Lecture 11

Graph mining

Golnaz Taheri, PhD
Senior Lecturer, Stockholm University

Stockholms universitet

What is a graph?

- A data structure that consists of a set of nodes (*vertices*) and a set of edges that relate the nodes to each other
- The set of edges describes relationships among the vertices

- A graph G is defined as follows:

$$G = (V, E)$$

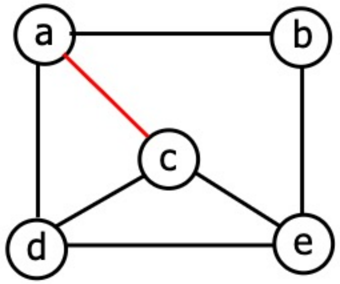
$V(G)$: a finite, nonempty set of vertices

$E(G)$: a set of edges (pairs of vertices)

Stockholms universitet

Graph terminology

- Adjacent nodes: two nodes are adjacent if they are connected by an edge



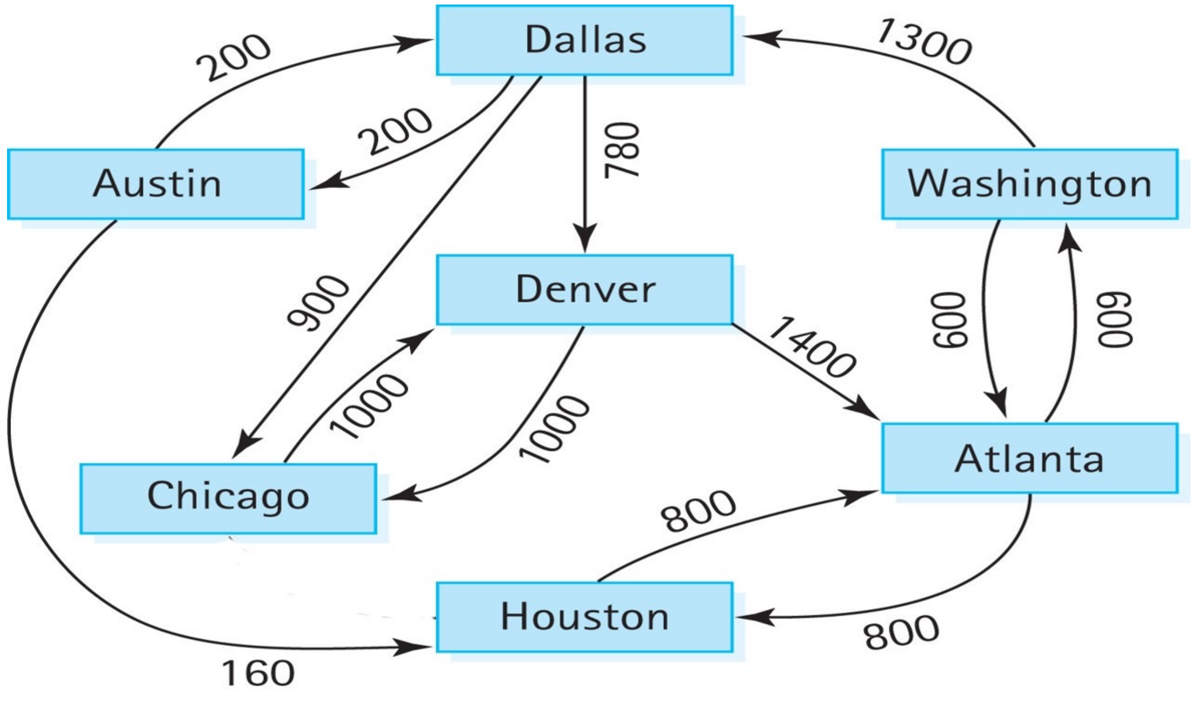
Vertex a is adjacent to c and
vertex c is adjacent to a

- Path: a sequence of vertices that connect two nodes in a graph

- Complete graph: a graph in which every vertex is directly connected to every other vertex

Graph terminology

- Weighted graph: a graph in which each edge carries a value



Stockholms universitet

Graph implementation

- Array-based implementation
  - 1D array is used to represent the vertices
  - 2D array (adjacency matrix) is used to represent the edges

numVertices: 7

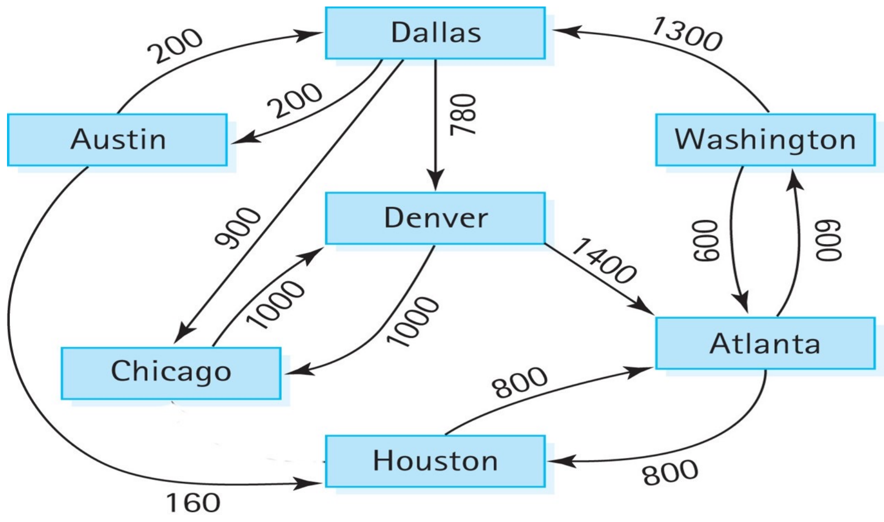
vertices

[0]	"Atlanta "
[1]	"Austin "
[2]	"Chicago "
[3]	"Dallas "
[4]	"Denver "
[5]	"Houston "
[6]	"Washington"
[7]	
[8]	
[9]	

edges

	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
[0]	0	0	0	0	0	800	600	•	•	•
[1]	0	0	0	200	0	160	0	•	•	•
[2]	0	0	0	0	1000	0	0	•	•	•
[3]	0	200	900	0	780	0	0	•	•	•
[4]	1400	0	1000	0	0	0	0	•	•	•
[5]	800	0	0	0	0	0	0	•	•	•
[6]	600	0	0	1300	0	0	0	•	•	•
[7]	•	•	•	•	•	•	•	•	•	•
[8]	•	•	•	•	•	•	•	•	•	•
[9]	•	•	•	•	•	•	•	•	•	•

(Array positions marked "•" are undefined)



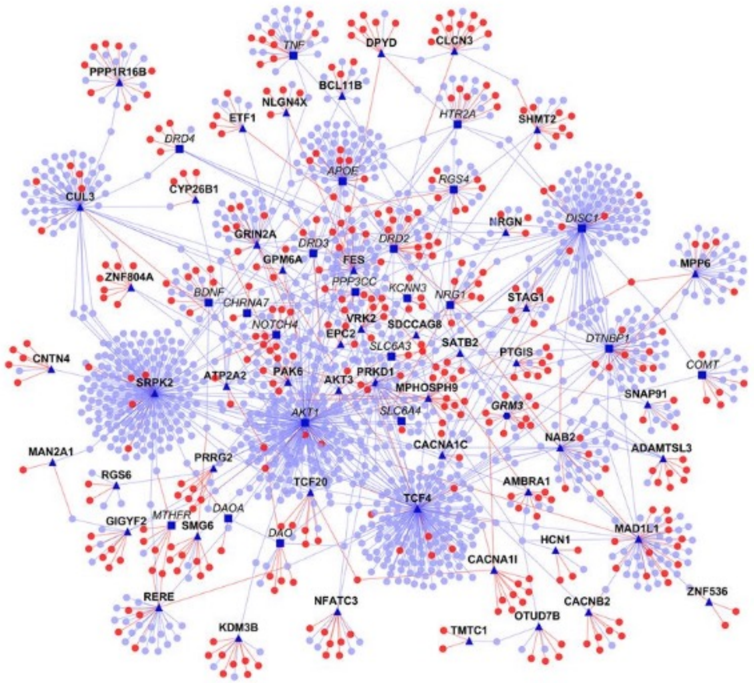
Stockholms universitet

What are we looking for

- Rank nodes for a particular query
  - Top k matches for “graph mining” from Google
  - Top k book recommendations for “Dan Brown” from Amazon
  - Top k websites matching “Pagerank algorithm”
  - Top k high frequency mutations in cancer.

Our model: a graph

- The underlying data is naturally a graph
  - WWW: Websites are linked with each other
  - Gene-gene interaction network.
  - Gene regulatory Network
  - Authors linked by co-authorship
  - Bipartite graph of customers and products
  - Friendship networks: who knows whom



How do search engines decide how to rank your query results?

- How does Google rank the query results?

- How would you do it?

Naïve ranking of query results

- Given query $\boldsymbol{q}$
- Rank the web pages $\boldsymbol{p}$ in your database based on some similarity measure $\boldsymbol{sim(p,q)}$

Stockholms universitet

Why Link Analysis?

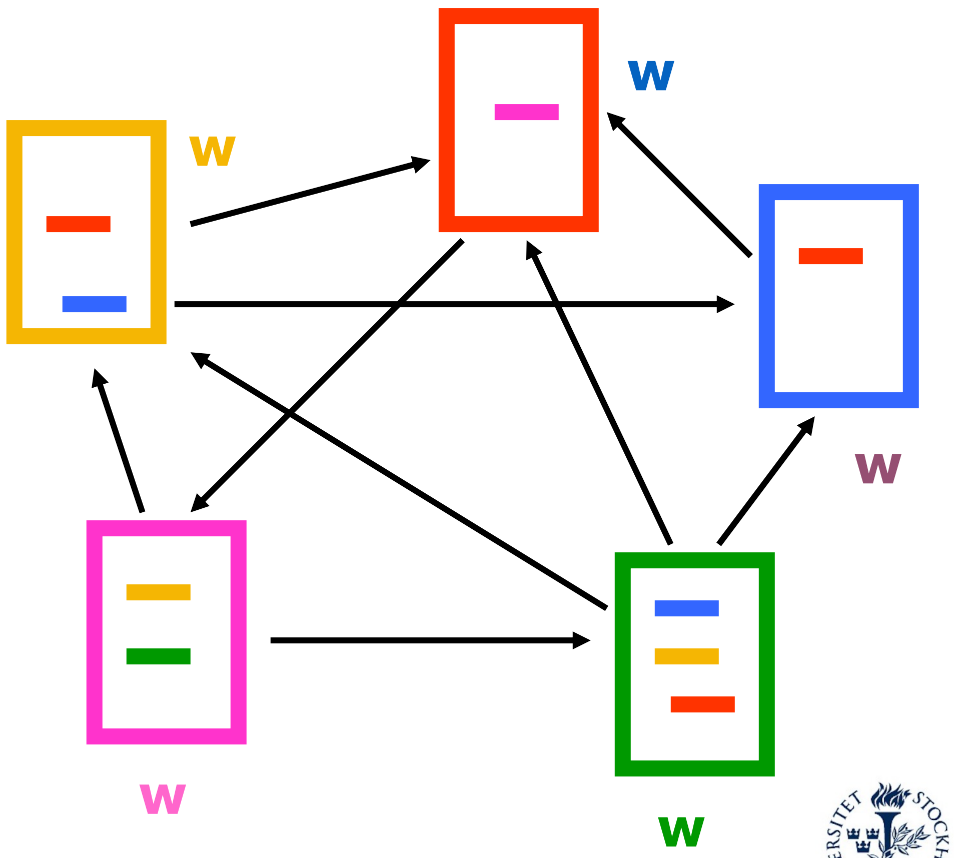
- First generation search engines
    - view documents as flat text files
    - could not cope with size or user needs
- Second generation search engines
    - ranking becomes critical
    - use of Web specific data: Link Analysis
    - shift from relevance to authoritativeness

Link Analysis: Intuition

- A link from page p to page q denotes endorsement
  - page p considers page q an authority on a subject
  - assign an authority value to every page
  - “mine” the web graph of recommendations

Link Analysis Ranking (LAR) Algorithms

- Start with a collection of web pages
- Extract the underlying hyperlink graph
- Run the LAR algorithm on the graph
- Output: an authority weight for each node



Stockholms universitet

Two Types of Algorithms

- **Query dependent**: rank a small subset of pages related to a specific query
  - HITS (Kleinberg 98) was proposed as query dependent

- **Query independent**: rank the whole Web
  - PageRank (Brin and Page 98) was proposed as query independent

Query-dependent LAR

- Given a query q, find a subset of web pages S that are related to q

- Rank the pages in S based on some ranking criterion

Properties of a good base set S

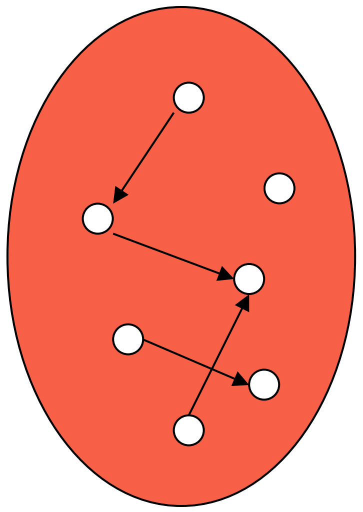
- S is relatively small
- S is rich in relevant pages
- S contains most (or many) of the strongest authorities

How to construct a good seed set $\boldsymbol{S}$

- Given a query $\boldsymbol{q}$:
  - collect the $\boldsymbol{t}$ highest-ranked pages for $\boldsymbol{q}$ from a text-based search engine to form set $\boldsymbol{\Gamma}$
  - $\boldsymbol{S} = \boldsymbol{\Gamma}$
  - add to $\boldsymbol{S}$ all the pages pointing to $\boldsymbol{\Gamma}$
  - add to $\boldsymbol{S}$ all the pages that pages from $\boldsymbol{\Gamma}$ point to

Query-dependent input

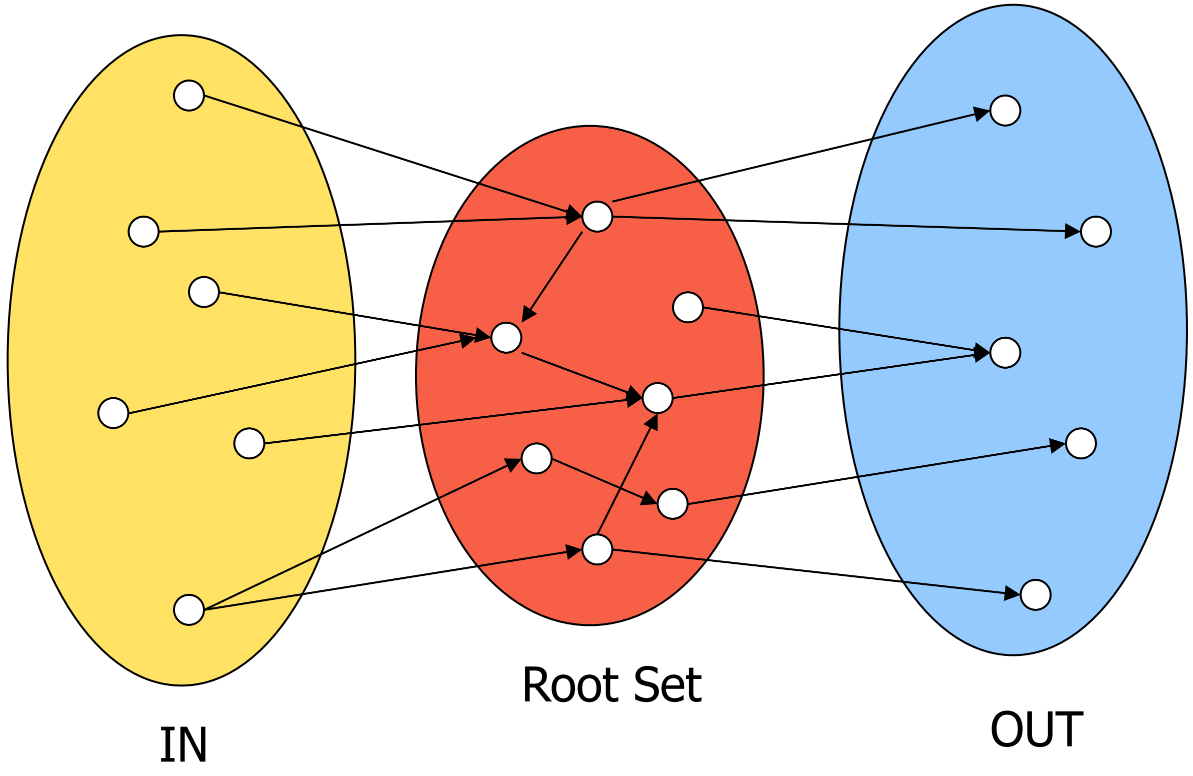
Root set: includes pages relevant to the query q



Stockholms universitet

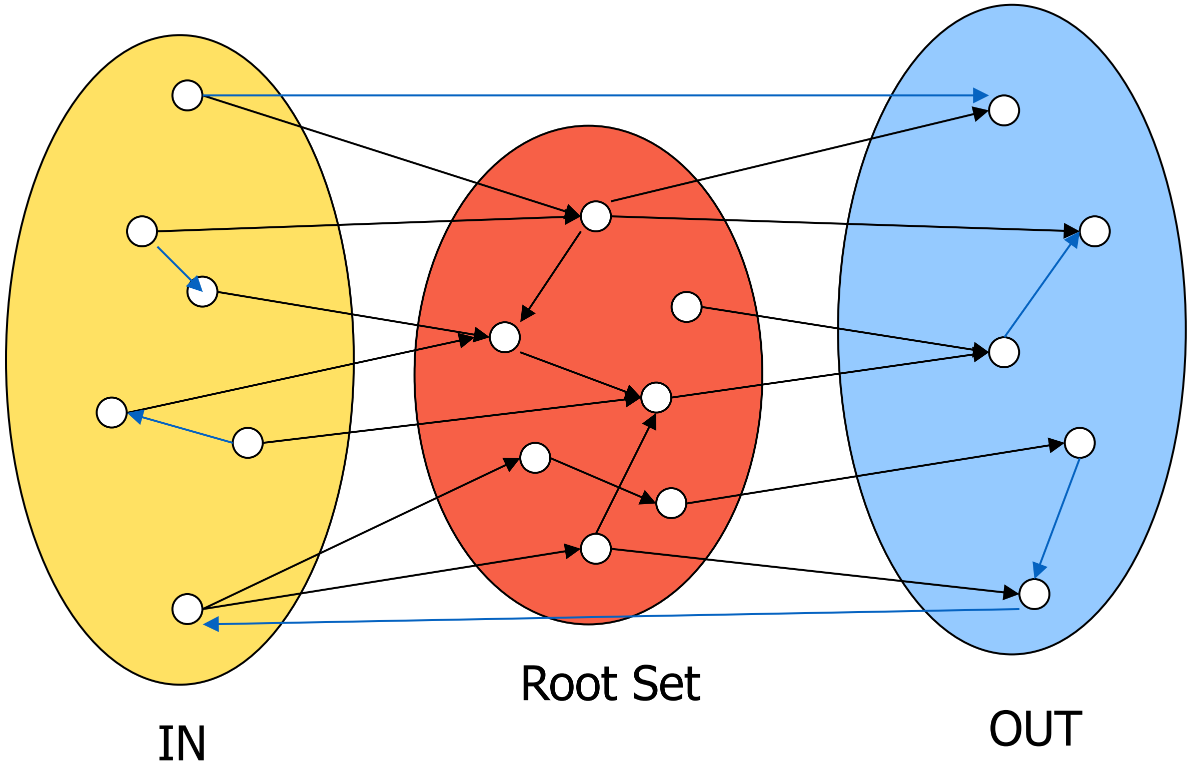
Query-dependent input

Root set: includes pages relevant to the query ***q***

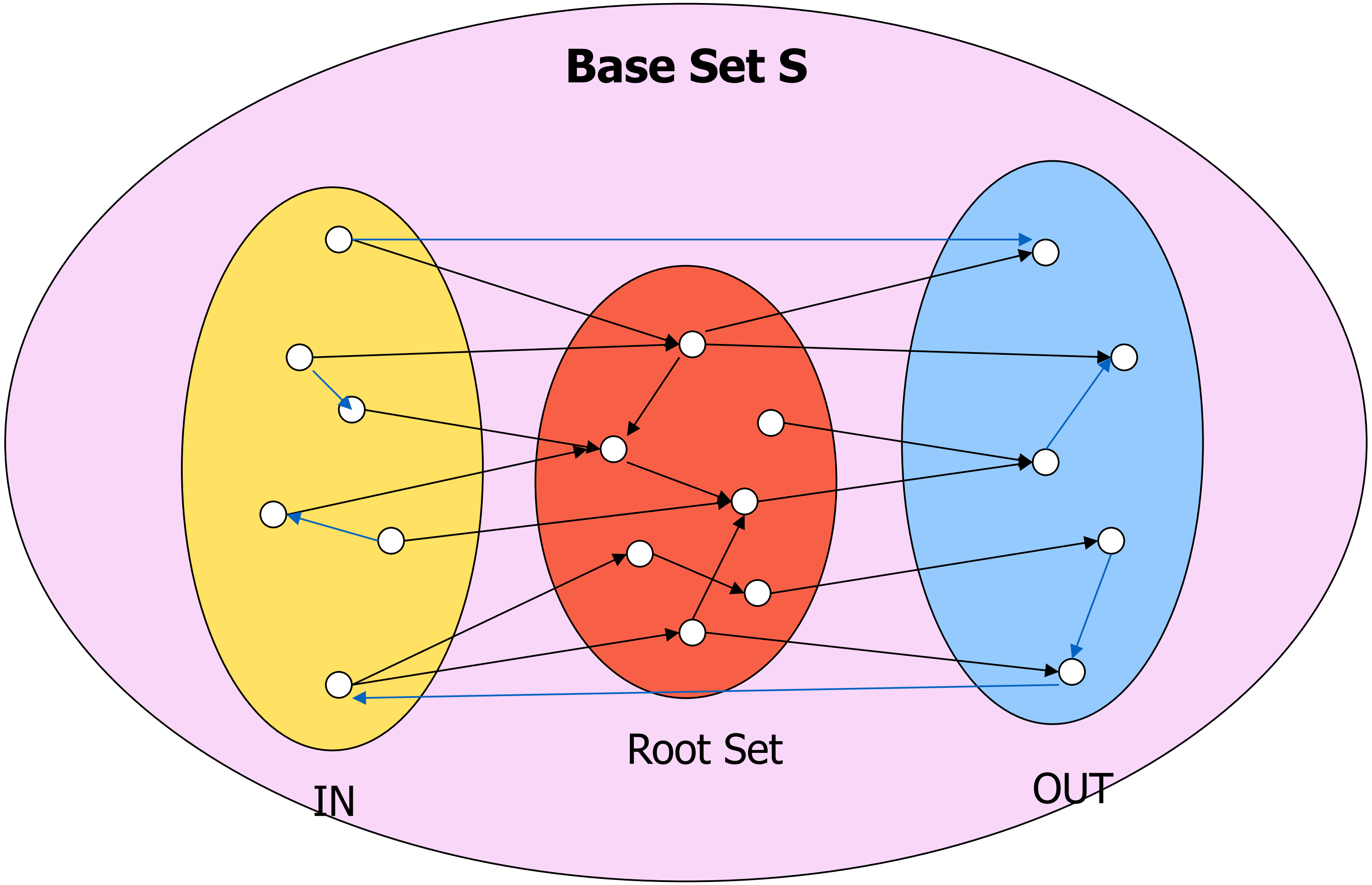


Query-dependent input

Root set: includes pages relevant to the query q



Query-dependent input



Link Filtering

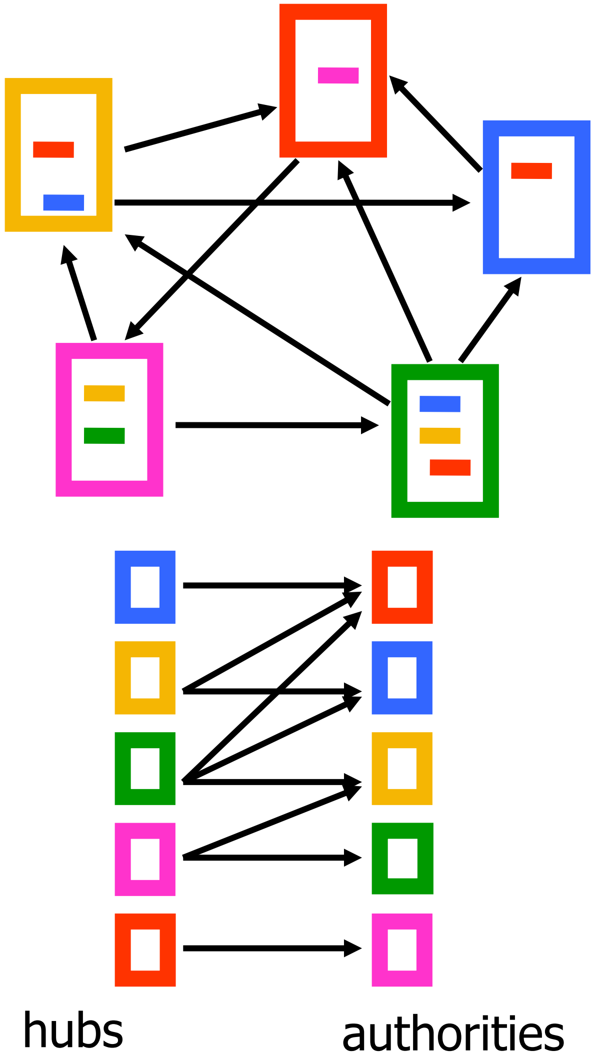
- Navigational links: serve the purpose of moving within a site (or to related sites)
    - `www.espn.com` → `www.espn.com/nba`
    - `www.yahoo.com` → `www.yahoo.it`
- Filter out navigational links
    - same domain name
    - same IP address

Hubs and Authorities [K98]

- Based on the relationship between
  - **authorities** (pages) for a topic and
  - **hubs** (pages) that link to many related authorities
- *We observe that:*
  - *a certain natural type of equilibrium exists between* hubs and authorities in the graph defined by the link structure
- We exploit this equilibrium to develop an algorithm that identifies both types of pages simultaneously

Hubs and Authorities [K98]

- Authority is not necessarily transferred directly between authorities
- Pages have double identity
  - hub identity
  - authority identity
- Good hubs point to good authorities
- Good authorities are pointed by good hubs



Stockholms universitet

HITS Algorithm

- Initialize all weights to 1
- Repeat until convergence
  - O operation : hubs collect the weight of the authorities

$$h_i = \sum_{j:i \to j} a_j$$

  - I operation: authorities collect the weight of the hubs

$$a_i = \sum_{j:j \to i} h_j$$

  - Normalize weights under some norm

Stockholms universitet

Strengths and weaknesses of HITS

- Strength: its ability to rank pages according to the query topic, which may be able to provide more relevant authority and hub pages

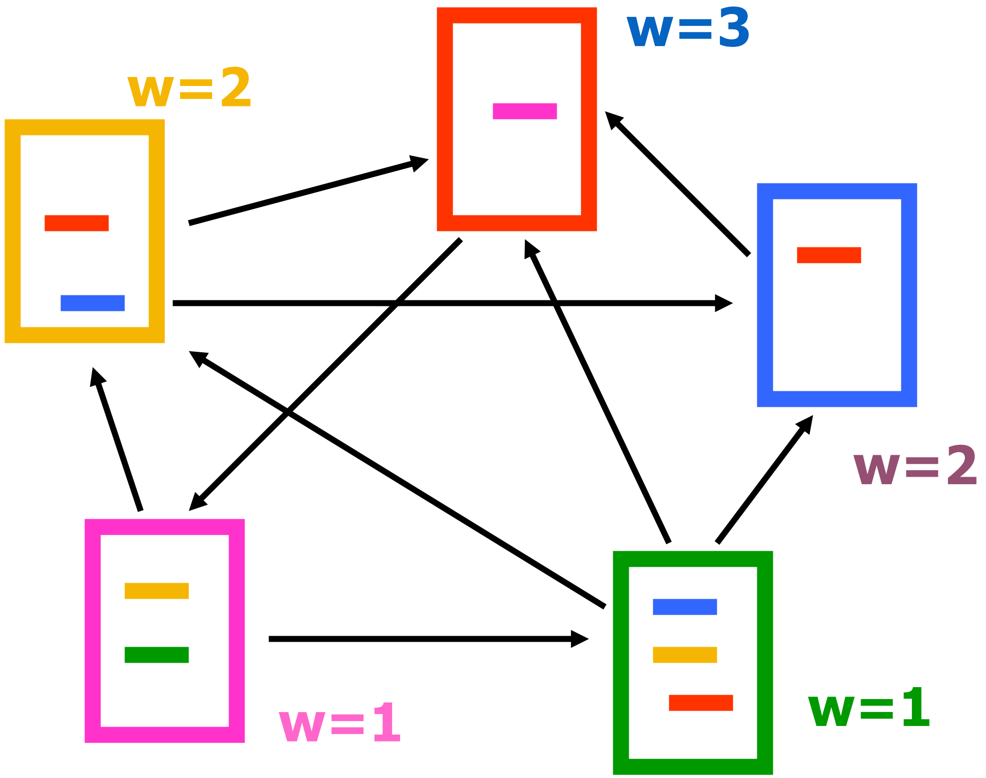
- Weaknesses:
  - One can easily cheat: adding out-links in one's own page
  - Topic drift: many pages in the expanded set may not be on topic
  - Inefficiency at query time: collecting the root set, expanding it, and performing eigenvector computation are all expensive operations

Query-independent LAR

- Have an apriori ordering of the web pages
- $\boldsymbol{Q}$: Set of pages that contain the keywords in the query $\boldsymbol{q}$
- Present the pages in $\boldsymbol{Q}$ ordered according to order $\boldsymbol{\pi}$

InDegree algorithm

- Rank pages according to in-degree
  - $w_i = |B(i)|$ where $B(i)$ is the number of incoming links to node i



1. Red Page
2. Yellow Page
3. Blue Page
4. Purple Page
5. Green Page

Stockholms universitet

Query-independent LAR

- In-degree is a local measure
- All links to a page are considered equal, regardless of where they come from
- Two pages with the same in-degree are considered equally important, even if one is cited by more prestigious sources than the other

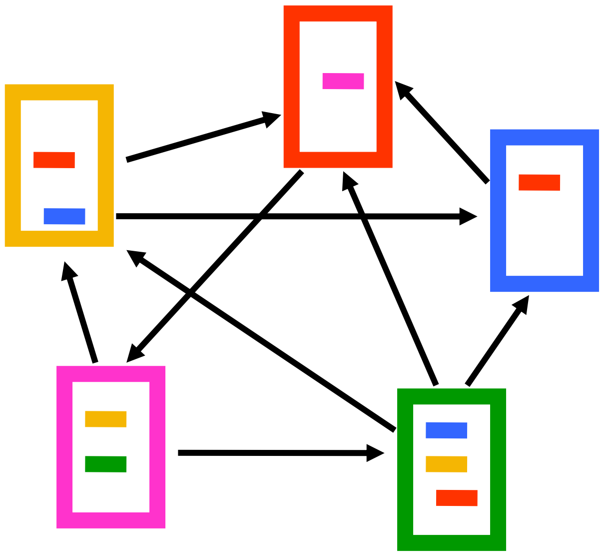
PageRank algorithm [BP98]

- Good authorities should be pointed by good authorities
- Random walk on the web graph
  - pick a page at random
  - with probability $1 - \alpha$ jump to a random page
  - with probability α follow a random outgoing link

- PageRank of page p:

$$PR(p) = \alpha \sum_{q \to p} \frac{PR(q)}{|F(q)|} + (1-\alpha)\frac{1}{n}$$

- $F(q)$: the set of outgoing links of page q



1. **Red Page**
2. **Purple Page**
3. **Yellow Page**
4. **Blue Page**
5. **Green Page**

Stockholms universitet

Convergence

- You can think of Pagerank as a random walk:
  - From each node you choose to move to a neighbouring node with some probability *p*
  - After you “walk” for a long time, the Pagerank of each node is the proportion of time you visited that node
- Pagerank will converge if
  - The graph has no cycles (loops where you can get stuck)
  - If you can reach any node from any node
- Both properties are enforced by the Pagerank formula

Pagerank vs InDegree

- Pagerank is better suited for directed graphs:
  - Directed graph: a graph where edges have a direction

- If the graph is not directed (e.g., it is bidirectional or direction is of no interest), then Pagerank is proportional to InDegree

- In other words:
  - In graphs where the edge direction is not indicated, PageRank and InDegree with produce the same ranking for the nodes

Huge Matrix Computation

- Computing PageRank:
  - can be done via matrix multiplication
  - matrix may have **3 billion rows and columns**
- Good news:
  - matrix is **sparse**
  - average number of out-links is small
- Setting $\alpha = 0.85$ or more requires at most 100 iterations to convergence
- Researchers still trying to speed-up the computation

Stockholms universitet

Personalized PageRank

- **Main idea:**
- Similar to Pagerank
- The only difference is that we use a non-uniform teleportation distribution, i.e.,
  - at any time step teleport to a set of webpages
  - assign weights to the pages depending on their relevance/topics you are interested in

Research on PageRank

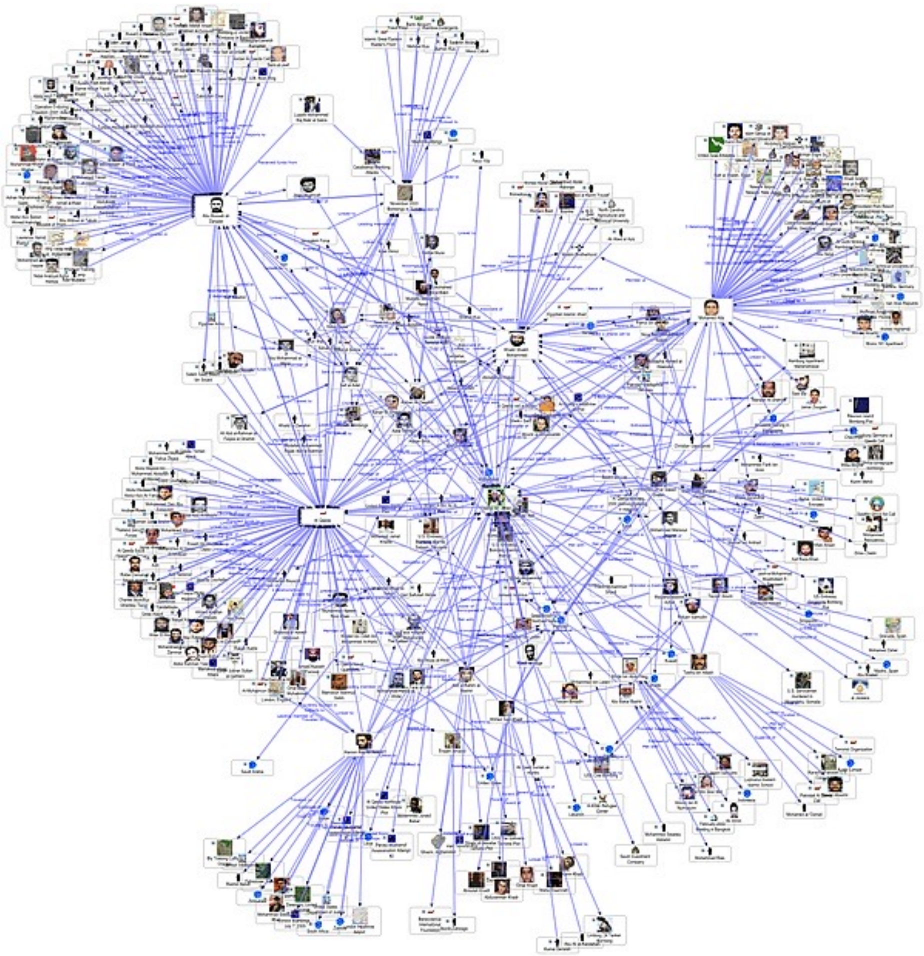
- Topic-sensitive PageRank
  - compute many PageRank vectors, one for each topic
  - estimate relevance of query with each topic
  - produce final PageRank as a weighted combination
- Updating PageRank [Chien et al 2002]
- Fast computation of PageRank
  - numerical analysis tricks
  - node aggregation techniques

New Generation of Search Engines

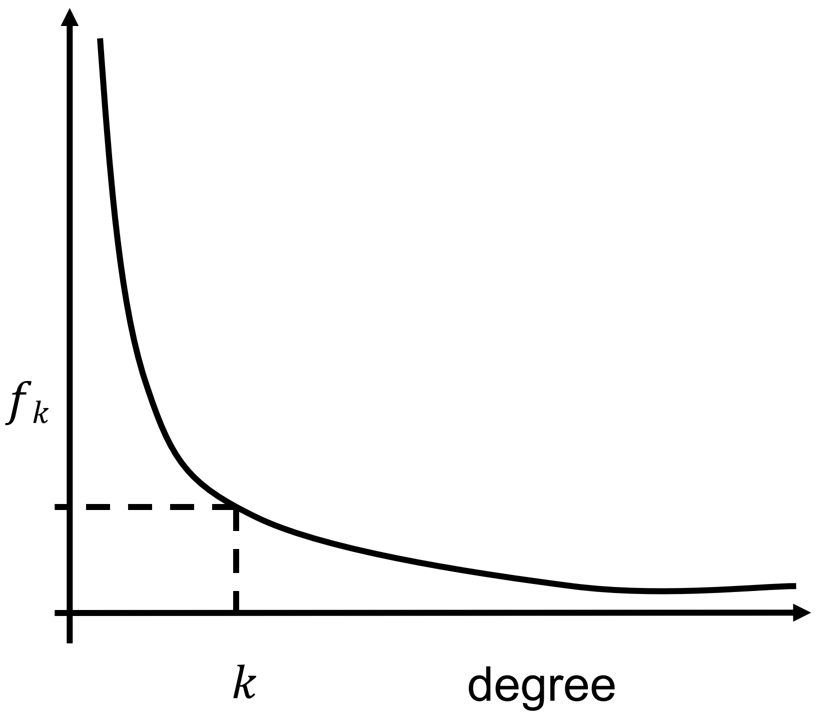
- 3rd generation of search engines:
  - machines collect our searches and analyze them
- For example:
  - we will tell the search engine
    - where and what we click
    - how long we stay on a page
    - how many pages we view
  - these actions are understood and learned by the search engines
  - they are used to improve the result for the next user who wants to find answers to a similar term

What is a social network?

- Facebook, LinkedIn, etc
- The network of your friends and acquaintances
- Social network is a graph $\boldsymbol{G = (V, E)}$
  - V: set of users
  - E: connections among users



Measuring networks: Degree distributions



f_k = fraction of nodes with degree >= k
= probability of a randomly selected node to have degree >= k

- Problem: find the probability distribution that best fits the observed data

Power-law distributions

- The degree distributions of most real-life networks follow a power law

$$p(k) = Ck^{-\alpha}$$

- *In other words, a quantity varies as a power of another…*

- Right-skewed/Heavy-tail distribution
  - there is a non-negligible fraction of nodes that has very high degree (hubs)
  - scale-free: no characteristic scale, average is not informative

- The probability that any node is connected to k other nodes is proportional to $1/ka$

Power-law distributions

- The degree distributions of most real-life networks follow a power law

$$p(k) = 1/k^2$$

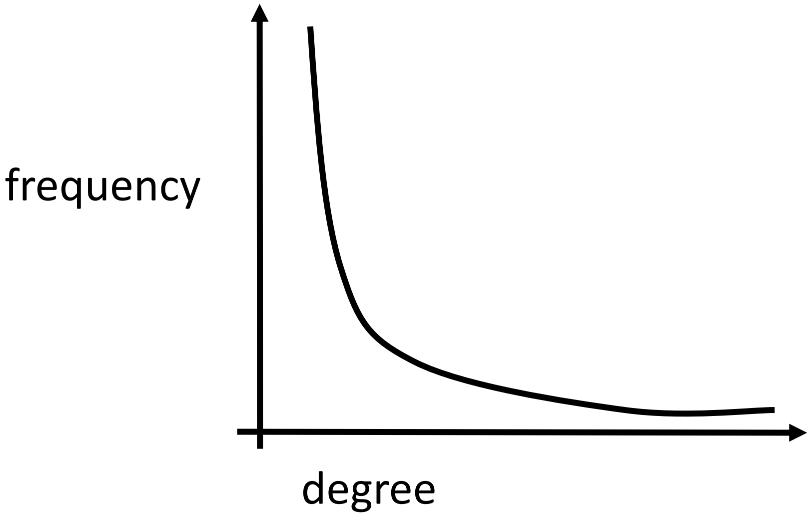
$$P(1) = 1$$
$$P(2) = 1/4$$
$$P(3) = 1/8$$

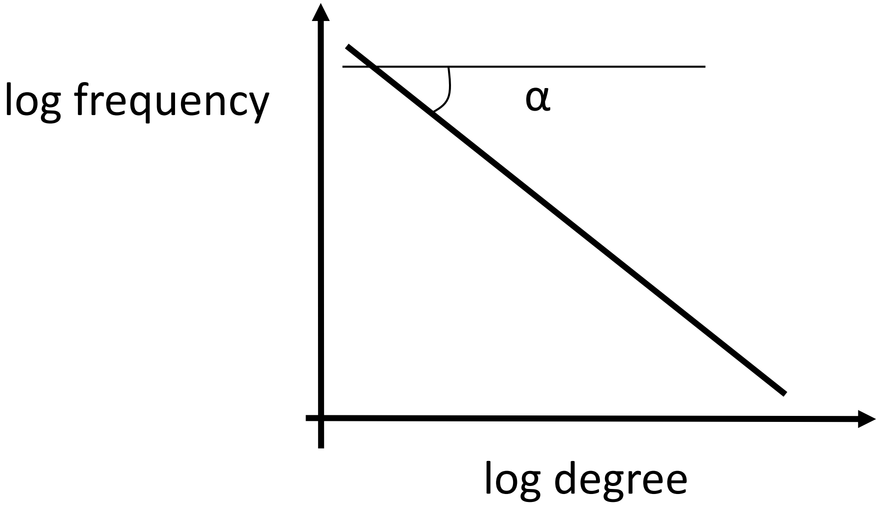
$$P(4) = 1/16$$
$$P(5) = 1/25$$
$$\dots$$
$$P(8) = 1/64$$

Power-law signature

- Power-law distribution gives a line in the log-log plot

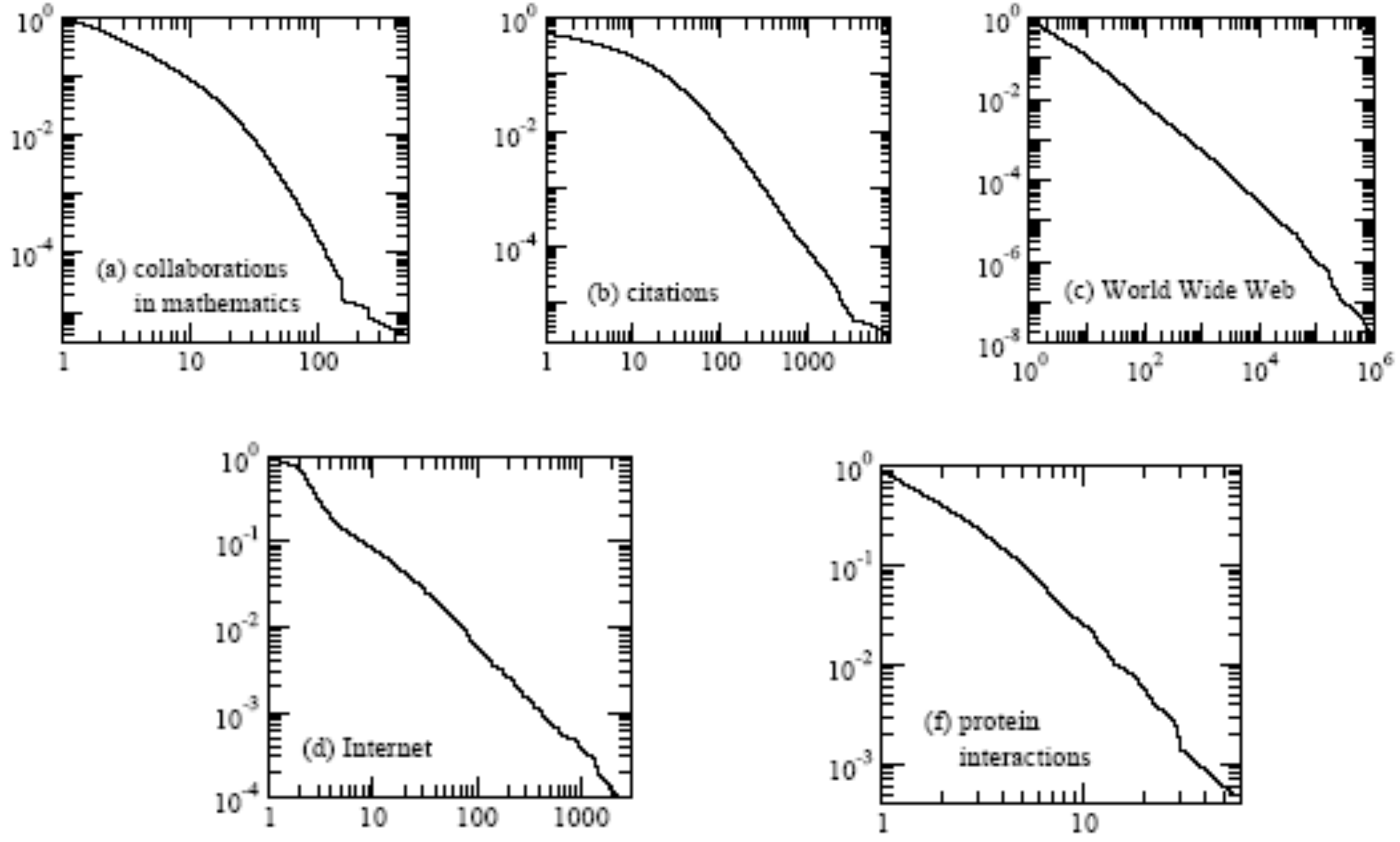
$$\log p(k) = -\alpha \, log k + log C$$





- α: power-law exponent (typically 2 ≤ α ≤ 3)

Examples



Taken from [Newman 2003]

Stockholms universitet

The small-world experiment

- Milgram 1967
- Picked 296 people at random from Omaha, Nebraska, Wichita, and Kansas
- Asked them to get a letter to a stockbroker in Boston
- Rule: they could bypass the letter through friends they knew on a first-name basis
- How many steps does it take?

Stockholms universitet

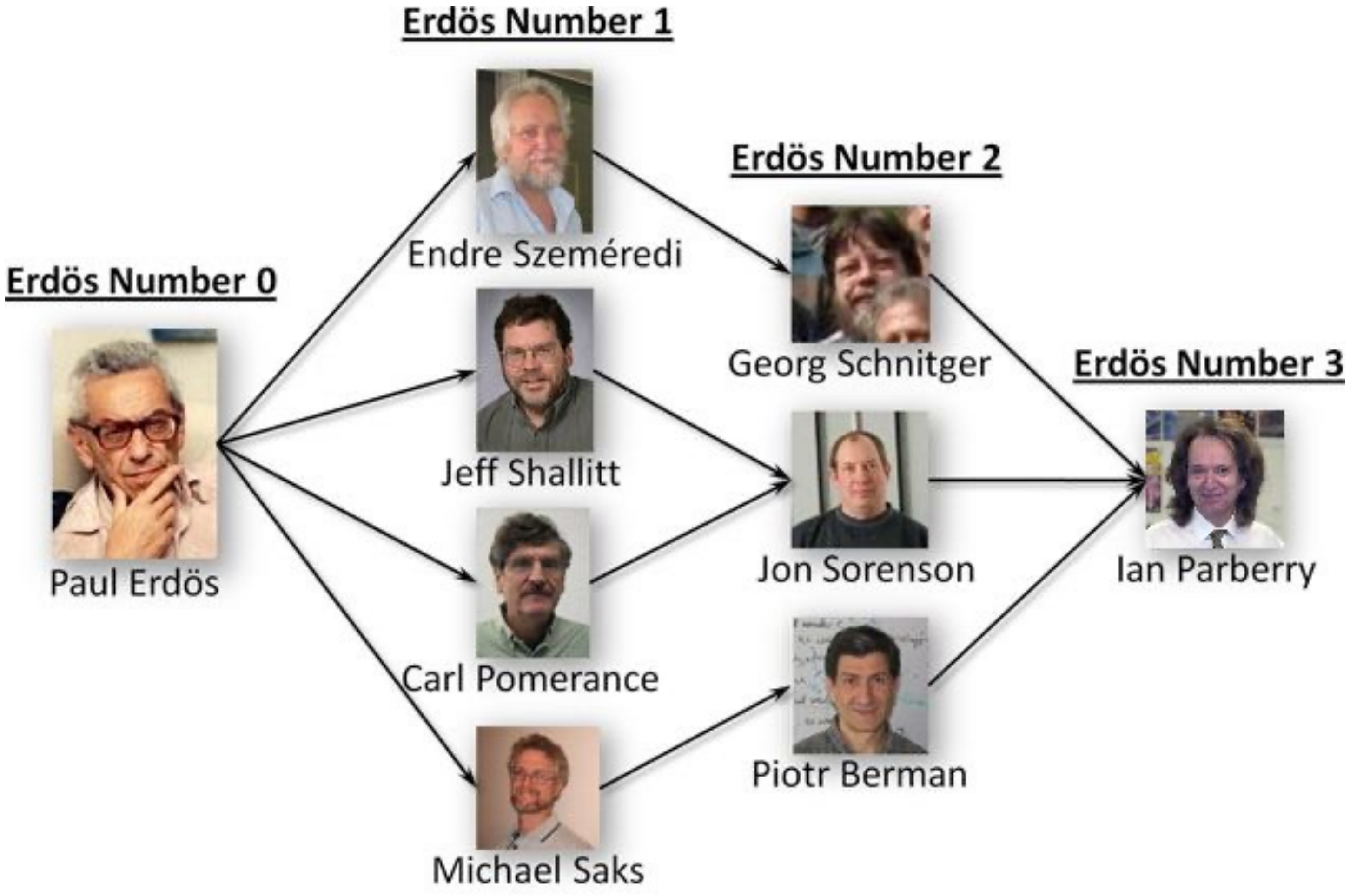
The small-world experiment

- 64 chains completed
  - 6.2 average chain length (thus "six degrees of separation")
- Critique
  - Several times people refused to forward the package
  - People had no knowledge of the topology of the network, hence the package may not follow the shortest path

Six Degrees of Kevin Bacon

- Bacon number:
  - Create a network of Hollywood actors
  - Connect two actors if they co-appeared in some movie
  - Bacon number: number of steps to Kevin Bacon
- As of Sep 2019, the highest (finite) Bacon number reported is 7
- Only approx 12% of all actors cannot be linked to Bacon

Erdős number



https://ianparberry.com/numbers.html

Measuring the small world phenomenon

- d_{ij} = shortest path between i and j

- Diameter:

$$d = \max_{i,j} d_{ij}$$

- Characteristic path length:

$$\ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- Harmonic mean:

$$\ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1}$$

What is a network model?

- Informally, a network model is a process (randomized or deterministic) for generating a graph
- Models of static graphs
  - input: a set of parameters Π and the size of the graph n
  - output: a graph $G(\Pi, n)$
- Models of evolving graphs
  - input: a set of parameters Π and an initial graph G_0
  - output: a graph G_t for each time t

Stockholms universitet

Families of random graphs

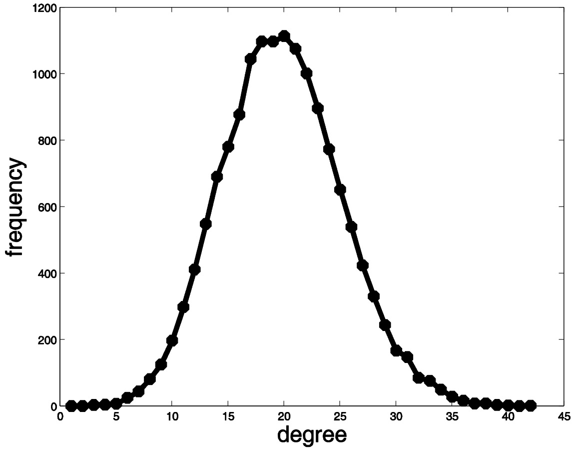
- A deterministic model D defines a single graph for each value of n (or t)

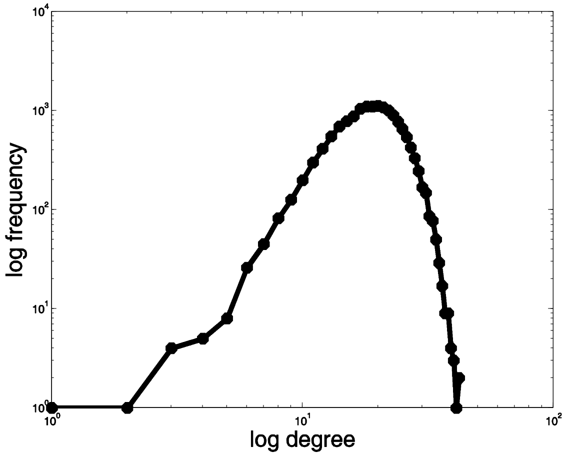
- A randomized model R defines a probability space $\langle G_n, P\rangle$ where G_n is the set of all graphs of size n, and P a probability distribution over the set G_n (similarly for t)
    - we call this a family of random graphs R, or a random graph R

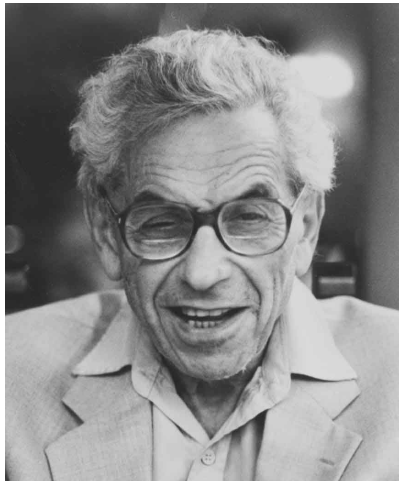
Stockholms universitet

Erdös-Renyi Random Graphs

- The $G_{n,p}$ model
  - input: number of vertices n, and parameter p, $0 \leq p \leq 1$
  - process: for each pair (i,j), generate the edge (i,j) independently with probability p
- The $G_{n,m}$ random model:
  - process: select m edges uniformly at random







Random graphs and real life

- A beautiful and elegant theory studied exhaustively
- Random graphs had been used as idealized network models
- Unfortunately, they don’t capture reality…

Barabasi-Albert model

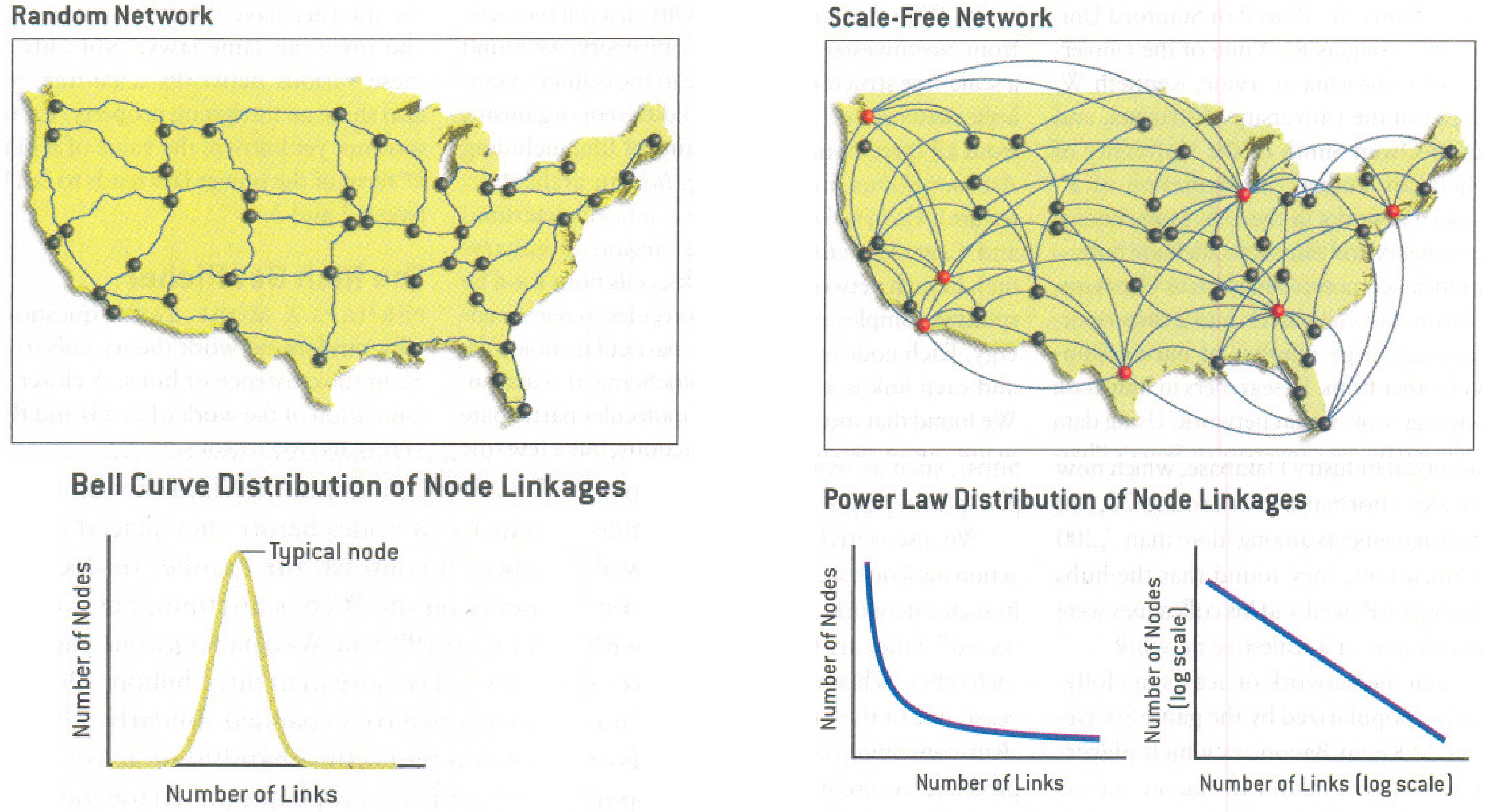
- The BA model (undirected graph)
  - input: some initial subgraph G_0 and m: number of edges per new node
  - the process:
    - nodes arrive one at the time
    - each node connects to m other nodes selecting them with probability proportional to their degree
    - if $[d_1, \dots, dt]$ is the degree sequence at time t, then node $t + 1$ links to node i with probability

$$\frac{\mathrm{d_i}}{\sum_{\mathrm{i}} \mathrm{d_i}}$$

- Results in power-law with exponent $\alpha = 3$

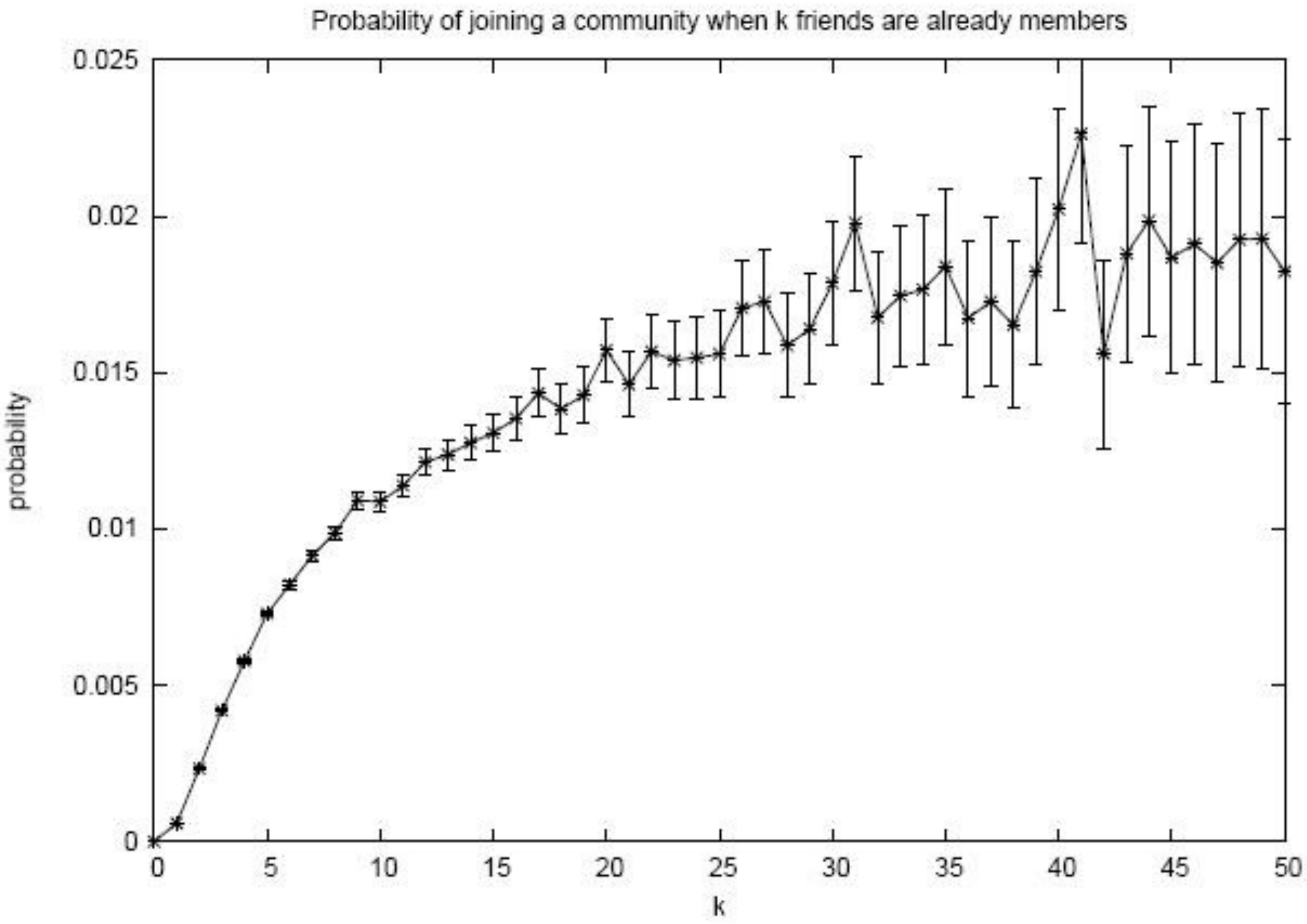
Stockholms universitet

Barabasi-Albert model



Stockholms universitet

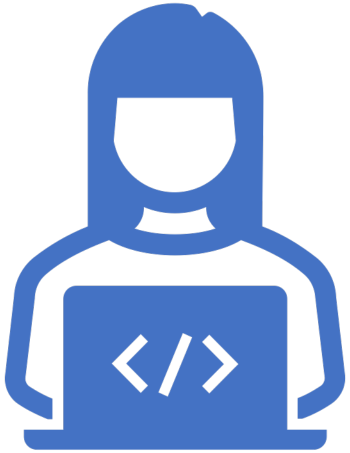
Example: Join an online group



Stockholms universitet

Models of Influence

- We saw that often decision is correlated with the number/fraction of friends
- The higher the number of friends, the higher the influence
- Graph models to capture that behavior:
  - Linear threshold model
  - Independent cascade model



Thanks!

golnaz.taheri@dsv.su.se