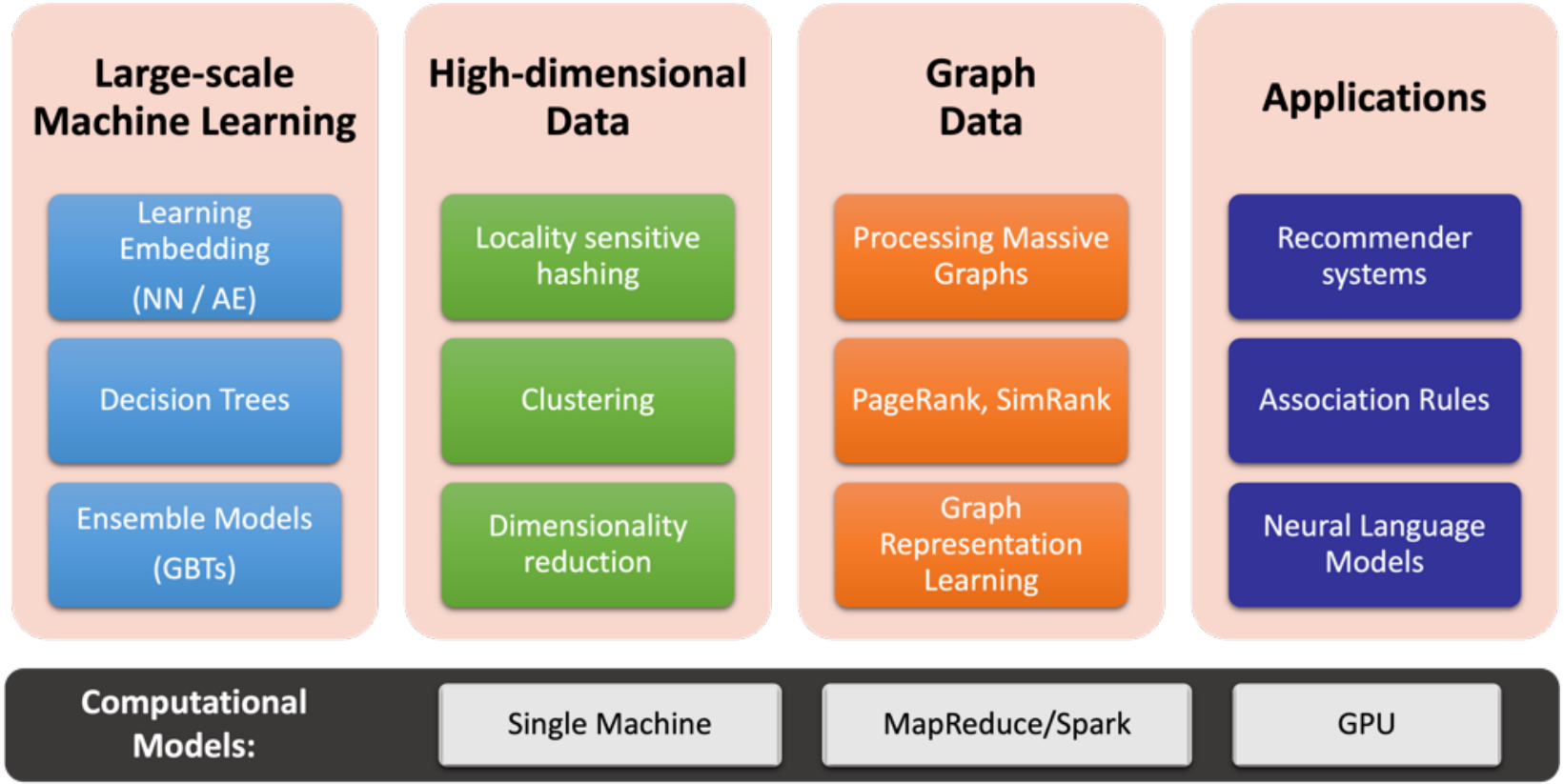
MIE524 Data Mining

PageRank Extensions

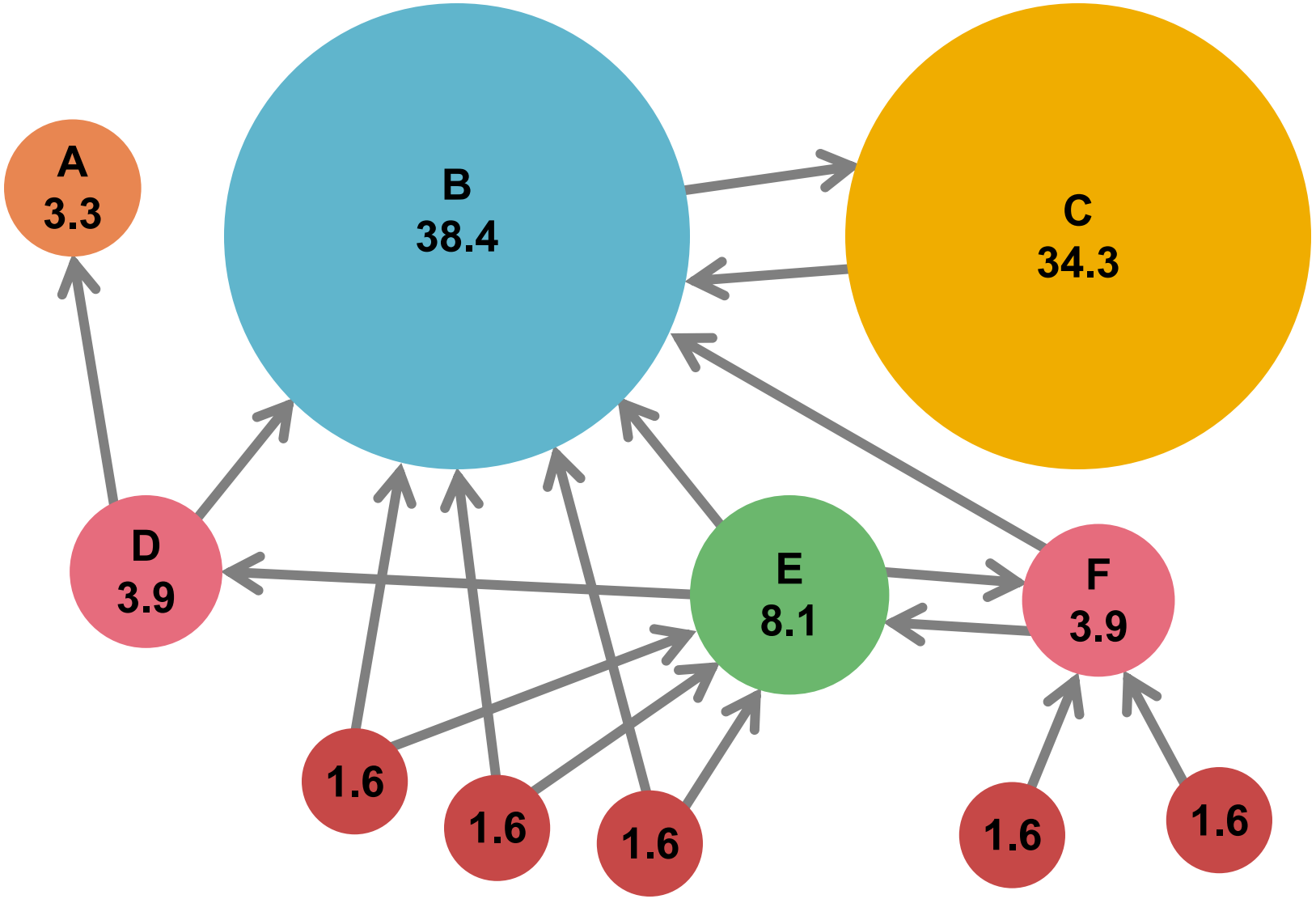
Slides Credits:
Slides from Leskovec, Rajaraman, Ullman (http://www.mmds.org), Leskovec & Ghashami

MIE524: Course Topics (Tentative)



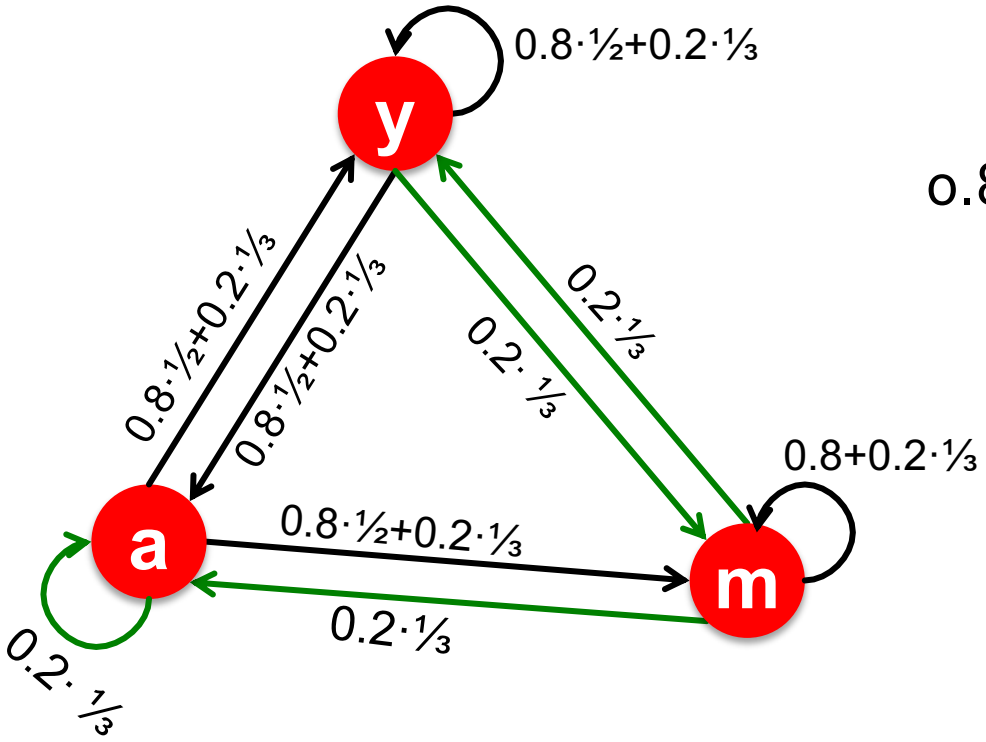
Example: PageRank Scores

Reminder



Random Teleports ($\beta = 0.8$)

Reminder



M

$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$[1/N]_{NxN}$

$$\begin{array}{c} y \\ a \\ m \end{array} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix}$$

A

$$\begin{array}{c} y \\ a \\ m \end{array} = \begin{array}{ccccc} 1/3 & 0.33 & 0.28 & 0.26 & \\ 1/3 & 0.20 & 0.20 & 0.18 & \ldots \\ 1/3 & 0.46 & 0.52 & 0.56 & \end{array} \begin{array}{c} 7/33 \\ 5/33 \\ 21/33 \end{array}$$

r = A r

PageRank: The Complete Algorithm

Reminder

- **Input: Graph G and parameter β**
  - Directed graph G (can have **spider traps** and **dead ends**)
  - Parameter β
- **Output: PageRank vector r**
  - **Set:** $r_j^{(0)} = \frac{1}{N}, t = 1$
  - **Do:** $\forall j$: $r'_j = \sum_{i \to j} \beta \frac{r_i^{(t-1)}}{d_i}$

    $r'_j = 0$ if in-degree of j is 0
    - **Now re-insert the leaked PageRank:**

      $\forall j$: $r_j^{(t)} = r'_j + \frac{1-S}{N}$ **where:** $S = \sum_j r'_j$
    - $t = t + 1$
  - **while** $\sum_j \left| r_j^{(t)} - r_j^{(t-1)} \right| < \varepsilon$

If the graph has no dead-ends then the amount of leaked PageRank is **1-β**. But since we have dead-ends the amount of leaked PageRank may be larger. We have to explicitly account for it by computing **S**.

Topic-Specific PageRank

- **Instead of generic importance, can we measure importance within a topic?**
- **Goal:** Evaluate Web pages not just according to their importance, but also by how close they are to a particular topic, e.g. "sports" or "history"
- **Allows search queries to be answered based on the interests of a user**
  - **Example:** Query "Trojan" wants different pages depending on whether you are interested in sports, history, or computer security

Topic-Specific PageRank

- Random walker has a small probability of teleporting at any step
- **Teleport can go to:**
  - **Standard PageRank: Any page with equal probability**
    - To avoid dead-end and spider-trap problems
  - **Topic Specific PageRank: A topic-specific set of "relevant" pages (teleport set)**
- **Idea: Bias the random walk**
  - When the walker teleports, she picks a page from a set ***S***
  - ***S*** contains only pages that are relevant to the topic
    - E.g., Open Directory (DMOZ) pages for a given topic/query
  - For each teleport set ***S***, we get a different vector r_S

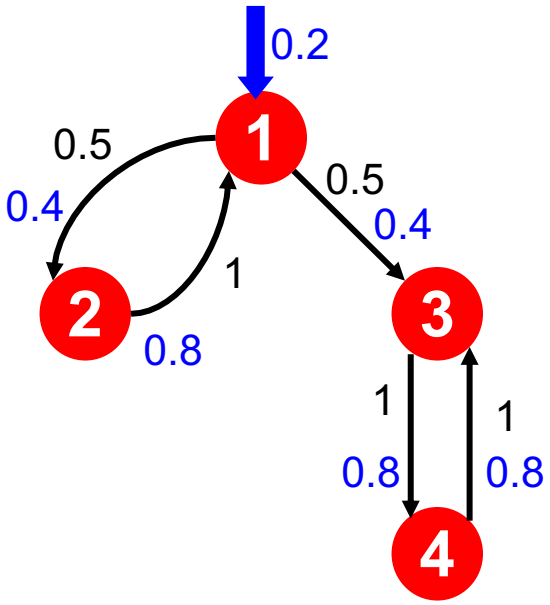
Matrix Formulation

- **To make this work all we need is to update the teleportation part of the PageRank formulation:**

$$A_{ij} = \begin{cases} \beta M_{ij} + (1-\beta)/|S| & \text{if } i \in S \\ \beta M_{ij} + 0 & \text{otherwise} \end{cases}$$

  - A is a stochastic matrix!
- We weighted all pages in the teleport set S equally
  - **Could also assign different weights to pages!**
- **Compute as for regular PageRank:**
  - Multiply by M, then add a vector of $(1-\beta)/|S|$
  - Maintains sparseness

Example: Topic-Specific PageRank



Suppose $S = \{1\}$, $\beta = 0.8$

Node	Iteration				
	0	1	2	…	stable
1	0.25	0.4	0.28		0.294
2	0.25	0.1	0.16		0.118
3	0.25	0.3	0.32		0.327
4	0.25	0.2	0.24		0.261

S={1}, β=0.9:
r=[0.17, 0.07, 0.40, 0.36]
S={1} , β=0.8:
r=[0.29, 0.11, 0.32, 0.26]
S={1}, β=0.7:
r=[0.39, 0.14, 0.27, 0.19]

S={1,2,3,4}, β=0.8:
r=[0.13, 0.10, 0.39, 0.36]
S={1,2,3} , β=0.8:
r=[0.17, 0.13, 0.38, 0.30]
S={1,2} , β=0.8:
r=[0.26, 0.20, 0.29, 0.23]
S={1} , β=0.8:
r=[0.29, 0.11, 0.32, 0.26]

Discovering the Topic Vector S

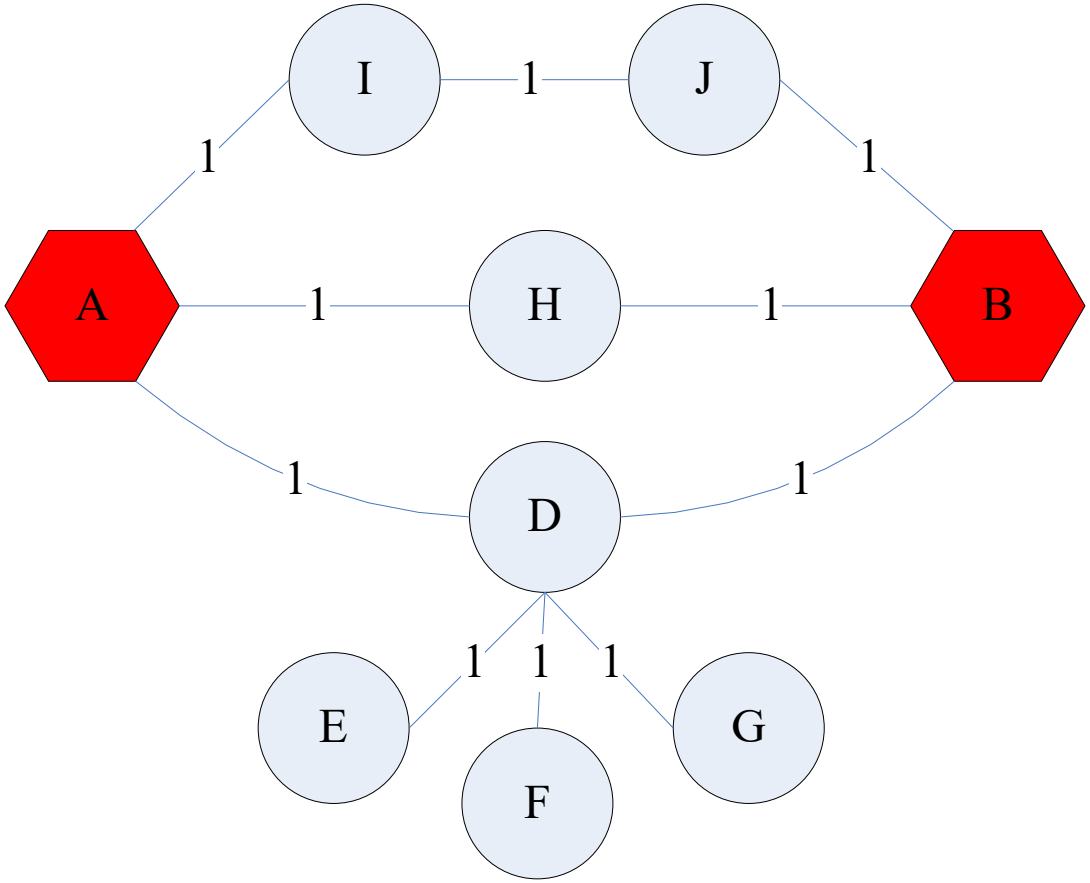
- **Create different PageRanks for different topics**
  - The 16 DMOZ top-level categories:
    - Arts, Business, Sports,…
- **Which topic ranking to use?**
  - User can pick from a menu
  - Classify query into a topic
  - Can use the **context** of the query
    - E.g., query is launched from a web page talking about a known topic
    - History of queries e.g., “basketball” followed by “Jordan”
  - User context, e.g., user’s bookmarks, …

Application to Measuring Proximity in Graphs

Random Walk with Restarts: Set *S* is a single node

[Tong-Faloutsos, '06]

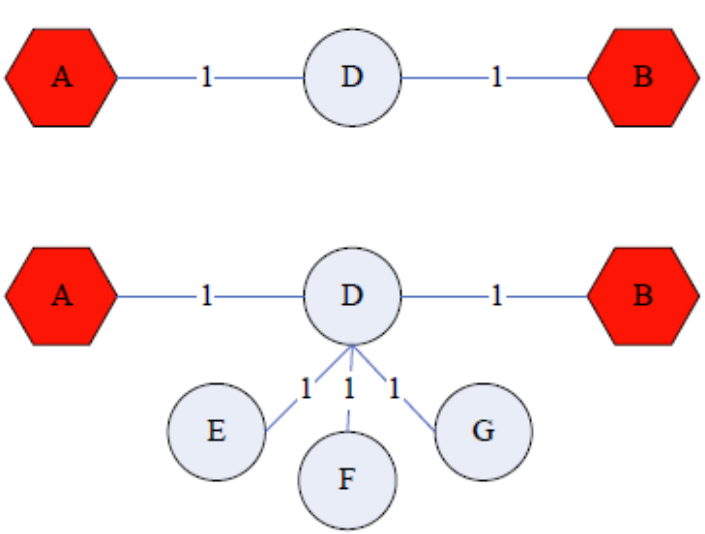
Proximity on Graphs

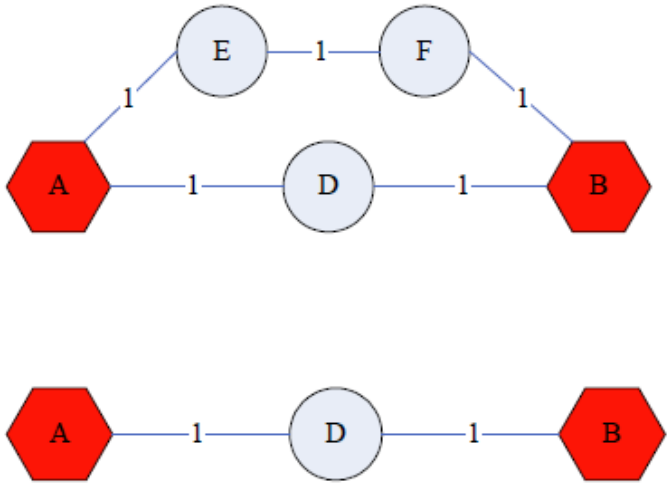


a.k.a.: Relevance, Closeness, 'Similarity'…

Good proximity measure?

- **Shortest path is not good:**

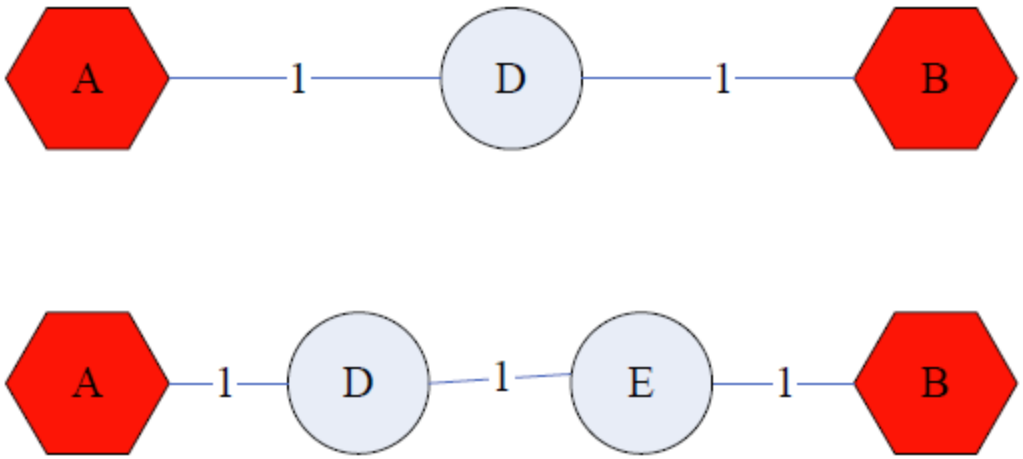




- **No effect of degree-1 nodes (E, F, G)!**
- Multi-faceted relationships

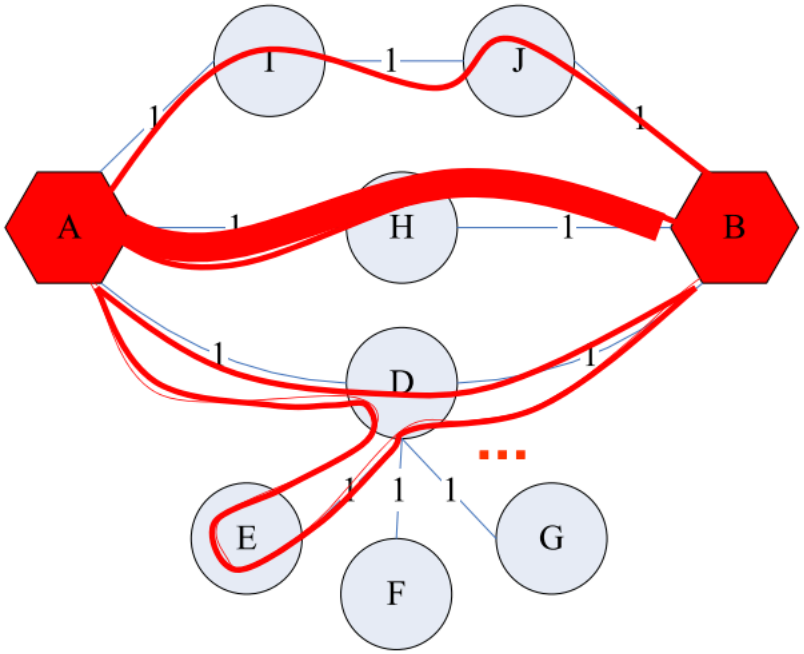
Good proximity measure?

- **Network flow is not good:**



- **Does not punish long paths**

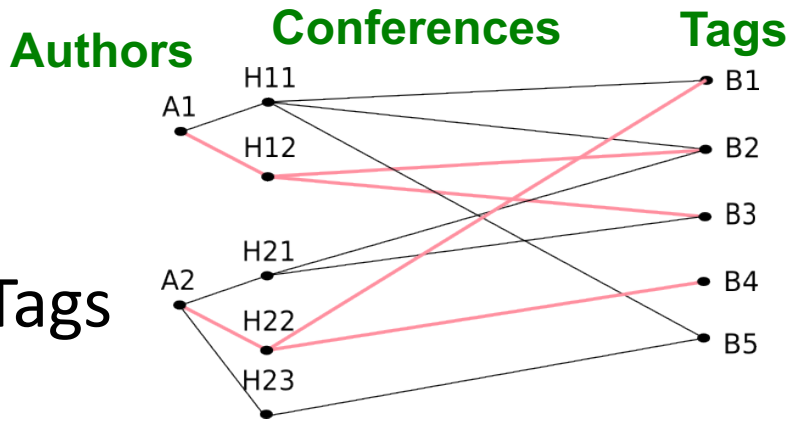
What is a good notion of proximity?



- **Need a method that considers:**
  - Multiple connections
  - Multiple paths
  - Direct and indirect connections
  - Degree of the node

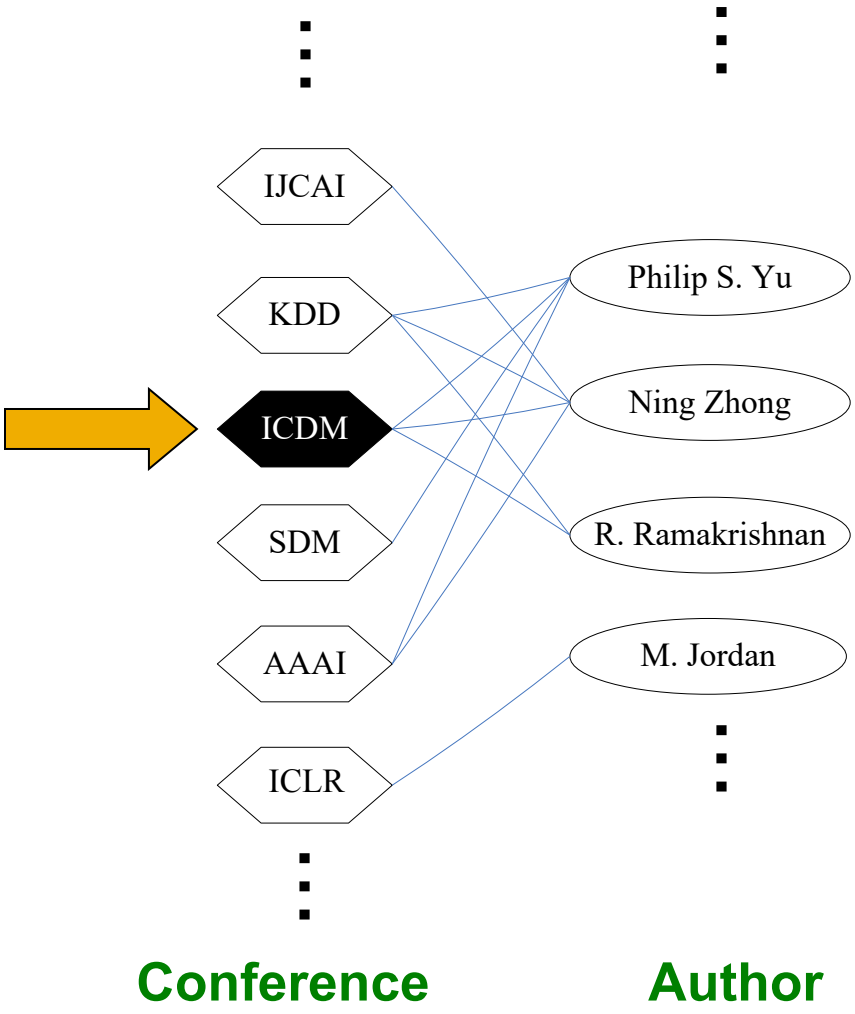
SimRank: Idea

- **SimRank:** Random walks from a **fixed node** on ***k*-partite graphs**
- **Setting: *k*-partite graph** with ***k*** types of nodes
  - E.g.: Authors, Conferences, Tags



- **Topic Specific PageRank** from node ***u***: **teleport set $S = \{u\}$**
- Resulting scores measure similarity/proximity to node ***u***
- **Problem:**
  - Must be done once for each node ***u***
  - Only suitable for sub-Web-scale applications

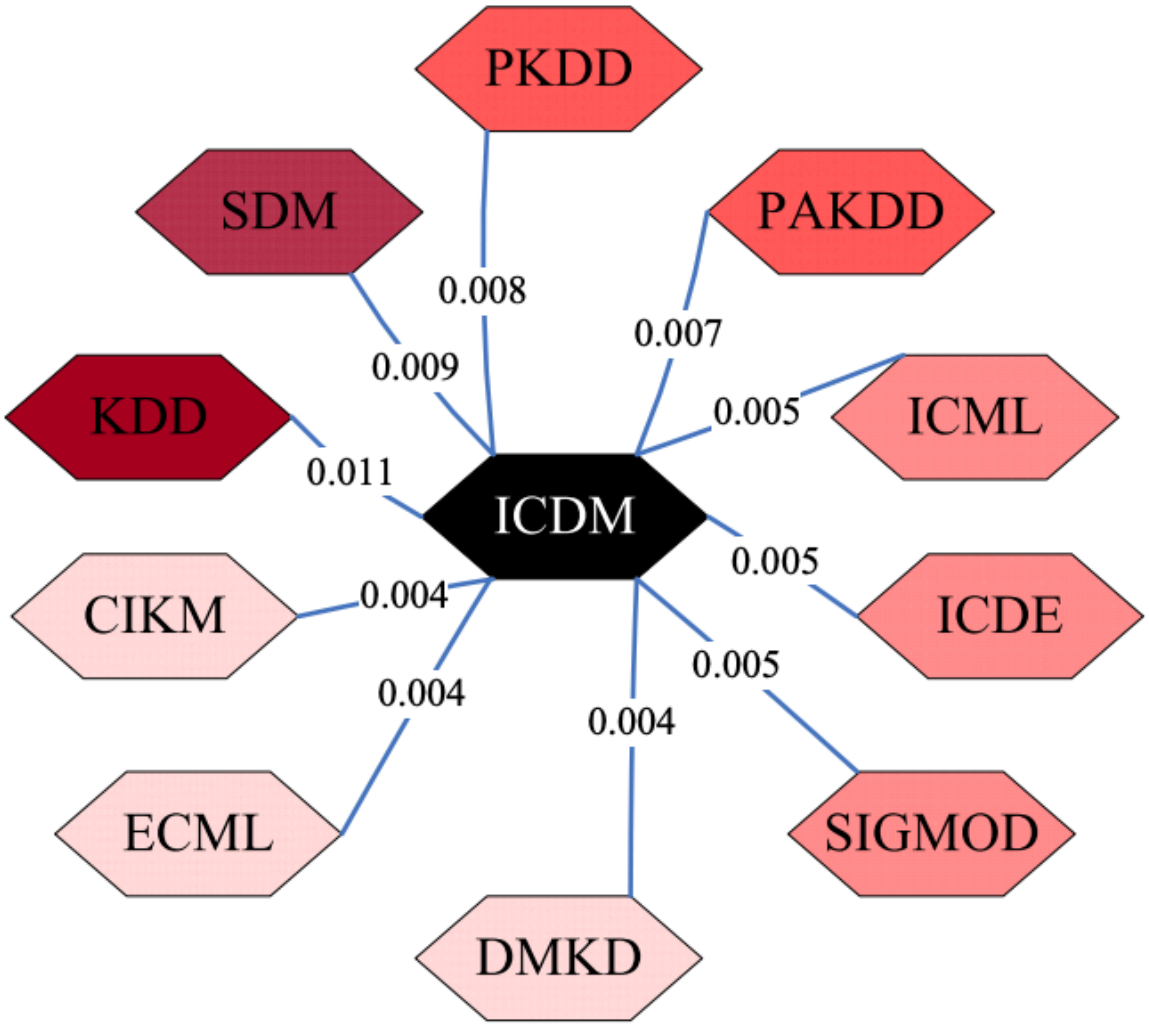
SimRank: Example



Q: What is the most related conference to **ICDM**?

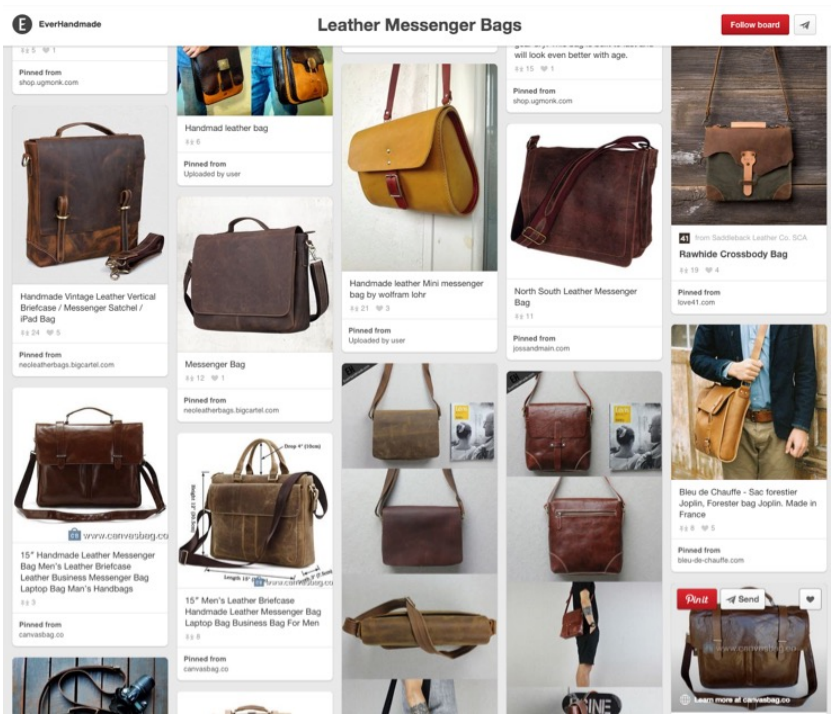
A: Topic-Specific PageRank with teleport set S={ICDM}

SimRank: Example



Pinterest: Pins and Boards

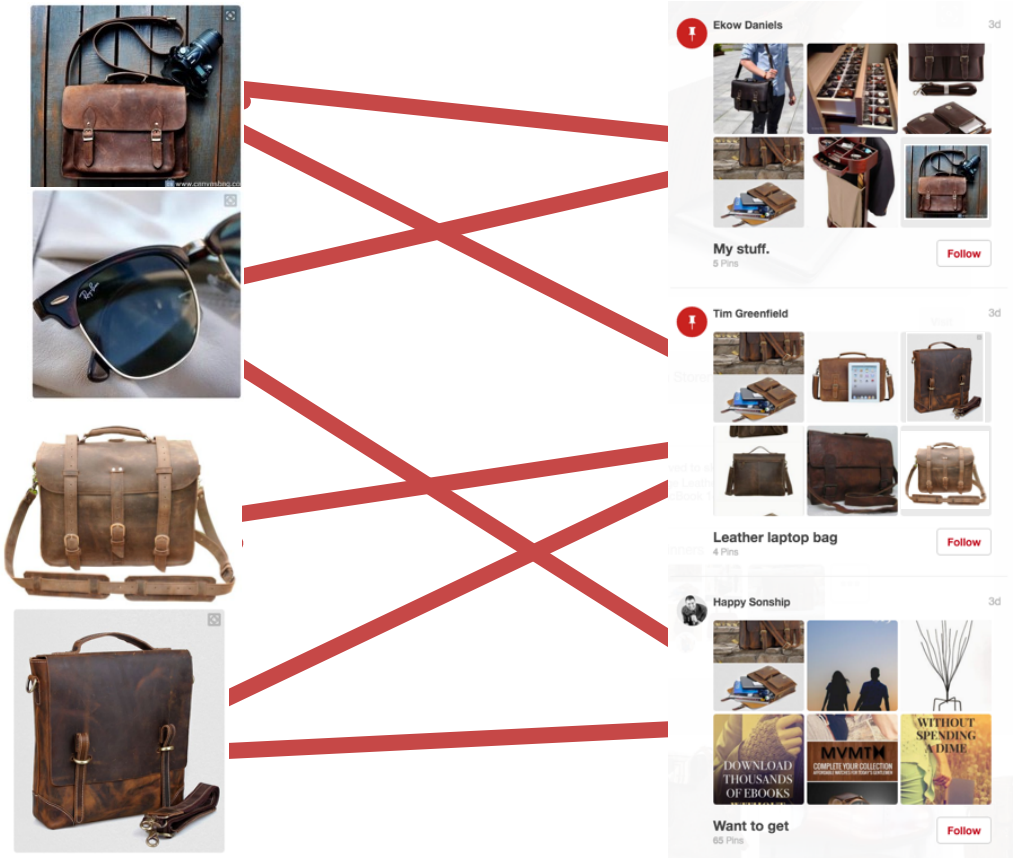
Pin



Board

Pinterest is a Giant Bipartite Graph

- Pins belong to Boards



Pins to Pins Recommendations

Input:

Pins to Pins Recommendations

Input: Recommendations:

Chocolate Strawberry Shake 249

This healthier chocolate strawberry shake is like sipping a…

One Lovely Life

Danielle Benzaia
Strawberries

Chocolate Dipped Strawberry Smoothie 5.3k

Chocolate Dipped Strawberry Smoothie. Just in time for…

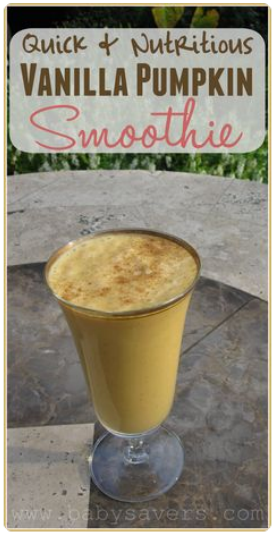
Be Whole. Be You.

Ed Todd
Drinks- Smoothies

Easy Breezy Tropical Orange Smoothie 80.1k

8 Staple Smoothies You Should Know How to Make 5.2k

8 Staple Smoothies That



The Perfect Vanilla Pumpkin Smoothie: A Quick &… 11.4k

The perfect vanilla pumpkin smoothie recipe. Quick, easy and…

BabvSavers

Marybeth @ Bab…
Best Comfort Fo…

Spinach-Pear-Celery Smoothie 60

drink this daily and watch the pounds come off without fuss…

areenreset.com

Spring Stutzman
R - Drink Up

Pins to Pins Recommendations

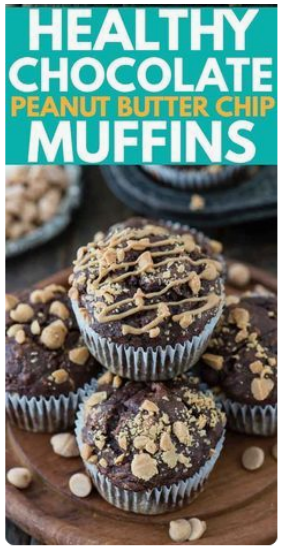
Input:

Chocolate Strawberry Shake 249

This healthier chocolate strawberry shake is like sipping a…

One Lovely Life

Danielle Benzaia
Strawberries



Healthy Chocolate Peanut Butter Chips Muffins 119

Healthy Chocolate Peanut Butter Chip Muffins made with greek…

The First Year

Katie - You Brew …
Healthy Recipes

The Ultimate Healthy Soft & Chewy Chocolate Chip Cookies 221

The ULTIMATE Healthy Chocolate Chip Cookies -- so buttery…

Amy's Healthy Baking

Robin Guertin
healthy cooking

Pins to Pins Recommendations

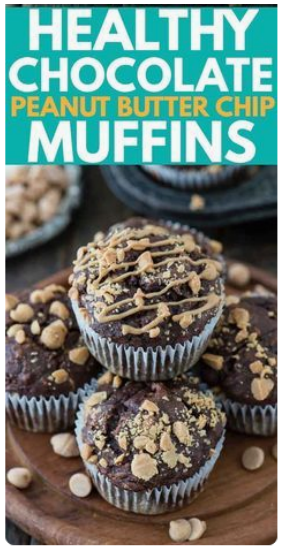
Input:

Chocolate Strawberry Shake 249

This healthier chocolate strawberry shake is like sipping a…

One Lovely Life

Danielle Benzaia
Strawberries



Healthy Chocolate Peanut Butter Chips Muffins 119

Healthy Chocolate Peanut Butter Chip Muffins made with greek…

The First Year

Katie - You Brew … Healthy Recipes

The Ultimate Healthy Soft & Chewy Chocolate Chip Cookies 221

The ULTIMATE Healthy Chocolate Chip Cookies -- so buttery…

Amy's Healthy Baking

Robin Guertin
healthy cooking

Recommendations:

30 minute Skinny Banana Chocolate Chip Muffins 2.3k

Almost fat free, healthy banana muffins with chocolate chips…

Ambitious Kitchen
Rosie Paterson
Dessert

6 Ridiculously Healthy But Delicious 3-Ingredient Treats… 204

Listotic
Rita Pittmon
Foodies

Healthy Peanut Butter Chocolate Chip Oatmeal Bars 5.4k

Healthy Peanut Butter Chocolate Chip Oatmeal Bars

Live Well Bake Often
Live Well. Bake O… Best Comfort Fo…

Freezer Fruit Treats : frozen greek yogurt covered… 96

These Frozen Greek Yogurt Covered Strawberries are…

TrendHunter.com
Gina @ Kierwort… Food, Drink & Al…

Chocolate Dipped Strawberry Smoothie 5.3k

Chocolate Dipped Strawberry Smoothie. Just in time for…

Be Whole. Be You.
Ed Todd
Drinks- Smoothies

Dark Chocolate Sea Salt Almonds 2.3k

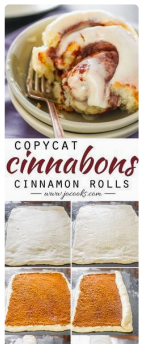
A simple, gluten free, healthy chocolate treat to feel good…

Sallys Baking Addiction
Jarena Campbell
clean eating

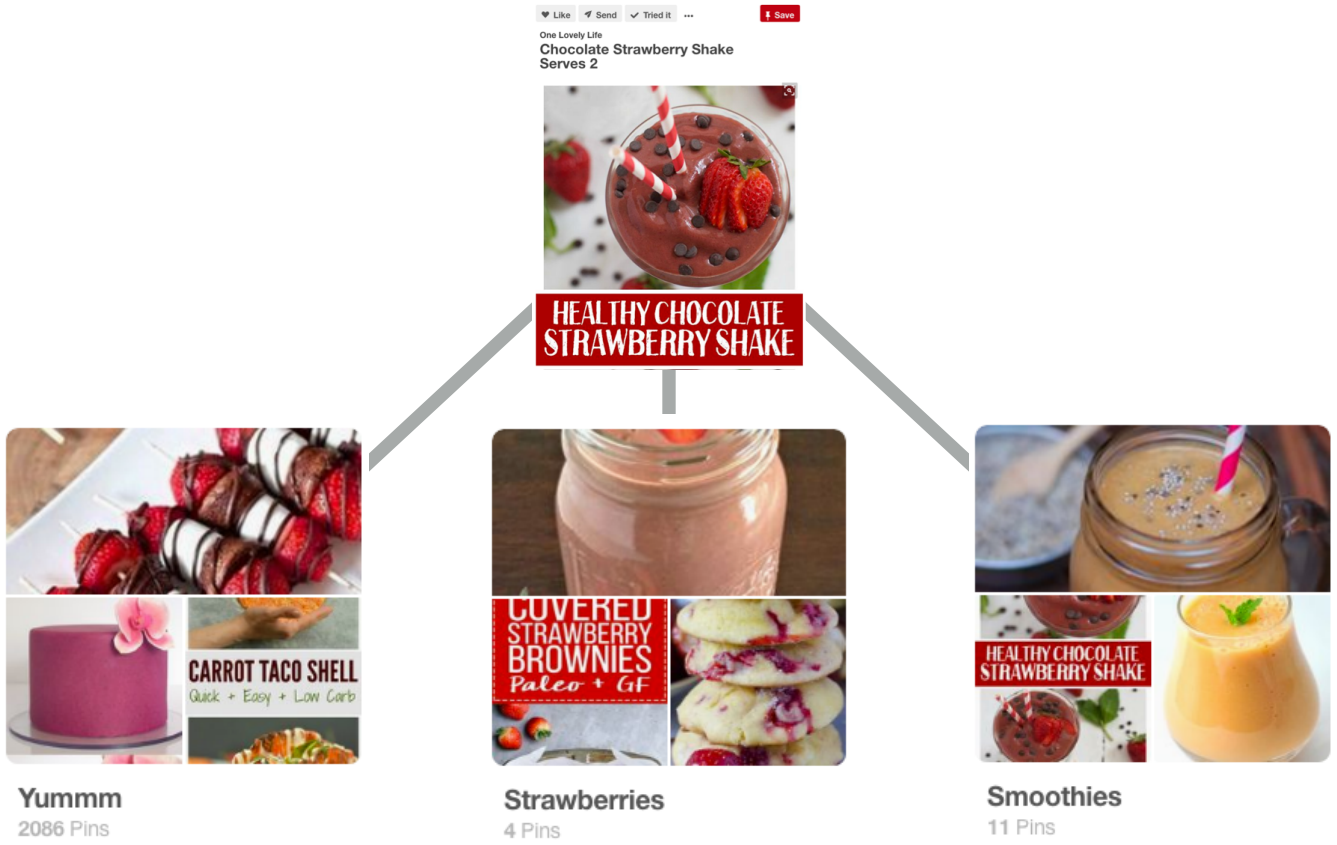
Healthy Chocolate Chip Cookie Dough Blizzard 108.3k

Healthy Chocolate Chip Cookie Dough Blizzard

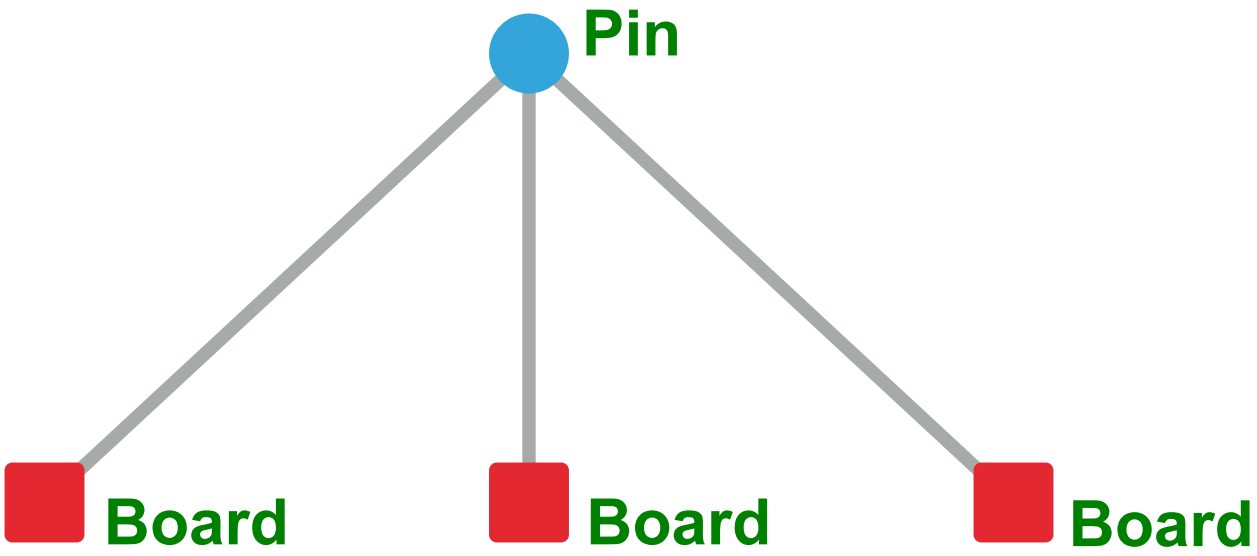
NeuroticMommy
NeuroticMommy
NeuroticMommy…



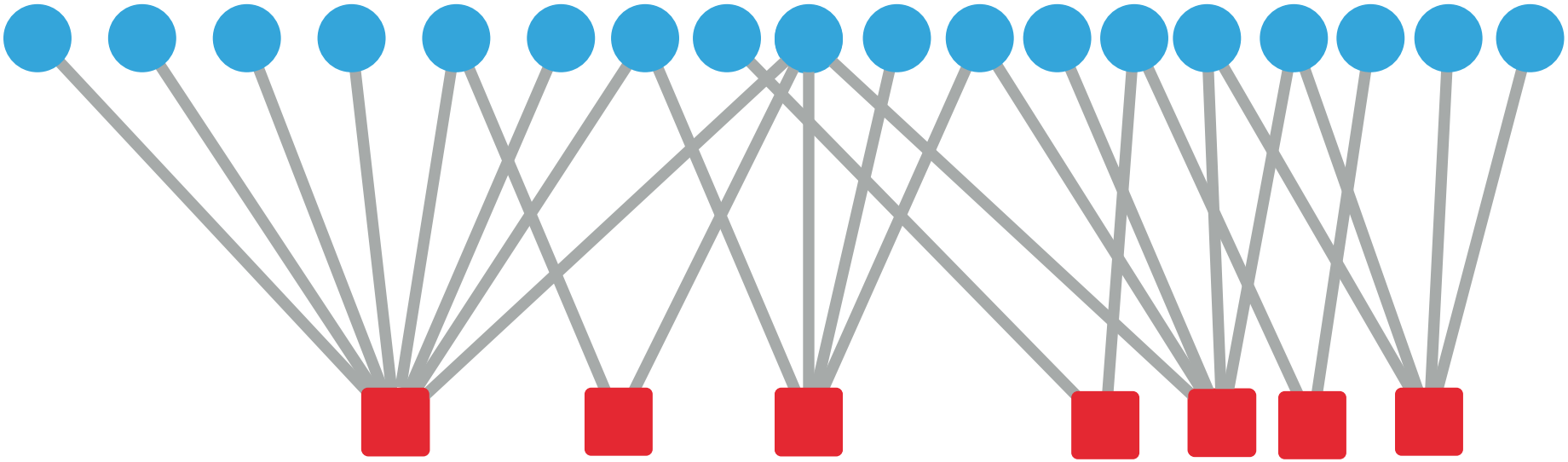
Bipartite Pin And Board Graph



Bipartite Pin And Board Graph

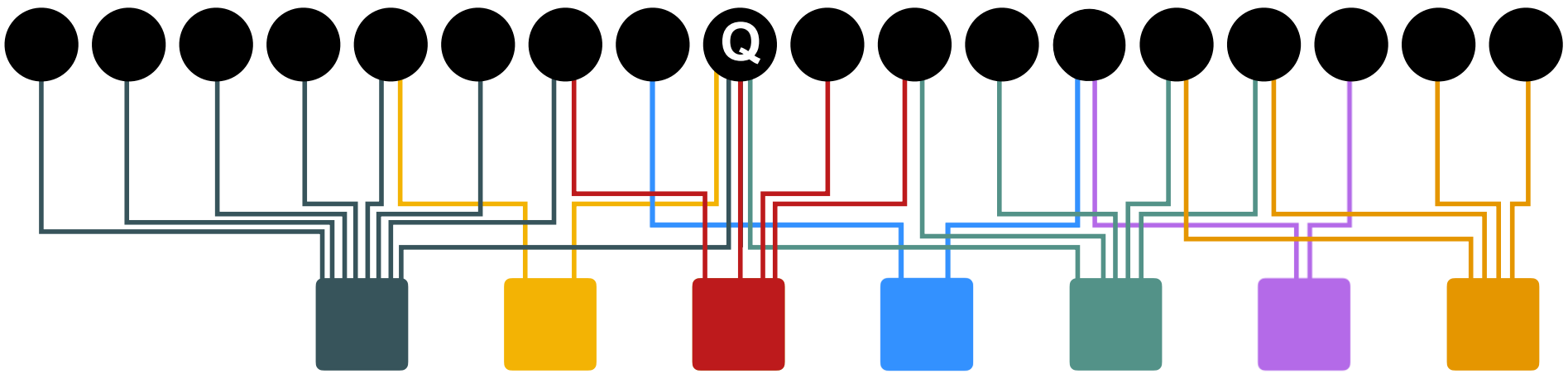


Bipartite Pin And Board Graph



Pixie Random Walks

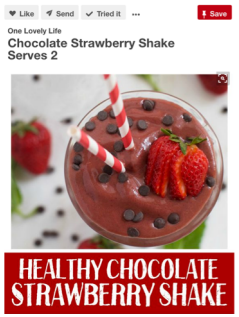
- **Idea:**
  - Every node has some importance
  - Importance gets evenly split among all edges and pushed to the neighbors
- Given a set of QUERY NODES Q, simulate a random walk:



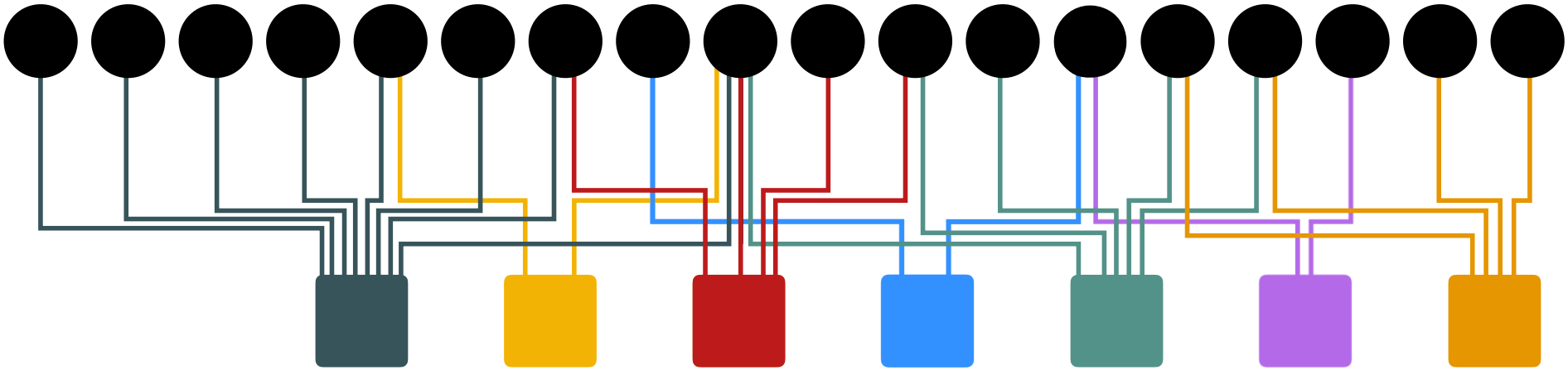
Pixie Random Walk Algorithm

- Proximity to query node(s) ***Q***:

```
ALPHA = 0.5
QUERY_NODES = {
```



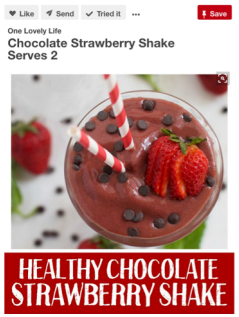

```
}
pin_node = QUERY_NODES.sample_by_weight()
for i in range(N_STEPS):
    board_node = pin_node.get_random_neighbor()
    pin_node = board_node.get_random_neighbor()
    pin_node.visit_count += 1
    if random() < ALPHA:
        pin_node = QUERY_NODES.sample_by_weight()
```



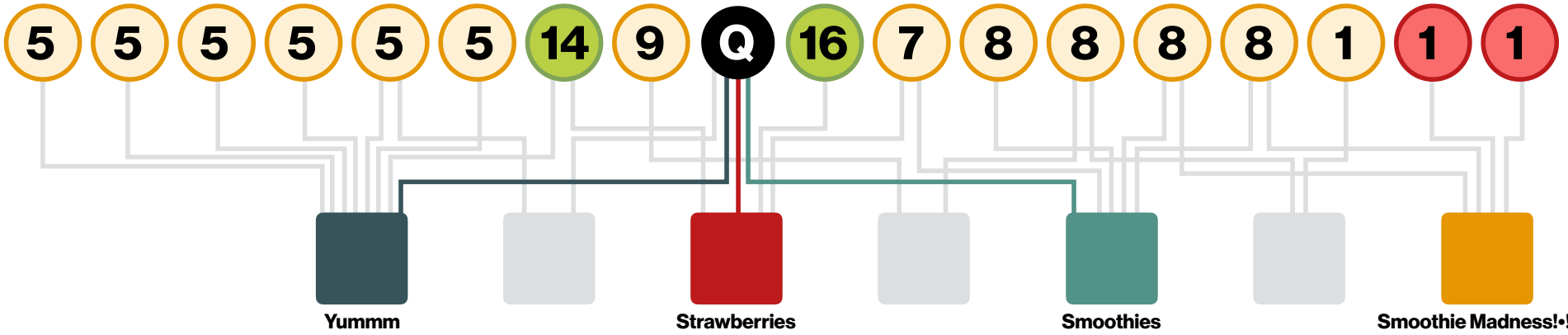
Pixie Random Walk Algorithm

- Proximity to query node(s) ***Q***:

```
ALPHA = 0.5
QUERY_NODES = {
```




```
}
pin_node = QUERY_NODES.sample_by_weight()
for i in range(N_STEPS):
    board_node = pin_node.get_random_neighbor()
    pin_node = board_node.get_random_neighbor()
    pin_node.visit_count += 1
    if random() < ALPHA:
        pin_node = QUERY_NODES.sample_by_weight()
```



Pixie Recommendations

- **Pixie:**
  - **Outputs top 1k pins with highest visit count**

Extensions:

- **Weighted edges:**
  - The walk prefers to traverse certain edges:
    - Edges to pins in your local language
- **Early stopping:**
  - Don’t need to walk a fixed big number of steps
  - Walk until 1k-th pin has at least 20 visits

Graph Cleaning/Pruning

- **Pinterest graph has 200B edges**
- We don't need all of them!
  - Super popular pins are pinned to millions of boards
    - **Not useful:** When the random walk hits the pin, the signal just disperses.
- **What we did:** Keep only good boards for pins
  - Compute the similarity between pin's topic vector and each of its boards. Only keep edges with high similarity.

Data Type	Number	Size	Memory
Pin Nodes	3 Billion	8 Bytes	24 GiB
Board Nodes	2 Billion	8 Bytes	16 GiB
Undirected Edges	20 Billion	8 Bytes	160 GiB
			208 GiB

Benefits of Pixie

- **Benefits:**
  - Blazingly fast: Given Q, we can output top 1k in 50ms (after doing 100k steps of the random walk)
  - Single machine can run 1500 walks in parallel! (1500 recommendation requests per second)
  - Can fit entire graph in RAM (17B edges, 3B nodes)
  - Can scale it by just adding more machines
- **Today about 60% of all the pins you see at Pinterest are recommended by random walks**

To learn more read: https://cs.stanford.edu/people/jure/pubs/pixie-www18.pdf

PageRank: Summary

- **"Normal" PageRank:**
  - Teleports uniformly at random to any node
  - All nodes have the same probability of surfer landing there: **S =** [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]
- **Topic-Specific PageRank** also known as **Personalized PageRank:**
  - Teleports to a topic specific set of pages
  - Nodes can have different probabilities of surfer landing there: **S =** [0.1, 0, 0, 0.2, 0, 0, 0.5, 0, 0, 0.2]
- **Random Walk with Restarts:**
  - Topic-Specific PageRank where teleport is always to the same node. S=[0, 0, 0, 0, **1**, 0, 0, 0, 0, 0, 0]