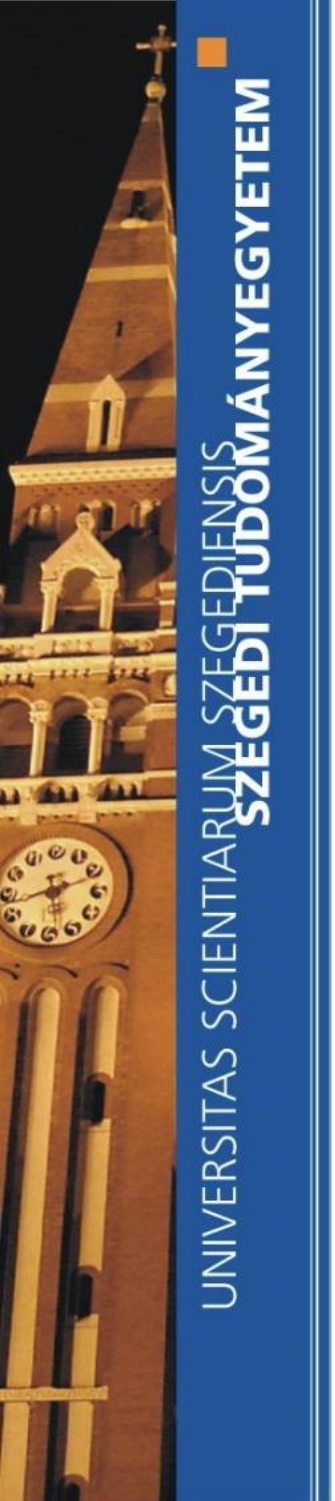


Hogy keres a Google?

Kevei Péter

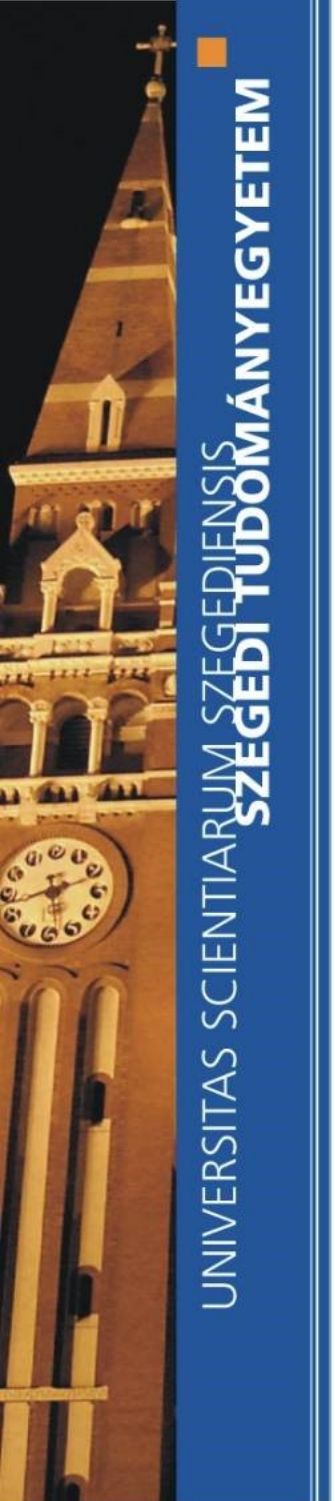
SZTE Bolyai Intézet

Kutatók Éjszakája
2018. szeptember 28.

WWW

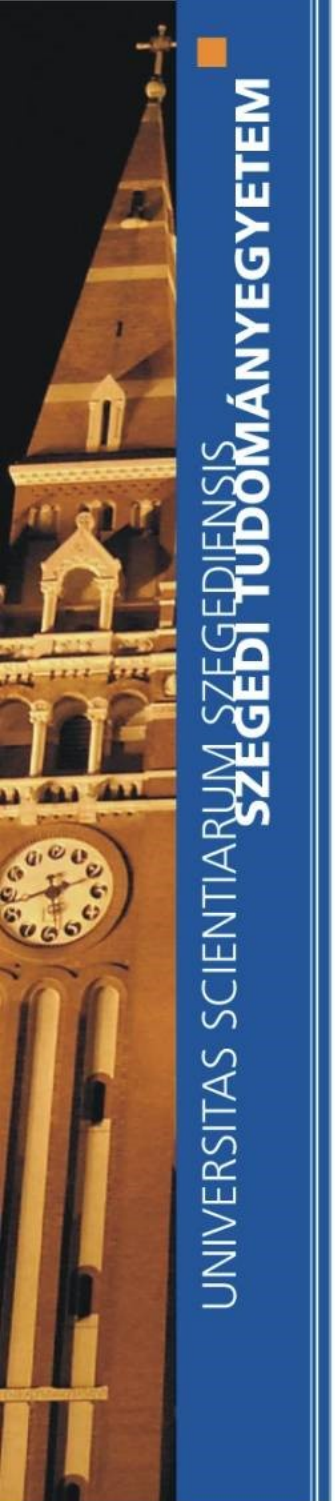
- Könyvtár 25 milliárd ($25 \cdot 10^9$) dokumentummal, és nincs könyvtáros (a Somogyi Könyvtárban 900 000, a Bolyai könyvtárában kb. 20 000 könyv van).
- Bárki bármikor feltölthet dokumentumokat.
- A dokumentumok nagy része hosszú (több 10 000 szó).

Somogyi Könyvtár 2. emelet (http://www.sk-szeged.hu)



Keresőprogramok feladata

- Feltérképezi a nyilvános elérésű weboldalakat.
- Olyan formába rendszerezi ezeket, hogy kereshető legyen (kulcsszó).
- Mi a fontos, mi nem? Rangsorolja a találatokat. Hogyan rangsorol?



Keresőprogramok

1994: WebCrawler, go.com, Lycos, Infoseek
1995: Yahoo, AltaVista, Excite, SAPO

„A 1997 novemberében a 4 legnépszerűbb keresőprogram közül csak *egy* találja meg saját magát.”
- Brin, Page: The Anatomy of a Large-Scalce Hypertextual Web Search Engine (1998)

1998: Google

Google

- 1998-ban alapította Sergey Brin és Larry Page (PhD hallgatók a Stanfordon)

By Joi Ito, via Wikimedia Commons

By Marcin Mycielski, from Wikimedia Commons

- keresőprogram, lényegében a PageRank algoritmus
- googol 10^{100} (hatalmas adathalmazban keres)

Google 1998

Google! BETA

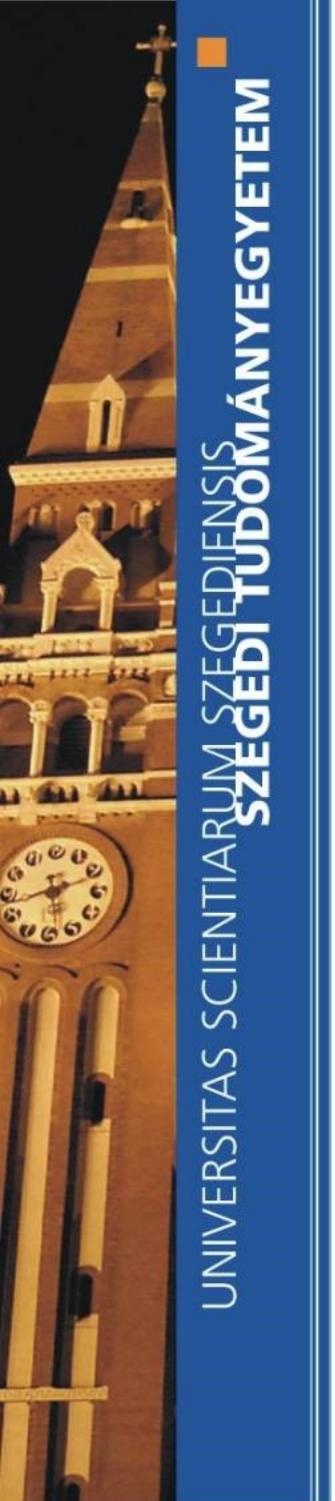
Search the web using Google!

Google Search | I'm feeling lucky

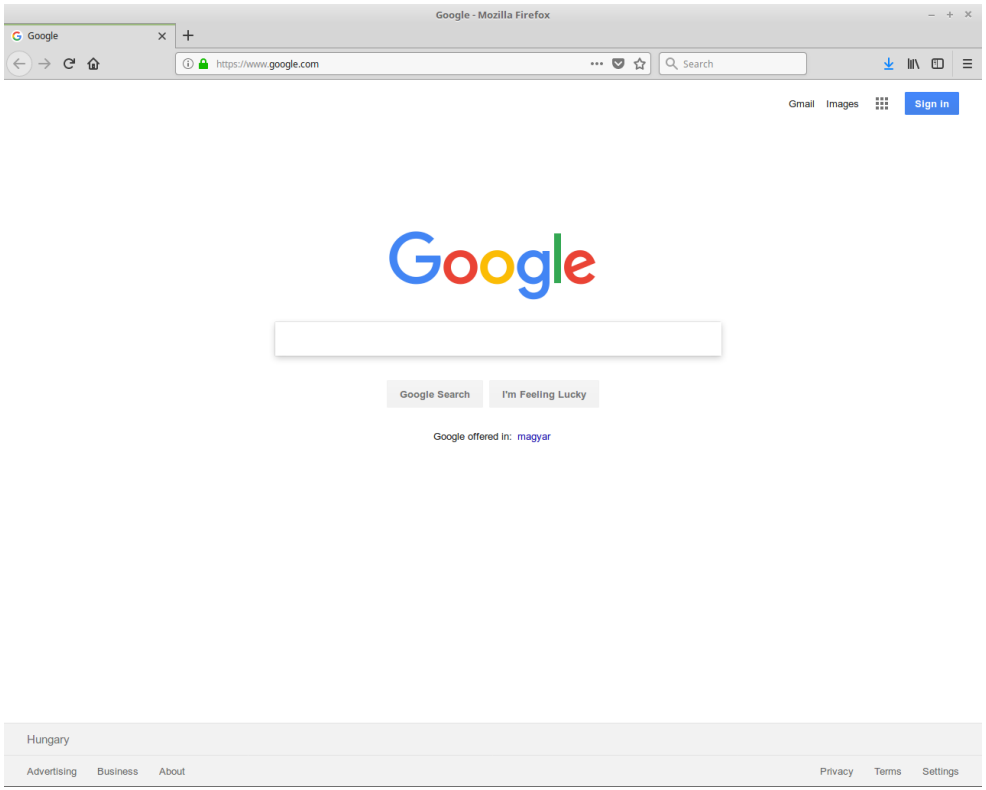
Special Searches
Stanford Search
Linux Search

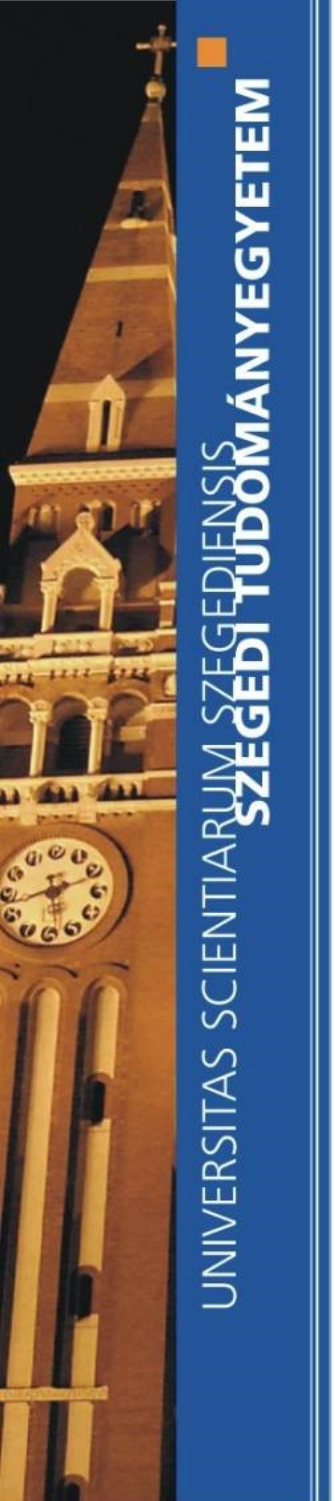
Help!
About Google!
Company Info
Google! Logos

Get Google! updates monthly:
your e-mail
Subscribe
Archive


UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM

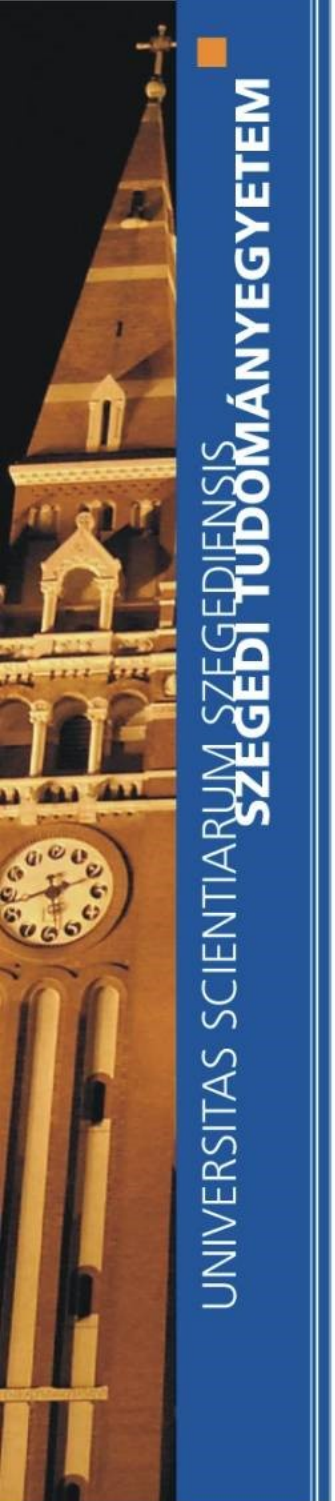
Google 2018


Google - Mozilla Firefox
Google
https://www.google.com
Search
Gmail
Images
Sign in
Google
Google Search
I'm Feeling Lucky
Google offered in: magyar
Hungary
Advertising
Business
About
Privacy
Terms
Settings

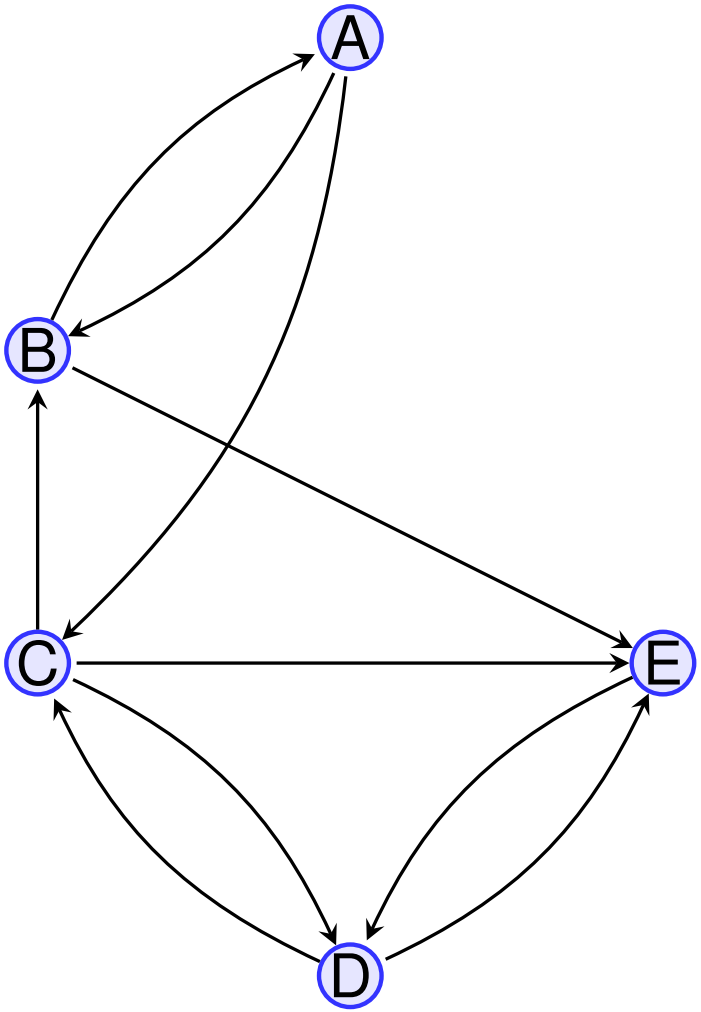


Rangsorolás

- Fontosság definiálása
- Objektív rangsor: a világháló szerkezetéből adódjon a rangsor

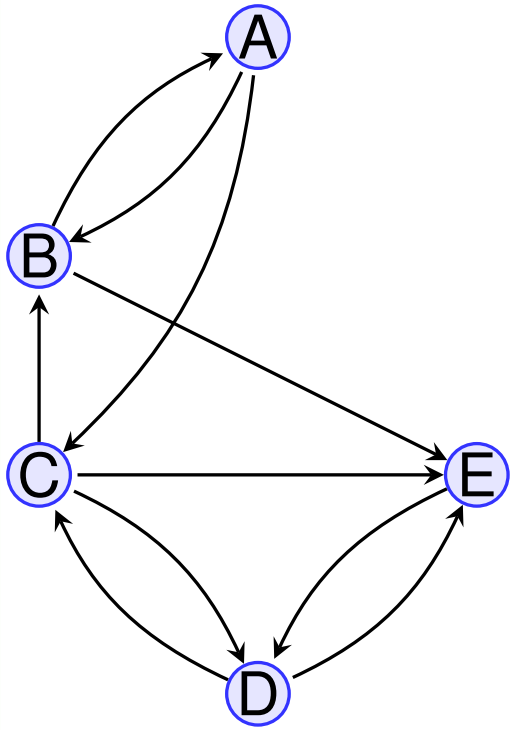


WWW modell



- Hivatkozó hiperlinkek száma? Nem jó, mert számít, hogy honnan jön a hivatkozás.
- $Pr(A) = Pr(B)$ mert A-ra van hiperlink B-ből? Nem jó, mert B fontossága többször szerepel.
- Osszuk el a fontosságot azok közt az oldalak közt, ahova van hiperlink!

Rangsor



$$Pr(A) = \frac{1}{2} Pr(B)$$

$$Pr(B) = \frac{1}{2} Pr(A) + \frac{1}{3} Pr(C)$$

$$Pr(C) = \frac{1}{2} Pr(A) + \frac{1}{2} Pr(D)$$

$$Pr(D) = \frac{1}{3} Pr(C) + 1 \cdot Pr(E)$$

$$Pr(E) = \frac{1}{2} Pr(B) + \frac{1}{3} Pr(C) + \frac{1}{2} Pr(D)$$

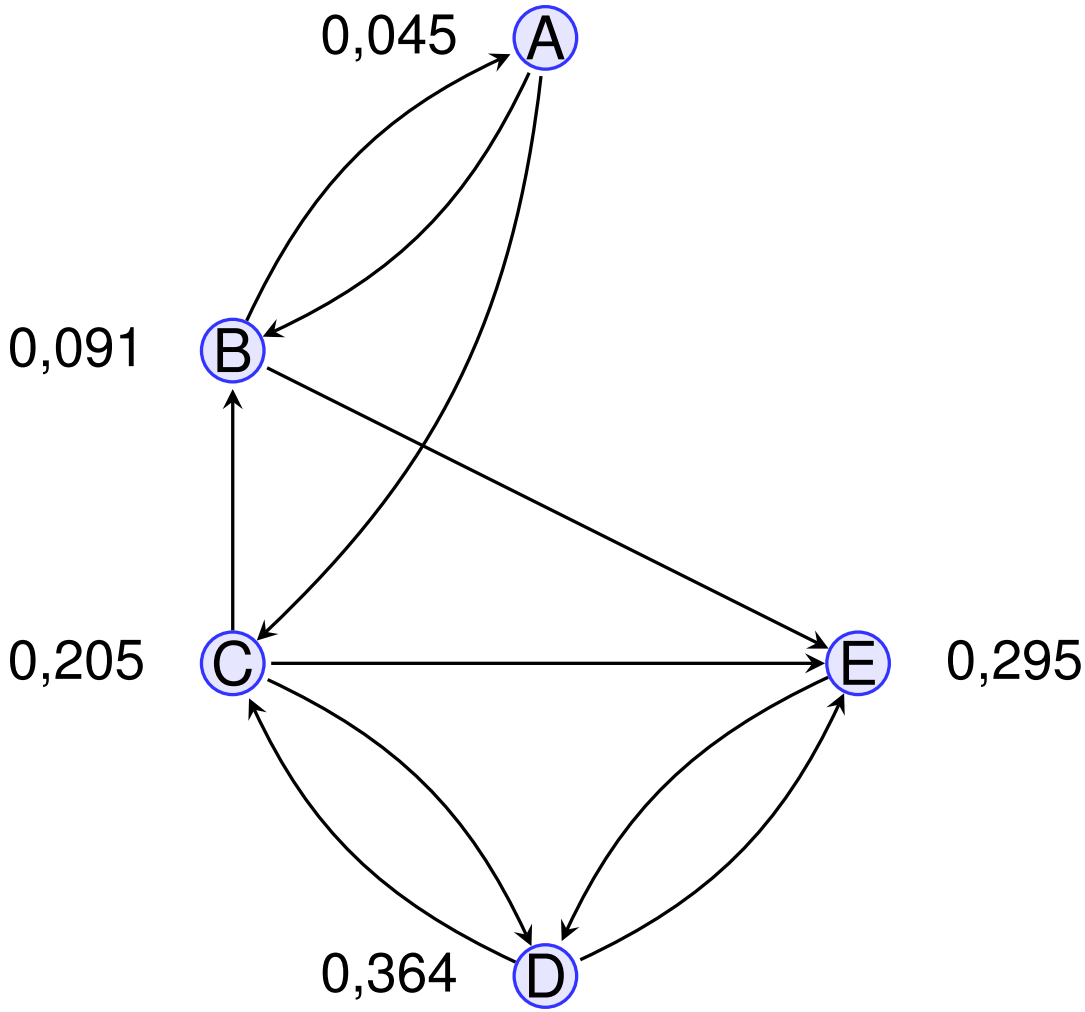
Egyenletrendszer - sajátérték, sajátvektor

$$\begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} Pr(A) \\ Pr(B) \\ Pr(C) \\ Pr(D) \\ Pr(E) \end{pmatrix} = \begin{pmatrix} Pr(A) \\ Pr(B) \\ Pr(C) \\ Pr(D) \\ Pr(E) \end{pmatrix}$$

Egy megoldás (ha mindet megszorzom 3-mal, akkor is megoldást kapok): $Pr(A) = 0,045$, $Pr(B) = 0,091$; $Pr(C) = 0,205$; $Pr(D) = 0,364$; $Pr(E) = 0,295$.

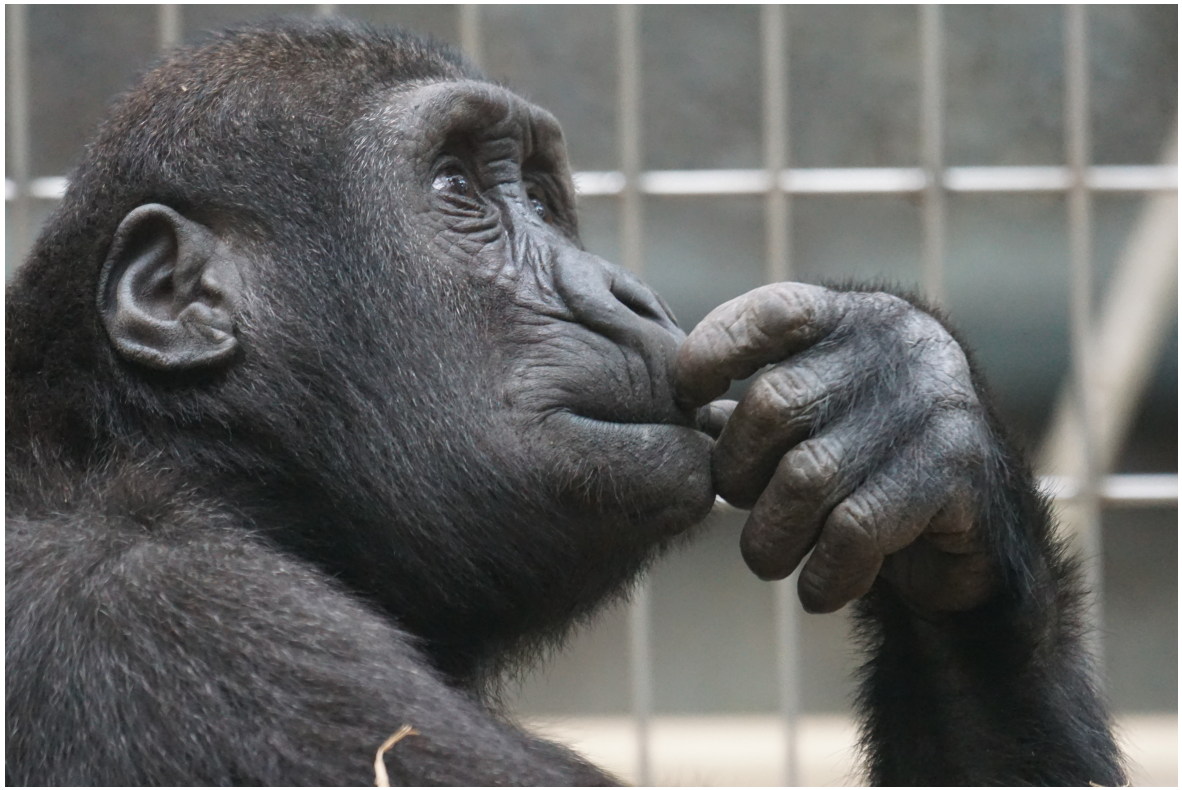

UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM

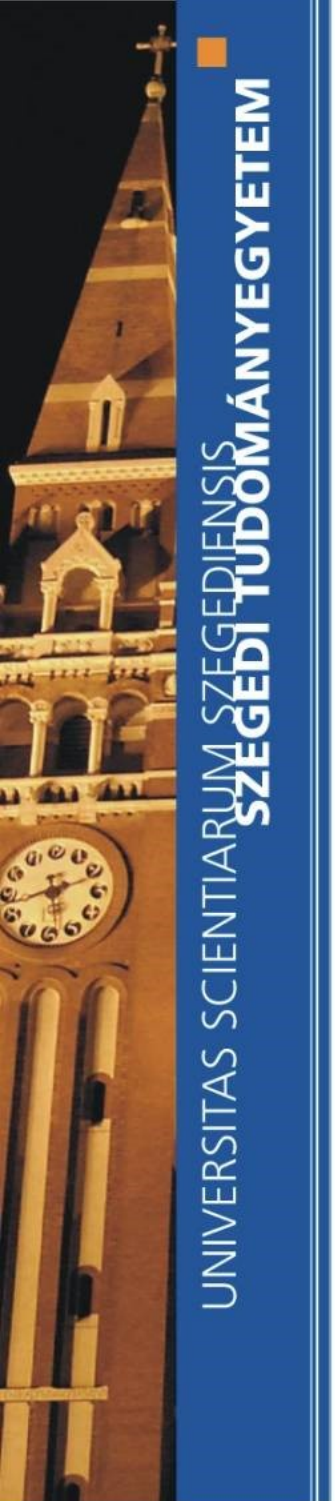
Rangsor


0,045
A
0,091
B
0,205
C
E
0,295
0,364
D

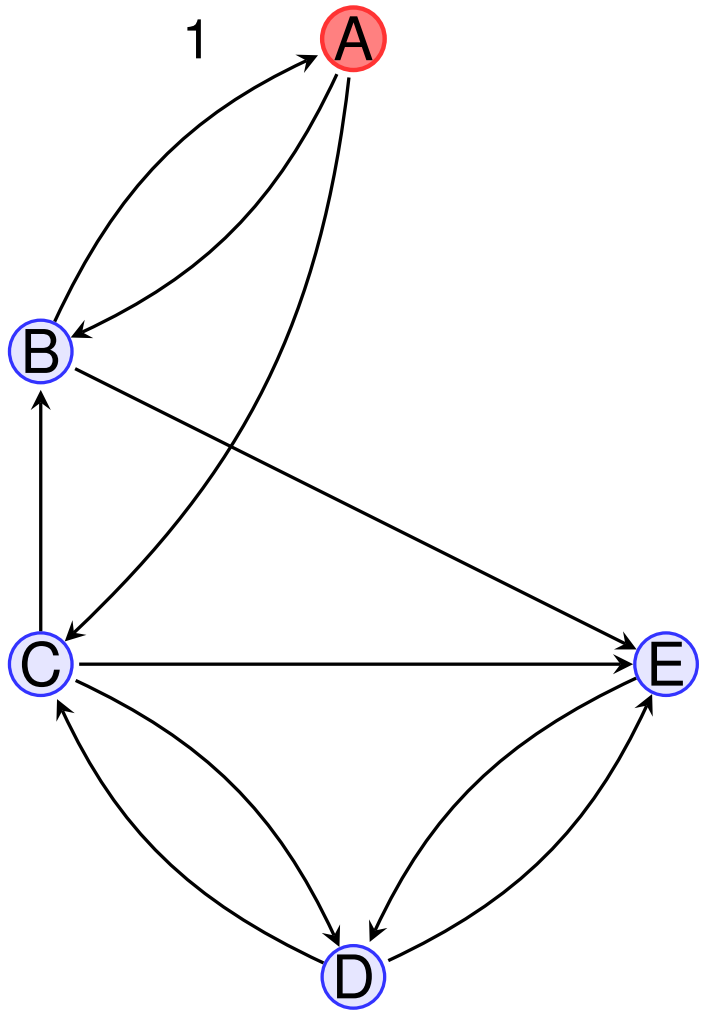
??

Na akkor most oldjunk meg egy 25 milliárd ismeretlenes egyenletrendszert?

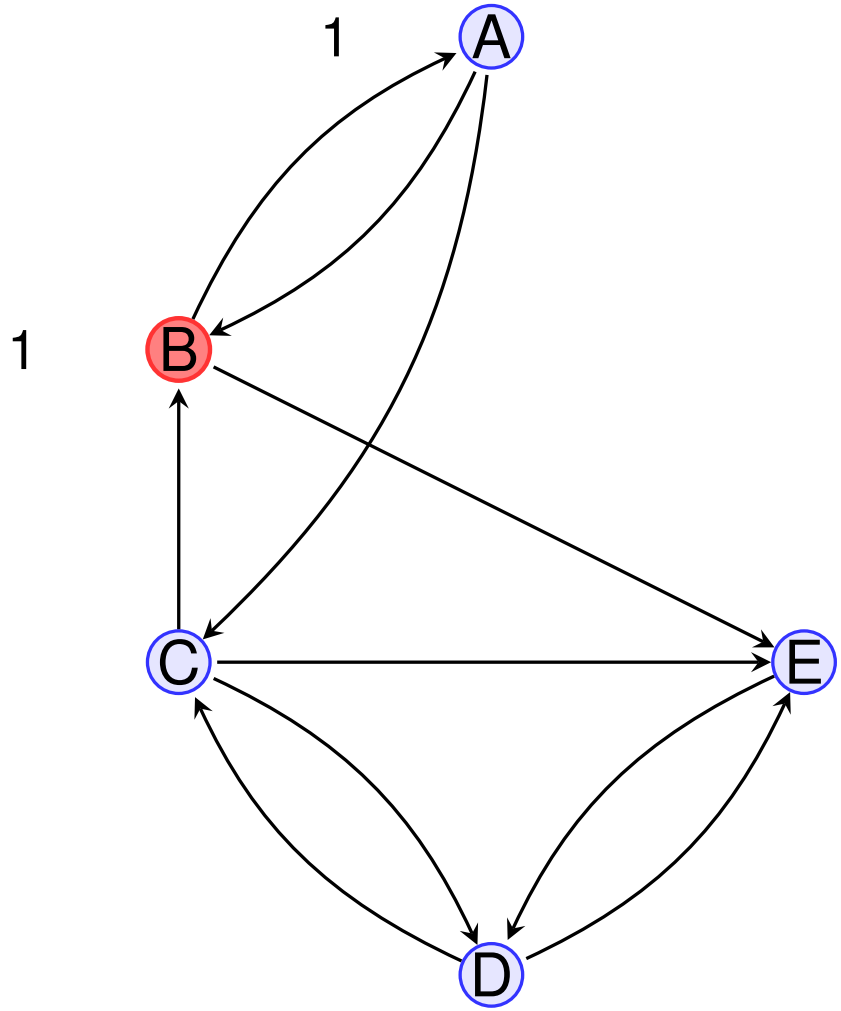




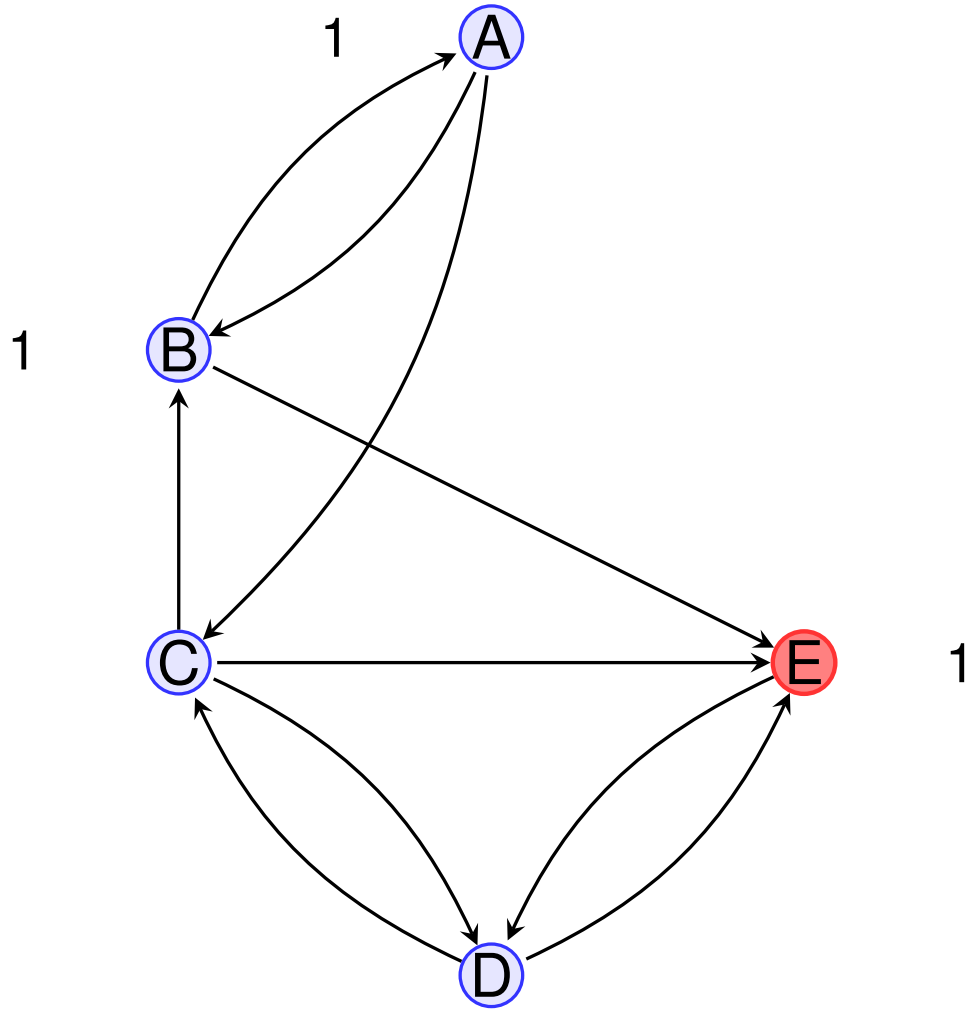
Véletlen szörfös - 1

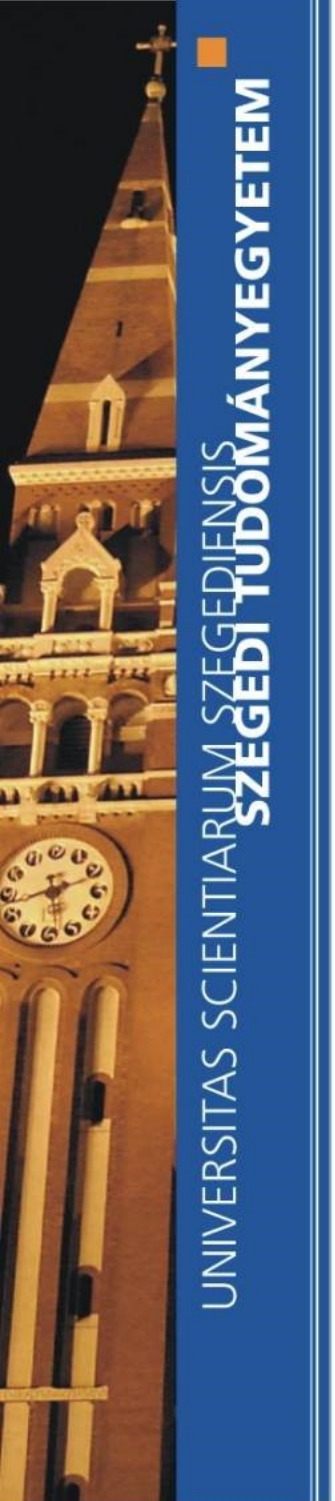


Véletlen szörfös - 2

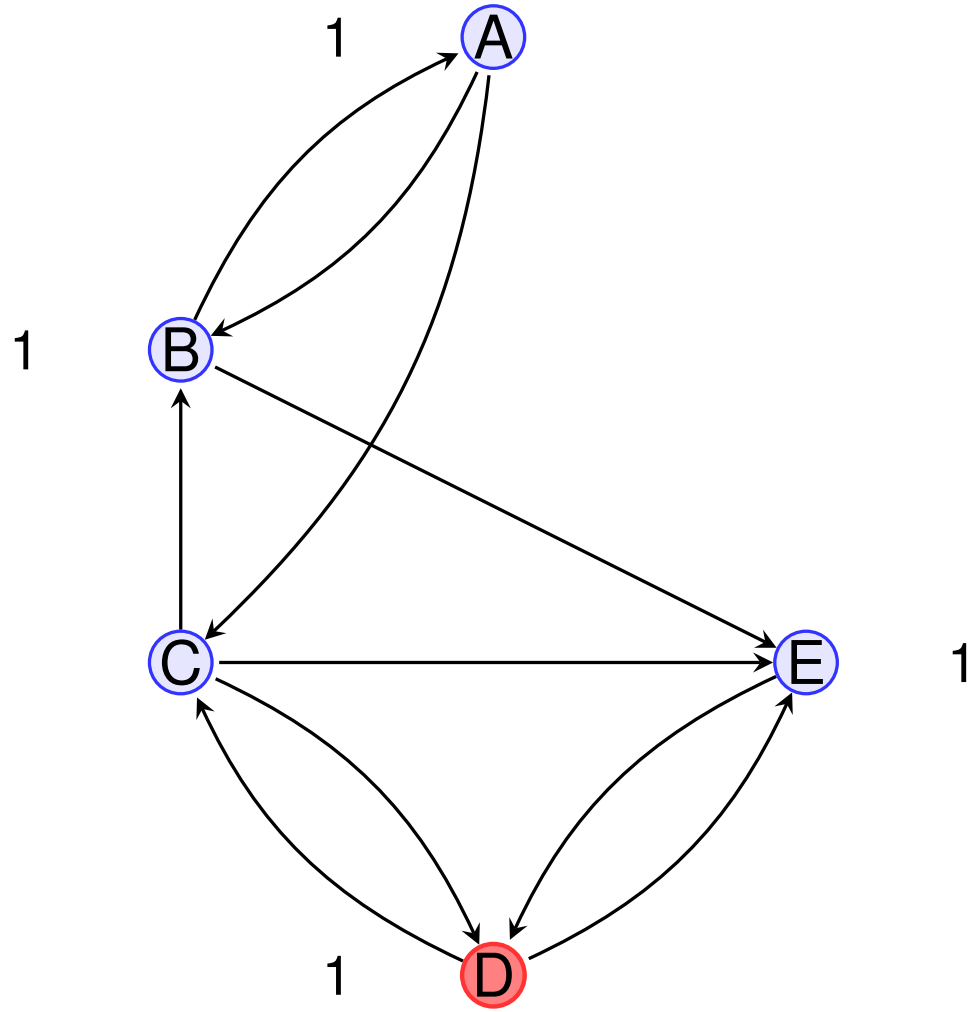


Véletlen szörfös - 3

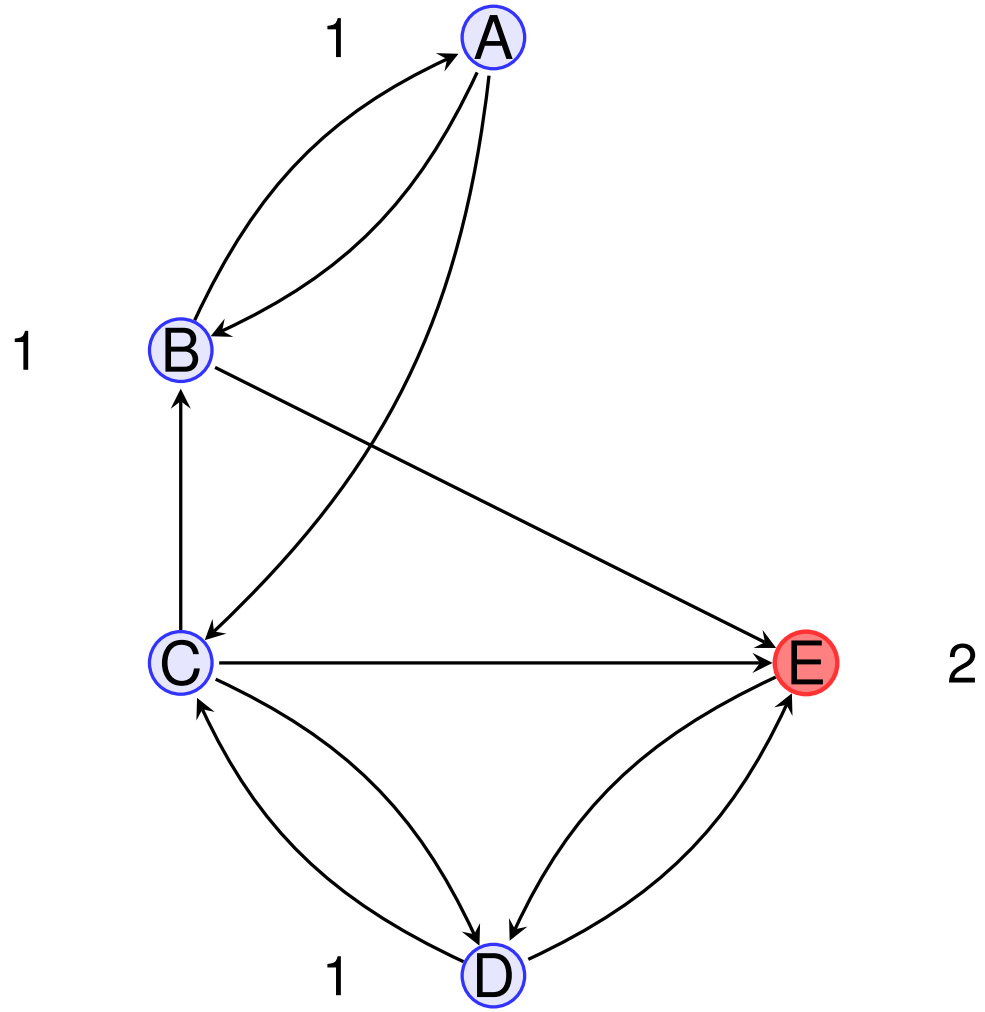




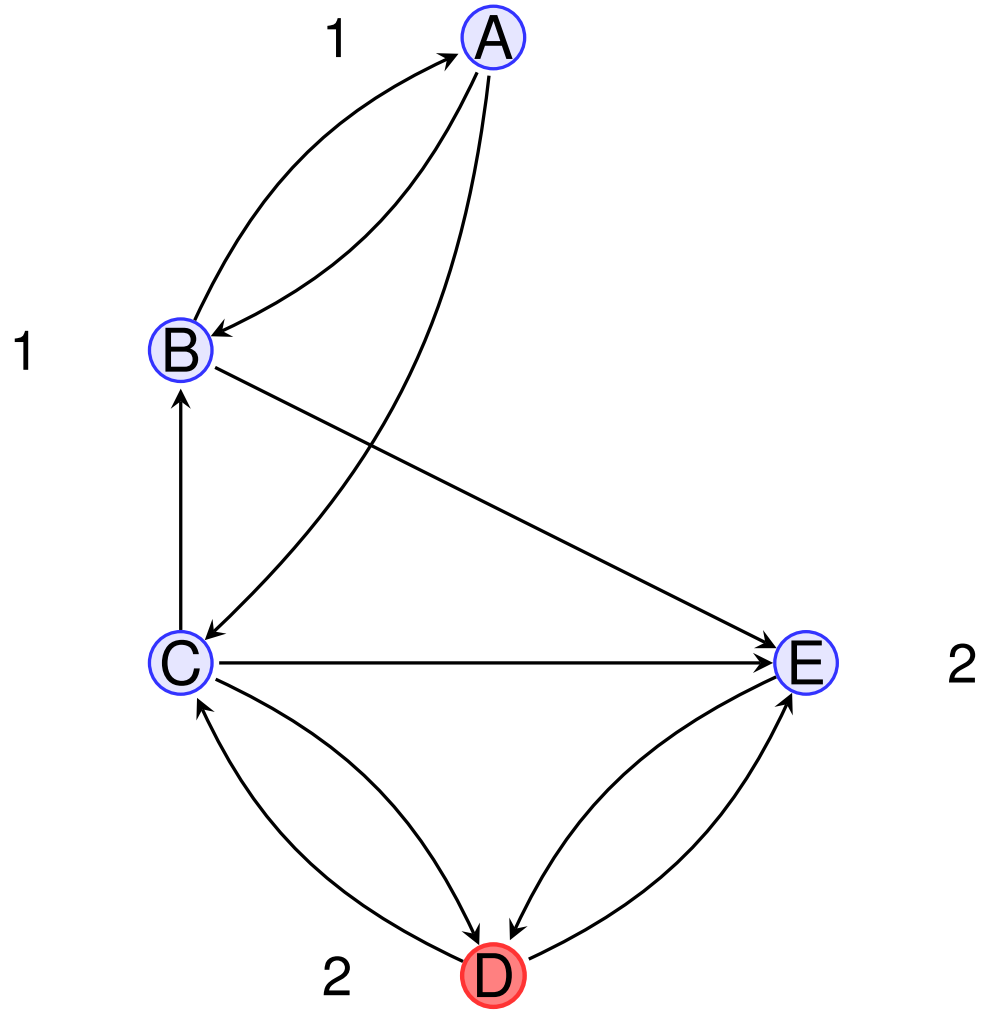
Véletlen szörfös - 4

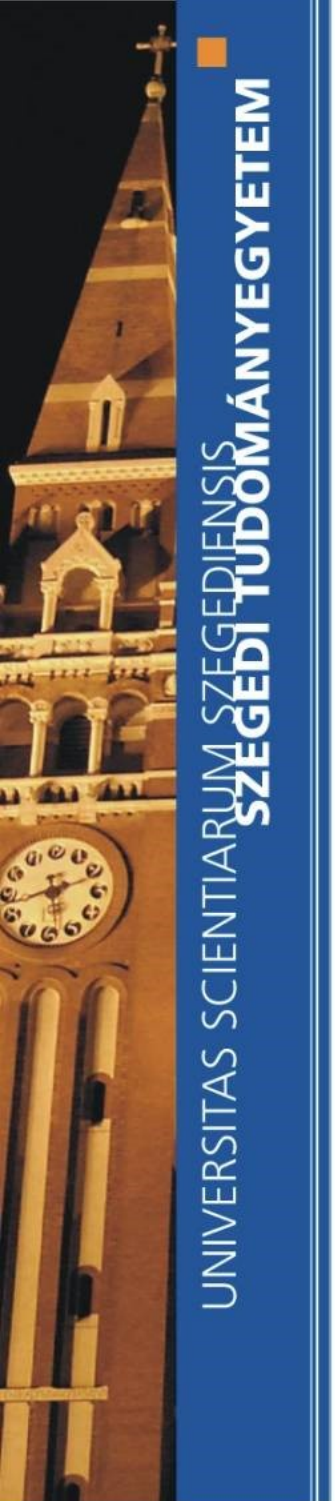


Véletlen szörfös - 5

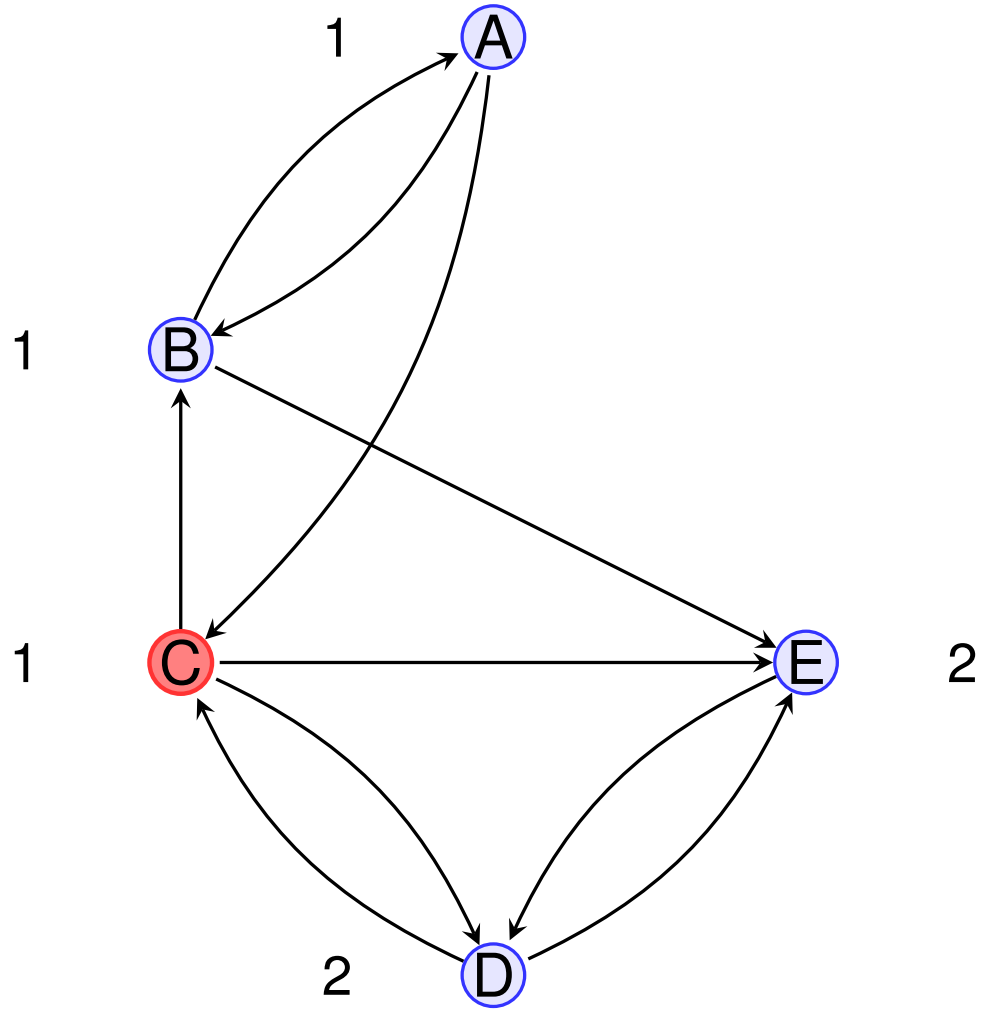


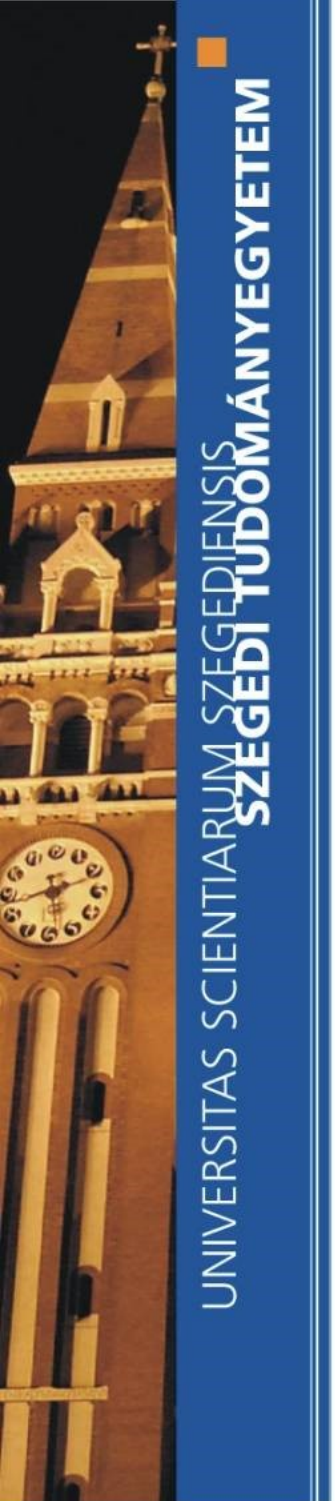
Véletlen szörfös - 6



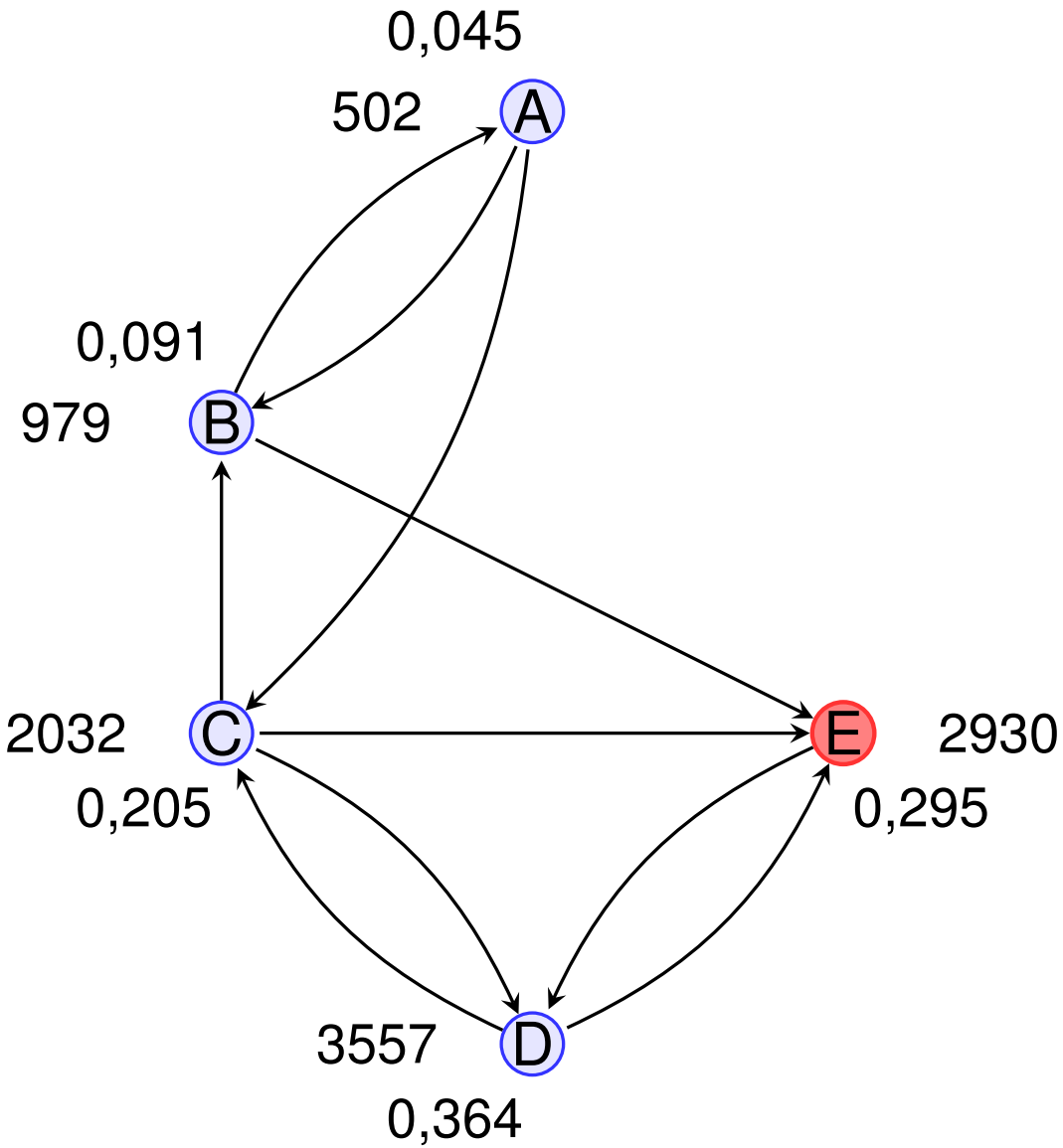


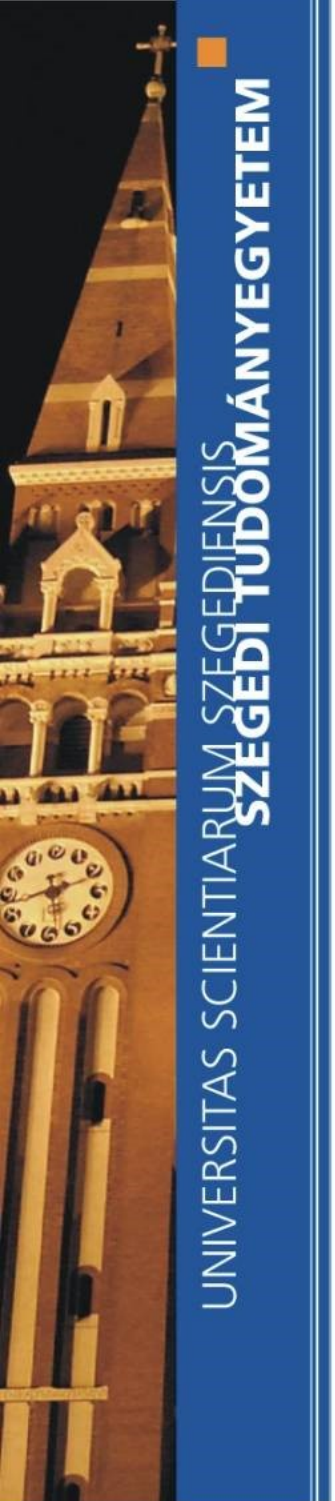
Véletlen szörfös - 7





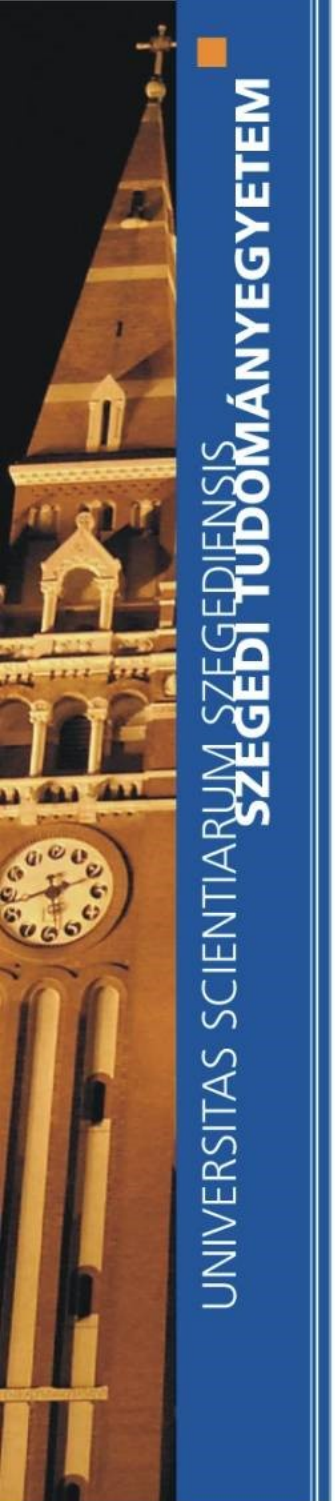
Véletlen szörfös - 10000





Markov-láncok

- A véletlen szörfös modell egy Markov-lánc, azaz a következő lépés véletlen, de csak attól függ, hogy éppen hol van a szörfös.
- Ha sokáig követjük a folyamatot (a szörföst), akkor beáll egy *stacionárius eloszlás*, ami azt mutatja, hogy az idő milyen részében volt az egyes állapotokban (weboldalakon). Pontosan ez az egyes weboldalak fontossága.
- A stacionárius eloszlás független attól, hogy honnan indult az első lépésben. Azaz a folyamat *ergodikus*.
- A PageRank iteratív módon is kiszámolható. Valójában csak így számolható ki. 50 iteráció elég (Brin & Page, 1998).



Iteráció - 1

Kezdeti érték: $a_0 = 1$, $b_0 = c_0 = d_0 = e_0 = 0$.

$$a_1 = \frac{1}{2} b_0 = 0$$

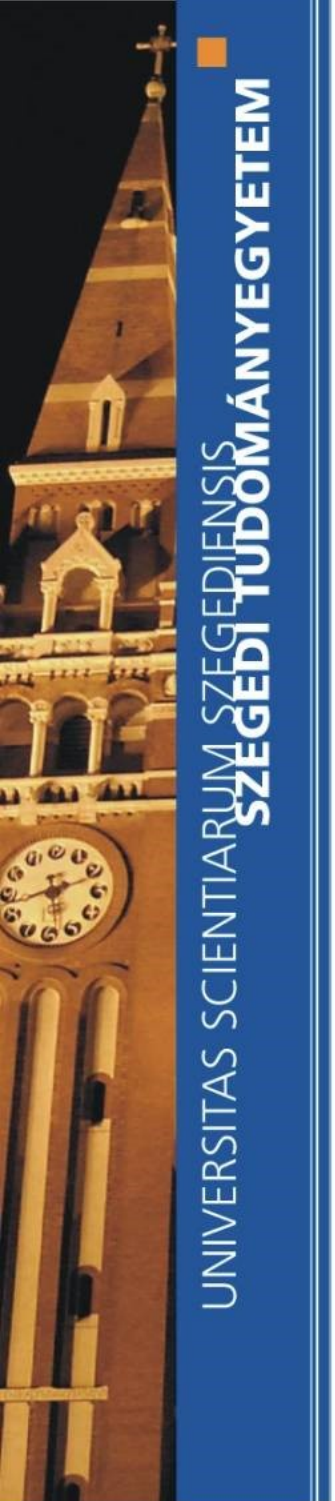
$$b_1 = \frac{1}{2} a_0 + \frac{1}{3} c_0 = 0.5$$

$$c_1 = \frac{1}{2} a_0 + \frac{1}{2} d_0 = 0.5$$

$$d_1 = \frac{1}{3} c_0 + 1 \cdot e_0 = 0$$

$$e_1 = \frac{1}{2} b_0 + \frac{1}{3} c_0 + \frac{1}{2} d_0 = 0$$

A megoldás: $(0,045, 0,091, 0,205, 0,364, 0,295)$.

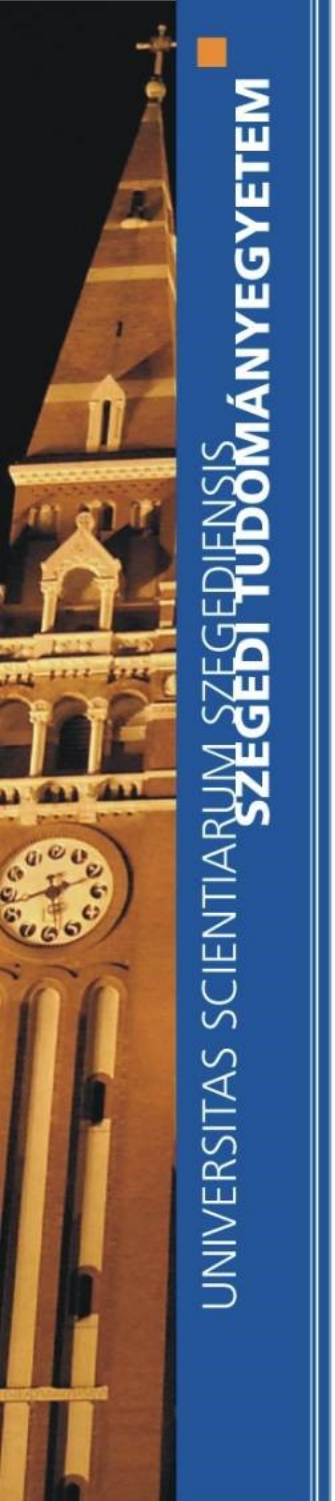


Iteráció - 2

Kezdeti érték: $a_1 = 0$, $b_1 = 0.5$, $c_1 = 0.5$, $d_1 = e_1 = 0$.

$$a_2 = \frac{1}{2} b_1 = 0,25$$
$$b_2 = \frac{1}{2} a_1 + \frac{1}{3} c_1 = 0,167$$
$$c_2 = \frac{1}{2} a_1 + \frac{1}{2} d_1 = 0$$
$$d_2 = \frac{1}{3} c_1 + 1 \cdot e_1 = 0,167$$
$$e_2 = \frac{1}{2} b_1 + \frac{1}{3} c_1 + \frac{1}{2} d_1 = 0,417$$

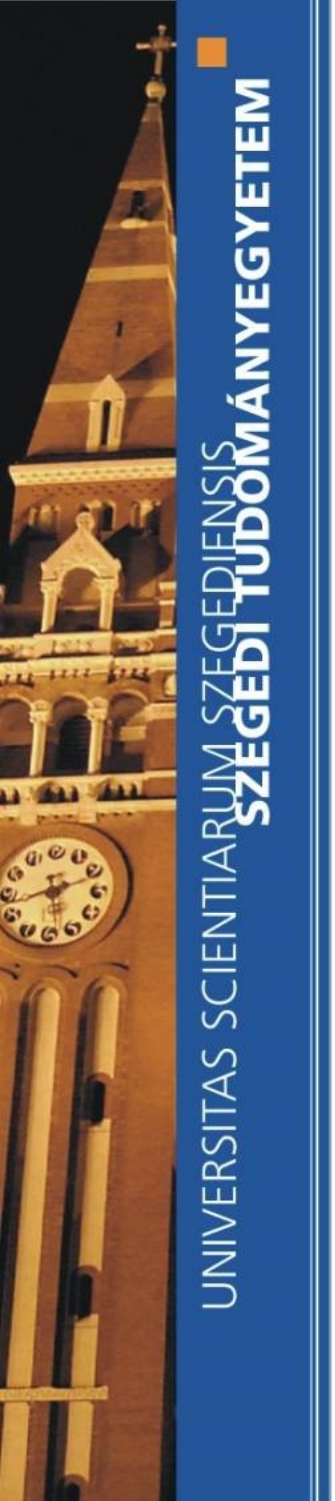
A megoldás: $(0,045, 0,091, 0,205, 0,364, 0,295)$.



Iteráció - 3

$$a_3 = \frac{1}{2} b_2 = 0,083$$
$$b_3 = \frac{1}{2} a_2 + \frac{1}{3} c_2 = 0,125$$
$$c_3 = \frac{1}{2} a_2 + \frac{1}{2} d_2 = 0,208$$
$$d_3 = \frac{1}{3} c_2 + 1 \cdot e_2 = 0,417$$
$$e_3 = \frac{1}{2} b_2 + \frac{1}{3} c_2 + \frac{1}{2} d_2 = 0,167$$

A megoldás: $(0,045, 0,091, 0,205, 0,364, 0,295)$.



Iteráció - 10

$$a_{10} = \frac{1}{2} b_9 = 0,005$$

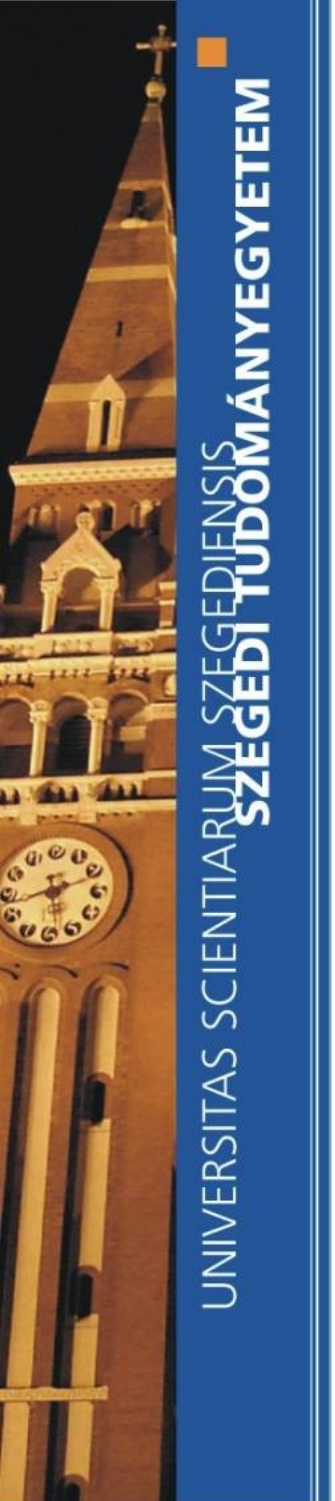
$$b_{10} = \frac{1}{2} a_9 + \frac{1}{3} c_9 = 0,084$$

$$c_{10} = \frac{1}{2} a_9 + \frac{1}{2} d_9 = 0,214$$

$$d_{10} = \frac{1}{3} c_9 + 1 \cdot e_9 = 0,344$$

$$e_{10} = \frac{1}{2} b_9 + \frac{1}{3} c_9 + \frac{1}{2} d_9 = 0,308$$

A megoldás: $(0,045, 0,091, 0,205, 0,364, 0,295)$.



Iteráció - 20

$$a_{20} = \frac{1}{2} b_{19} = 0,045$$

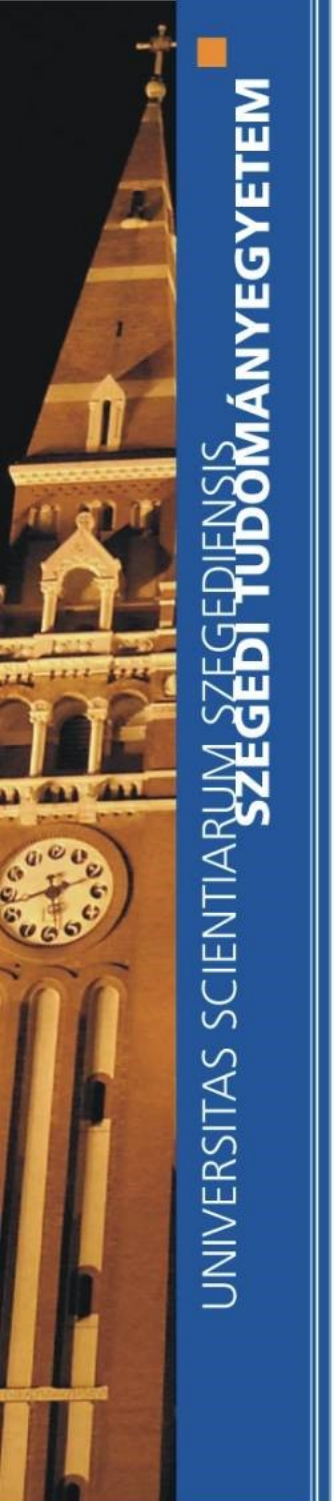
$$b_{20} = \frac{1}{2} a_{19} + \frac{1}{3} c_{19} = 0,091$$

$$c_{20} = \frac{1}{2} a_{19} + \frac{1}{2} d_{19} = 0,205$$

$$d_{20} = \frac{1}{3} c_{19} + 1 \cdot e_{19} = 0,363$$

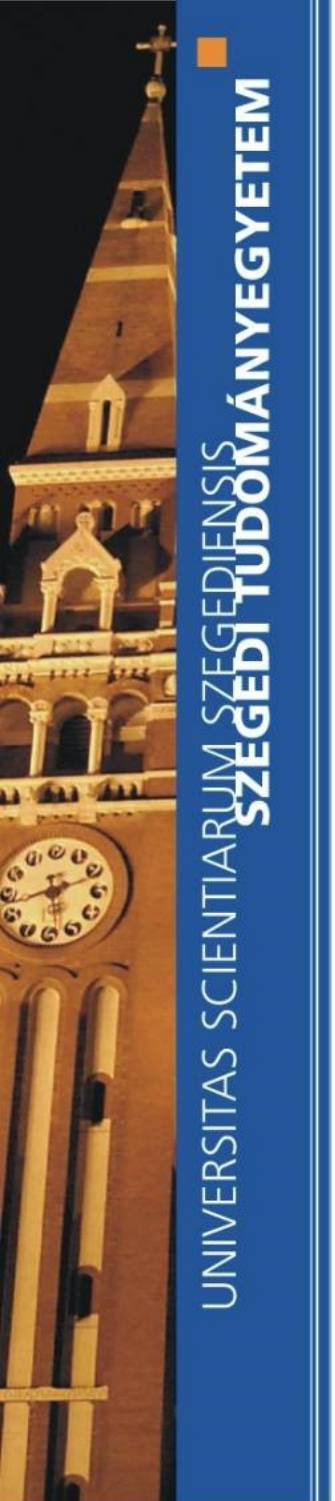
$$e_{20} = \frac{1}{2} b_{19} + \frac{1}{3} c_{19} + \frac{1}{2} d_{19} = 0,296$$

A megoldás: $(0,045, 0,091, 0,205, 0,364, 0,295)$.

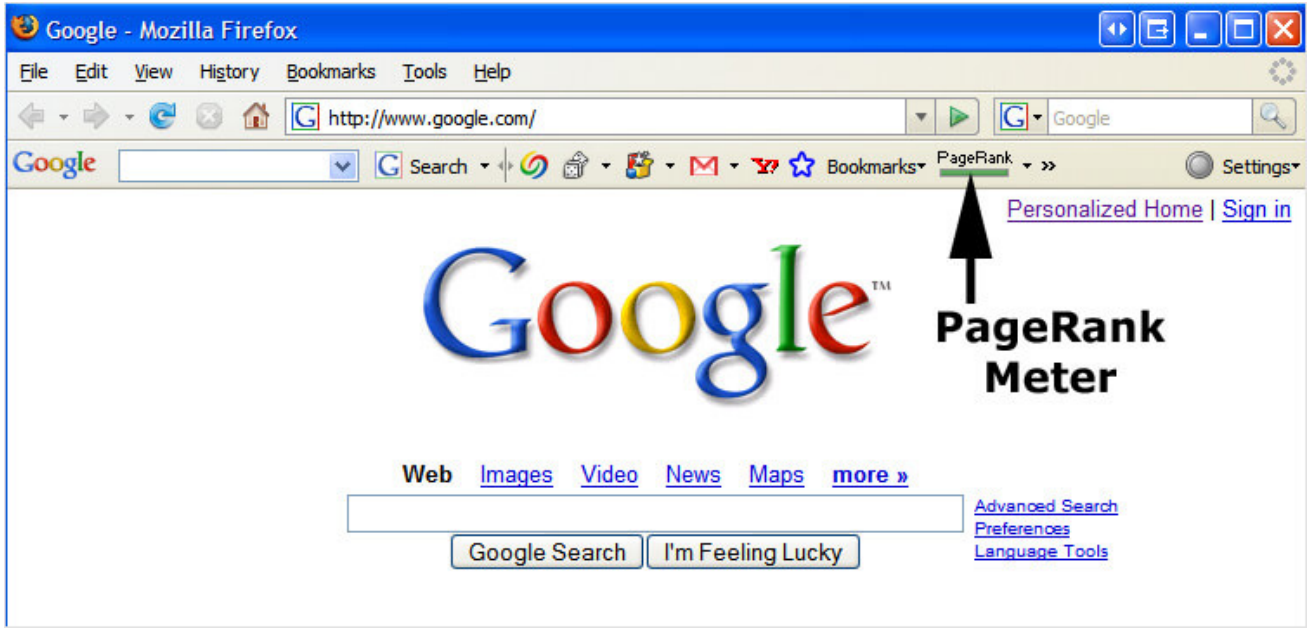


Csak a PageRank!

- Minél fontosabb egy oldal, annál nagyobb a PageRank értéke.
- Minél nagyobb a PageRank, annál fontosabb az oldal.
- Google Toolbar 1.0 (2000. december) az oldalak PageRank értéke (valamiféle közelítése) nyilvános.
- Hogy lehet becsapni az algoritmust?



Google Toolbar - PageRank



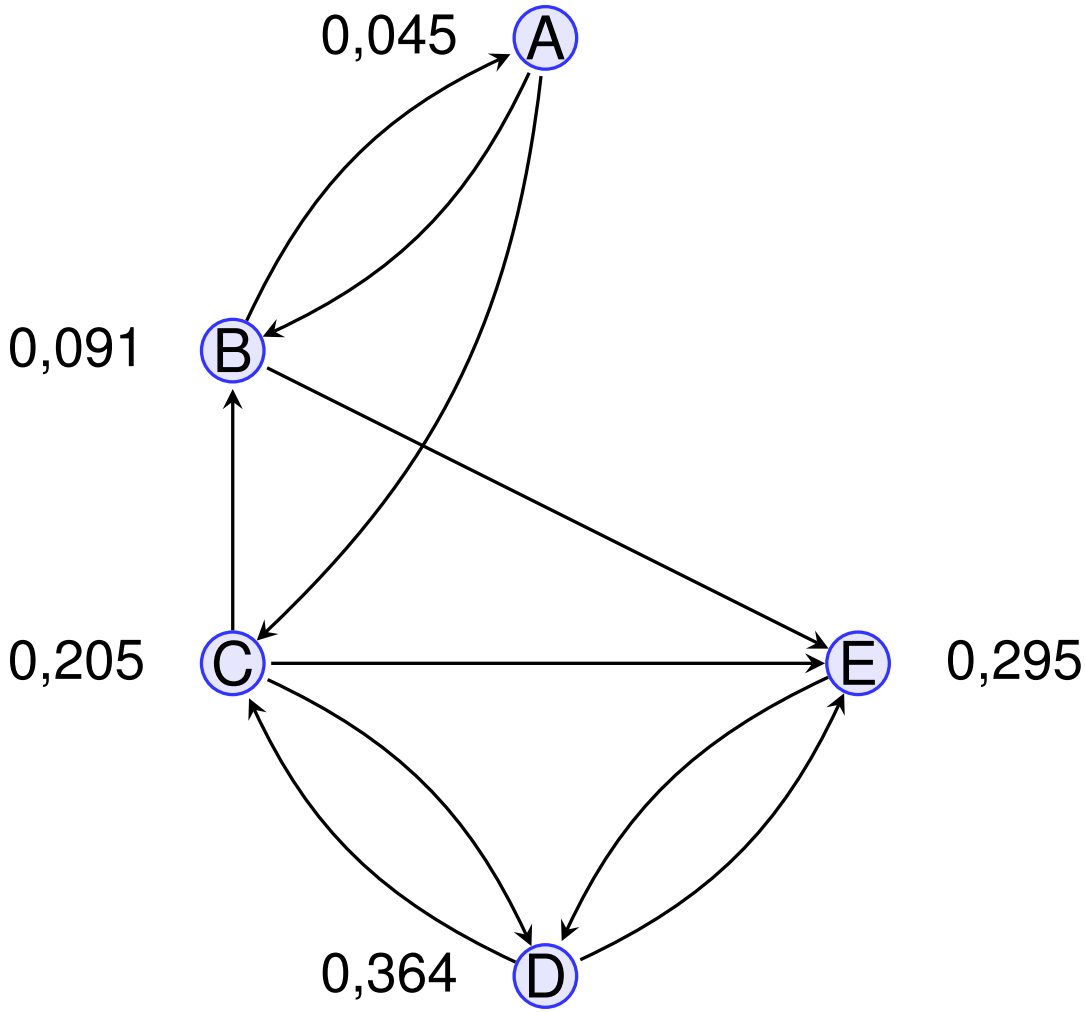
2000. december (Google Toolbar 1.0)

1 - 10 (logaritmikus) skálán mutatja a PageRank értéket

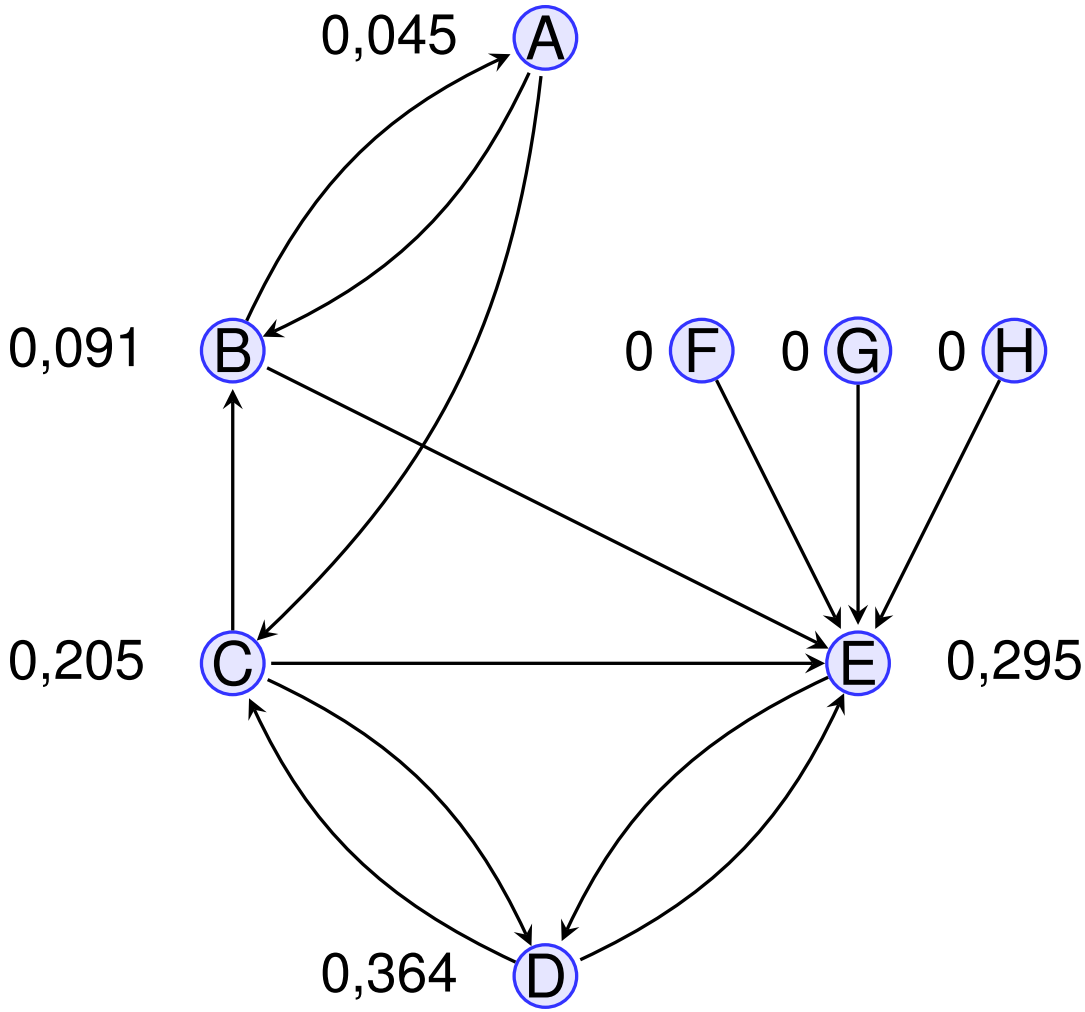
??


UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM

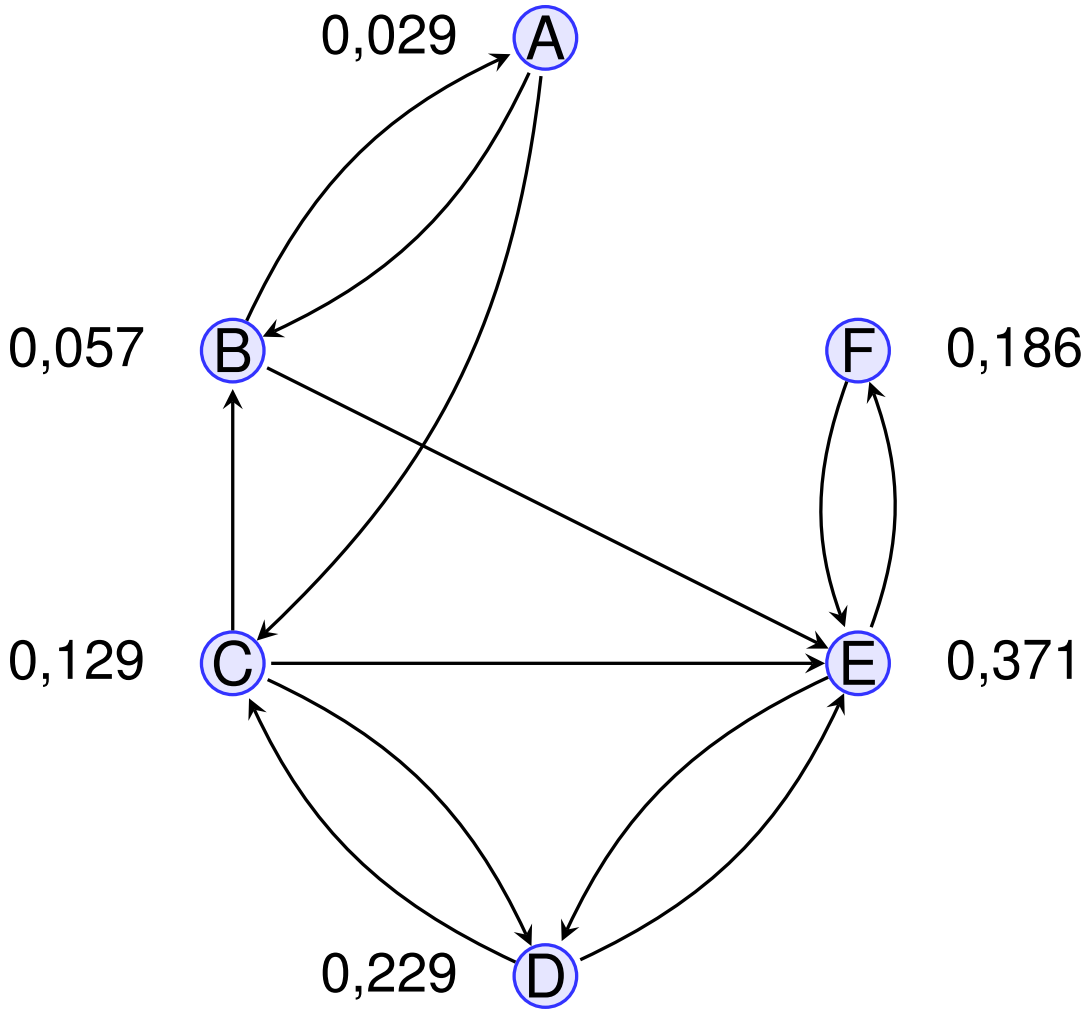
Rangsor

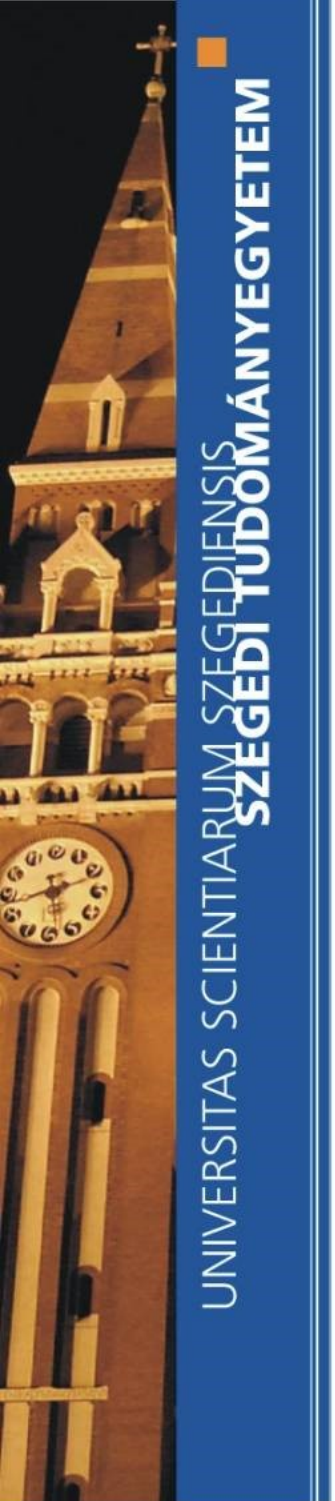

0,045
A
0,091
B
0,205
C
E
0,295
0,364
D

Rangsor2 - E újragondolja



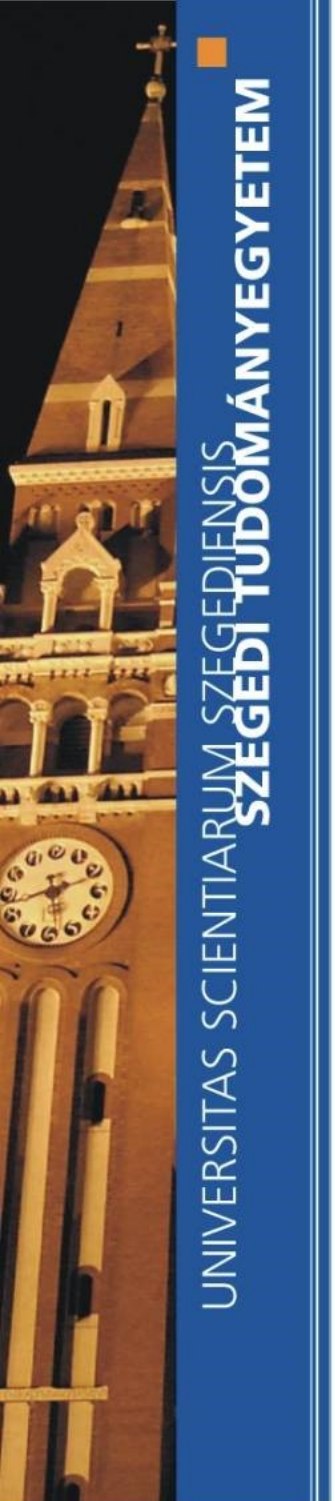
Rangsor3 - E jobban újragondolja





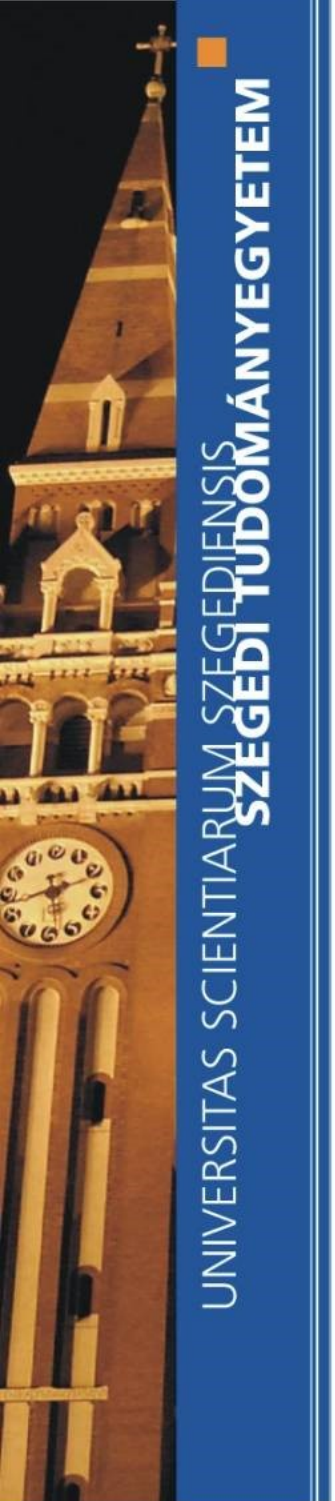
Manipulálás vs. objektív rangsor

- Linkfarmok (egymásra mutató oldalak, melyeken nincs információ).
- Spam linkek (megjegyzésekben szereplő linkek).
- Ezeket szűri (nofollow), bünteti az algoritmus.
- Google Toolbarban 2016. áprilisától nincs PageRank mérő.
- A PageRank algoritmust továbbra is használják.



PageRank általános algoritmus gráfokon, sok alkalmazással

- Twitter
- neuronhálózatok (orvostudomány)
- proteinhálózatok
- úthálózatok
- nyelvészet (szavak közötti kapcsolatok)
- tudománymetria (újságok fontossági sorrendje): SCImago Journal Ranking.



Irodalomjegyzék

Sergey Brin, Lawrence Page:
The Anatomy of a Large-Scale Hypertextual Web Search Engine.
Seventh International World-Wide Web Conference (WWW 1998)

Kurt Bryan, Tanya Leise:
The $25,000,000,000 eigenvector. The linear algebra behind Google.

David Austin:
How Google Finds Your Needle in the Web's Haystack?
AMS Feature Column, 2006.