Priveľa informácie škodí

Michal „Mišof“ Forišek

Department of Theoretical Computer Science
Faculty of Mathematics, Physics and Informatics
Comenius University
Bratislava, Slovakia

10. decembra 2010

Vyhľadávanie v bežnom texte

Bežná webstránka má do 10 kB *textu*.
Ctrl-F (find) v prehliadači funguje takmer okamžite.

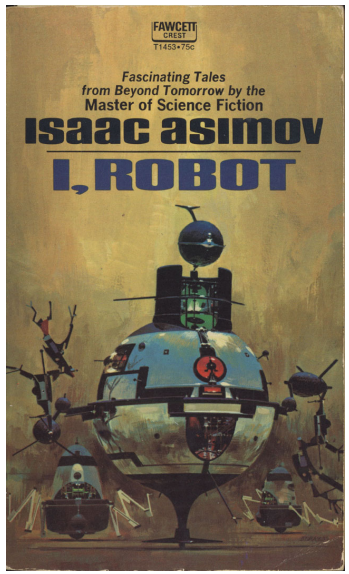
Kniha *Isaac Asimov: Ja, robot* má 425 kB.
Vyhľadanie všetkých výskytov
slova „`robot`": 0.007 s.

Kým máme jedného človeka a málo textu,
všetko funguje ako má.

Vyhľadávanie v bežnom texte

Bežná webstránka má do 10 kB *textu*.
Ctrl-F (find) v prehliadači funguje takmer okamžite.

Kniha *Isaac Asimov: Ja, robot* má 425 kB.
Vyhľadanie všetkých výskytov
slova „`robot`": 0.007 s.

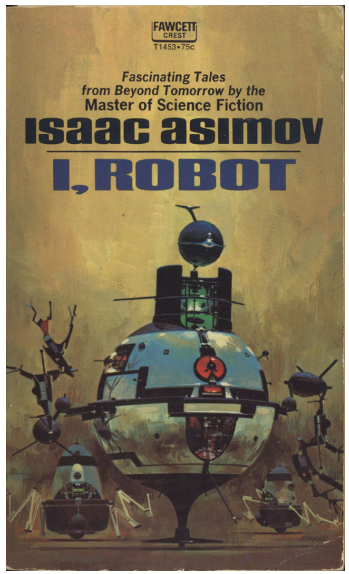


Kým máme jedného človeka a málo textu,
všetko funguje ako má.

Vyhľadávanie v bežnom texte

Bežná webstránka má do 10 kB *textu*.
Ctrl-F (find) v prehliadači funguje takmer okamžite.

Kniha *Isaac Asimov: Ja, robot* má 425 kB.
Vyhľadanie všetkých výskytov
slova „`robot`": 0.007 s.



Kým máme jedného človeka a málo textu,
všetko funguje ako má.

Množstvo informácie rastie

- Disketa: posledné bežné mali **1.44 MB**
- CD: typicky do **700 MB**
  (plné auto kníh)
- DVD: najrozšírenejšie majú **4.7 GB**
- BluRay disky: bežne **50 GB**
  Video v kvalite HD: **28 Mb/s**

  Inými slovami: každú sekundu 3.5 megabajtu údajov.
  (Každé 4 desatiny sekundy meniť disketu?)

- Nový pevný disk: **1 TB**

Množstvo informácie rastie

- Disketa: posledné bežné mali **1.44 MB**
- CD: typicky do **700 MB**
  (plné auto kníh)
- DVD: najrozšírenejšie majú **4.7 GB**
- BluRay disky: bežne **50 GB**
  Video v kvalite HD: **28 Mb/s**

  Inými slovami: každú sekundu 3.5 megabajtu údajov.
  (Každé 4 desatiny sekundy meniť disketu?)

- Nový pevný disk: **1 TB**

Množstvo informácie rastie

- Disketa: posledné bežné mali **1.44 MB**
- CD: typicky do **700 MB**
  (plné auto kníh)
- DVD: najrozšírenejšie majú **4.7 GB**
- BluRay disky: bežne **50 GB**
  Video v kvalite HD: **28 Mb/s**

  Inými slovami: každú sekundu 3.5 megabajtu údajov.
  (Každé 4 desatiny sekundy meniť disketu?)

- Nový pevný disk: **1 TB**

Množstvo informácie rastie

- Disketa: posledné bežné mali **1.44 MB**
- CD: typicky do **700 MB**
  (plné auto kníh)
- DVD: najrozšírenejšie majú **4.7 GB**
- BluRay disky: bežne **50 GB**
  Video v kvalite HD: **28 Mb/s**

  Inými slovami: každú sekundu 3.5 megabajtu údajov.
  (Každé 4 desatiny sekundy meniť disketu?)

- Nový pevný disk: **1 TB**

Množstvo informácie rastie

- Disketa: posledné bežné mali **1.44 MB**
- CD: typicky do **700 MB**
  (plné auto kníh)
- DVD: najrozšírenejšie majú **4.7 GB**
- BluRay disky: bežne **50 GB**
  Video v kvalite HD: **28 Mb/s**

  Inými slovami: každú sekundu 3.5 megabajtu údajov.
  (Každé 4 desatiny sekundy meniť disketu?)

- Nový pevný disk: **1 TB**

Množstvo informácie rastie

- Disketa: posledné bežné mali **1.44 MB**
- CD: typicky do **700 MB**
  (plné auto kníh)
- DVD: najrozšírenejšie majú **4.7 GB**
- BluRay disky: bežne **50 GB**
  Video v kvalite HD: **28 Mb/s**

  Inými slovami: každú sekundu 3.5 megabajtu údajov.
  (Každé 4 desatiny sekundy meniť disketu?)

- Nový pevný disk: **1 TB**

Množstvo informácie rastie ešte viac

americká Library of Congress: **50 TB** informácií

Množstvo informácie rastie ešte viac

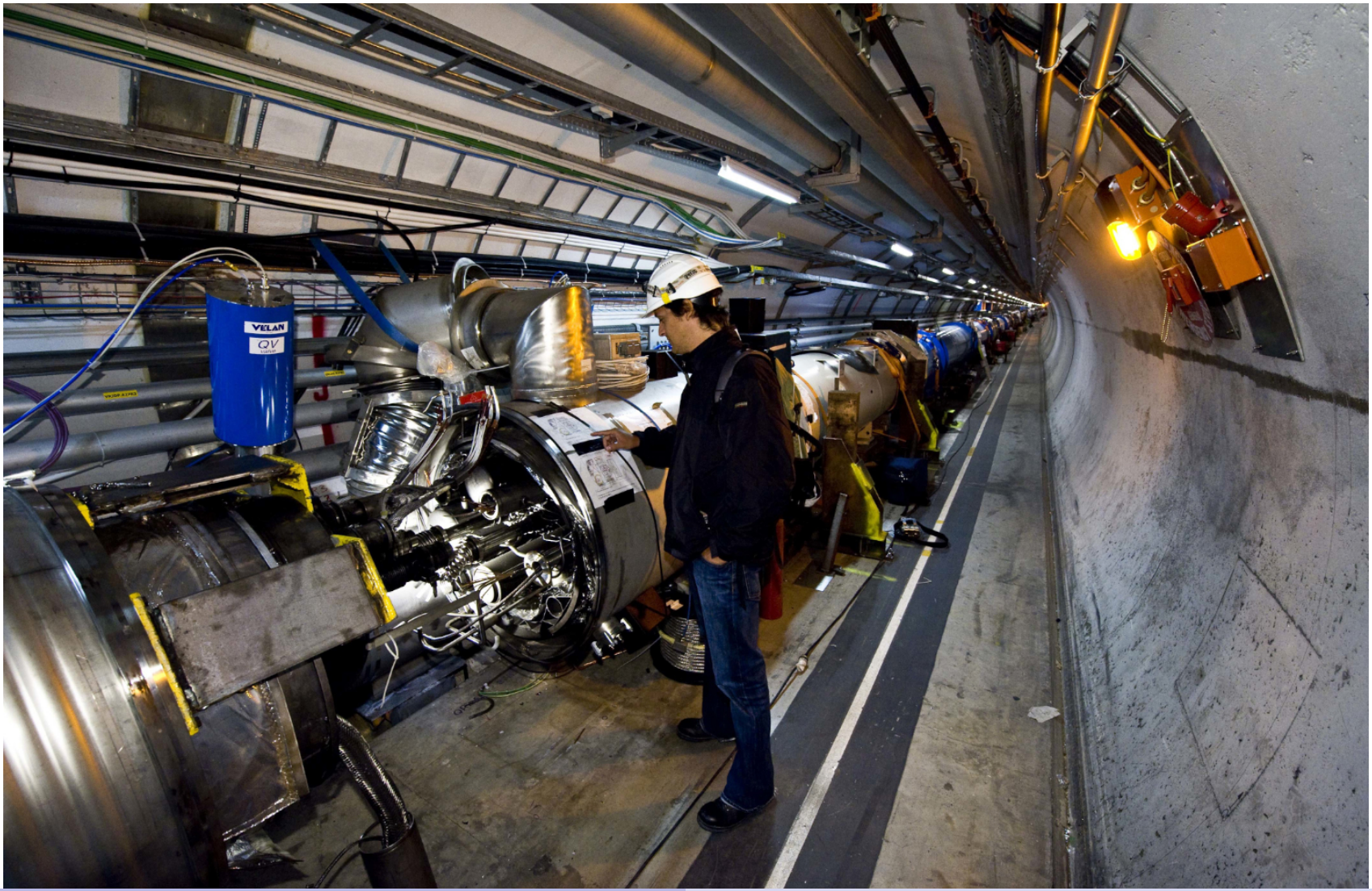
špeciálne efekty pre film *Avatar*: **1 PB** = 1000 TB

Množstvo informácie rastie ešte viac

hra *World of Warcraft*: **1.3 PB**

Množstvo informácie rastie ešte viac

fyzikálny experiment LHC: získame **15 PB** dát za rok



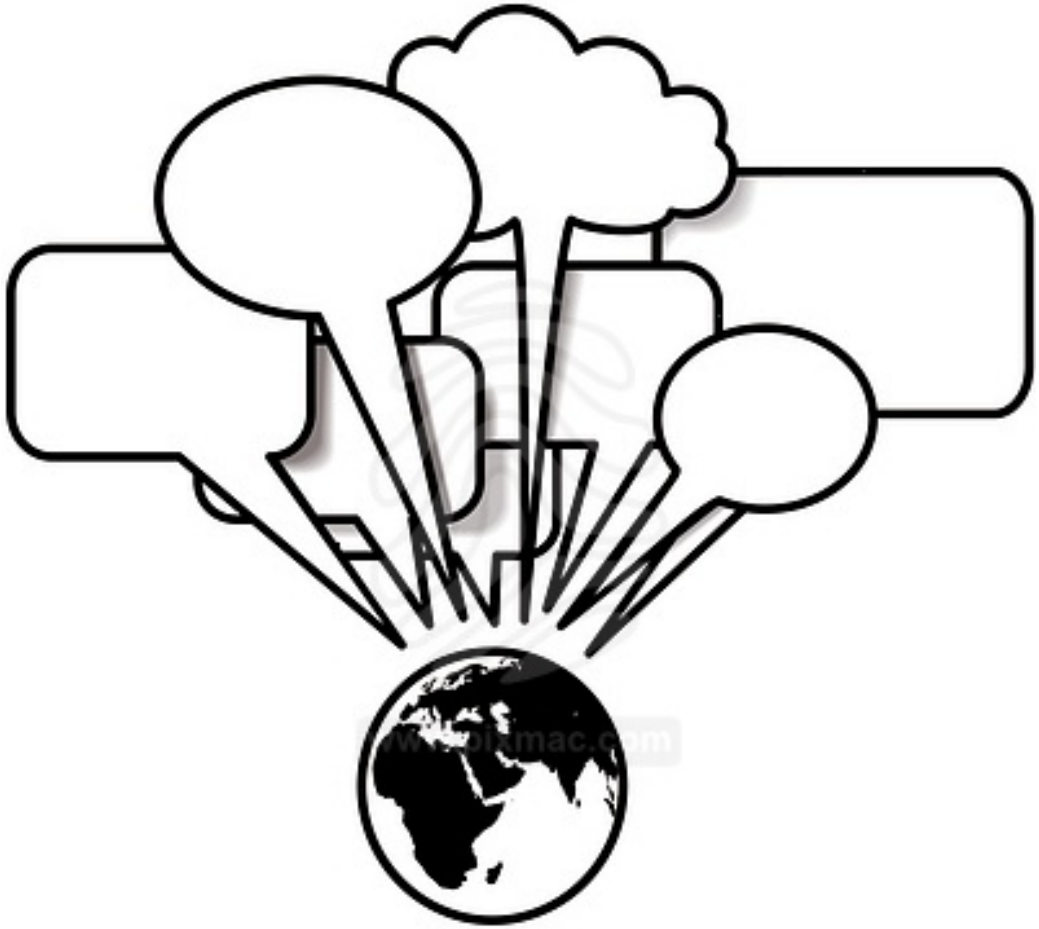
Množstvo informácie rastie ešte viac

Google: spracuje **24 PB** dát denne

(celé dáta používané pri vyhľadávaní majú vyše 100 PB)

Množstvo informácie rastie ešte viac

Všetky slová, ktoré kto kedy povedal: **5 EB = 5000 PB**



Web rastie

- 1998: Založené Google, pozná 26 miliónov stránok
- 2008: Google pozná **bilión** (10^{12}) rôznych URL, zodpovedajú odhadom **20 miliardám** unikátnych webstránok.

Ako dlho by trvalo nájsť výskyty slova „`robot`" na celom webe?

rátajme spolu:
20 miliárd stránok $\times$ trebárs 1 kB $=$ 20 TB textu.
to je zhruba 50 miliónov krát viac ako Asimovova kniha
bude to trvať 50 miliónov krát dlhšie
$\Rightarrow$ niekoľko dní

Web rastie

- 1998: Založené Google, pozná 26 miliónov stránok
- 2008: Google pozná **bilión** (10^{12}) rôznych URL, zodpovedajú odhadom **20 miliardám** unikátnych webstránok.

Ako dlho by trvalo nájsť výskyty slova „`robot`" na celom webe?

rátajme spolu:
20 miliárd stránok $\times$ trebárs 1 kB $=$ 20 TB textu.
to je zhruba 50 miliónov krát viac ako Asimovova kniha
bude to trvať 50 miliónov krát dlhšie
$\Rightarrow$ niekoľko dní

Web rastie

- 1998: Založené Google, pozná 26 miliónov stránok
- 2008: Google pozná **bilión** (10^{12}) rôznych URL, zodpovedajú odhadom **20 miliardám** unikátnych webstránok.

Ako dlho by trvalo nájsť výskyty slova „`robot`" na celom webe?

rátajme spolu:
20 miliárd stránok $\times$ trebárs 1 kB = 20 TB textu.
to je zhruba 50 miliónov krát viac ako Asimovova kniha
bude to trvať 50 miliónov krát dlhšie
$\Rightarrow$ niekoľko dní

Web rastie

- 1998: Založené Google, pozná 26 miliónov stránok
- 2008: Google pozná **bilión** (10^{12}) rôznych URL, zodpovedajú odhadom **20 miliardám** unikátnych webstránok.

Ako dlho by trvalo nájsť výskyty slova „`robot`" na celom webe?

rátajme spolu:
20 miliárd stránok $\times$ trebárs 1 kB $=$ 20 TB textu.
to je zhruba 50 miliónov krát viac ako Asimovova kniha
bude to trvať 50 miliónov krát dlhšie
$\Rightarrow$ niekoľko dní

Web rastie ešte viac

Realita: **0.25 sekundy**
Niekoľko dní čakať na výsledok nechceme!
Riešenie: paralelizácia
„nech si Google nakúpi viac počítačov :-)“

Na čo zabúdame?

Na ostatných ľudí!
Denne Google spracuje **300 miliónov** vyhľadávaní.

To by bolo 1 a pol miliardy dní výpočtu.
Ak by to chceli za deň stihnúť, treba miliardu a pol počítačov.

Web rastie ešte viac

Realita: **0.25 sekundy**
Niekoľko dní čakať na výsledok nechceme!
Riešenie: paralelizácia
„nech si Google nakúpi viac počítačov :-)“

Na čo zabúdame?

Na ostatných ľudí!
Denne Google spracuje **300 miliónov** vyhľadávaní.

To by bolo 1 a pol miliardy dní výpočtu.
Ak by to chceli za deň stihnúť, treba miliardu a pol počítačov.

Web rastie ešte viac

Realita: **0.25 sekundy**
Niekoľko dní čakať na výsledok nechceme!
Riešenie: paralelizácia
„nech si Google nakúpi viac počítačov :-)“

Na čo zabúdame?

Na ostatných ľudí!
Denne Google spracuje **300 miliónov** vyhľadávaní.

To by bolo 1 a pol miliardy dní výpočtu.
Ak by to chceli za deň stihnúť, treba miliardu a pol počítačov.

Potrebujeme lepšie algoritmy

Prvý nápad:
pre každé slovo budeme mať zoznam stránok, ktoré ho obsahujú.

Na jednej strane žiadna výhra (veľa pamäte; čo zložitejšie otázky?).
Na druhej strane ani toto nestačí.

Potrebujeme lepšie algoritmy

Prvý nápad:
pre každé slovo budeme mať zoznam stránok, ktoré ho obsahujú.

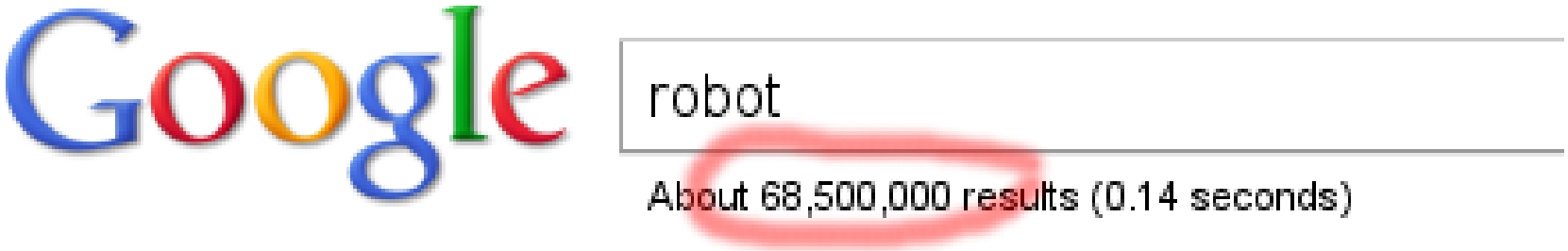
Na jednej strane žiadna výhra (veľa pamäte; čo zložitejšie otázky?).
Na druhej strane ani toto nestačí.

Potrebujeme lepšie algoritmy

Prvý nápad:
pre každé slovo budeme mať zoznam stránok, ktoré ho obsahujú.

Na jednej strane žiadna výhra (veľa pamäte; čo zložitejšie otázky?).
Na druhej strane ani toto nestačí.

Problém?



Ktoré z nich ukázať užívateľovi?
Ako spoznať, ktoré sú relevantné?

PageRank

Nový algoritmus pri zrode Googlu.
Autori Larry Page (podľa neho sa volá) a Sergej Brin.
Vyhodnocuje, ktoré stránky sú dôležité a ktoré nie.

Myšlienka: dôležité stránky sú tie, ktoré ľudia navštevujú.
Problém: ale my nevieme, ktoré to sú...

Riešenie: oni nám to povedia – spravia linky!

PageRank

Nový algoritmus pri zrode Googlu.
Autori Larry Page (podľa neho sa volá) a Sergej Brin.
Vyhodnocuje, ktoré stránky sú dôležité a ktoré nie.

Myšlienka: dôležité stránky sú tie, ktoré ľudia navštevujú.
Problém: ale my nevieme, ktoré to sú...

Riešenie: oni nám to povedia – spravia linky!

PageRank

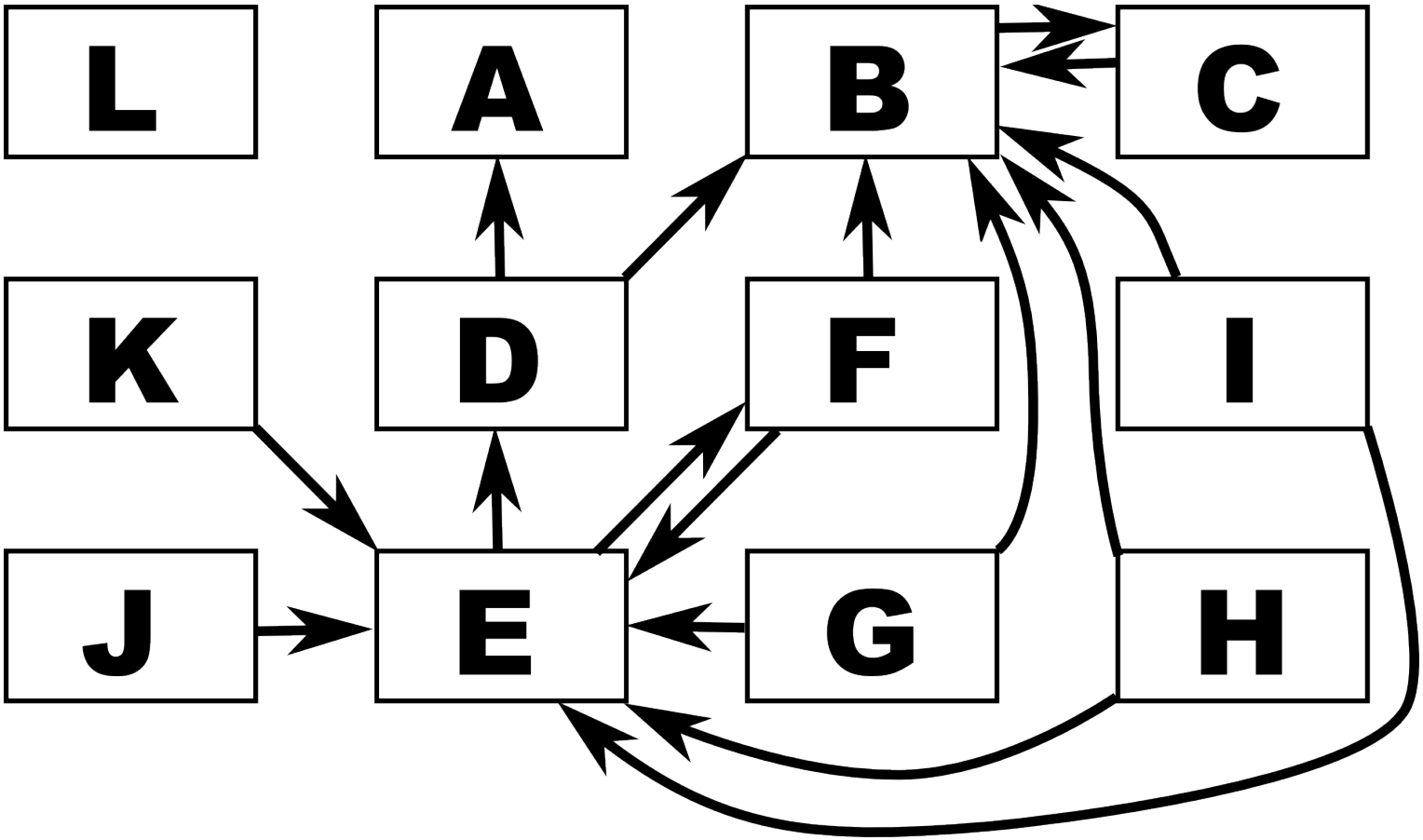
Nový algoritmus pri zrode Googlu.
Autori Larry Page (podľa neho sa volá) a Sergej Brin.
Vyhodnocuje, ktoré stránky sú dôležité a ktoré nie.

Myšlienka: dôležité stránky sú tie, ktoré ľudia navštevujú.
Problém: ale my nevieme, ktoré to sú...

Riešenie: oni nám to povedia – spravia linky!

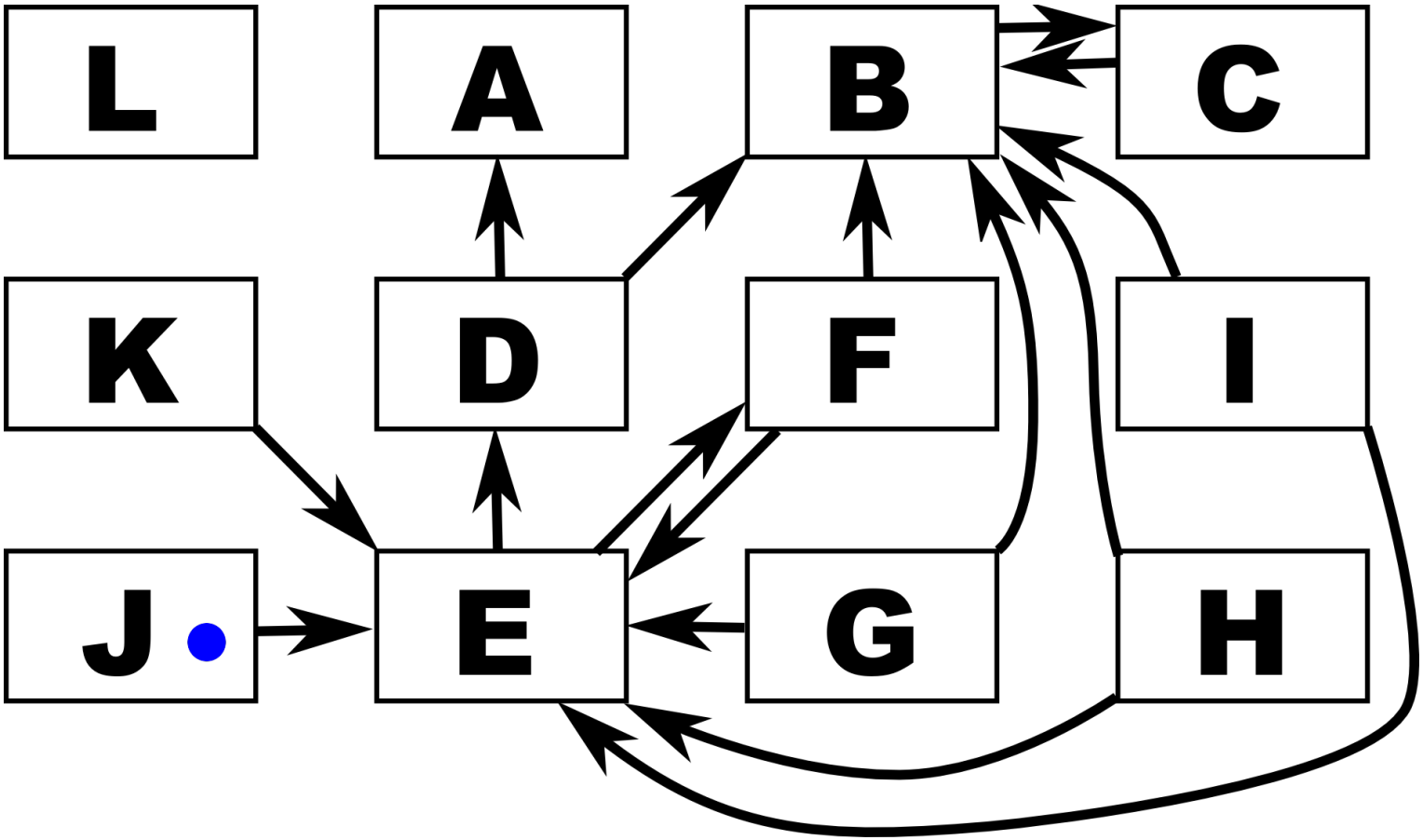
Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



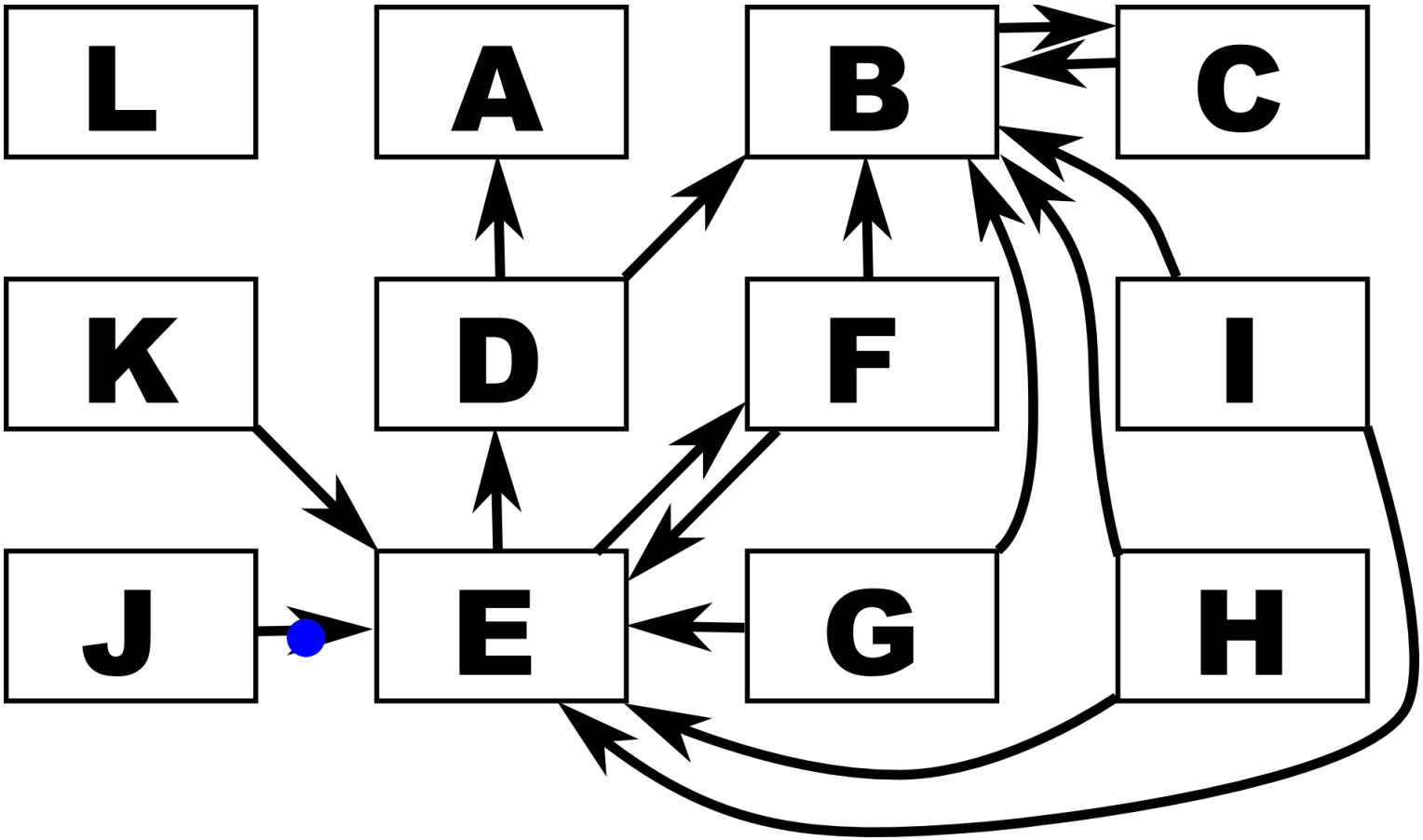
Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



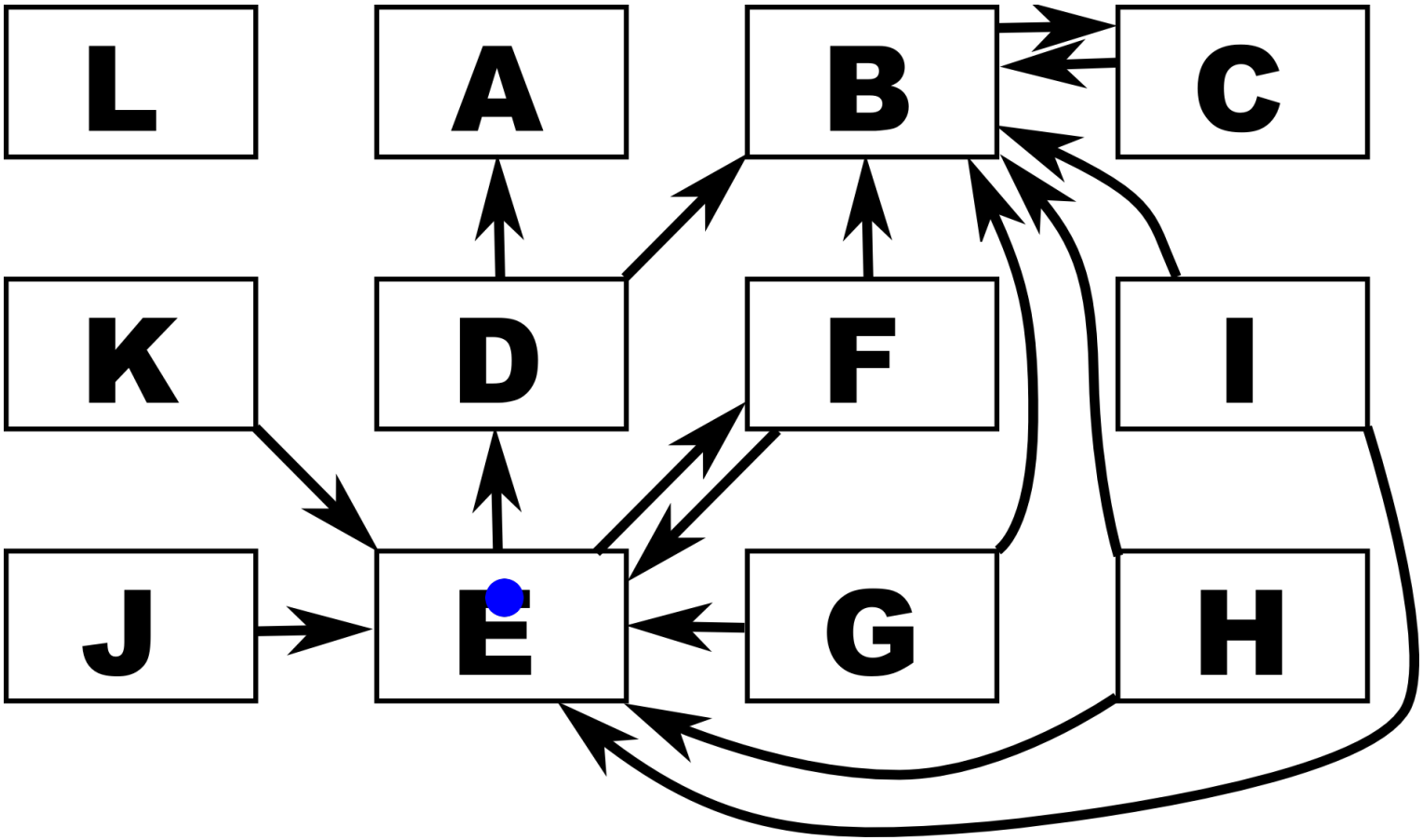
Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



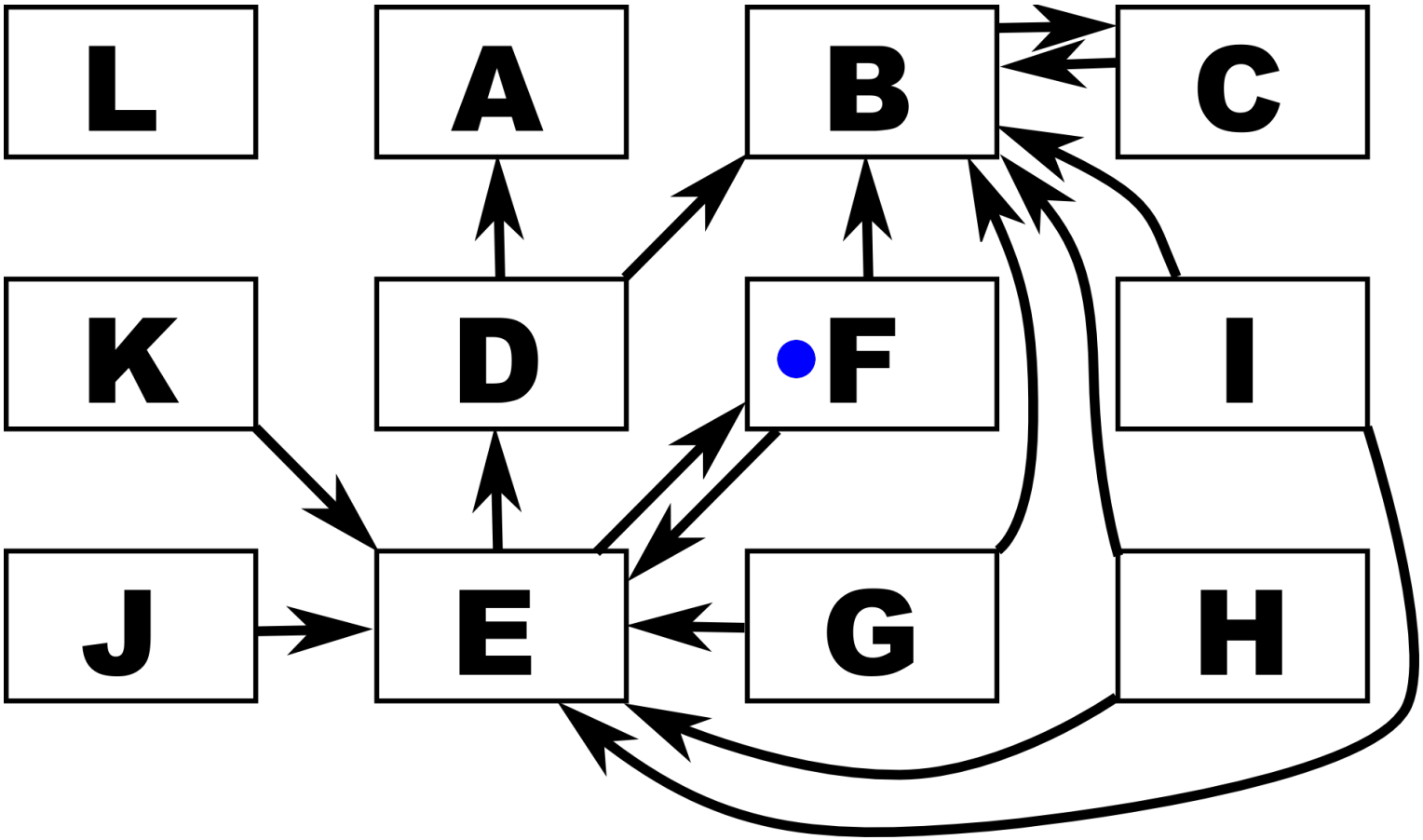
Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



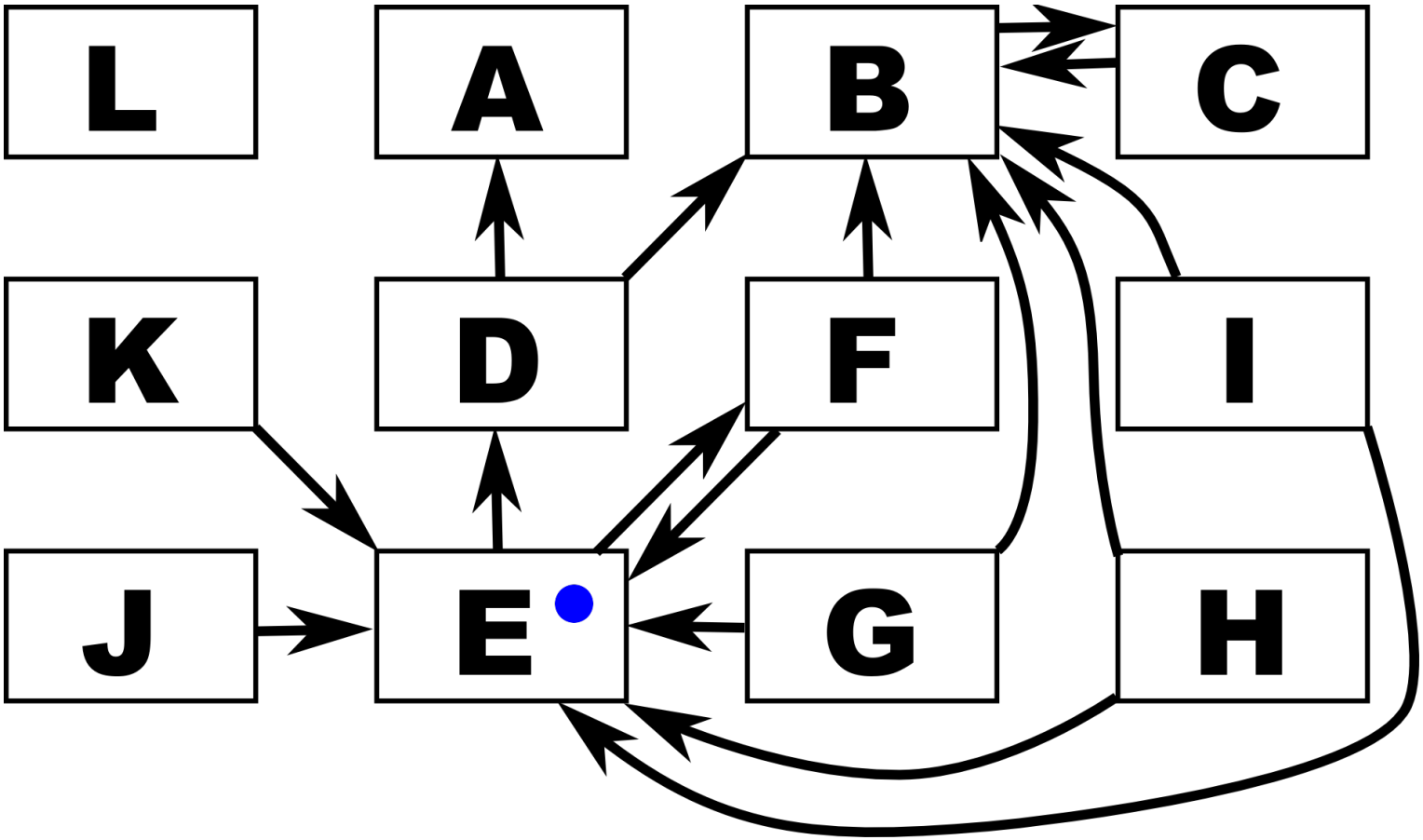
Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



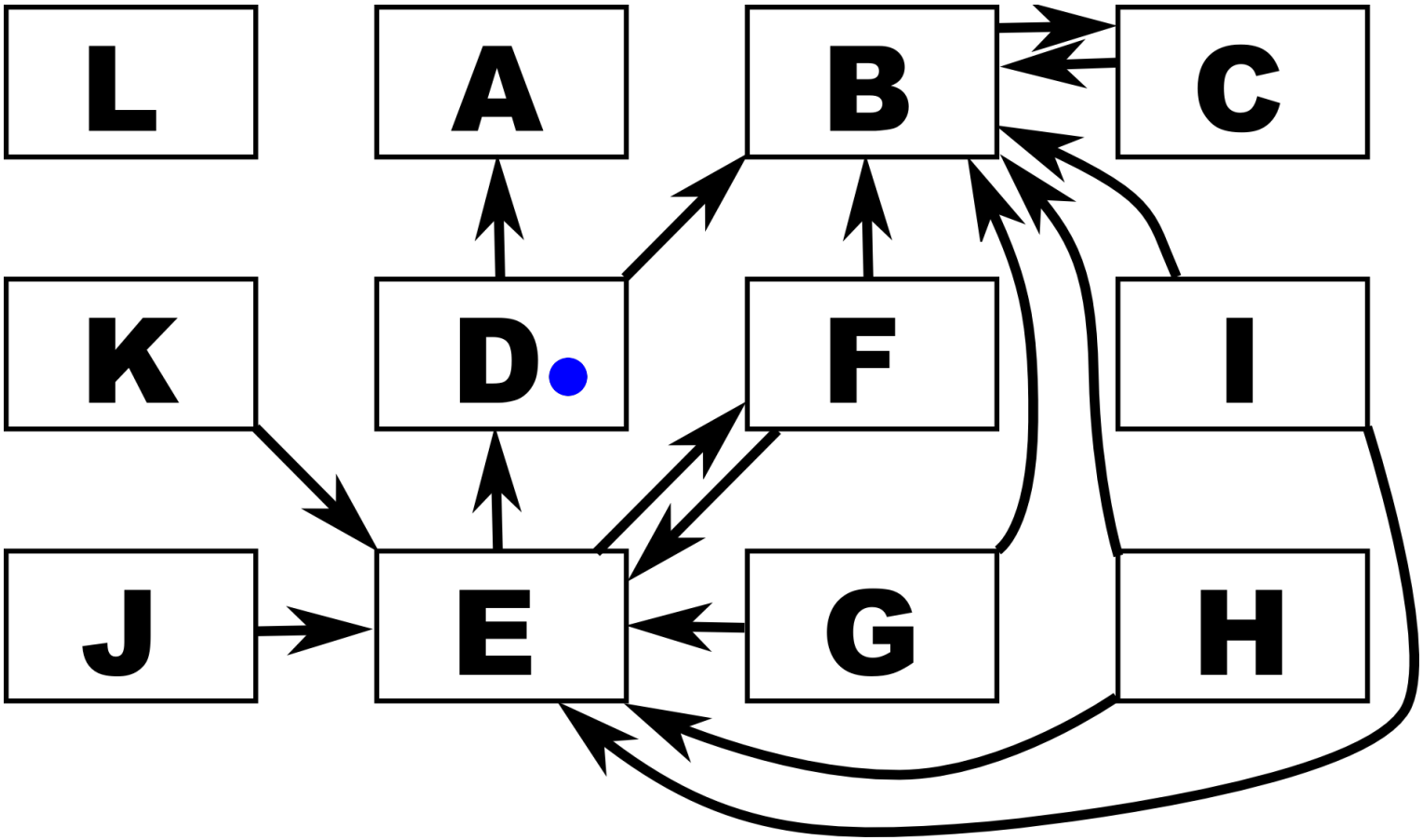
Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



Základná myšlienka PageRanku

Človek náhodne kliká po linkách.



Nie je náhodné klikanie blbosť?

Človek predsa nekliká po linkách *náhodne*!

Jeden človek nie – ale všetci dokopy áno!

Presnejší model správania sa ľudí:

- 85% prípadov: klikne na linku
- 15% prípadov: odíde na nesúvisiacu stránku

Nie je náhodné klikanie blbosť?

Človek predsa nekliká po linkách *náhodne*!

Jeden človek nie – ale všetci dokopy áno!

Presnejší model správania sa ľudí:

- 85% prípadov: klikne na linku
- 15% prípadov: odíde na nesúvisiacu stránku

Nie je náhodné klikanie blbosť?

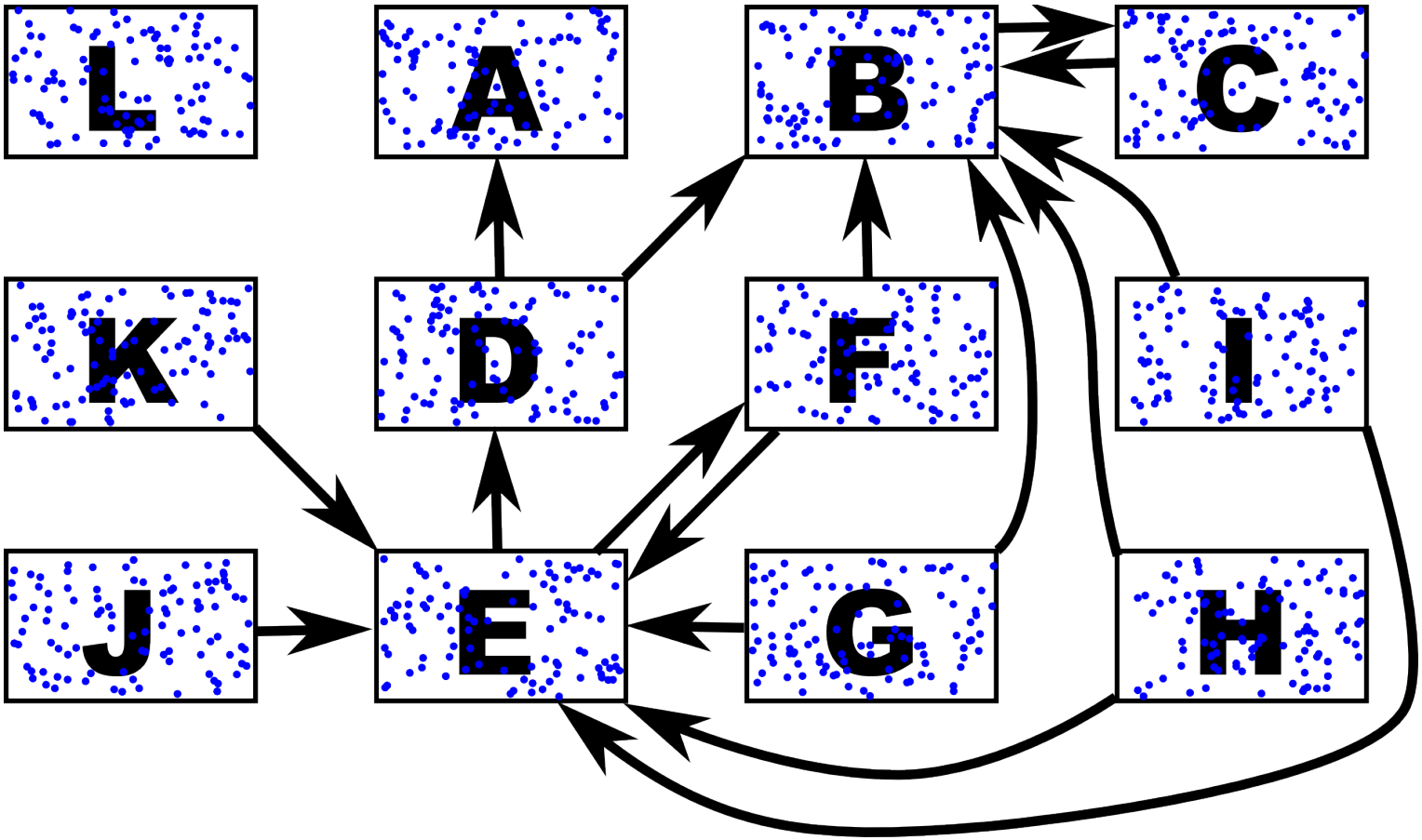
Človek predsa nekliká po linkách *náhodne*!

Jeden človek nie – ale všetci dokopy áno!

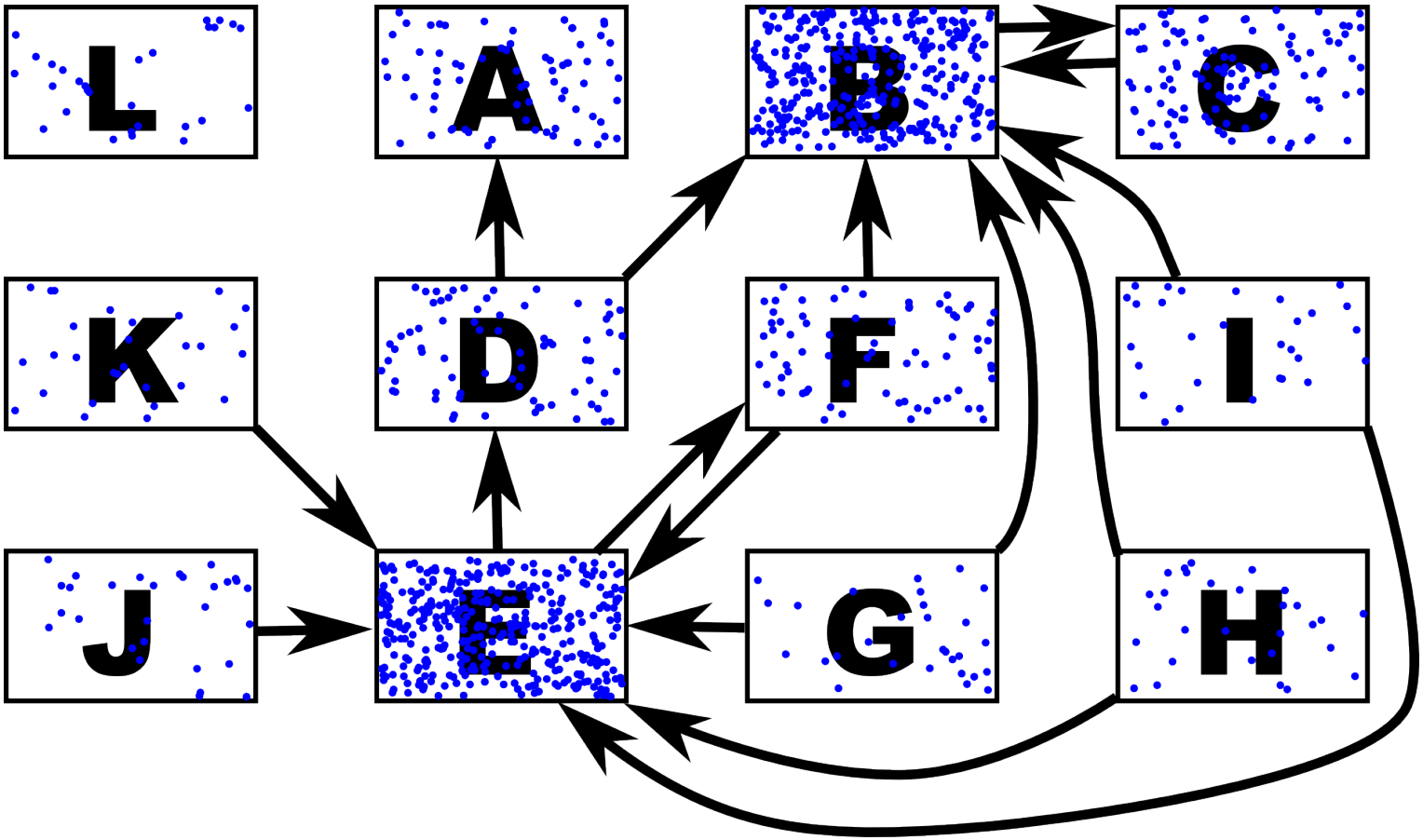
Presnejší model správania sa ľudí:

- 85% prípadov: klikne na linku
- 15% prípadov: odíde na nesúvisiacu stránku

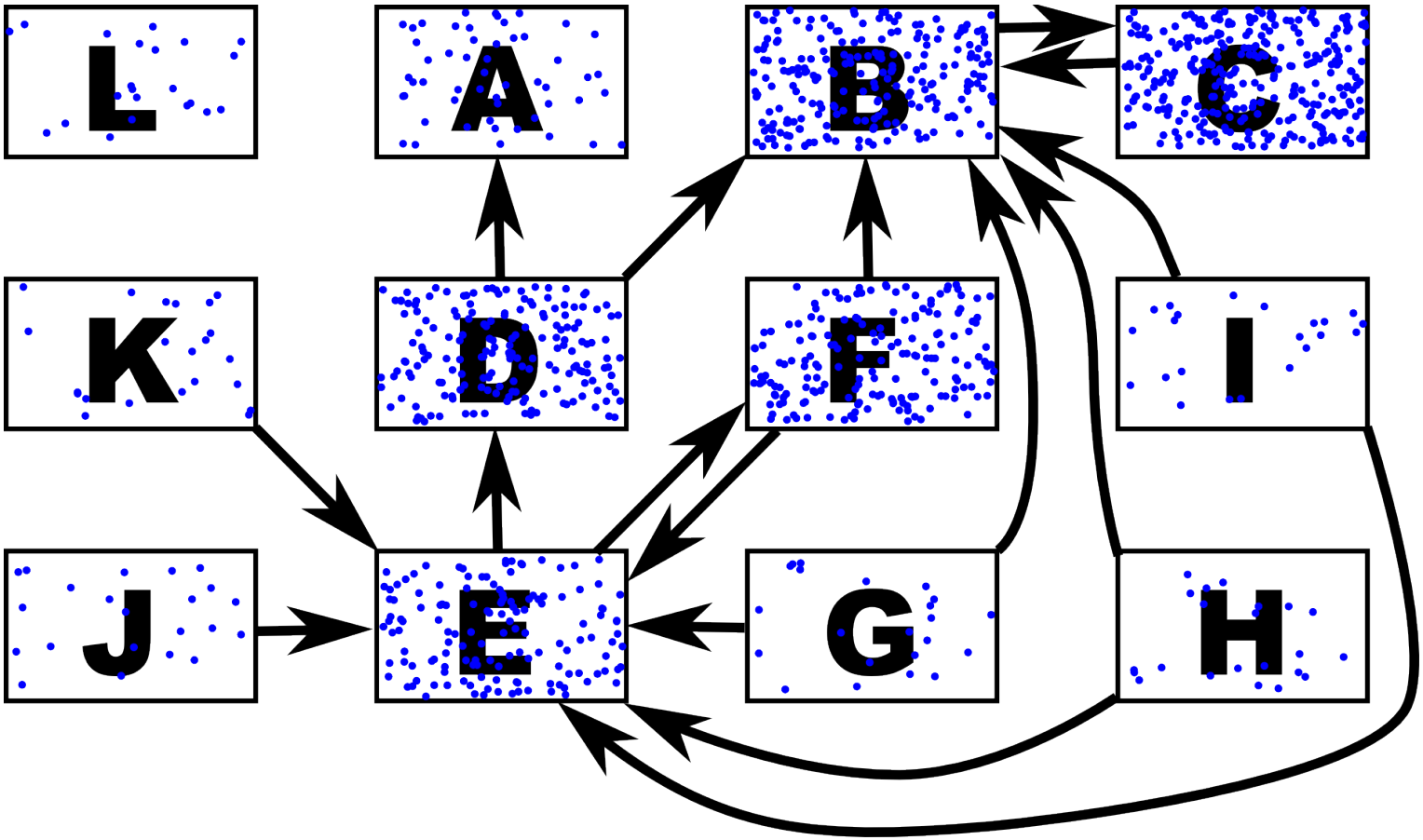
Simulácia pre veľa ľudí



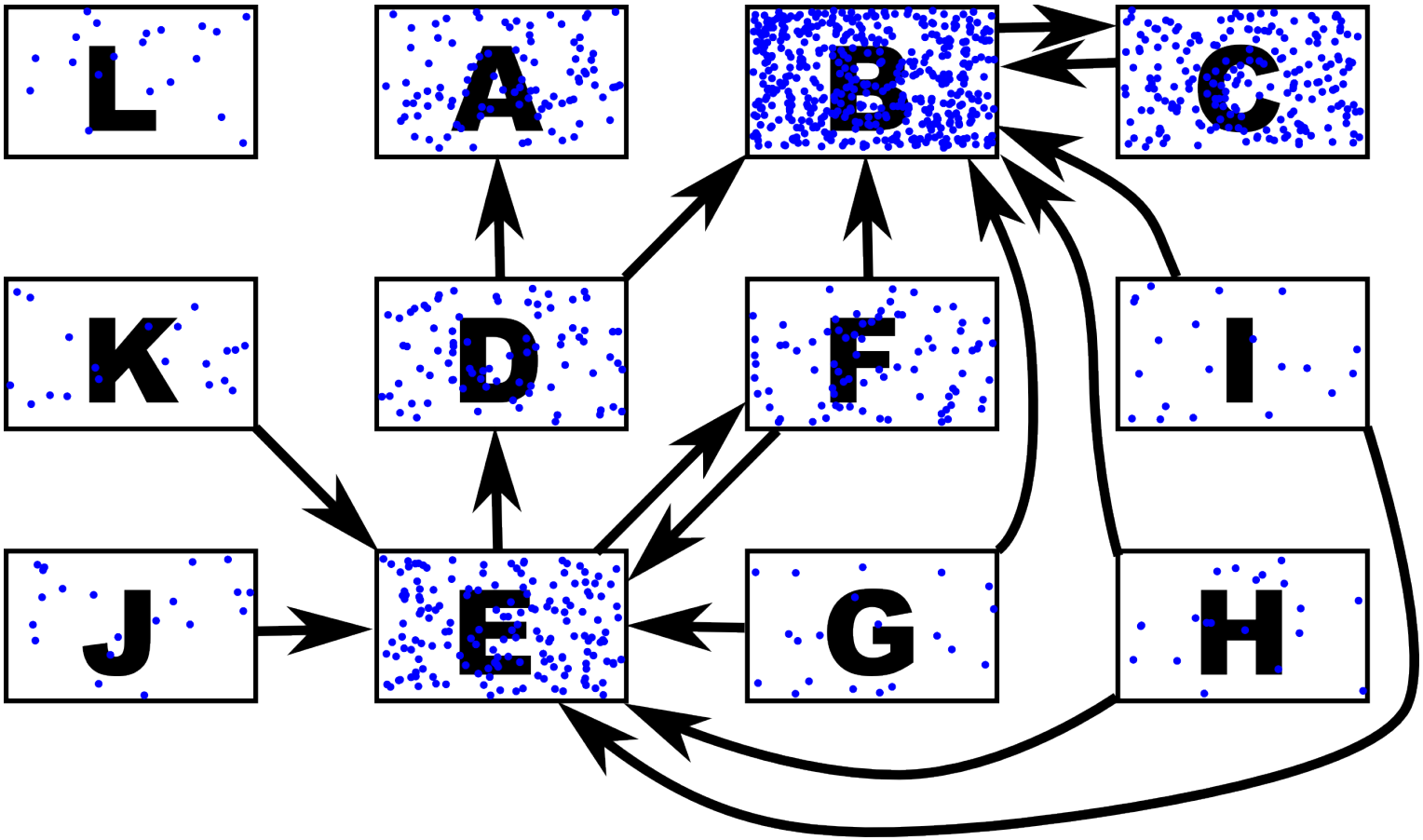
Simulácia pre veľa ľudí



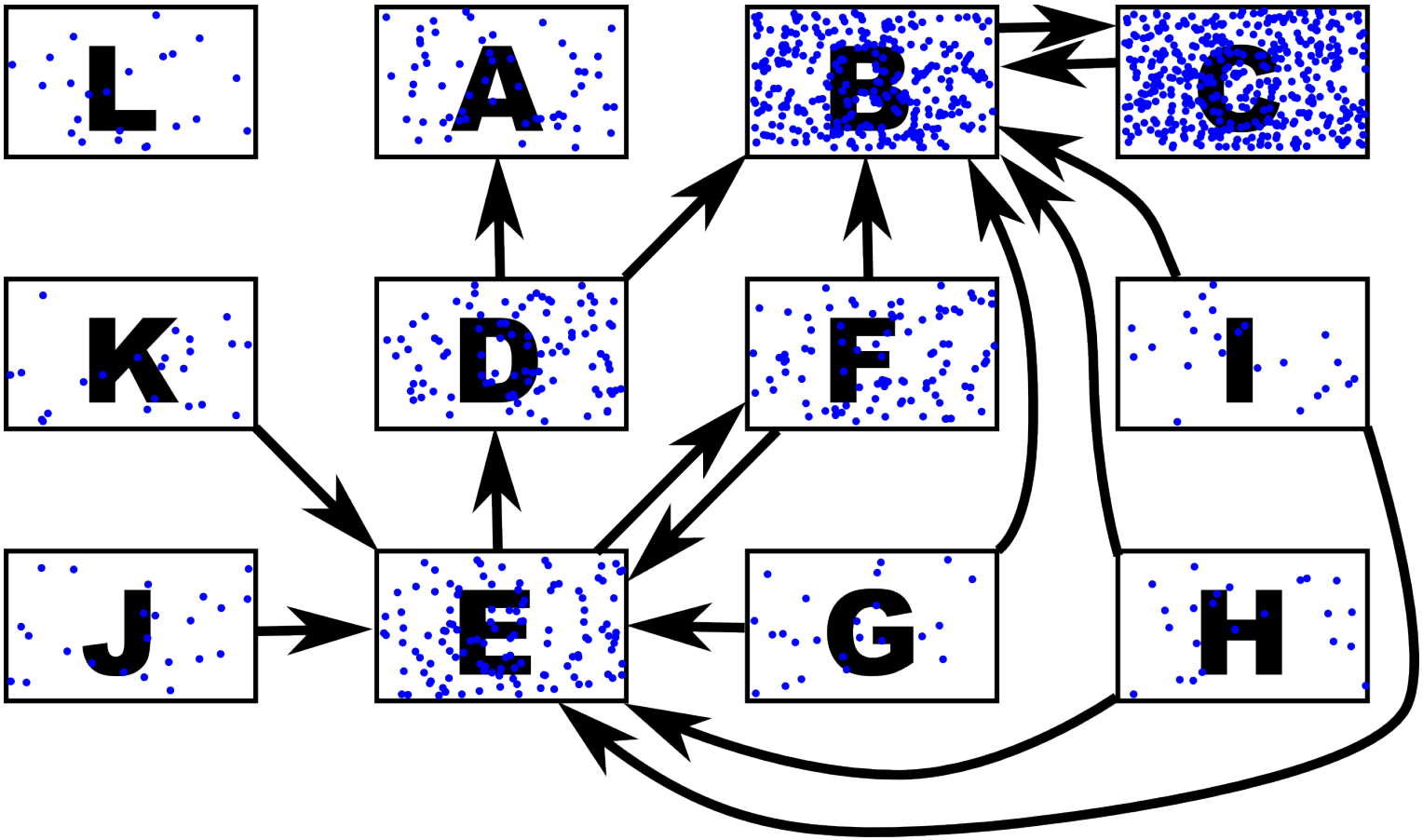
Simulácia pre veľa ľudí



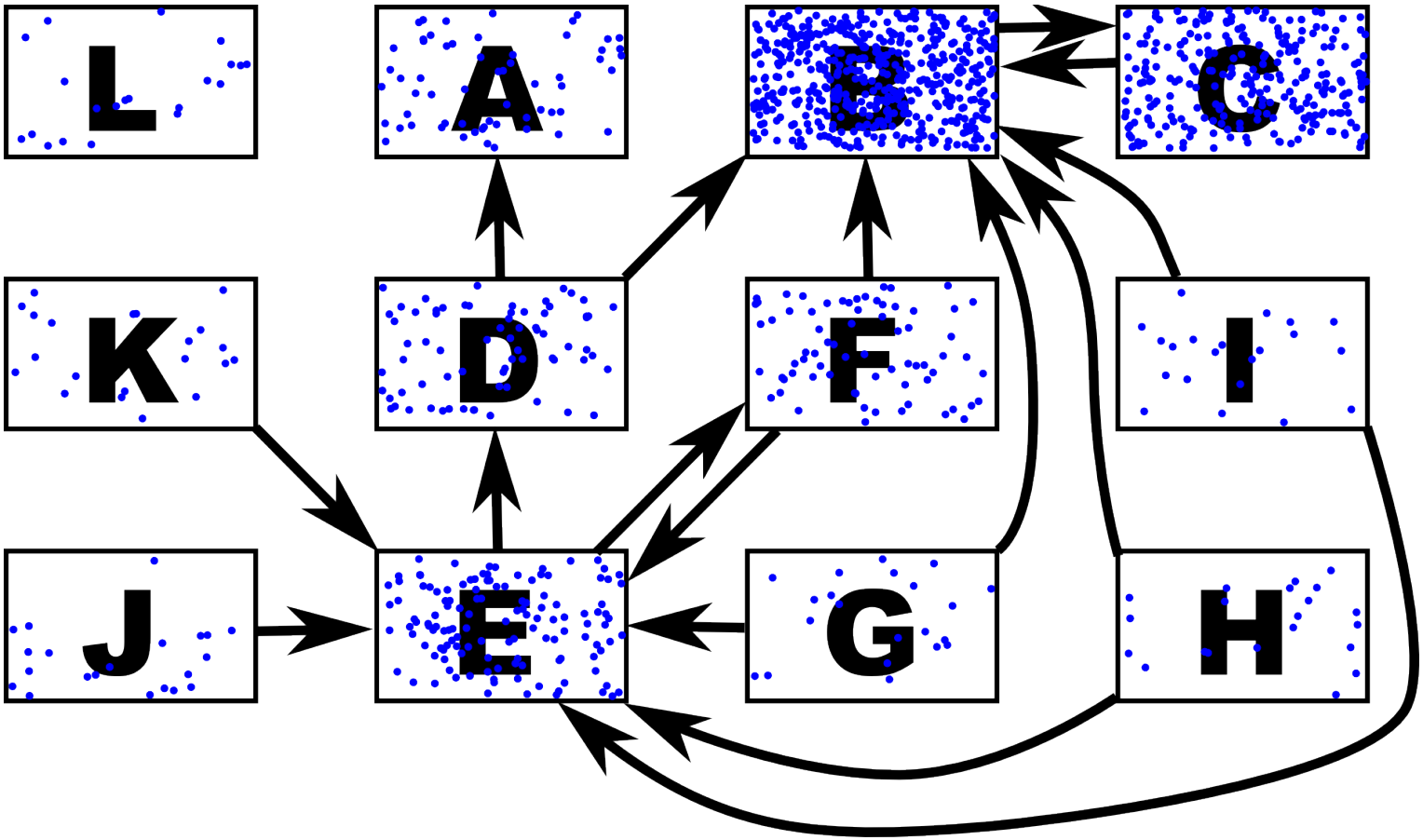
Simulácia pre veľa ľudí



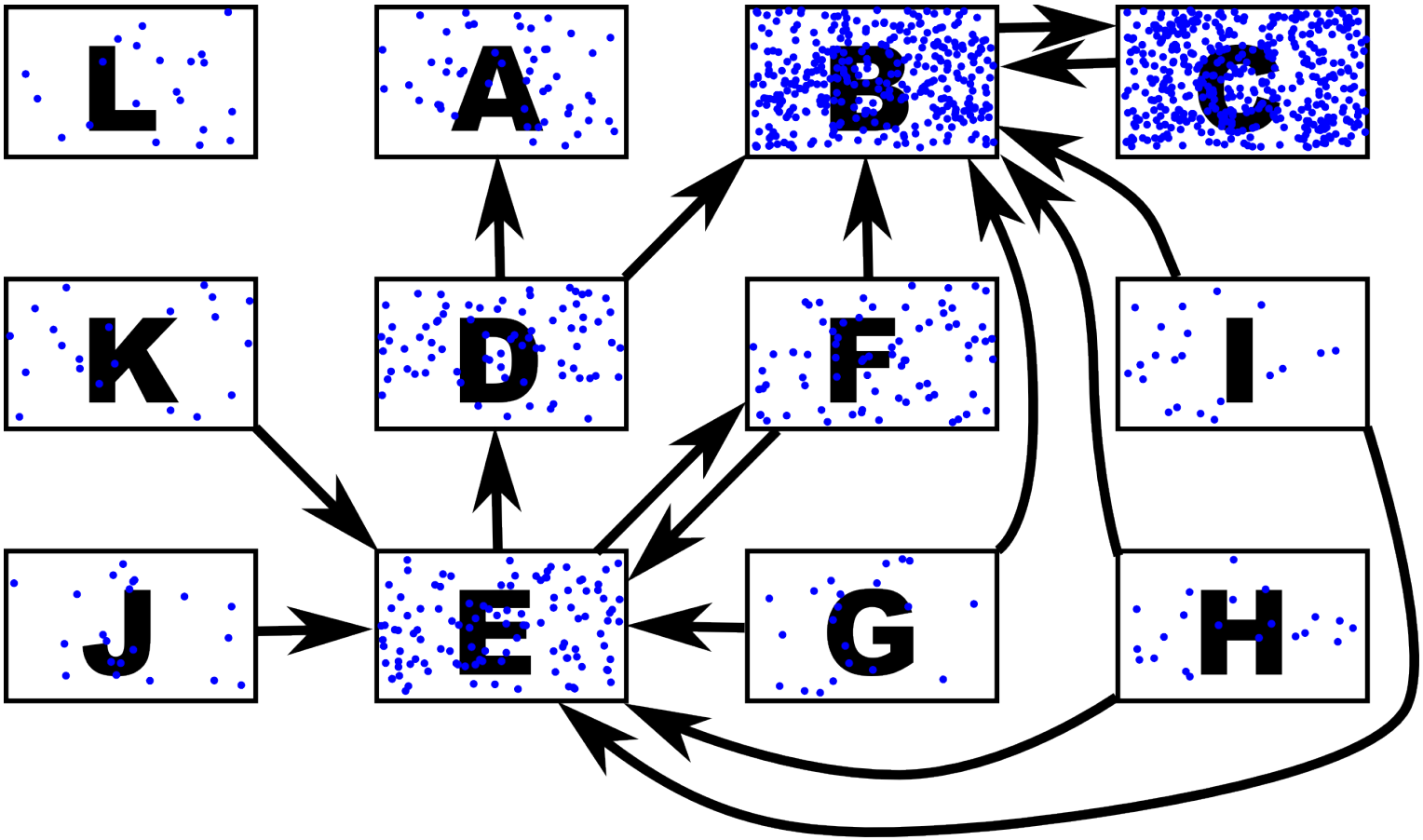
Simulácia pre veľa ľudí



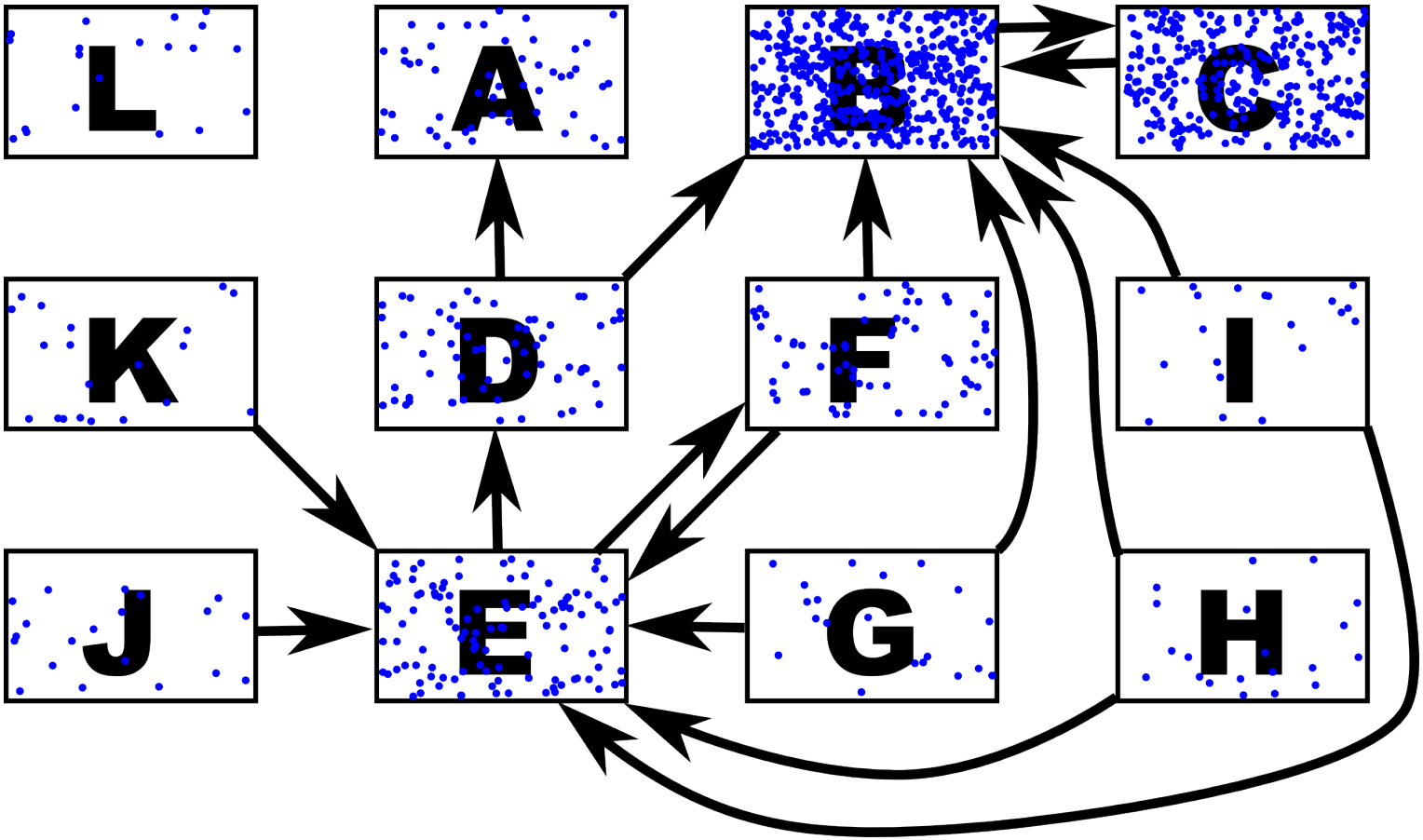
Simulácia pre veľa ľudí



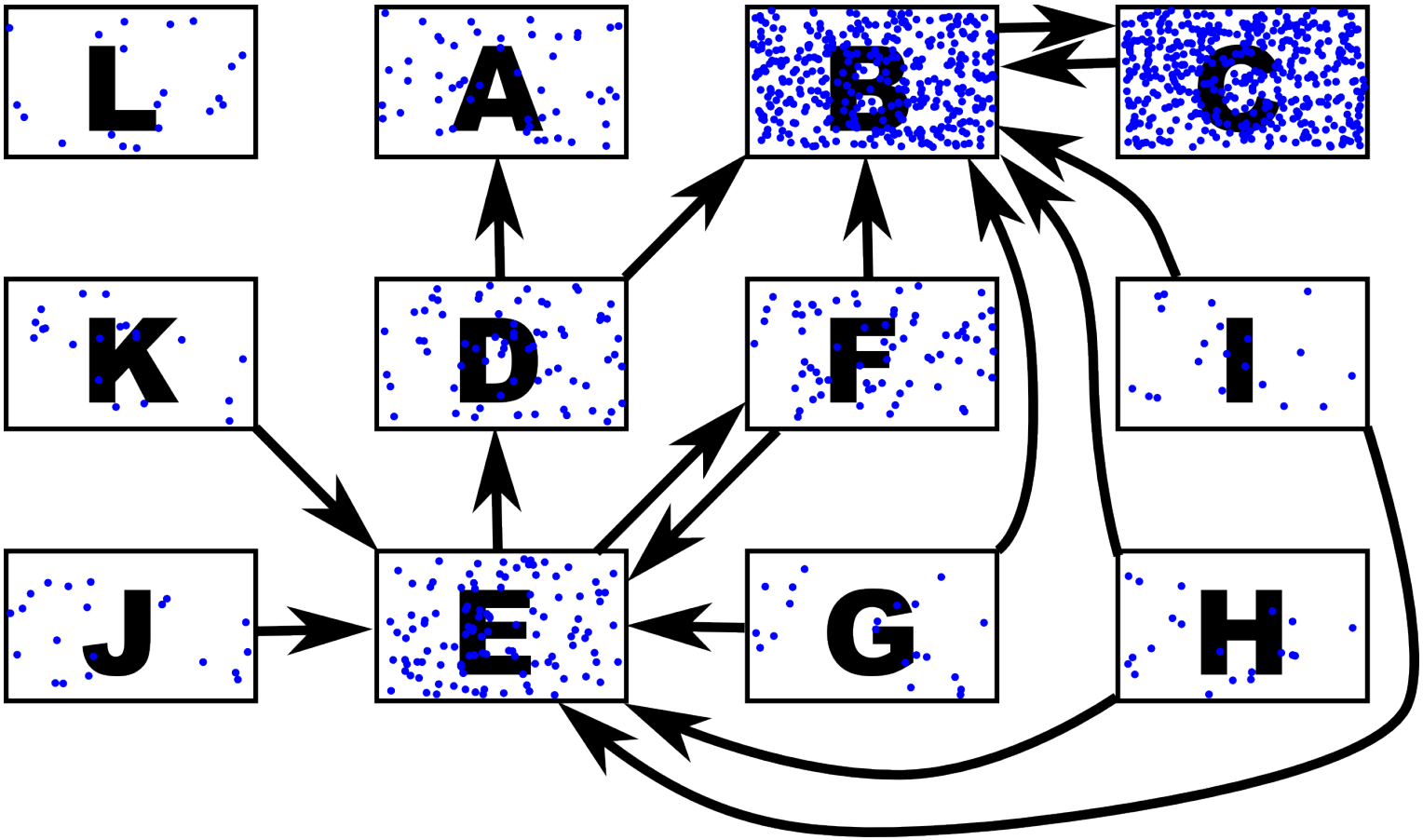
Simulácia pre veľa ľudí



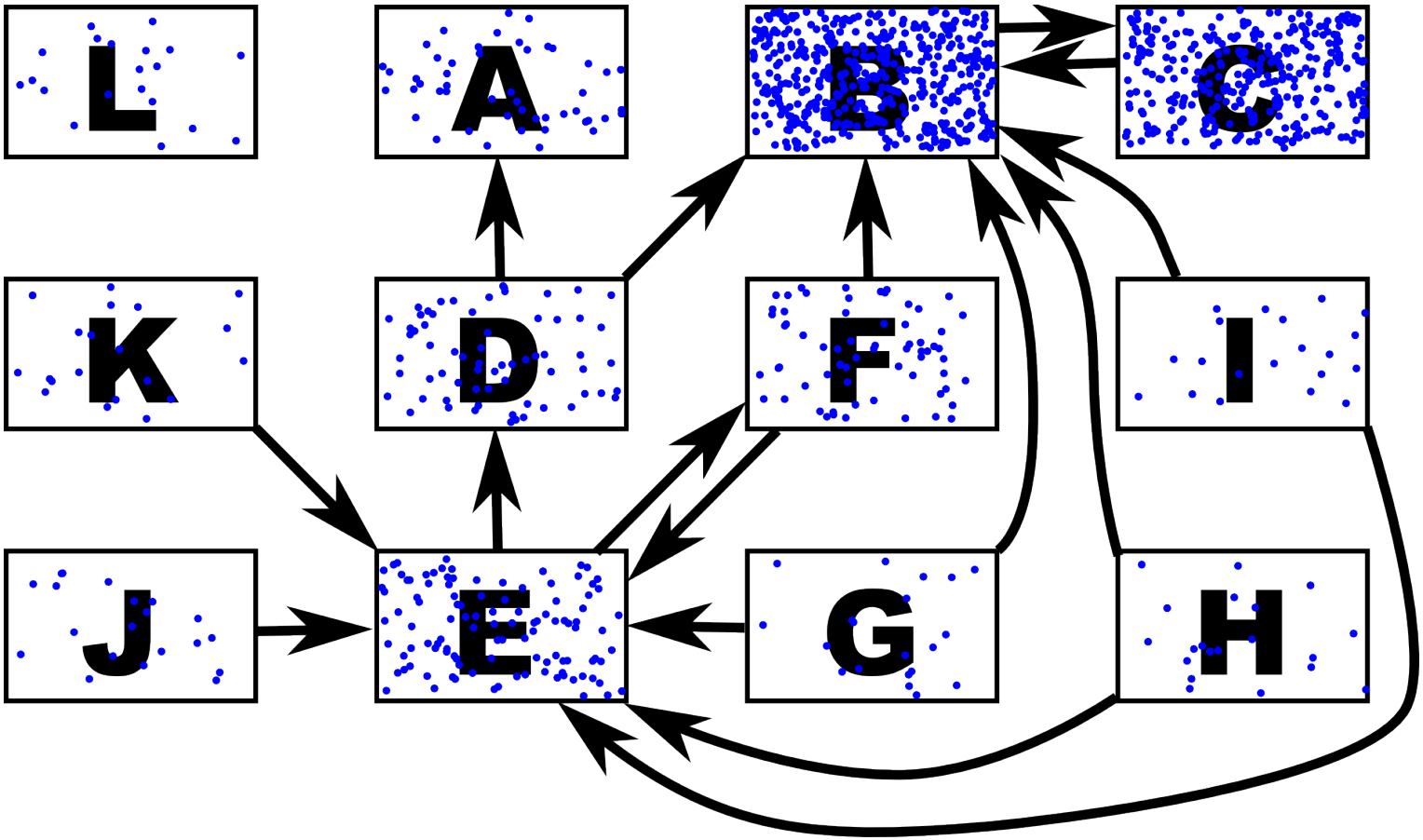
Simulácia pre veľa ľudí



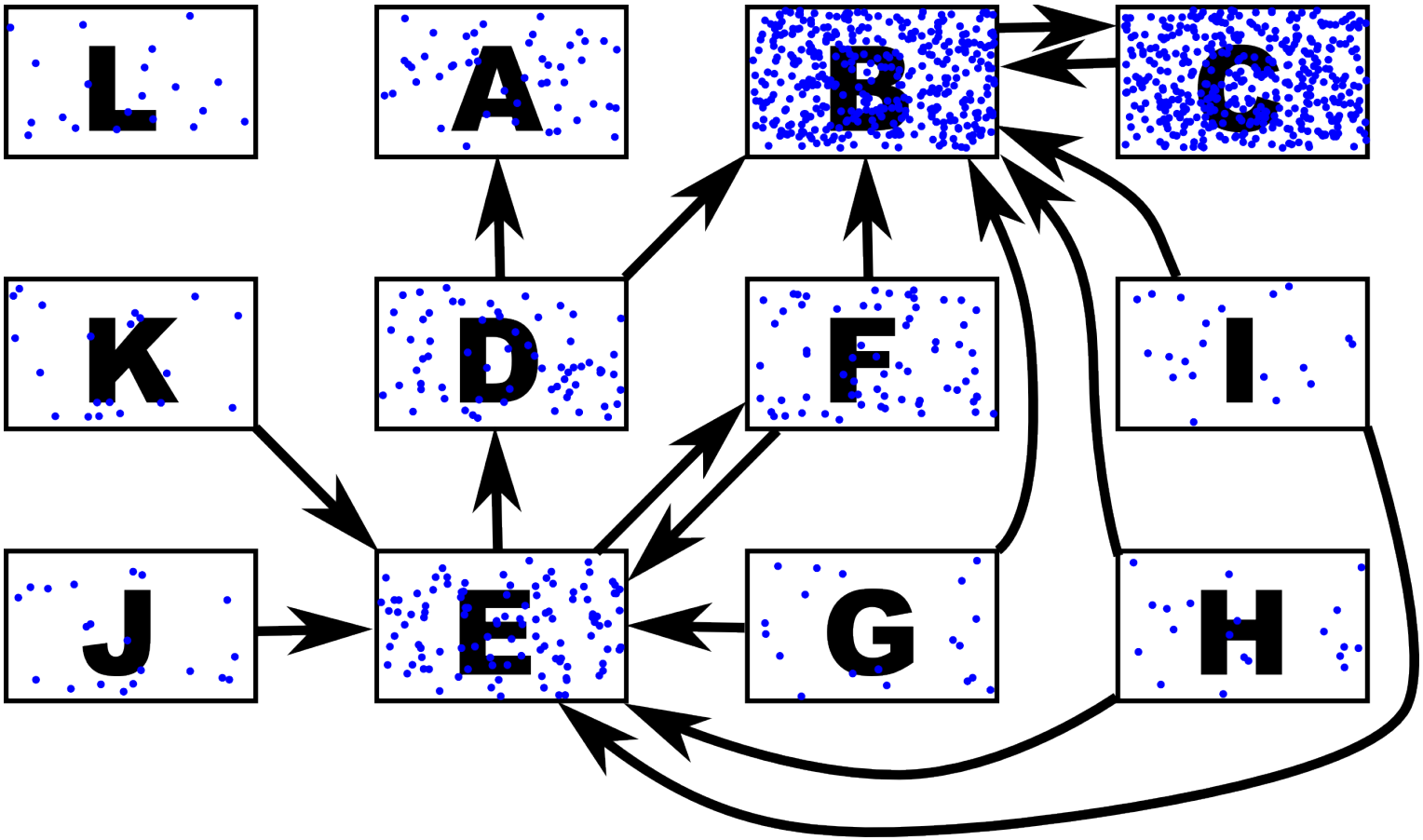
Simulácia pre veľa ľudí



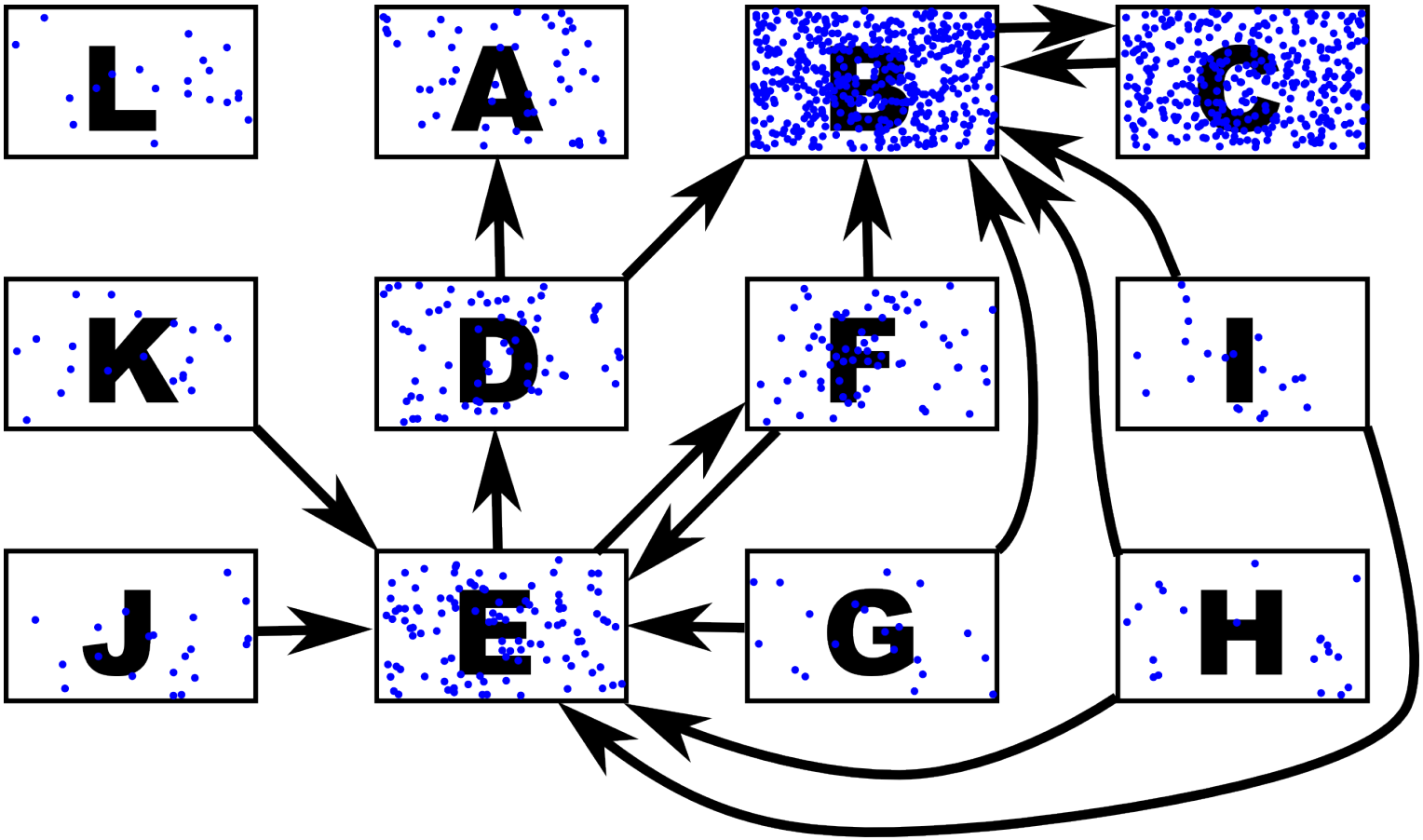
Simulácia pre veľa ľudí



Simulácia pre veľa ľudí

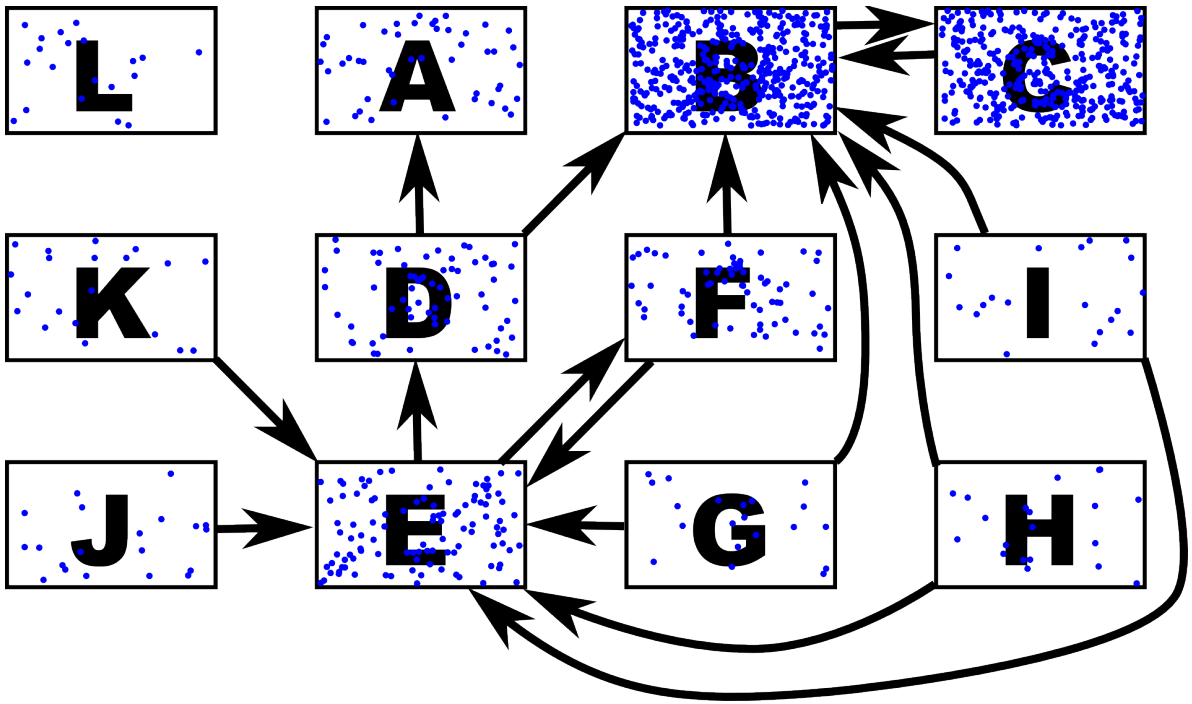


Simulácia pre veľa ľudí



Sústava lineárnych rovníc

Ako zistiť, k čomu to celé smeruje?

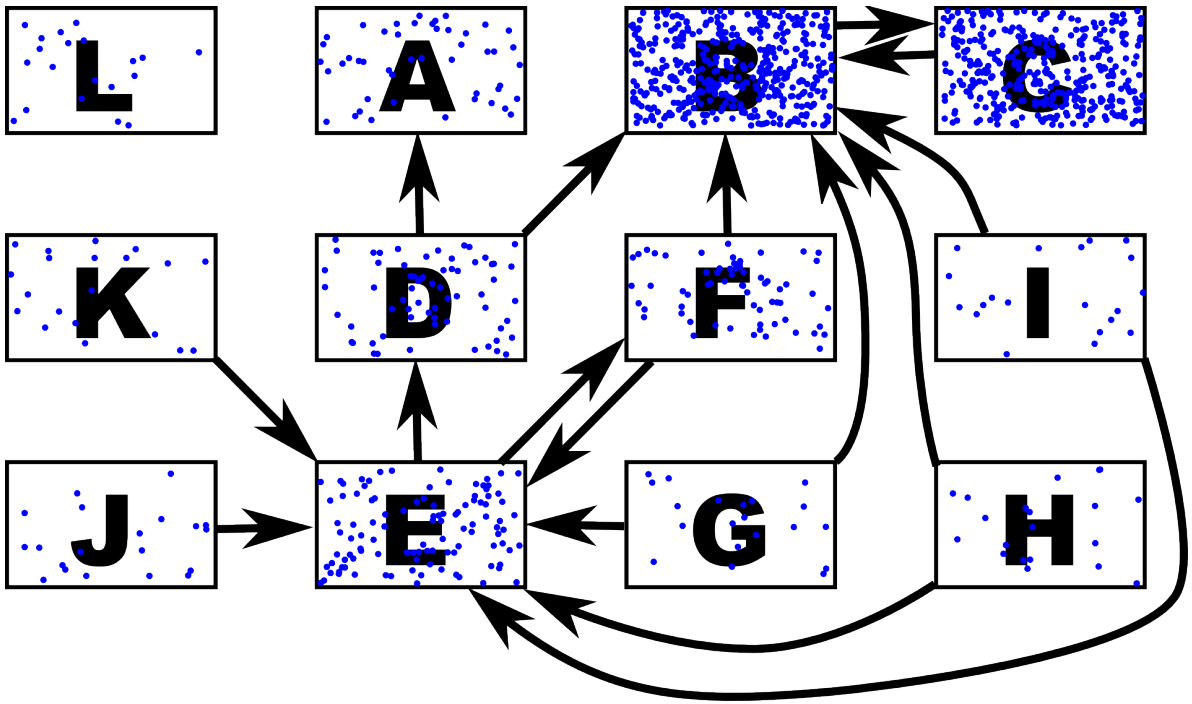


Rovnice!

$$rank_E = 0.15 \cdot \frac{1}{12} + 0.85 \cdot \left(\frac{rank_F}{2} + \frac{rank_G}{2} + \cdots + rank_K \right)$$

Sústava lineárnych rovníc

Ako zistiť, k čomu to celé smeruje?



Rovnice!

$$rank_E = 0.15 \cdot \frac{1}{12} + 0.85 \cdot \left(\frac{rank_F}{2} + \frac{rank_G}{2} + \cdots + rank_K \right)$$

Sústava lineárnych rovníc

Gaussova eliminačná metóda:

$$\begin{aligned} x + 2y - z &= -1 \\ -x - 3y + 5z &= 10 \\ 3x + y + 4z &= 16 \end{aligned}$$

Sústava lineárnych rovníc

Gaussova eliminačná metóda:

$$\begin{aligned} x + \quad 2y - \quad z &= -1 \\ -y + \quad 4z &= 9 \\ -5y + \quad 7z &= 19 \end{aligned}$$

Sústava lineárnych rovníc

Gaussova eliminačná metóda:

$$\begin{aligned} x + \quad 2y - \quad z &= -1 \\ -y + \quad 4z &= 9 \\ 13z &= 26 \end{aligned}$$

a postupným dosadzovaním máme $z = 2$, $y = -1$, $x = 3$

Sústava lineárnych rovníc

Gaussova eliminačná metóda:

$$\begin{aligned} x + \quad 2y - \quad z &= -1 \\ -y + \quad 4z &= 9 \\ 13z &= 26 \end{aligned}$$

a postupným dosadzovaním máme $z = 2$, $y = -1$, $x = 3$

Sústava lineárnych rovníc

Gaussova eliminačná metóda je algoritmus.

n premenných $\Rightarrow$ rádovo n^3 operácií

Problém:
miliardy webstránok $\Rightarrow$ miliardy premenných
miliardy premenných $\Rightarrow$ miliardy miliárd miliárd operácií
$\Rightarrow$ stotisíc miliárd rokov výpočtu?!

A samotnú sústavu rovníc tvorí miliarda miliárd čísel
$\Rightarrow$ exobajty pamäte?!

Sústava lineárnych rovníc

Gaussova eliminačná metóda je algoritmus.

n premenných $\Rightarrow$ rádovo n^3 operácií

Problém:
miliardy webstránok $\Rightarrow$ miliardy premenných
miliardy premenných $\Rightarrow$ miliardy miliárd miliárd operácií
$\Rightarrow$ stotisíc miliárd rokov výpočtu?!

A samotnú sústavu rovníc tvorí miliarda miliárd čísel
$\Rightarrow$ exobajty pamäte?!

Sústava lineárnych rovníc

Gaussova eliminačná metóda je algoritmus.

n premenných $\Rightarrow$ rádovo n^3 operácií

Problém:
miliardy webstránok $\Rightarrow$ miliardy premenných
miliardy premenných $\Rightarrow$ miliardy miliárd miliárd operácií
$\Rightarrow$ stotisíc miliárd rokov výpočtu?!

A samotnú sústavu rovníc tvorí miliarda miliárd čísel
$\Rightarrow$ exobajty pamäte?!

Sústava lineárnych rovníc

Gaussova eliminačná metóda je algoritmus.

n premenných $\Rightarrow$ rádovo n^3 operácií

Problém:
miliardy webstránok $\Rightarrow$ miliardy premenných
miliardy premenných $\Rightarrow$ miliardy miliárd miliárd operácií
$\Rightarrow$ stotisíc miliárd rokov výpočtu?!

A samotnú sústavu rovníc tvorí miliarda miliárd čísel
$\Rightarrow$ exobajty pamäte?!

Sústava lineárnych rovníc

Riešenie problému: nové algoritmy.

Šikovne využijeme to, že naša sústava rovníc je „riedka“:
Každá webstránka ukazuje len na pár iných.

Prototyp Googlu: jeden počítač za niekoľko hodín spočítal PageRank pre „celý internet“ (vtedy pár miliónov stránok)

Súčasnosť: PageRank je jedným z mnohých faktorov stále používaných Googlom.

Sústava lineárnych rovníc

Riešenie problému: nové algoritmy.

Šikovne využijeme to, že naša sústava rovníc je „riedka“:
Každá webstránka ukazuje len na pár iných.

Prototyp Googlu: jeden počítač za niekoľko hodín spočítal PageRank pre „celý internet“ (vtedy pár miliónov stránok)

Súčasnosť: PageRank je jedným z mnohých faktorov stále používaných Googlom.

Sústava lineárnych rovníc

Riešenie problému: nové algoritmy.

Šikovne využijeme to, že naša sústava rovníc je „riedka“:
Každá webstránka ukazuje len na pár iných.

Prototyp Googlu: jeden počítač za niekoľko hodín spočítal PageRank pre „celý internet“ (vtedy pár miliónov stránok)

Súčasnosť: PageRank je jedným z mnohých faktorov stále používaných Googlom.

Podobnosť vkusu

Čo máme: webserver s novinami

Čo chceme: odporúčať návštevníkom, čo majú čítať

Ako to dosiahnuť:
odporučiť to, čo si otvorili ľudia s podobným vkusom.

Množstvo iných aplikácií:
napr. odporúčanie filmov (Netflix), kníh (Amazon)

Podobnosť vkusu

Čo máme: webserver s novinami

Čo chceme: odporúčať návštevníkom, čo majú čítať

Ako to dosiahnuť:
odporučiť to, čo si otvorili ľudia s podobným vkusom.

Množstvo iných aplikácií:
napr. odporúčanie filmov (Netflix), kníh (Amazon)

Podobnosť vkusu

Čo máme: webserver s novinami

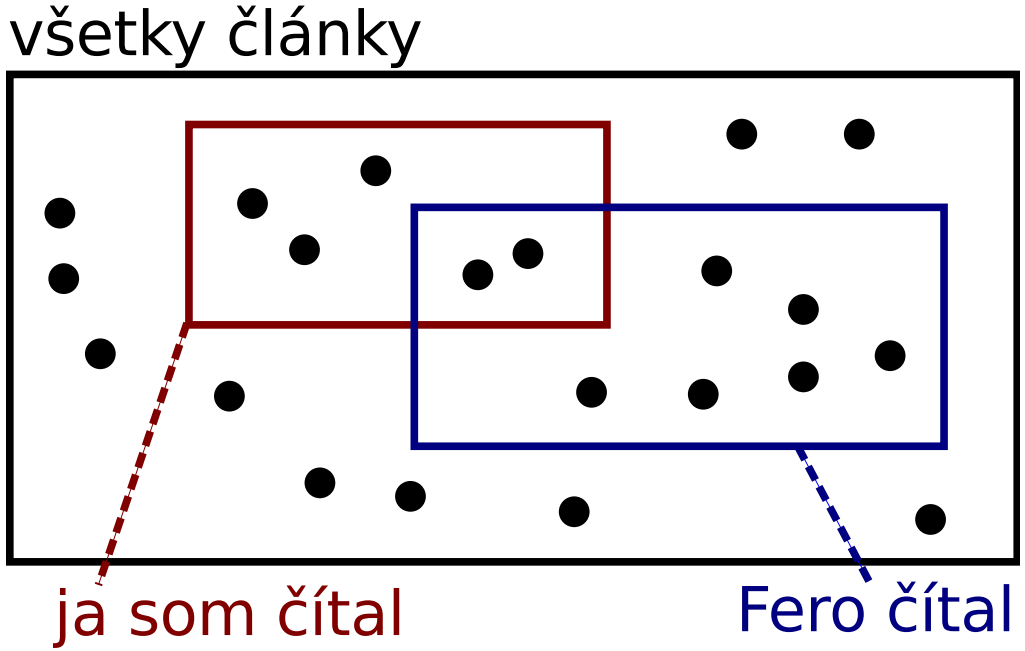
Čo chceme: odporúčať návštevníkom, čo majú čítať

Ako to dosiahnuť:
odporučiť to, čo si otvorili ľudia s podobným vkusom.

Množstvo iných aplikácií:
napr. odporúčanie filmov (Netflix), kníh (Amazon)

Podobnosť vkusu

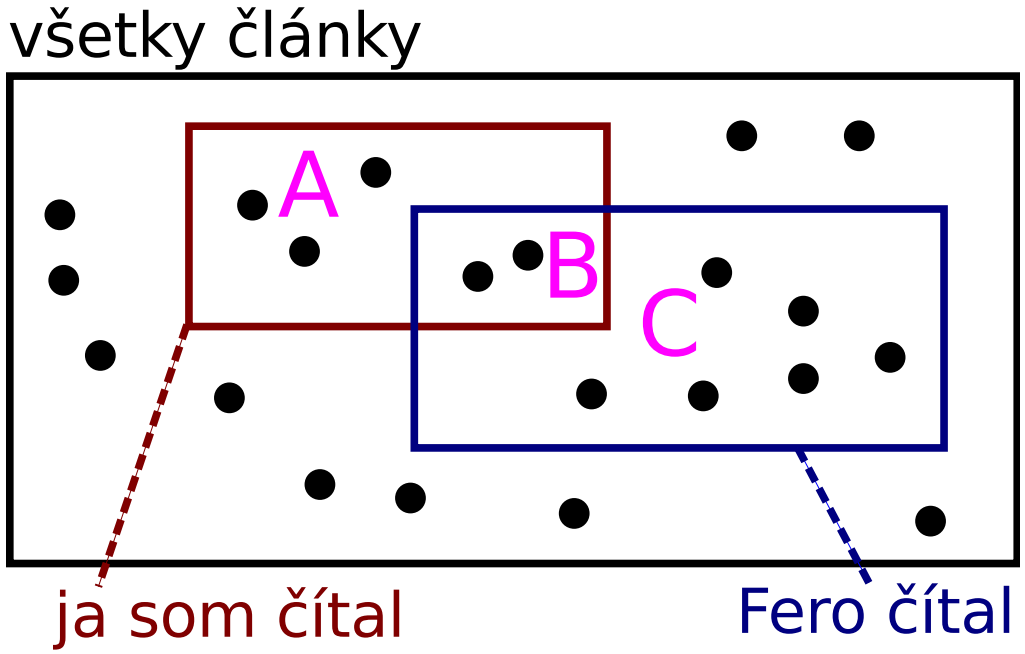
Kto sú ľudia s podobným vkusom?



Kedy máme ja a Fero podobný vkus?

Podobnosť vkusu

Kto sú ľudia s podobným vkusom?

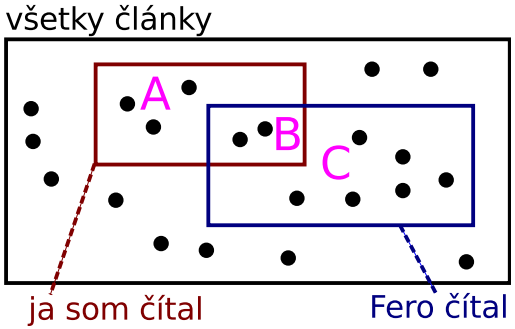


Dobré: veľké B, malé A a C.

Podobnosť vkusu

Jednoduchý vzťah:

$$\text{podobnosť}(ja, Fero) = \frac{B}{A + B + C}$$

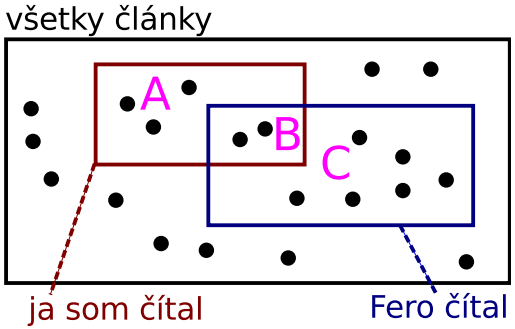


Pravdepodobnostný pohľad:
vyberieme náhodný článok, ktorý niekto z nás čítal
aká je pravdepodobnosť, že sme ho čítali obaja?

Podobnosť vkusu

Jednoduchý vzťah:

$$\text{podobnosť}(ja, Fero) = \frac{B}{A + B + C}$$



Pravdepodobnostný pohľad:
vyberieme náhodný článok, ktorý niekto z nás čítal
aká je pravdepodobnosť, že sme ho čítali obaja?

Ako nájsť ľudí s podobným vkusom?

Jednoduchý trik:

- náhodne očíslujeme články
- pre každého človeka si zapamätáme *najmenšie číslo* článku, ktorý čítal (teda akoby jeden náhodný článok)
- *Kolegovia* človeka budú všetci, ktorým sme vybrali ten istý článok ako jemu.

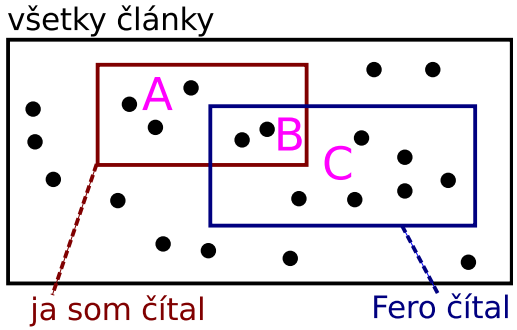
Dôležité pozorovanie:
podobnosť vkusu ľudí X a Y
je pravdepodobnosťou, že budú kolegovia!

Ako nájsť ľudí s podobným vkusom?

Jednoduchý trik:

- náhodne očíslujeme články
- pre každého človeka si zapamätáme *najmenšie číslo* článku, ktorý čítal (teda akoby jeden náhodný článok)
- *Kolegovia* človeka budú všetci, ktorým sme vybrali ten istý článok ako jemu.

Dôležité pozorovanie:
podobnosť vkusu ľudí X a Y
je pravdepodobnosťou, že budú kolegovia!

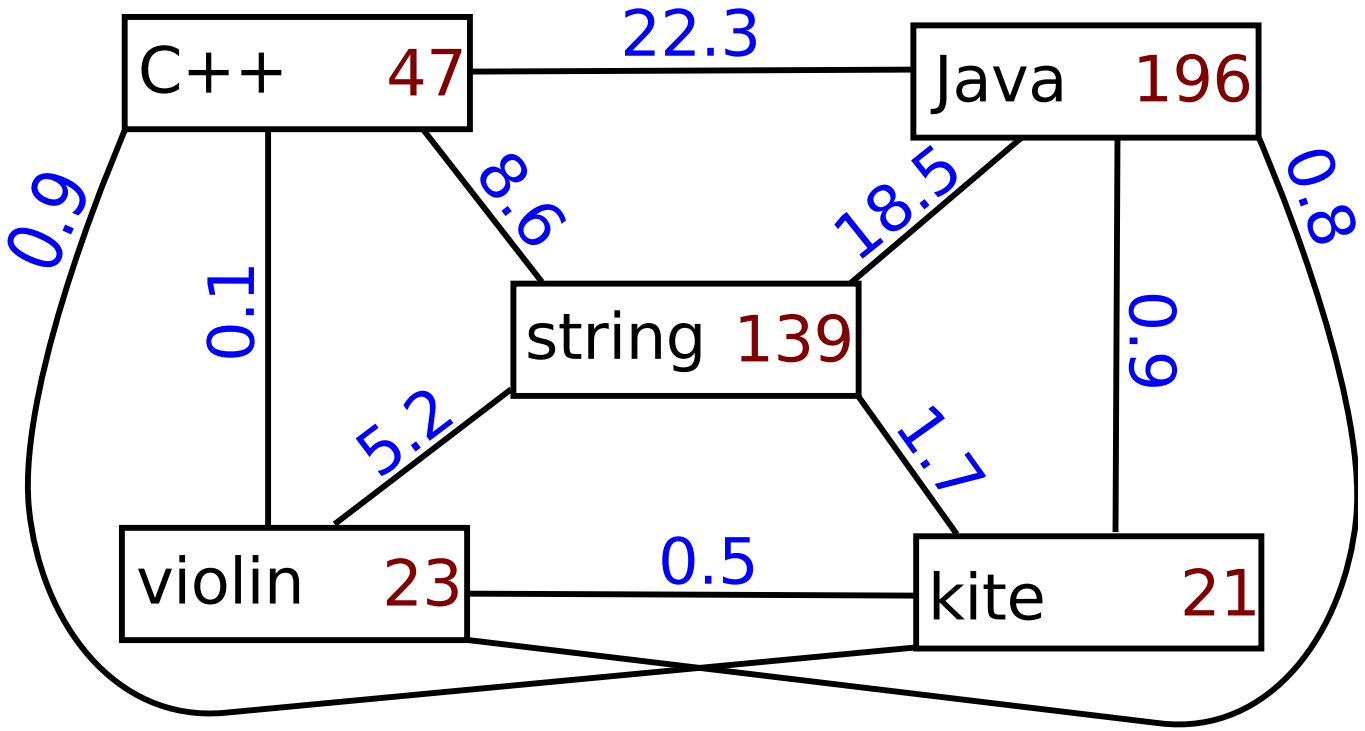


Efektívna výroba odporúčania

- čitateľ Č príde na webstránku
- nájdeme jeho kolegov
- náhodne vyberáme články, ktoré oni čítali, ale Č nie
- tie, ktoré sme vybrali najčastejšie, ponúkneme

Podobnosť významov slov

Štatistika: mocná zbraň pri skúmaní podobnosti slov



(milióny výskytov slov a dvojíc slov)

Koniec

Ďakujem za pozornosť!