L23 ~~L22:~~ PageRank

Jeff M. Phillips

April 11, 2018

Final Report

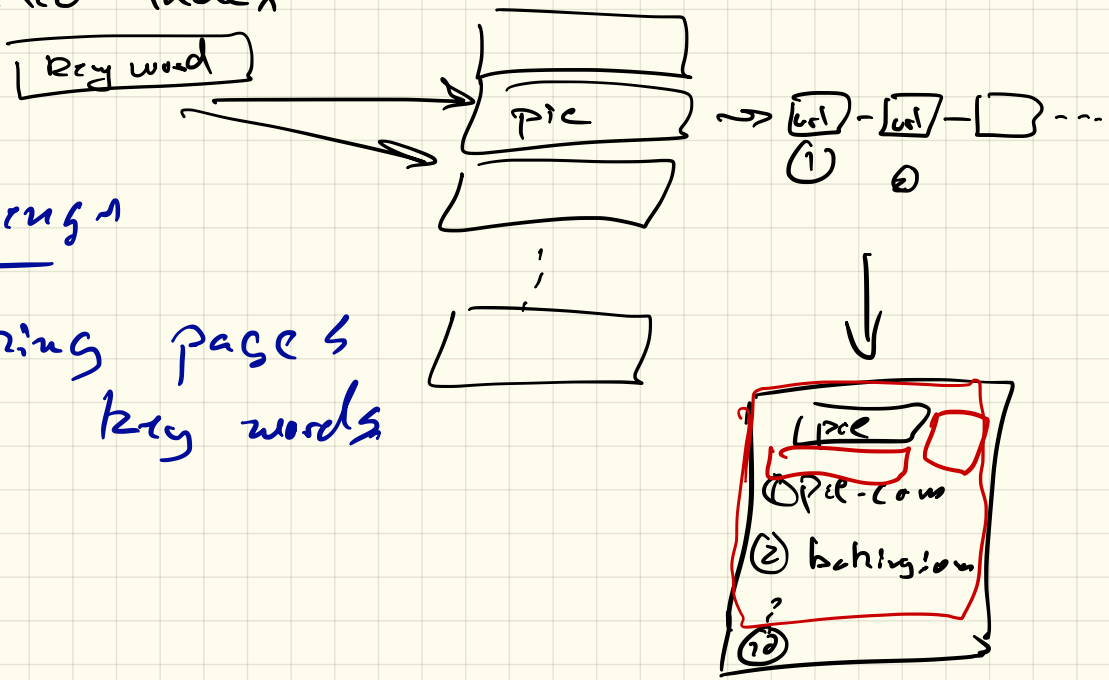
At most 4 pages/student. Don't cram in too much!

- Succinct title (and names)
- Problem definition and motivation.
- Explain your Data.
- **key idea**
- What did you do (which techniques, an implementation, a comparison, an extension)
- What did you learn? Artifacts (charts, plots, examples, math) and Intuition (in words, did it work?)

Page Rank ← key technology inside Google Search.

- Search Engine?

Inverted index



Big challenge

- Ranking pages for key words

Rank webpage on keyword?

→ pie

k-grams on text on page

+ Jaccard w/ {pie}

+ Minhashing

→ All Visits

---

cosine similarity $V_{pie} = (0, 0, 0, \ldots, 1, \ldots)$

bow page $(1, 2, 4_{pie}, \ldots, 7, 1, 0)$

What can go wrong?

↳ many copies of "pie"

Add more context.

search 

recipe

$V_{pie} = (0, 0, 0, \ldots, 0, 0, \ldots, 1, \ldots, 1, \ldots, 1, \ldots)$

delicious pie apple

what can break this?

↳ copy entire popular webpages

How did search engines know about pages?

→ Crawlers: goes to webpage, follows links, & records page.

hyperlinks text

text: very informative
↳ put into bow rep for page

page bow vec $(v_1, v_2, \ldots v_n)$

words on page

text hyperlinks $(v_1', v_2', \ldots v_m')$

$(\bar{v}_1, \bar{v}_2, \ldots \bar{v}_n)$

Index

pages hand-curated list of links

↳ Yahoo! LookSmart

1. Google
2. Youtube
3. Facebook
4. Baidu
5. Wikipedia
6. Reddit
7. Yahoo.
8. Google India
9. Tencent QQ
10. Amazon

Page Rank

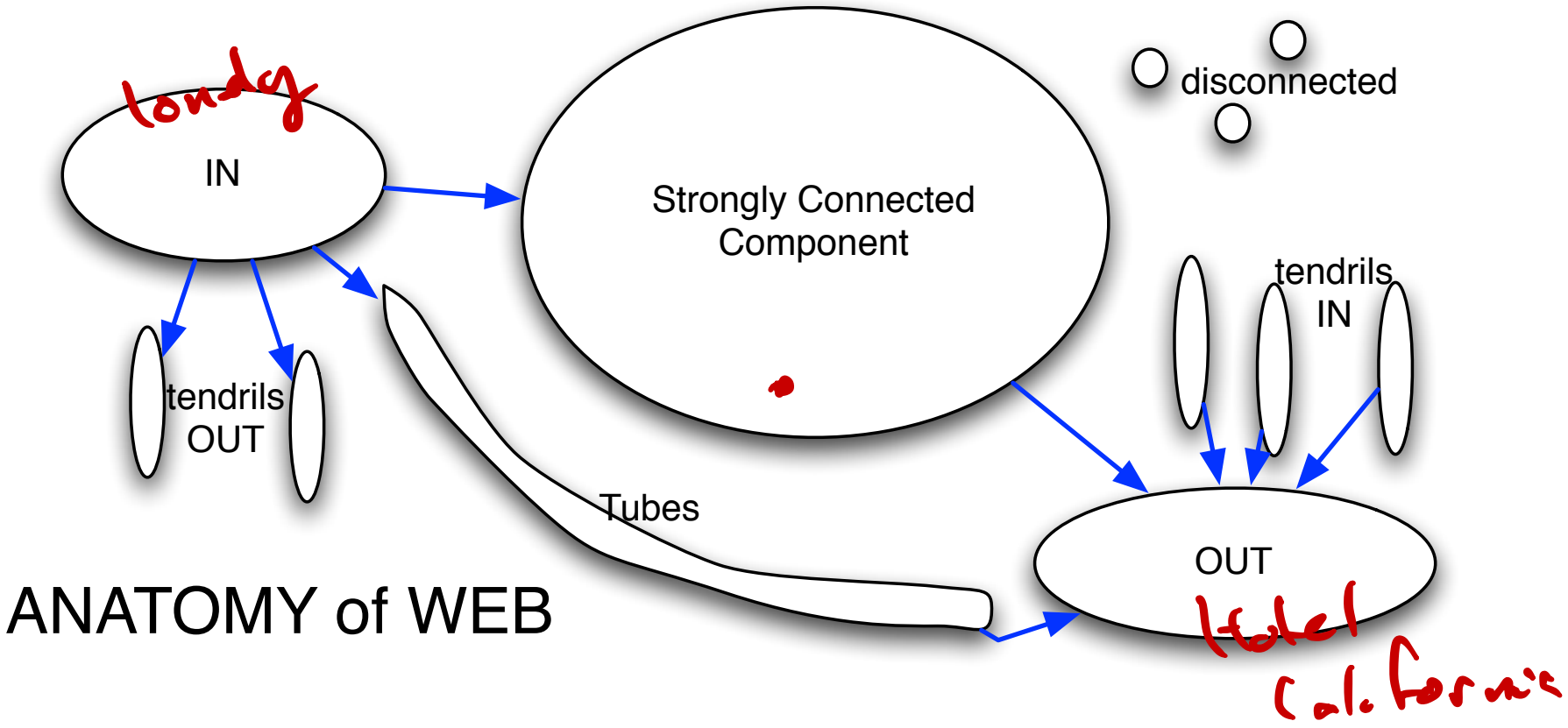
Idea #1. Pages are important if linked to by other important pages.

Idea #2 How likely a random surfer would find this page.

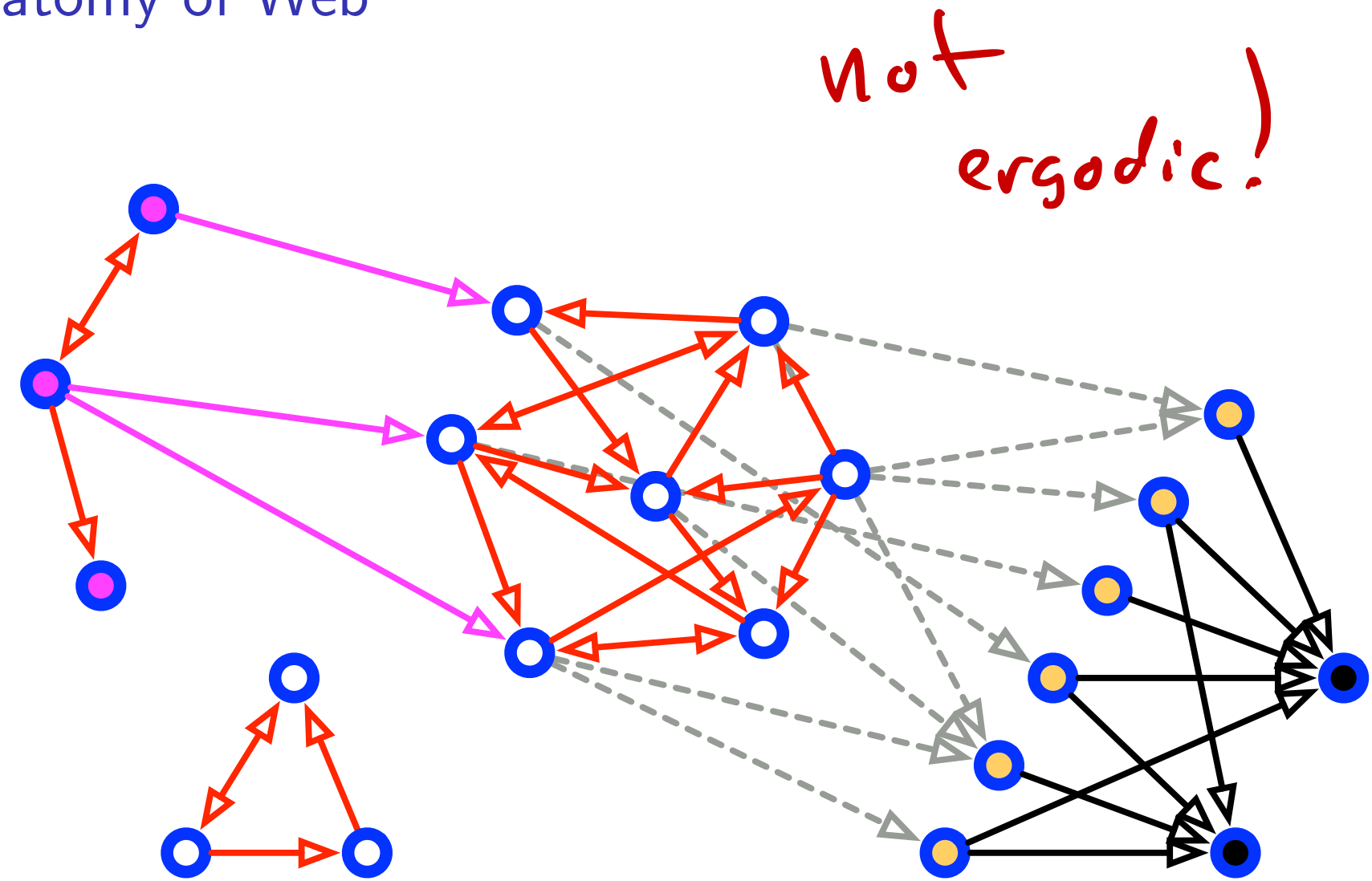
Markov Chain ← model of web graph.

ergodic ?

Anatomy of Web



Anatomy of Web



Teleportation (Taxation)

Idea: 15% steps, jump to random page.

$$q_* = P^n q_0$$

$$P_{mxm}$$

$$\beta = 0.15$$

$$q_{i+1} = \left((1-\beta)P + \beta Q \right) q_i$$

P' ergodic

$$q_* = (P')^n q_0$$

Page Rank Vec

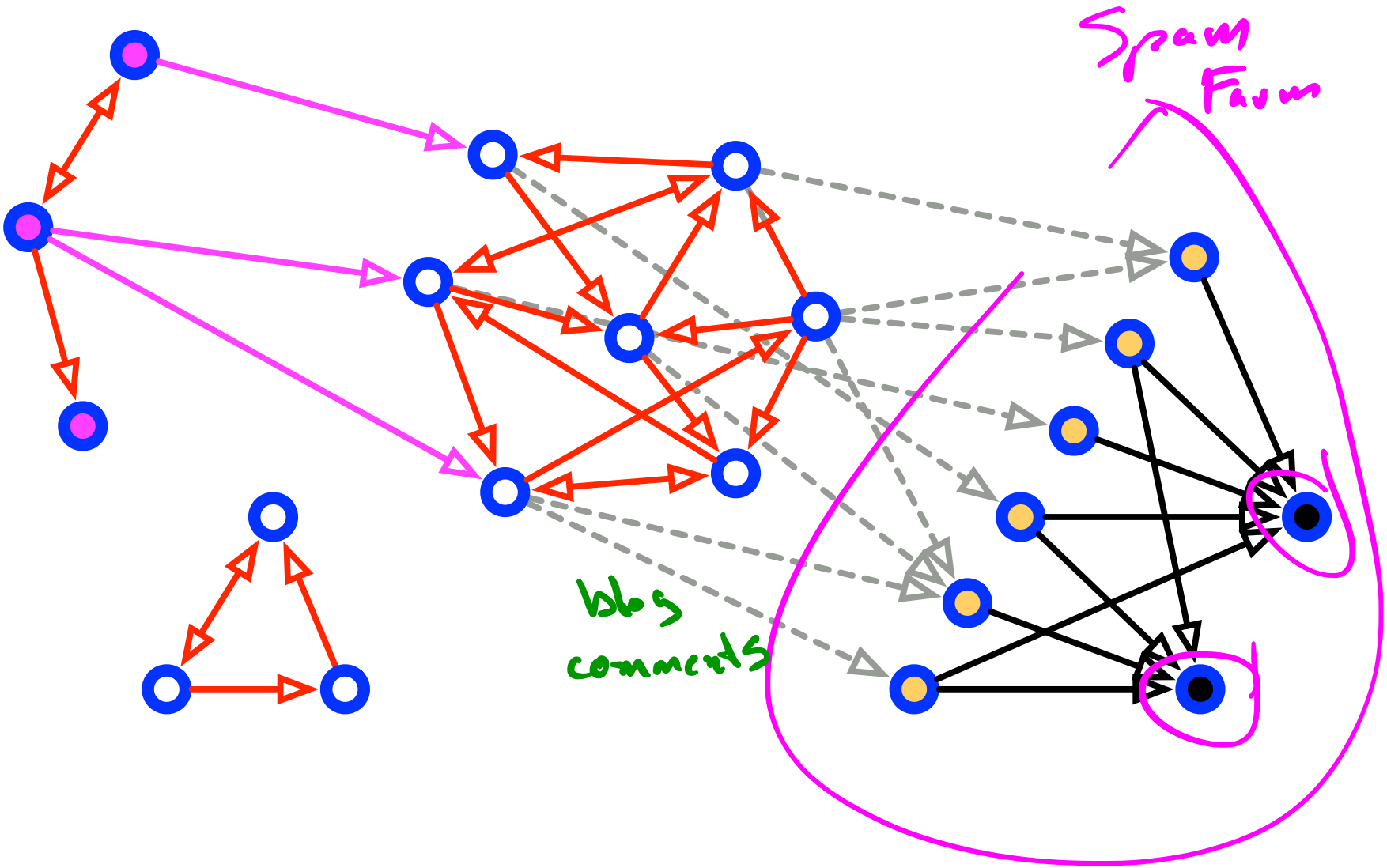
Q all $1/m$

$$\begin{pmatrix} 0.15/m \\ 0.15/m \\ \vdots \\ 0.15/m \end{pmatrix}$$

Data Mining

0
0
1/100
1/100
1/100
0
1/100
0
⋮

Spam Farms



$$eigs(M) = [V, L]$$

$$L = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

$$M = V L V^T$$

$$\sqrt{M} = V \sqrt{L} \qquad \sqrt{L} = \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_n} \end{pmatrix}$$