

# TEORETYCZNE PODSTAWY INFORMATYKI

4/12/2017

WFAiS UJ, Informatyka Stosowana  
I rok studiów, I stopień

# Wykład 9

2

Dane w postaci grafów

- Dane w postaci grafów
- Algorytm PageRanking

# Przykład: social network

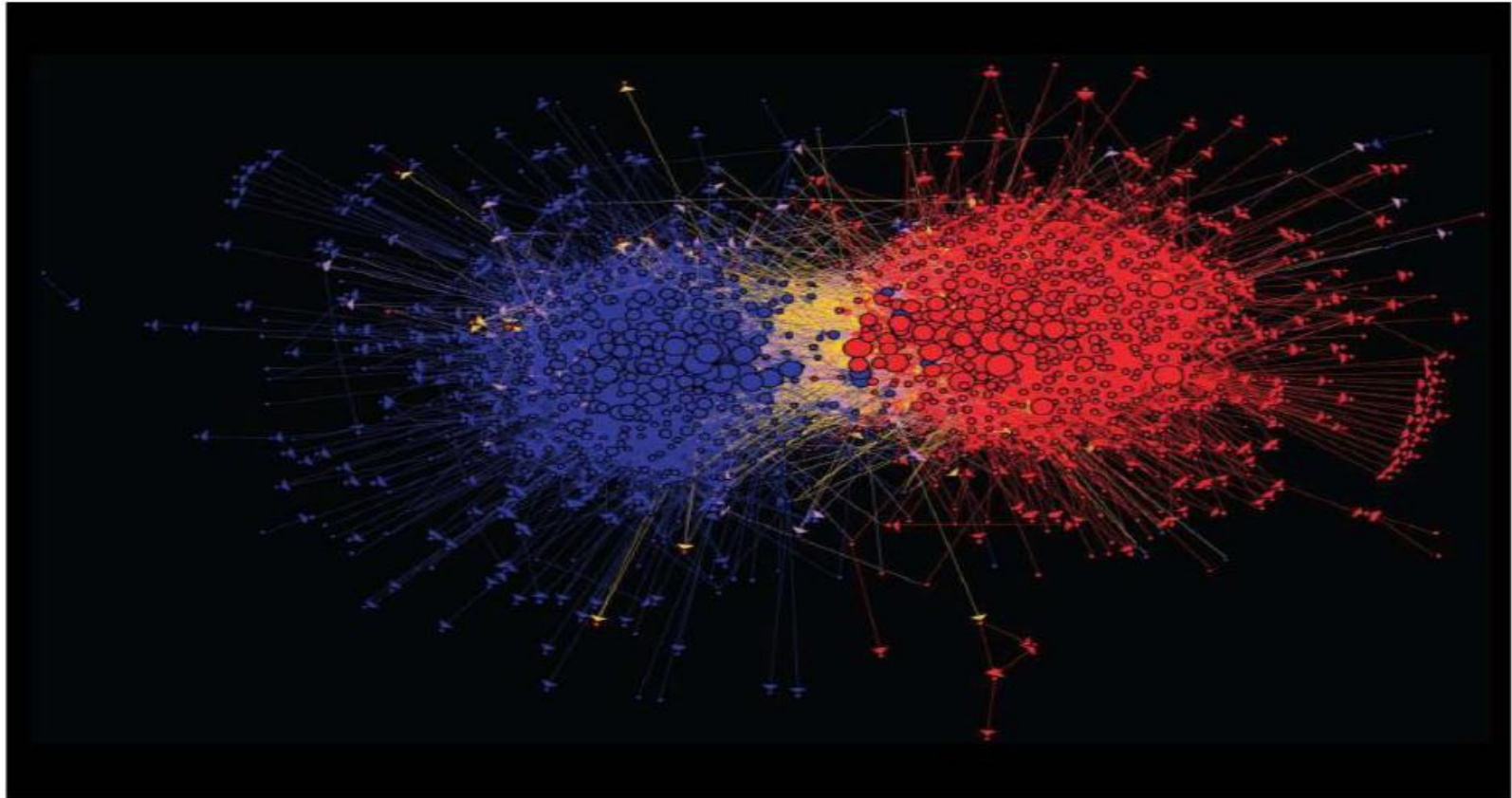
3



**Facebook social graph**

# Przykład: media network

4

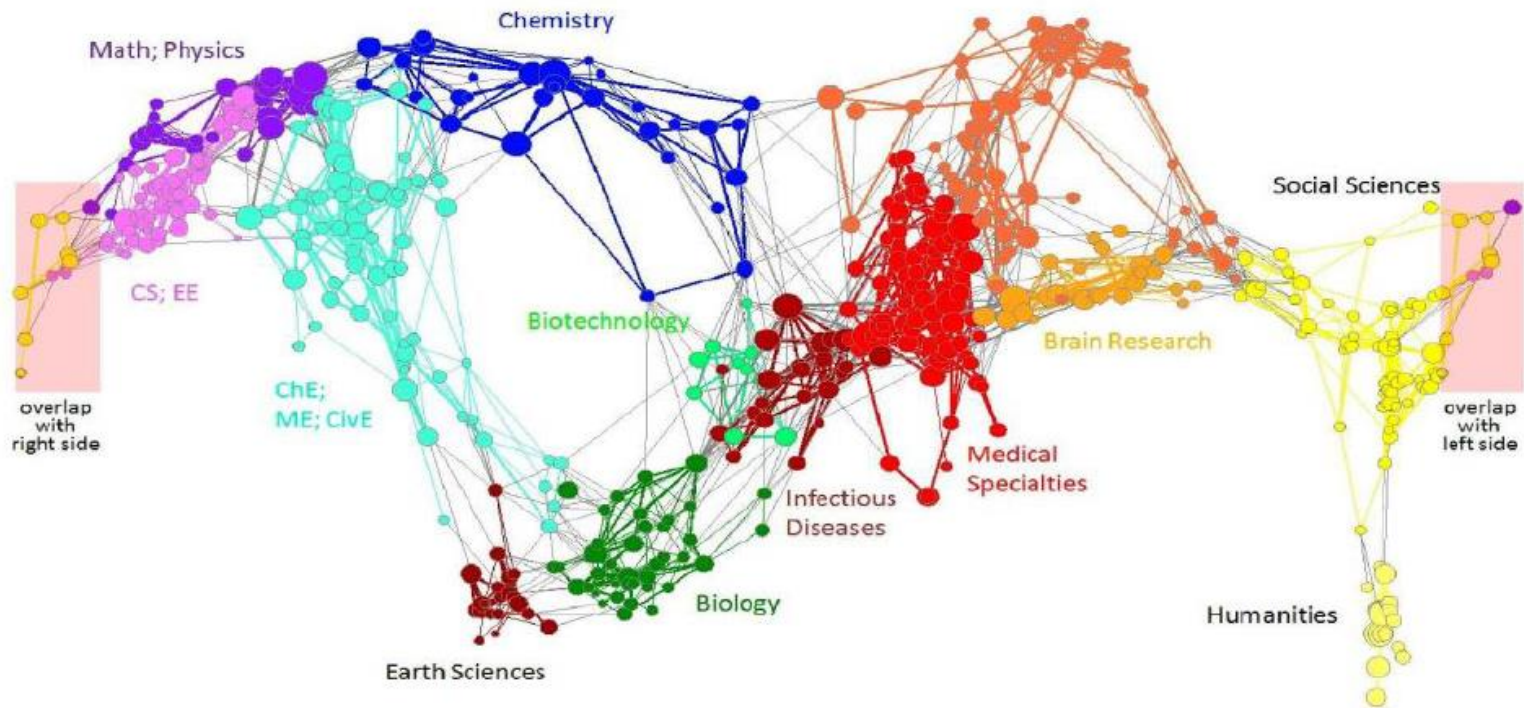


**Connections between political blogs**  
Polarization of the network [Adamic-Glance, 2005]



# Przykład: information network

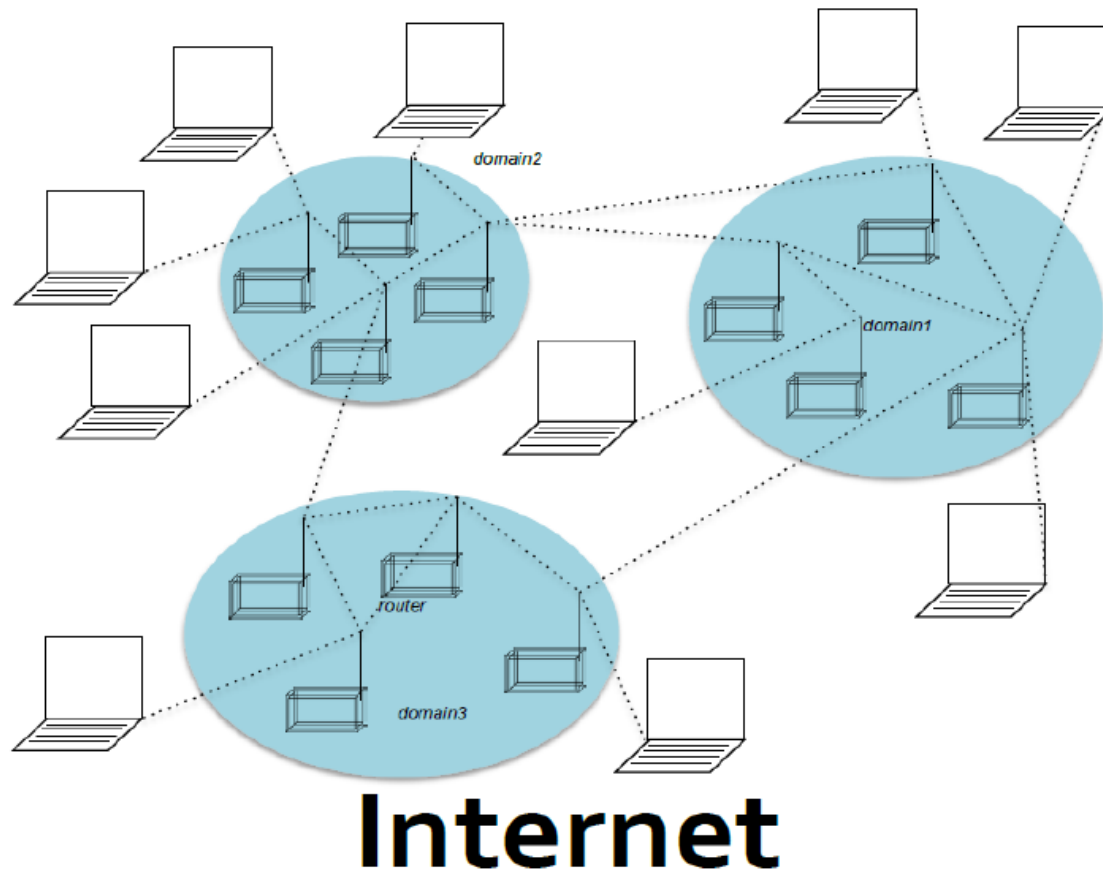
5



**Citation networks and Maps of science**  
[Börner et al., 2012]

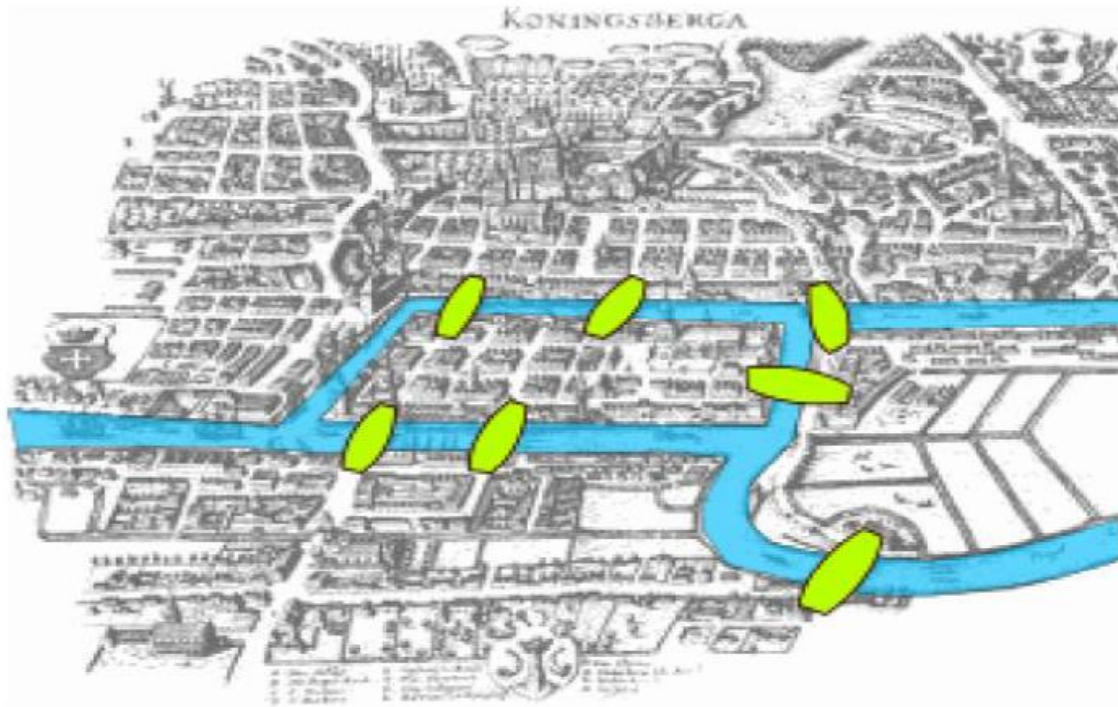
# Przykład: communication network

6



# Przykład: technologicalical network

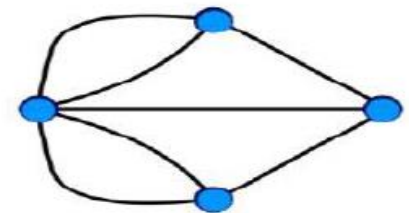
7



## Seven Bridges of Königsberg

[Euler, 1735]

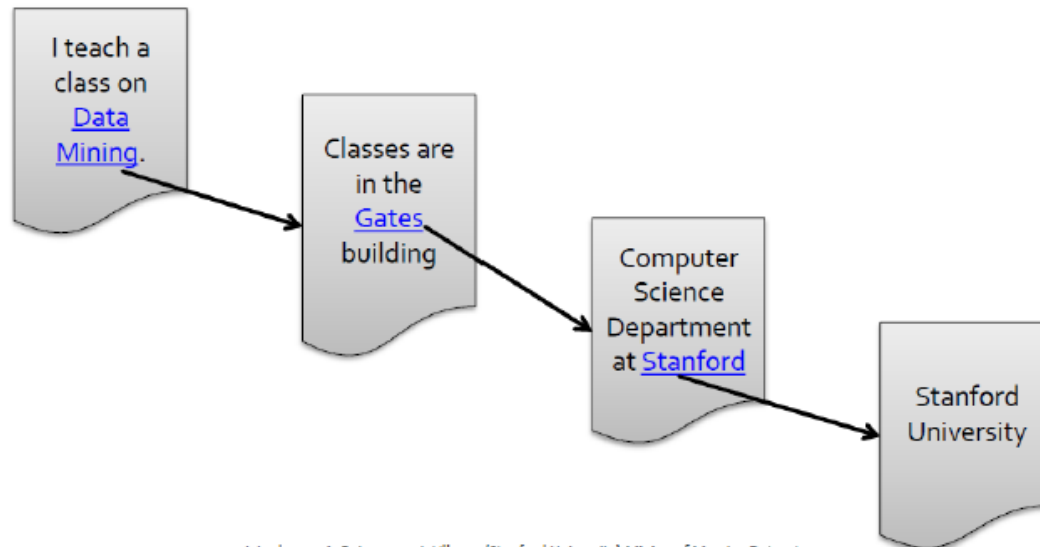
Return to the starting point by traveling each link of the graph once and only once.



# Web jako graf

8

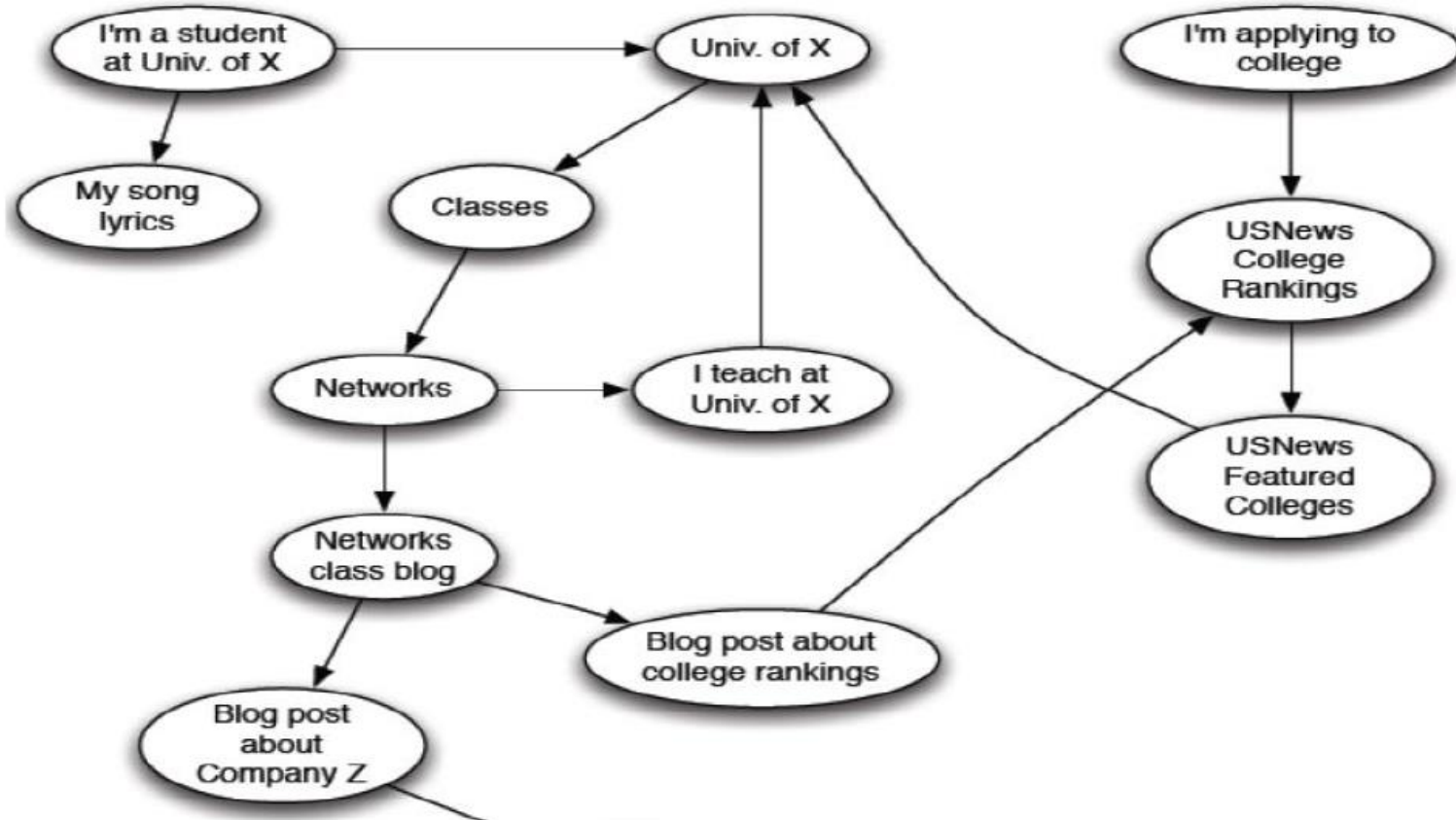
- Web jako skierowany graf
  - ▣ Węzły: strony internetowe
  - ▣ Strzałki: hyperlinki





# Web jako skierowany graf

9



# Jak organizować web

10

- Pierwsza próba: ręcznie tworzone katalogi
  - ▣ Yahoo, DMOZ, LookSmart
- Następna próba: WebSearch
  - ▣ Przeszukiwanie zawartości stron z małych i wiarygodnych podzbiorów: artykuły w gazetach, patenty, etc.
  - ▣ Ale: Web jest ogromny, wiarygodność stron trudna do weryfikacji, spamowanie web, itd.

# Podstawowe wyzwania

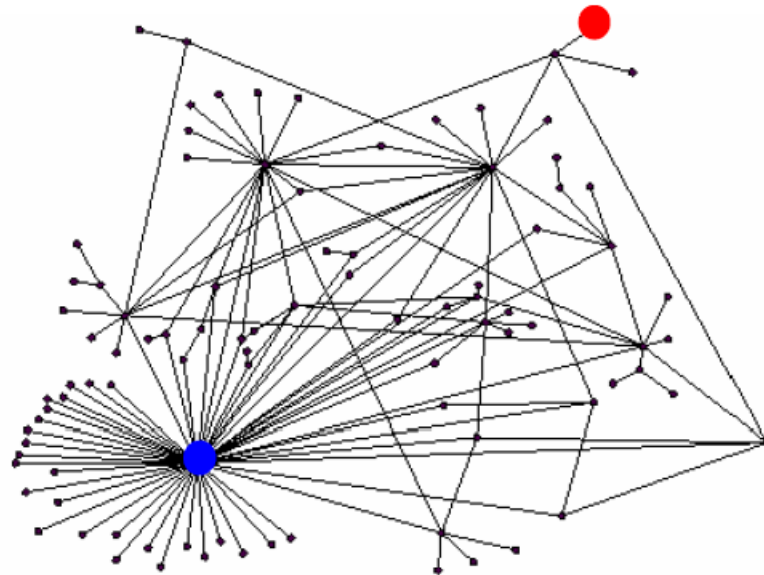
11

- W jaki sposób sprawdzać że strona jest wiarygodna?
- Jaka jest najlepsza odpowiedź na hasło „gazeta”?
  - Nie ma jednej najlepszej odpowiedzi
  - Strony które pokazują informacje na temat gazet mogą dotyczyć wielu różnych gazet

# Które strony są najważniejsze

12

- Nie wszystkie strony są tak samo ważne
- Zróbmy „ranking” stron ze względu na ilość linków



# Algorytmy do analizy linków

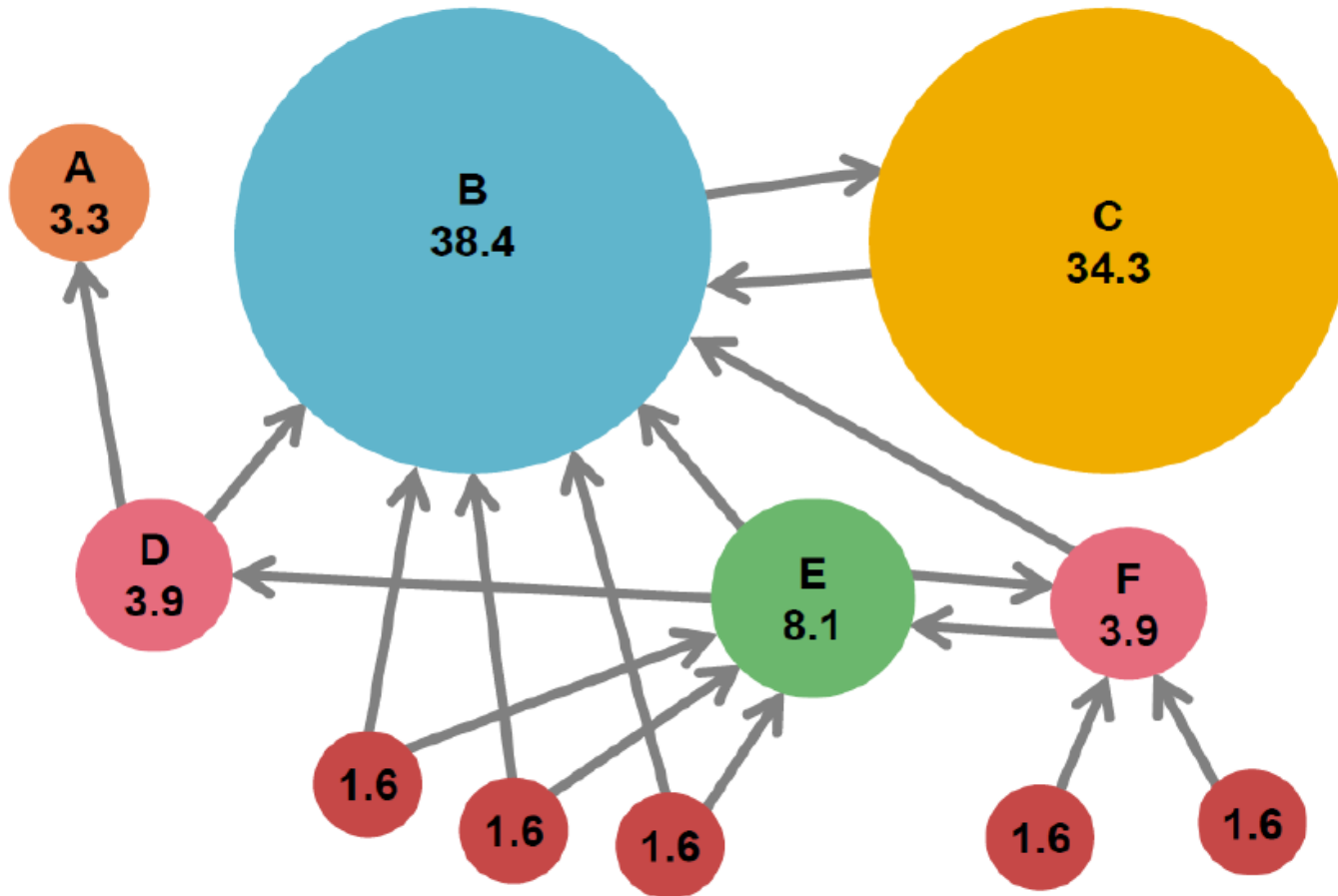
13

- Wiele różnych algorytmów zostało rozwiniętych
- Omówię algorytm PageRank używany przez Google.
  - Każdy link liczy się jako punkt, bardziej ważna strona to ta która ma więcej linków: Wchodzących? Wychodzących?
  - A linki a ważnych stron do danej strony powinny się liczyć bardziej! Hm.. To wygląda jak pytanie rekurencyjne..



# Przykład: PageRank punktacja

14

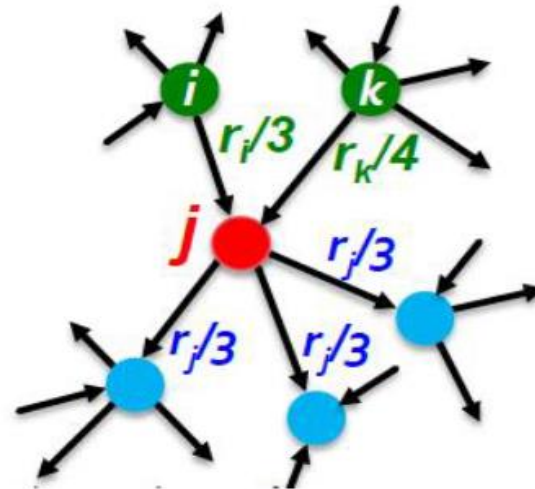


# Prosta rekurencyjna formuła

15

- Wartość każdego linku jest proporcjonalna do wartości strony z której wychodzi
- Jeżeli strona  $j$  z wartością  $r_j$  na  $n$ -wychodzących linków to każdy ma wartość  $r_j/n$ .
- Wartość strony  $j$  jest sumą wartości linków do niej wchodzących.

$$r_j = r_i/3 + r_k/4$$



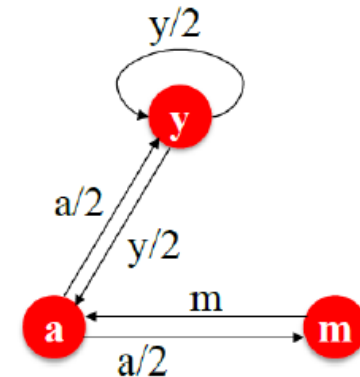
# Page Rank: flow model

16

- Link z „ważnej” strony jest więcej warty
- Strona jest „ważna” jeżeli jest wskazywana przez wiele innych stron.
- Zdefiniujemy „rank” strony

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

$d_i$  ... out-degree of node  $i$



“Flow” equations:

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

# Rozwiązanie dla „flow” równania

17

- 3 równania, 3 niewiadome, nie ma stałej
- Nie ma jednego rozwiązania
- Dodatkowy warunek dot. Normalizacji

Flow equations:

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

- $r_y + r_a + r_m = 1$

- **Solution:**  $r_y = \frac{2}{5}, r_a = \frac{2}{5}, r_m = \frac{1}{5}$

- Układ równań możemy rozwiązać np. metodą eliminacji Gaussa. Ale potrzebujemy czegoś lepszego

# Interpretacja macierzowa

18

- Ponieważ kolumny normalizowane do 1 => interpretacja prawdopodobieństwowa.
  - Page  $i$  ma  $d_i$  out-link
  - Jeżeli  $i \rightarrow j$ , to  $M_{ji} = 1/d_i$ , w pozostałych  $M_{ji} = 0$
  - Kolumny tej macierzy sumują się do 1
- Oznaczmy  $r_i = \text{rank strony}$
- Równanie „flow”

$$\sum_i r_i = 1$$

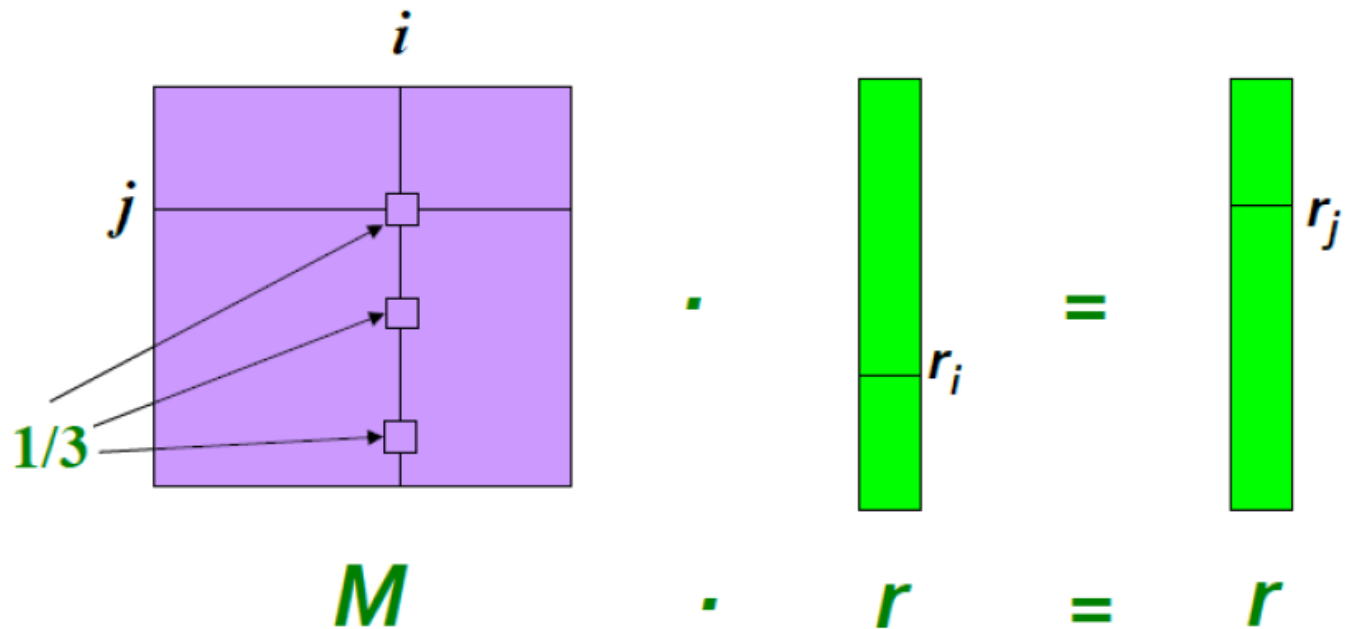
$$\mathbf{r} = \mathbf{M} \cdot \mathbf{r}$$

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$



# Przykład

19



# Power iteration

20

□ Jak efektywnie możemy rozwiązać takie równanie?

□ Metoda „power iteration”

$$\mathbf{r} = \mathbf{M} \cdot \mathbf{r}$$

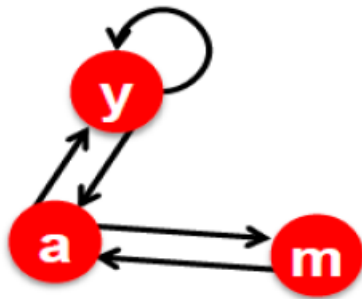
□ Wektor  $\mathbf{r}$  jest wektorem własnym macierzy prawdopodobieństwa.

**NOTE:**  $\mathbf{x}$  is an eigenvector with the corresponding eigenvalue  $\lambda$  if:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

# Przykład

21



$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

$$r = M \cdot r$$

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

# Metoda „power iteration”

22

- Mając dany graf z  $N$  węzłami, gdzie każdy węzeł to są strony a skierowane krawędzie to są hiperlinki
  - ▣ Zainicjalizuj  $r(0) = [1/N, \dots, 1/N]^T$
  - ▣ Iteruj:  $r^{(t+1)} = M r^{(t)}$
  - ▣ Zatrzymaj jeżeli  $\|r^{(t+1)} - r^{(t)}\|_1 < \varepsilon$

$\|x\|_1 = \sum_{1 \leq i \leq N} |x_i|$  is the  $L_1$  norm

$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

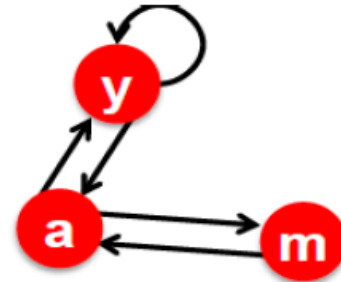
$d_i$  .... out-degree of node  $i$

# Przykład

23

## ■ Power Iteration:

- Set  $r_j = 1/N$
- **1:**  $r'_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$
- **2:**  $r = r'$
- If not converged: goto **1**



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

## ■ Example:

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{matrix} 1/3 & 1/3 & 5/12 & 9/24 & & 6/15 \\ 1/3 & 3/6 & 1/3 & 11/24 & \dots & 6/15 \\ 1/3 & 1/6 & 3/12 & 1/6 & & 3/15 \end{matrix}$$

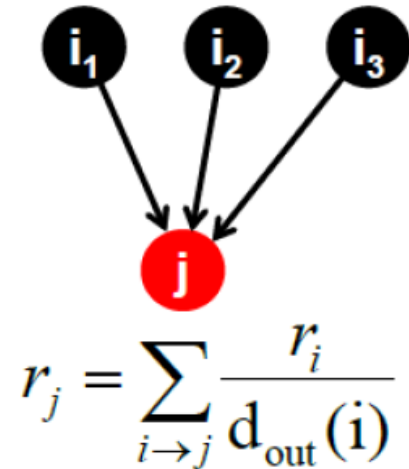
Iteration 0, 1, 2, ...



# Błądzenie przypadkowe

24

- Oglądamy przypadkowe strony internetowe
  - ▣ W momencie  $t$ , jesteśmy na stronie  $i$
  - ▣ W czasie  $(t+1)$  przechodzimy losowo na jedną ze stron podłączonych ze strony  $i$
  - ▣ W pewnym momencie kończymy na stronie  $j$
  - ▣ Powtarzamy ten krok nieskończoną ilość razy.



$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{\text{out}}(i)}$$

# Rozkład stacjonarny

25

- Na jakiej stronie jesteśmy w czasie  $(t+1)$ ?
  - ▣ Błądzimy losowo

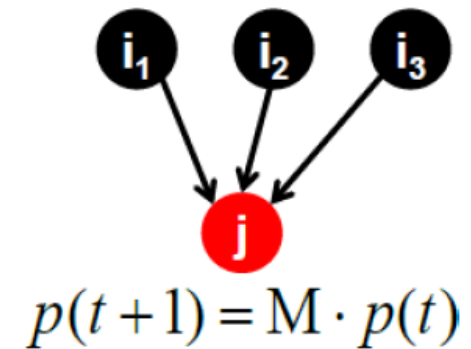
$$p(t+1) = M \cdot p(t)$$

- ▣ Załóżmy że rozwiązanie jest stacjonarne

$$p(t+1) = M \cdot p(t) = p(t)$$

- ▣ Nasze oryginalne równanie to było

$$r = M \cdot r$$



# Procesy Markowa

26

- **Dla grafów które spełniają pewne warunki rozwiązanie stacjonarne będzie osiągnięte niezależnie od warunków początkowych.**
- **To rozwiązanie jest jednoznaczne.**

# PageRank wg. Googla

27

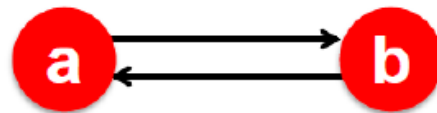
$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \quad \text{or equivalently} \quad r = Mr$$

- Trzy pytania:
  - ▣ Czy rozwiązanie zbieżne?
  - ▣ Czy zbieżne do rozwiązania które oczekujemy?
  - ▣ Czy rozwiązanie jest rozsądne?

# Czy rozwiązanie jest zbieżne?

28

- „Spider trap” problem



$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

- **Example:**

$$\begin{array}{l} r_a \\ r_b \end{array} = \begin{array}{cccc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}$$

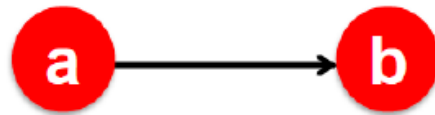
Iteration 0, 1, 2, ...



# Czy zbieżne do tego czego oczekujemy?

29

## □ „Dead end” problem



$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

### ■ Example:

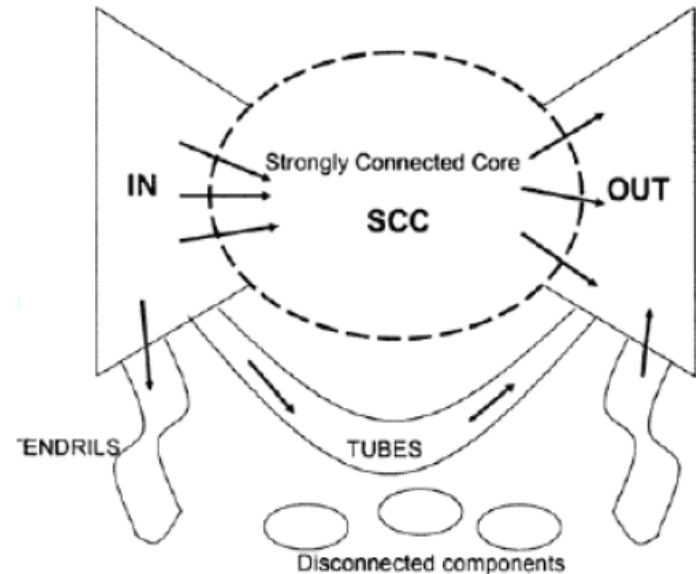
$$\begin{array}{l} r_a \\ r_b \end{array} = \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}$$

Iteration 0, 1, 2, ...

# PageRank: problemy

30

- Niektóre strony są „dead end” czyli nie mają wychodzących linków
  - ▣ Takie strony powodują „wyciekanie” informacji o ważności stron
- Niektóre grupy stron tworzą „spider traps” czyli wszystkie out-linki są do zamkniętej grupy stron.
  - ▣ I wtedy tak grupa stron zaasimiluje wszystkie punkty ważności stron.

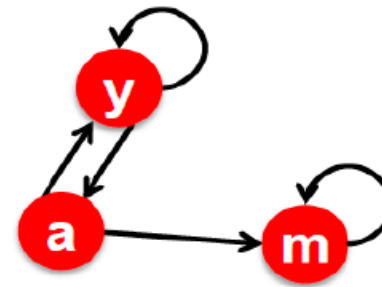


# Spider Traps

31

## ■ Power Iteration:

- Set  $r_j = 1$
- $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$ 
  - And iterate



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2$$

$$r_m = r_a/2 + r_m$$

## ■ Example:

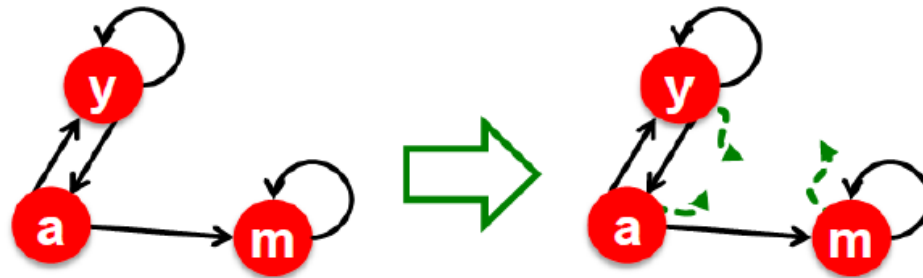
$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 3/6 & 7/12 & 16/24 & & 1 \end{matrix}$$

Iteration 0, 1, 2, ...

# Rozwiązanie: teleportacja

32

- W każdym kroku błądzenia można
  - ▣ z prawdopodobieństwem  $\beta$  pójść jedną z wychodzących ścieżek
  - ▣ z prawdopodobieństwem  $(1 - \beta)$  przeskoczyć na losowo wybraną stronę
  - ▣ Najczęściej stosowane wartości  $\beta = 0.8-0.9$

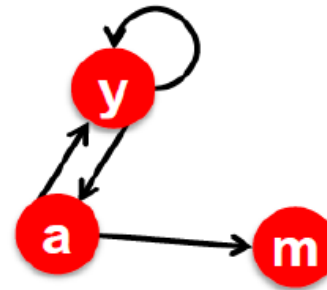


# Problem: dead end

33

## Power Iteration:

- Set  $r_j = 1$
- $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$ 
  - And iterate



	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	0

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2$$

$$r_m = r_a/2$$

## Example:

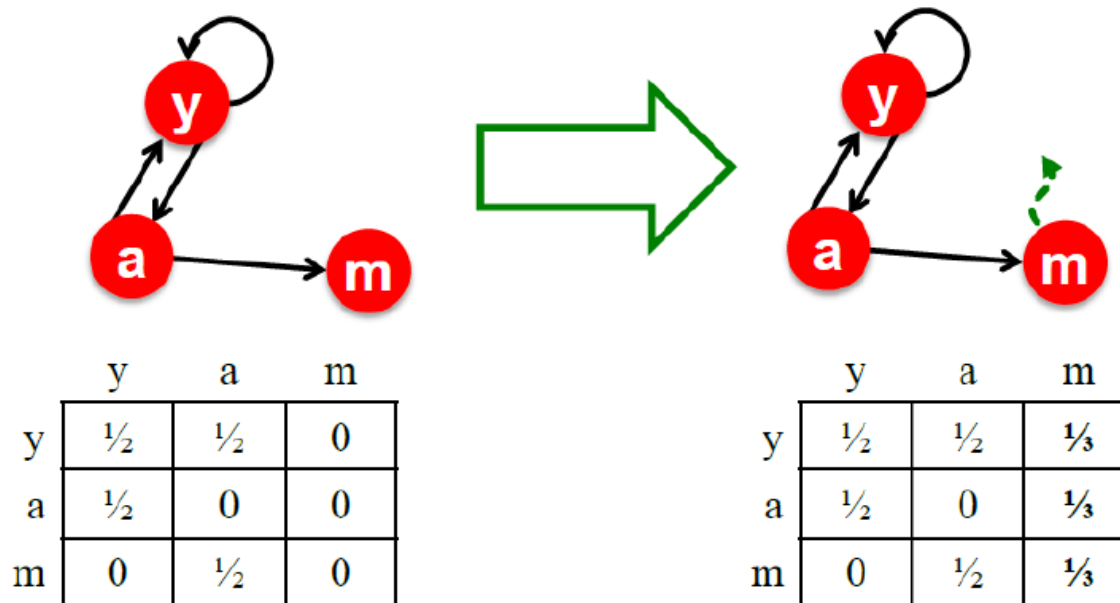
$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 1/6 & 1/12 & 2/24 & & 0 \end{matrix}$$

Iteration 0, 1, 2, ...

# Rozwiązanie teleportacja

34

- Zawsze przeskocz do losowo wybranej strony
- Odpowiednio zmodyfikuj macierz przejść



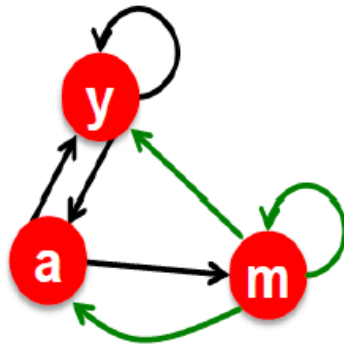
# Teleportacja

35

- Powoduje że macierz staje się stochastyczna

$$A = M + a^T \left( \frac{1}{n} e \right)$$

- $a_{i\dots} = 1$  if node  $i$  has out deg 0, =0 else
- $e$ ...vector of all 1s



	y	a	m
y	1/2	1/2	1/3
a	1/2	0	1/3
m	0	1/2	1/3

$$r_y = r_y/2 + r_a/2 + r_m/3$$

$$r_a = r_y/2 + r_m/3$$

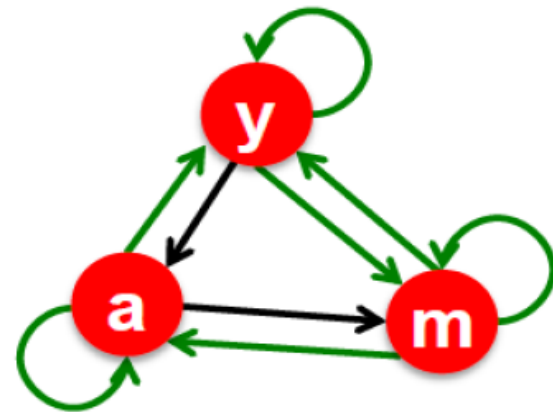
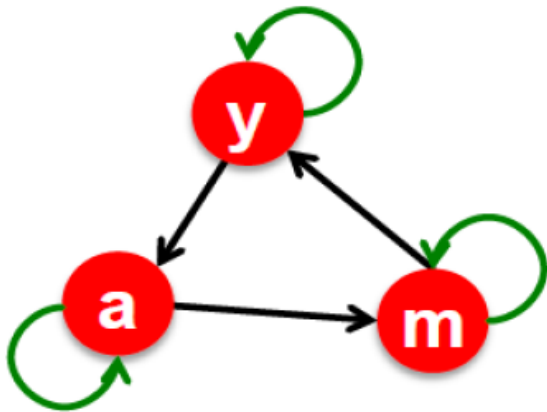
$$r_m = r_a/2 + r_m/3$$

# Teleportacja

36

- Powoduje że macierz staje się aperiodyczna i nieredukowalna

Niezerowe prawdopodobieństwo  
Przejścia z każdego stanu do  
każdego innego.





# Rozwiązanie Google

37

## PageRank equation [Brin-Page, 98]

$$r_j = \sum_{i \rightarrow j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{n}$$

$d_i$  ... out-degree  
of node  $i$

The above formulation assumes that  $M$  has no dead ends. We can either preprocess matrix  $M$  (**bad!**) or explicitly follow random teleport links with probability 1.0 from dead-ends.

# Rozwiązanie Google

38

- **PageRank equation** [Brin-Page, 98]

$$r_j = \sum_{i \rightarrow j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{n}$$

- **The Google Matrix A:**

$$A = \beta M + (1 - \beta) \frac{1}{n} \mathbf{e} \cdot \mathbf{e}^T$$

e...vector of all 1s

- **A is stochastic, aperiodic and irreducible, so**

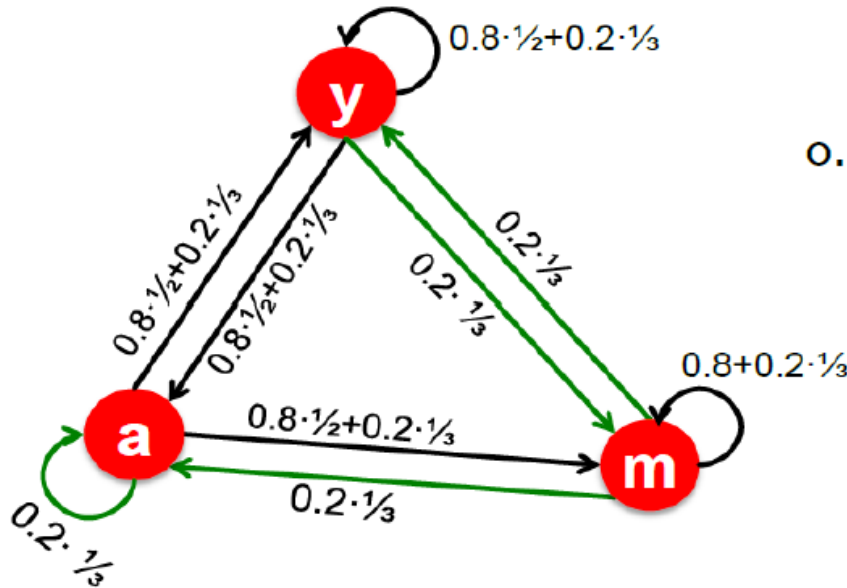
$$\mathbf{r}^{(t+1)} = \mathbf{A} \cdot \mathbf{r}^{(t)}$$

- **What is  $\beta$ ?**

- In practice  $\beta = 0.8, 0.9$  (make 5 steps and jump)

# Przykład

39



$$0.8 \begin{matrix} \mathbf{M} \\ \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \end{matrix} + 0.2 \begin{matrix} \mathbf{1/n \cdot 1 \cdot 1^T} \\ \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{matrix}$$

y	7/15	7/15	1/15
a	7/15	1/15	1/15
m	1/15	7/15	13/15

**A**

y	=	1/3	0.33	0.24	0.26	7/33
a		1/3	0.20	0.20	0.18	5/33
m		1/3	0.46	0.52	0.56	21/33

# PageRank: kompletny algorytm

40

## ■ Output: PageRank vector $r$

- **Set:**  $r_j^{(0)} = \frac{1}{N}, \quad t = 1$

- **do:**

- $\forall j: r'_j{}^{(t)} = \sum_{i \rightarrow j} \beta \frac{r_i^{(t-1)}}{d_i}$

- $r'_j{}^{(t)} = \mathbf{0}$  if in-deg. of  $j$  is  $\mathbf{0}$

- **Now re-insert the leaked PageRank:**

- $\forall j: r_j^{(t)} = r'_j{}^{(t)} + \frac{1-S}{N}$  where:  $S = \sum_j r'_j{}^{(t)}$

- $t = t + 1$

- **while**  $\sum_j |r_j^{(t)} - r_j^{(t-1)}| > \varepsilon$