University of BRISTOL | Department of Computer Science

PROGRAMMING and ALGORITHMS II

INTRODUCTION TO

PAGERANK

Dr Tilo Burghardt

Unit Code COMS10001

Algorithm Design Paradigms

DIVIDE & CONQUER

Break down a problem into **independent** sub-problems of related type, solve them separately and combine the solutions.

GREEDY APPROACH

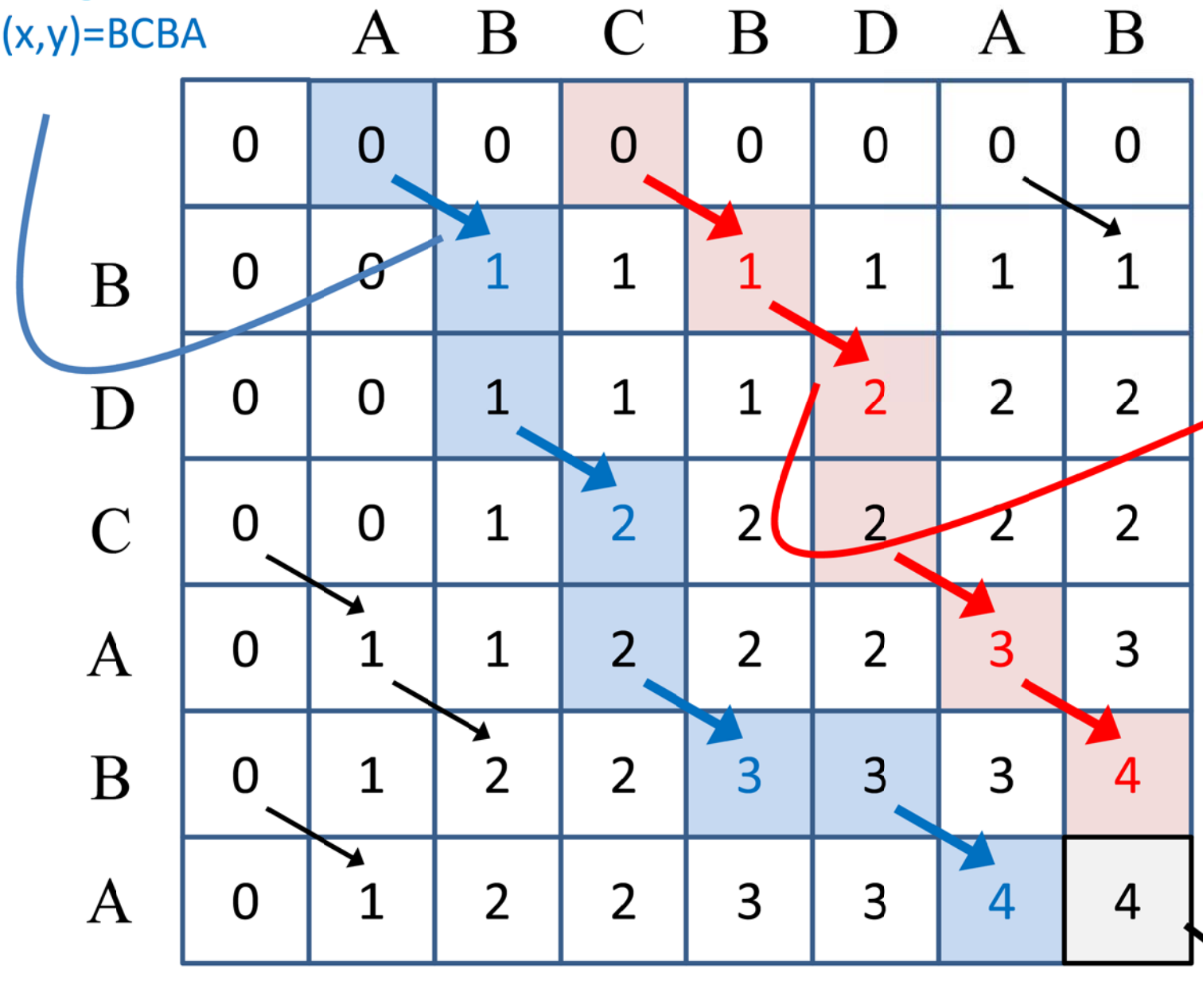
Use a sequence of **locally optimal** decisions incrementally to build up a solution.

DYNAMIC PROGRAMMING

Break down a problem into **overlapping** sub-problems of related type, build up solutions from larger and larger sub-solutions.

Recap: LCS

Matching alternative 1:
$LCS_1(x,y)$=BCBA

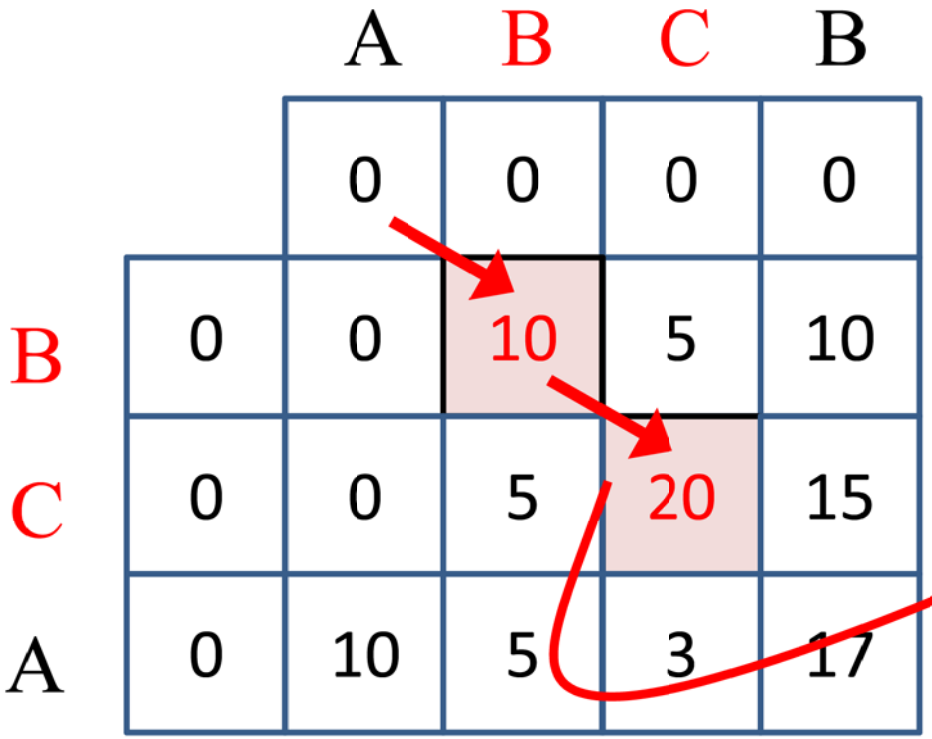


Matching alternative 2:
$LCS_2(x,y)$=BDAB

Overall, LCS has 4 characters in any case.

Recap: Local Alignment

University of BRISTOL | Department of Computer Science



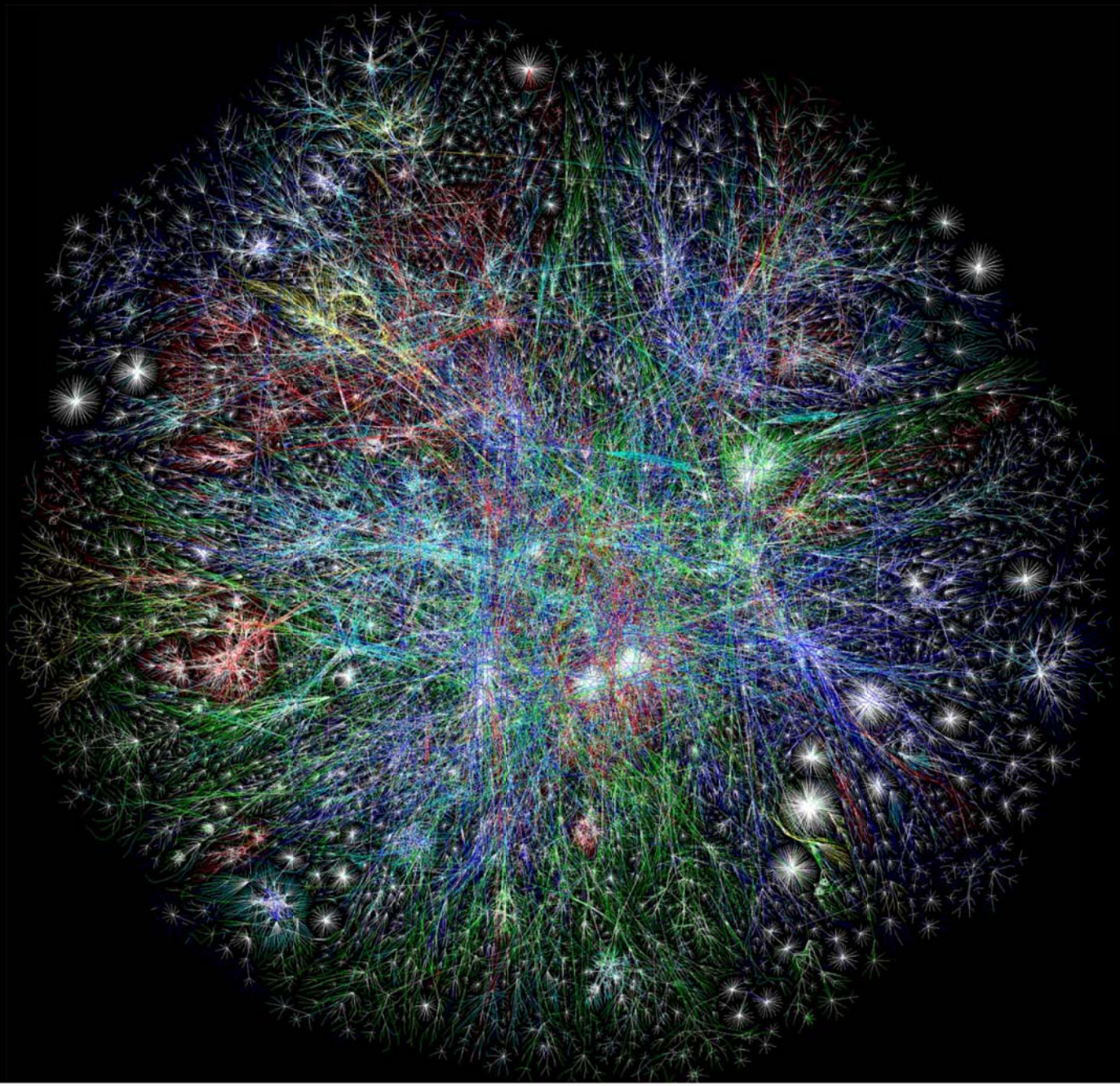
Result: local match of length 2 and score 20 found: BC

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + match\,score(x_i, y_j) \\ F(i-1, j) - gap\ penalty \\ F(i, j-1) - gap\ penalty \\ 0 \end{cases}$$

+10 for match, -2 for mismatch, -5 for space

Visualisation of the Internet

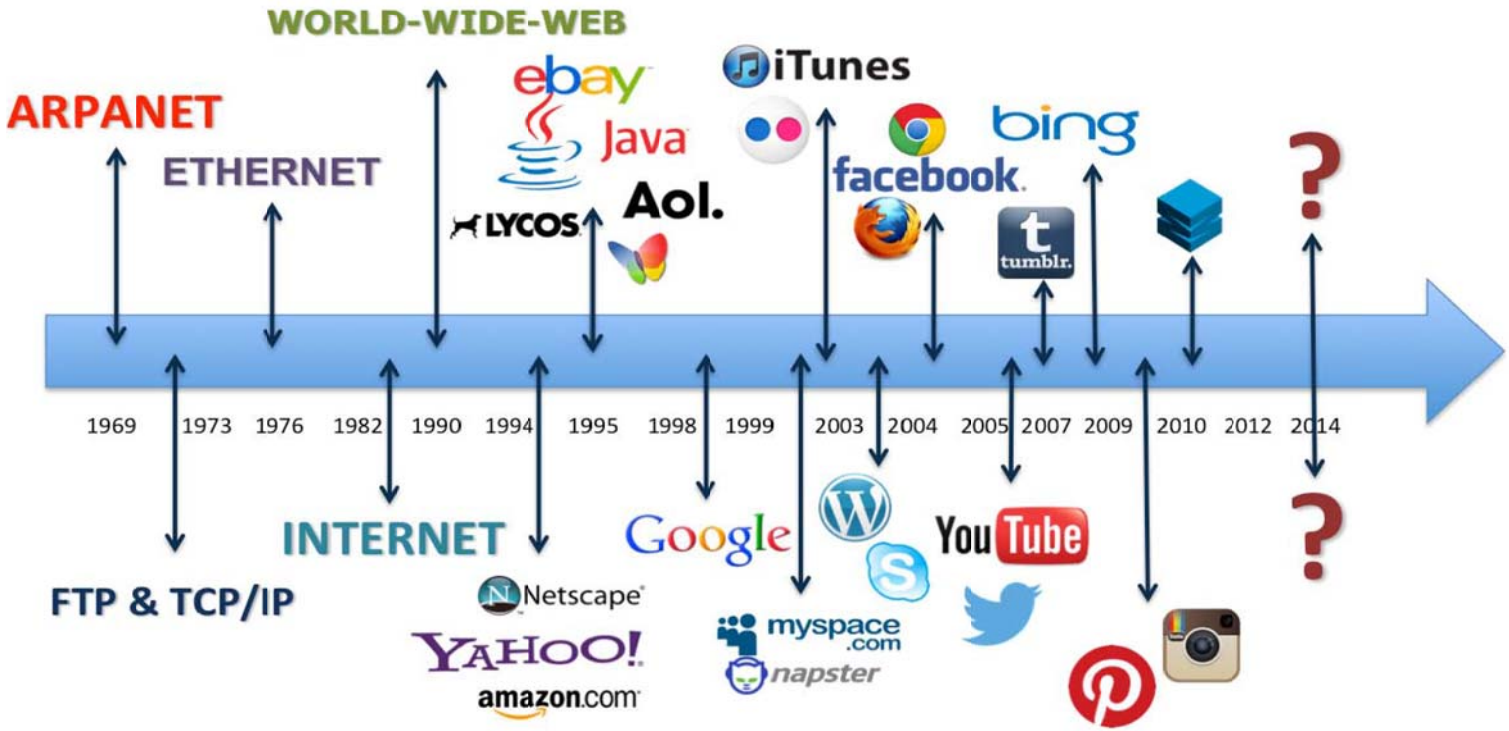
University of BRISTOL | Department of Computer Science



Early History of the Web

University of BRISTOL | Department of Computer Science

- **Early Search Engines:** direct comparison of query with content of indexed pages
- In the mid-90s it became clear that content similarity alone was not sufficient due to growth of the web



Google-Indexed Websites

University of BRISTOL | Department of Computer Science

Year	Pages Indexed
2014	67,000,000,000
2013	58,000,000,000
2012	50,000,000,000
2011	46,000,000,000
2010	29,000,000,000
2009	17,000,000,000
2008	11,000,000,000

PageRank – Key Idea

- **Key Idea:** pages that are pointed to by many other pages are likely to contain authoritative and prestigious information

- **PageRank** Algorithm published 1997/98:

Sergey Brin & Larry Page (Stanford University)

published PageRank at the Seventh International World Wide Web Conference (WWW7)

Concept of Rank Prestige

- Hyperlinks are an implicit conveyance of authority to the target page. A hyperlink from page x to page y is interpreted as a vote, by page x, for page y
- According to **Rank Prestige**, the importance P of page i (that is i's PageRank score) is the sum of the PageRank scores of all pages that point to i
- Since a page may point to many other pages, its score should be shared.

prestige of page i prestige of page j

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j},$$

presence of link from j to i

number of out-links of j

Overall Linear System

- Rank Prestige gives an overall system of n linear equations with n unknowns

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j},$$

- Lets define $\boldsymbol{P} = (P(1), P(2), \ldots, P(n))^T$ to be the n-dimensional column vector of Rank Prestige values

- Defining $\boldsymbol{A}$ as adjacency matrix we reach a compact form:

$$A_{ij} = \begin{cases} \frac{1}{O_i} & if\ (i,j) \in E \\ 0 & otherwise \end{cases} \quad \rightarrow \quad \boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P}$$

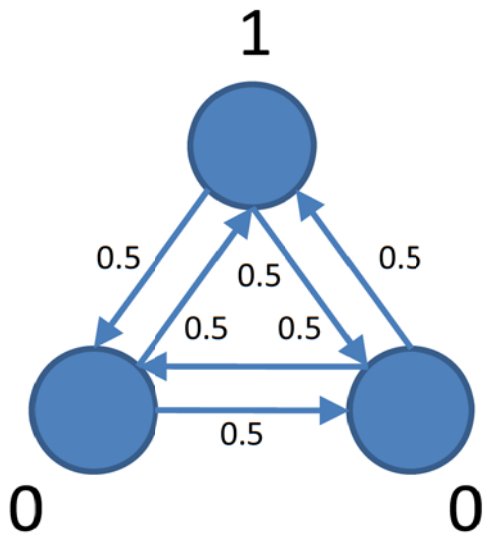
Resulting Eigensystem

- $\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P}$ is a characteristic equation of an **eigensystem**, where $\boldsymbol{P}$ is an **eigenvector** with the corresponding **eigenvalue** $\boldsymbol{1}$
- if $\boldsymbol{A}$ is **stochastic, irreducible** and **aperiodic**, then **power iteration** can be used to find $\boldsymbol{P}$
- for this case, $\boldsymbol{1}$ is the largest **eigenvalue** and the PageRank vector $\boldsymbol{P}$ is the **principal eigenvector**

→ **Problem**: the `Web graph' is not **stochastic, irreducible** and **aperiodic**

A Stochastic Model

- Web surfing modelled as a **stochastic process:**
  - each page in the Web graph is a state
  - each hyperlink is a transition with a probability
- each transition probability is $1/O_i$ if one assumes a Web surfer will click random hyperlinks on page i
- **Strong Simplification:**
  so far, no modelling of URL typing, back button or other 'jumps' outside forward hyperlinks

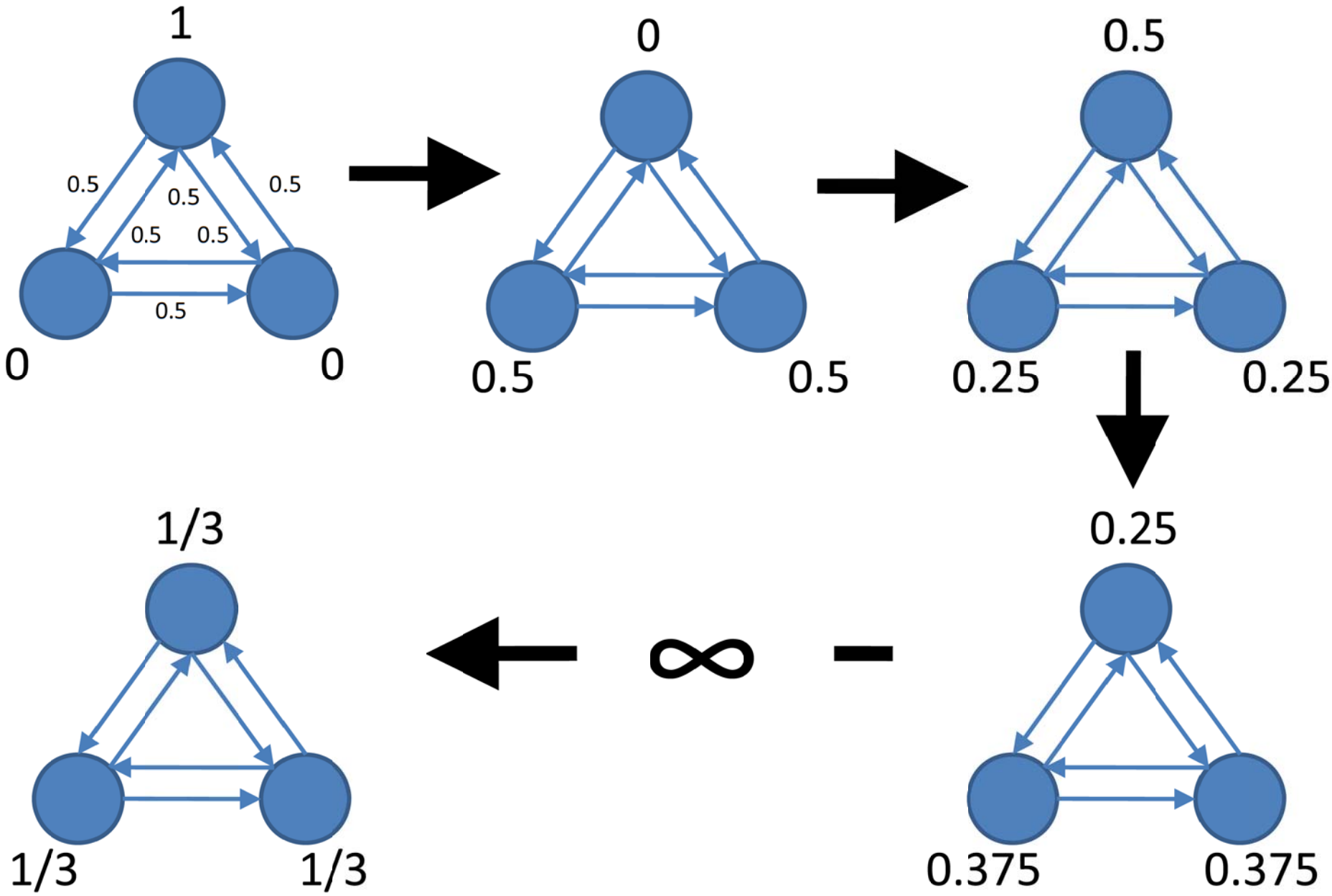


Andrey Markov
(1856-1922)

Three-Page Web Example:

Simple Markov Process

University of BRISTOL | Department of Computer Science



The Transition Matrix

- Set $\boldsymbol{A}$ to be the state transition probability matrix:

$$\boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} & . & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & . & A_{2n} \\ . & . & & & & . \\ . & . & & A_{ij} & & . \\ . & . & & & & . \\ A_{n1} & A_{n2} & . & . & . & A_{nn} \end{pmatrix}$$

A_{ij} represents the transition probability that a surfer on page i will move on to page j

`Stochastic' Matrix Property

- A should be a `stochastic' matrix to reflect that all probabilities of choices for a particular current page (i.e. the rows of the system) add up to 1:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdot & \cdot & \cdot & A_{1n} \\ A_{21} & A_{22} & \cdot & \cdot & \cdot & A_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ A_{n1} & A_{n2} & \cdot & \cdot & \cdot & A_{nn} \end{pmatrix} \qquad \sum_{j=1}^{n} A_{ij} = 1$$

System Evolution

Given $\boldsymbol{p_0}$ as an initial state distribution with

$$\sum_{i=1}^{n} p_0(i) = 1$$

what is the probability that m transitions later the Markov chain will be on page j?

System Evolution:

$$p_t(j) = \sum_{i=1}^{n} A_{ij}\, p_{t-1}(i)$$

Convergence to PageRank

- a finite Markov chain defined by a **stochastic matrix** $\boldsymbol{A}$ has a unique **stationary probability distribution** if $\boldsymbol{A}$ is **irreducible** and **aperiodic**.

- after a series of transitions $\boldsymbol{p}_t$ will converge to a **steady-state probability vector** $\boldsymbol{P}$ independent of the initial probability vector $\boldsymbol{p}_0$:

$$\lim_{t \to \infty} \boldsymbol{p}_t = \boldsymbol{P} \quad \rightarrow \quad \boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P}$$

Problem 1: Dangling Pages

University of BRISTOL | Department of Computer Science

Dangling Pages

- **Dangling Pages** are Web pages that have no outgoing links
- for them, the transition matrix $\boldsymbol{A}$ contains a row of $\boldsymbol{0}$'s according to:

$$A_{ij} = \begin{cases} \dfrac{1}{O_i} & if\,(i,j) \in E \\ 0 & otherwise \end{cases}$$

- the matrix $\boldsymbol{A}$ is, thus, no longer stochastic ☹, that is:

$$\sum_{j=1}^{n} A_{ij} \neq 1$$

Practical Options

1. do not consider dangling pages at all during the PageRank computation since such pages do not affect the ranking other pages directly
2. create outgoing links from each dangling page to all the pages on the Web → enforce a random page jump

$$A = \begin{pmatrix} 0 & 1/4 & 1/4 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix} \rightarrow A' = \begin{pmatrix} 0 & 1/4 & 1/4 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 0 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

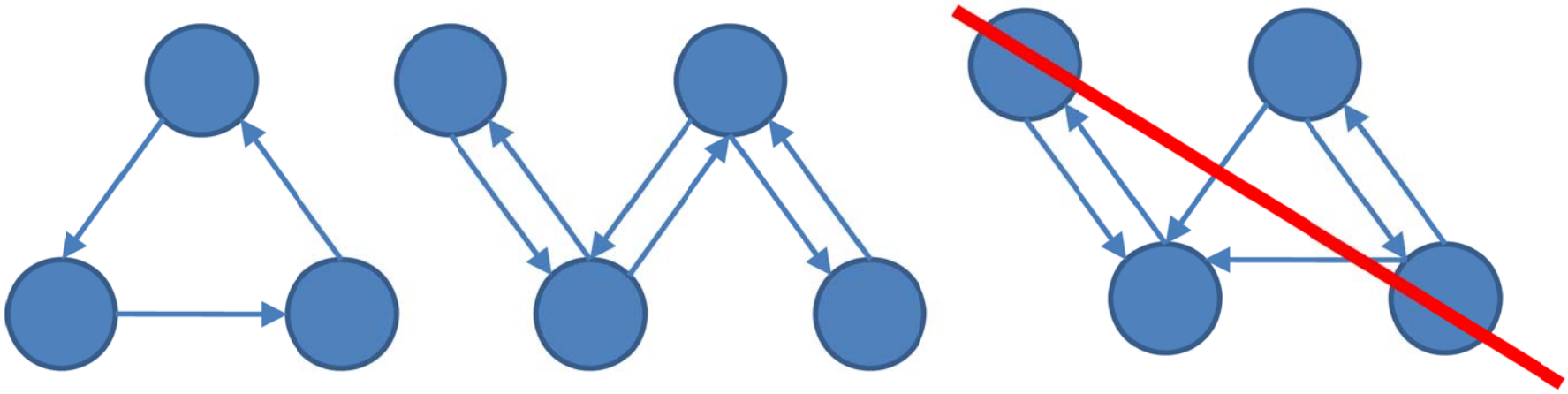
Problem 2: Weak Connectivity

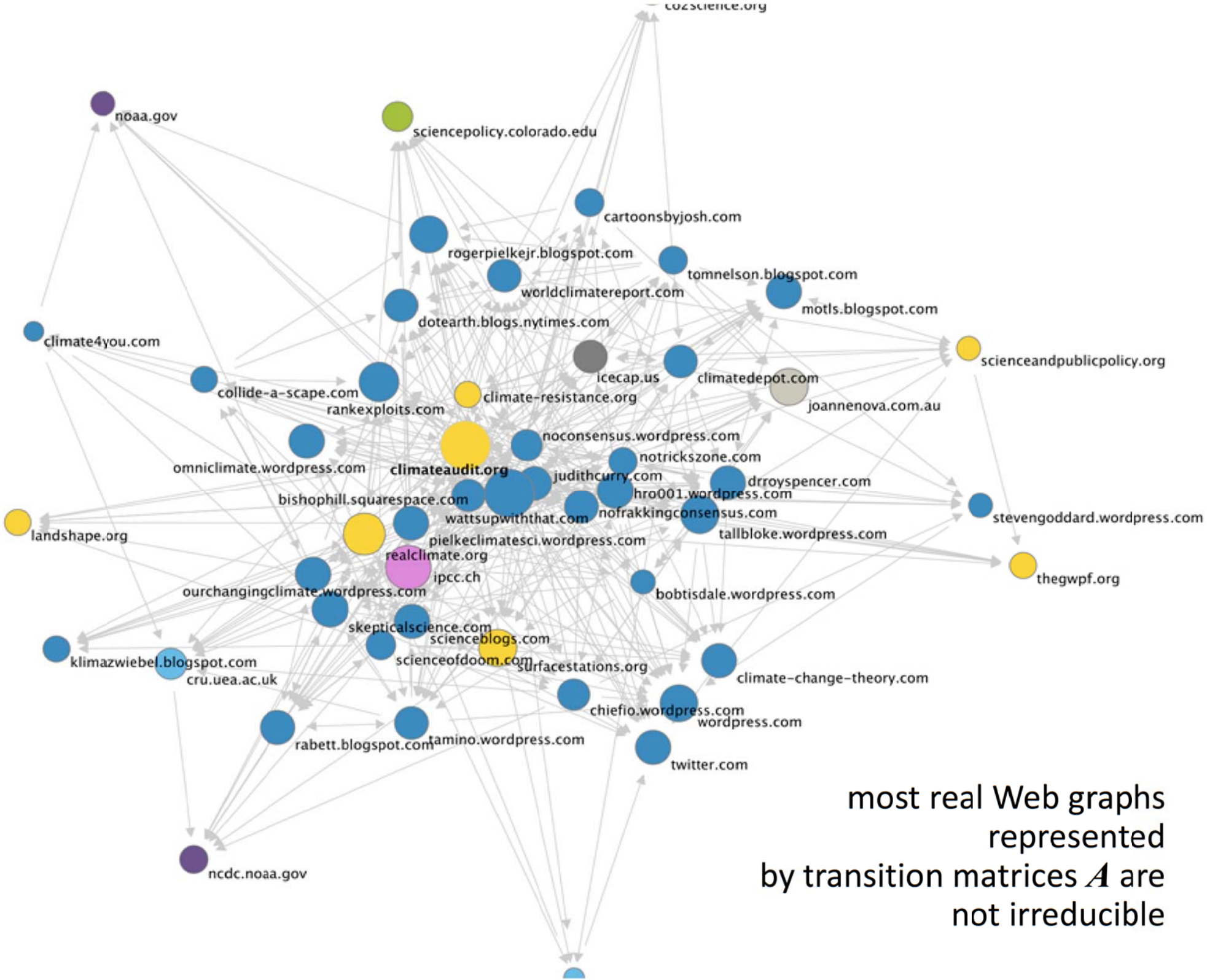
University of BRISTOL | Department of Computer Science

Strong Connectivity

A directed graph $G = (V, E)$ is **strongly connected** if its transition matrix is **irreducible**, that is for each pair of nodes $(u, v) \in V$, there is a path from u to v.

Examples:

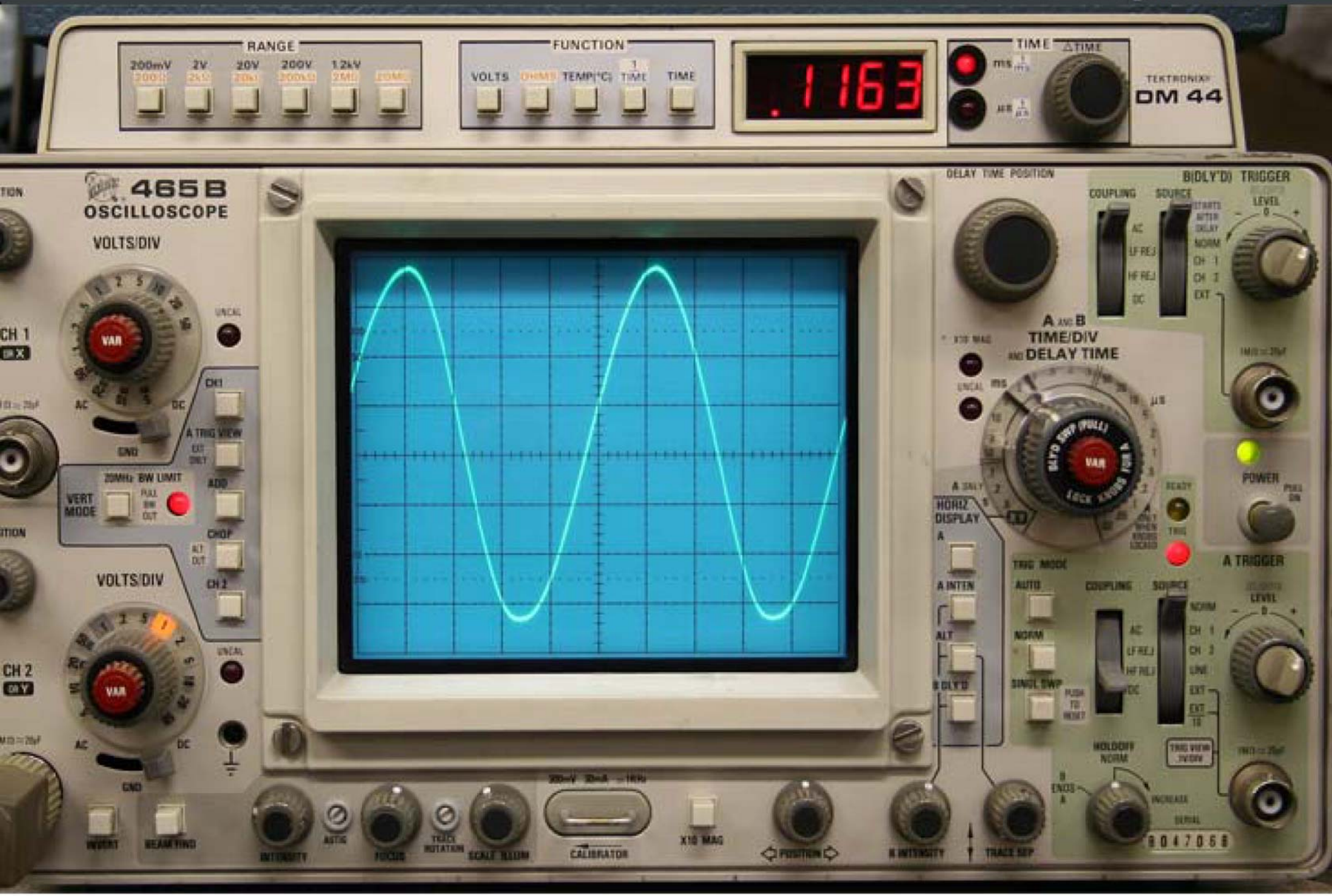




most real Web graphs
represented
by transition matrices A are
not irreducible

Problem 3: Periodicity

University of BRISTOL | Department of Computer Science



Periodicity

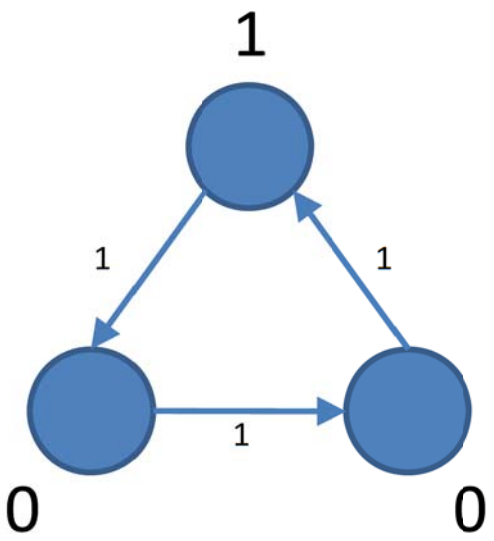
A state i is **periodic** with period $k > 1$ if k is the smallest number such that all paths leading from state i back to state i have a length that is a multiple of k. (A state i has **period** k if any return to state *i* must occur in multiples of k steps.)

- if a state is not periodic (i.e., $k = 1$), it is aperiodic
- a Markov chain is aperiodic if all states are aperiodic

Curse of Periodicity

Even if stochastic and irreducible, periodic graphs may not converge at all!

Example:



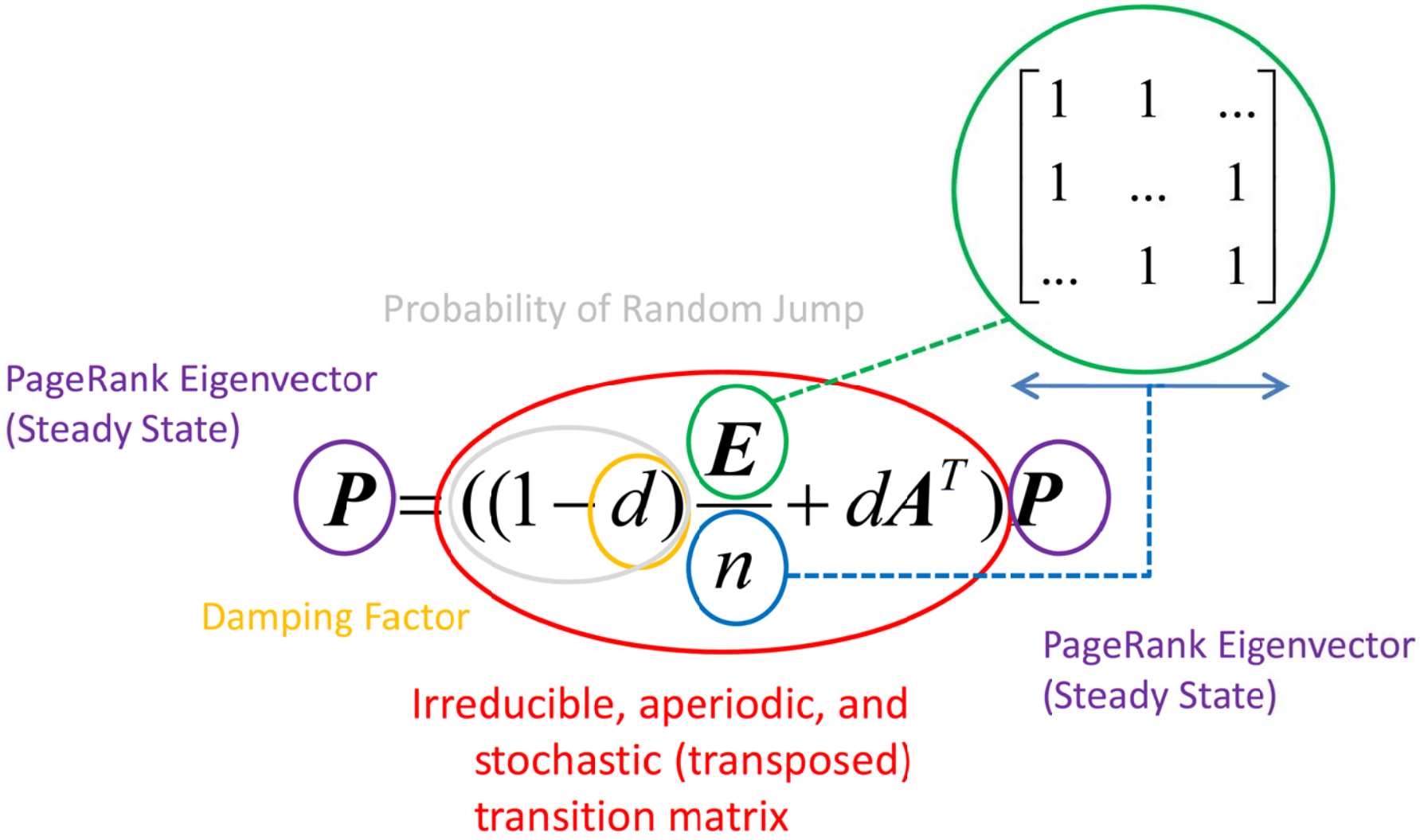
Random Jumps Solution

- **Solution Idea:** Add a link from each page to every page in **any** case. Give each link a small probability, say 1-d, of transition modelling random jumps to new webpages.

→ Using this extension, a random surfer now has two options for navigation:

- with probability *d* a random out-link is followed
- with probability 1-*d*, a random page is opened

Fabric of PageRank



Original PageRank

$$\boldsymbol{P} = ((1-d)\frac{\boldsymbol{E}}{n} + d\boldsymbol{A}^T)\boldsymbol{P}$$

... scaling with $\boldsymbol{e}^T\boldsymbol{P} = n$, yields...

$$\boldsymbol{P} = (1-d)\boldsymbol{e} + d\boldsymbol{A}^T\boldsymbol{P}$$

...which yields the originally published PageRank equation for a page...

$$P(i) = (1-d) + d\sum_{(j,i)\in E}\frac{P(j)}{O_j}$$

Pros & Cons of PageRank

- **Robust against Search Spam:**
  - it is difficult for Web page owner to generate incoming links into his/her site
- **PageRank is Global:**
  - PageRank values of all the pages are computed and saved off-line rather than at the query time
- **Limitation:**
  - query-independence

Google Search Statistics

University of BRISTOL | Department of Computer Science

Year	Annual Number of Google Searches	Average Searches Per Day
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000

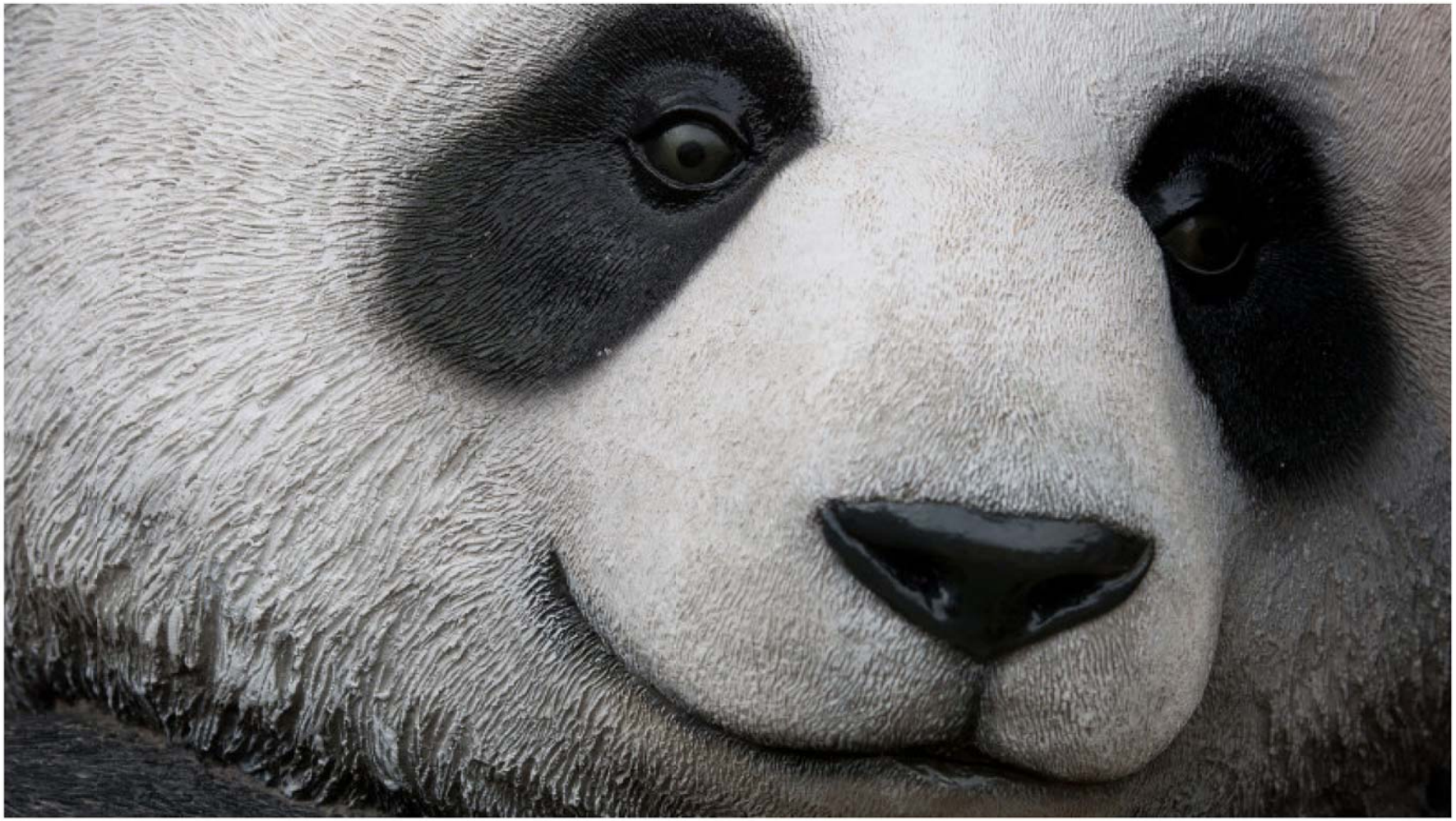
PageRank powers Google

- Google recalculates PageRank scores each time it builds its index
- increasing the number of documents in PageRank causes the initial approximation of PageRank to decrease
- PageRank is used by Twitter to present users with other accounts they may wish to follow
- in early 2005, Google implemented a `nofollow' attribute to combat search spam
- PageRank is now one of 200 ranking factors that Google uses to determine a page's popularity

Google cares about us! ☺

University of BRISTOL | Department of Computer Science

Google: Panda Update



What Is The Google Panda Update?