

Seek and Ye shall Find

The continuum of computer “intelligence”

2/28/2006

COS 116

Instructor: Sanjeev Arora

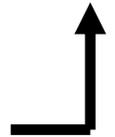
Recap: Binary Representation



Powers of 2

2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
1	2	4	8	16	32	64	128	256	512	1024

$$2^{10} = 1024 \approx 10^3$$



Fact: Every integer can be uniquely represented as a sum of powers of 2.

$$\begin{aligned} \text{Ex: } 25 &= 16 + 8 + 1 \\ &= 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \end{aligned}$$

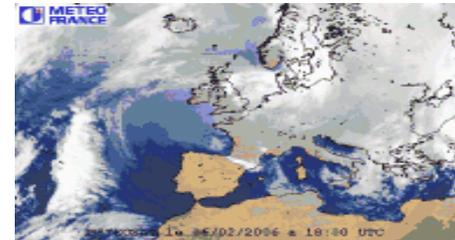
$$[25]_2 = 11001$$

Misconceptions about Computers

Just a calculator on steroids



Weather Forecast



Just maintains large amount of data



Airline Reservation System



Just does what programmer tells it



Yes, but ...

Various meanings of SEARCH



- Look up “Shirley Tilghman” in online phonebook.
- In consumer database, find “credit-worthy” consumers.
- Find web pages relevant to “computer music.”
- Among all cell phone conversations originating in Country X, identify suspicious ones.
- Search all religion and philosophy books of the world for meaning of life.

These are major scientific problems with many components

Engineering

Algorithms

Linguistics

Statistical
Modeling

Ethics, Policy,
Society

Electronic Phonebook

- **ASCII:** Agreed-upon convention for representing letters with numbers
- Example:

T	i	l	g	h	m	a	n	,	2	5	8	-	6	1	0	0
84	105	108	103	104	109	97	110	44	50	53	56	45	54	49	48	48

- Sorted Phonebook = sorted array of numbers
- Use binary search

33 !	65 A	97 a
34 "	66 B	98 b
35 #	67 C	99 c
36 \$	68 D	100 d
37 %	69 E	101 e
38 &	70 F	102 f
39 '	71 G	103 g
40 (72 H	104 h
41)	73 I	105 i
42 *	74 J	106 j
43 +	75 K	107 k
44 ,	76 L	108 l
45 -	77 M	109 m
46 .	78 N	110 n
47 /	79 O	111 o
48 0	80 P	112 p
49 1	81 Q	113 q
50 2	82 R	114 r
51 3	83 S	115 s
52 4	84 T	116 t
53 5	85 U	117 u
54 6	86 V	118 v
55 7	87 W	119 w
56 8	88 X	120 x
57 9	89 Y	121 y
58 :	90 Z	122 z
59 ;	91 [123 {
60 <	92 \	124
61 =	93]	125 }
62 >	94 ^	126 ~
63 ?	95 _	127 □
64 @	96 `	128 €

Rest of the lecture: Web Search

Web Desktop News Images Local (BETA) Encarta

shirley tilghman

+Search Builder Settings Help Español



* Were you looking for ['tilghman' near Shirley, NY](#)

Web Results

Page 1 of 33,281 results containing **shirley tilghman** (0.38 seconds)

[Google Press Center: Press Release](#)

Google Appoints **Shirley M. Tilghman**, Ph.D., to its Board of Directors MOUNTAIN VIEW, Calif. - October 5, 2005 - Google Inc. (NASDAQ

www.google.com/press/pressrel/tilghman_board.html [Cached page](#)

[Princeton University Office of the President - President's Biography](#)

Shirley M. Tilghman was elected Princeton University's 19th president on May 5, 2001, and assumed office on June 15, 2001. An exceptional teacher and a world-renowned scholar and leader in the field ...

www.princeton.edu/president/biography [Cached page](#)

[Princeton - News - Shirley Tilghman named University's 19th President](#)

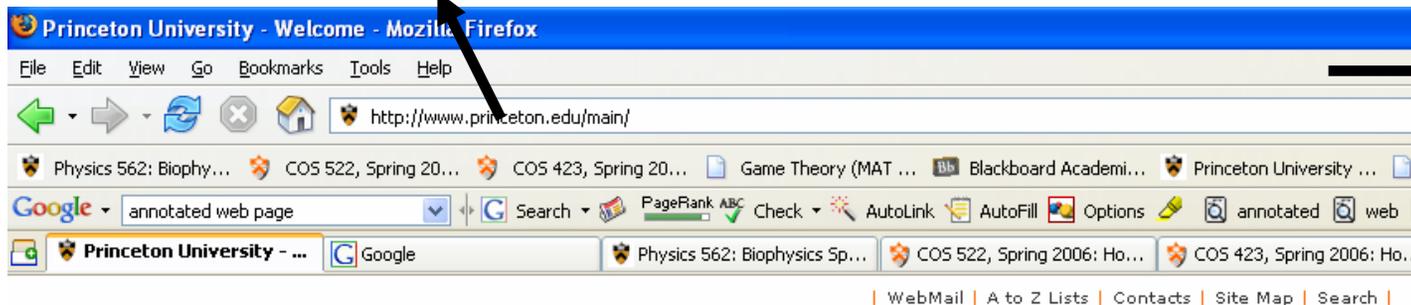
Shirley Tilghman named Princeton University's 19th President Princeton, N.J. -- **Shirley M. Caldwell Tilghman**, a member of the Princeton University faculty since 1986, an ...

www.princeton.edu/pr/news/01/q2/0505-tilghman.htm [Cached page](#)

World Wide Web (simplified view)



URL: Unique address for each document



Browser



Web Page

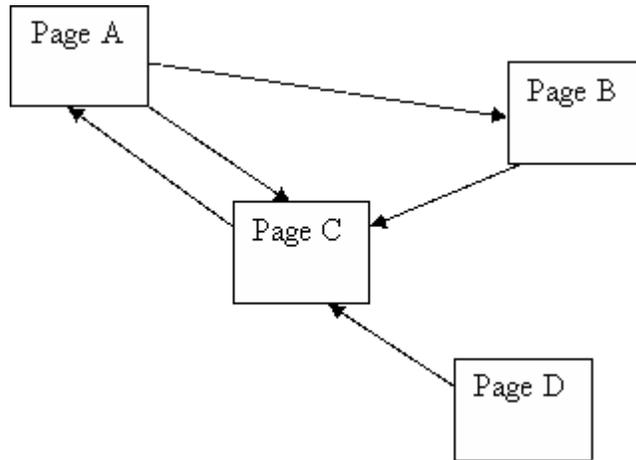
Hyperlink



Future lecture: Physical infrastructure of the Web

Routers, gateways, DNS, etc.

Logical Structure of the Web



“Directed graph”

“edges” = link from one node to another

- **Important:** This logical structure is created by independent actions of 100s of millions of users

1st step for search engines: create snapshot of the web



■ **Webcrawler:** Browser on autopilot

- Maintains array of web pages it has seen
- 2 types of pages: “visited”, “fully explored”
- Do forever

{

Pick any webpage marked “visited” from array.

Mark it “fully explored.”

Open all its linked pages in browser.

Save them in array and mark them “visited.”

}

Feasibility Calculation

- About 15 billion web pages today.
- Say 10 Kilobytes (10,000 bytes) of data per page
- 15×10^{13} bytes to store the web
- $\approx 150,000$ Gb
- ≈ 500 Hard Disks (about \$150,000)

Searching for “Computer Music”

Ideas?

- Identify all pages that contain “Computer Music.”
- Sort according to number of occurrences of “computer music” in the page.
- Human staff computes answers to all possible questions.



Some pitfalls

- “Spamming” by unscrupulous websites
- Synonymy
- Polysemy

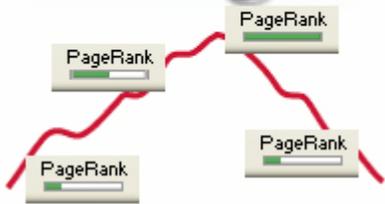
Solution



IBM's CLEVER – 1996



Google's PAGERANK – 1997



Take advantage of the link structure of the web

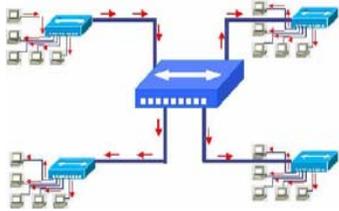
Web link confers “approval”

CLEVER



Authorities: Sites that are viewed “with respect” by many

- New York Times
- International Computer Music Association



Hubs: Clearinghouses of information

- “My favorite computer music links”

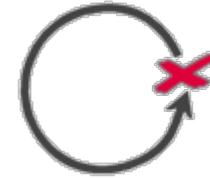
Typically Authorities point to hubs and hubs point to authorities

Circular Definition?

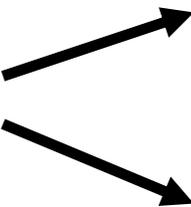


Circular Definition – *see* Definition, Circular

Breaking Circularity



- Iterative algorithm

- Start with  Pages containing “Computer music”
All pages they point to

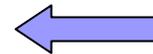
- At every step each page has:
 - “Hub Score”
 - “Authority Score” } Initially all 1

Score Calculation

- Do forever

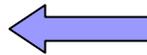
{

Next Hub Score for page



Sum of current Authority
Scores of pages that link
to it.

Next Authority Score for page



Sum of current Hub
Scores of pages that link
to it.

}

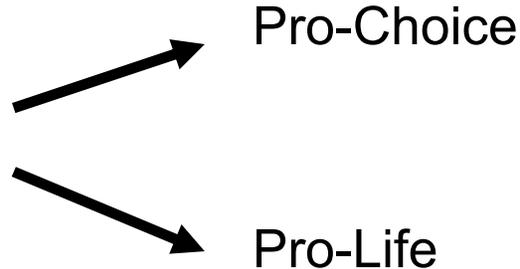
Fact The scores converge.

(Proof uses Linear Algebra,
Eigenvalues)

- By Product – Algorithm reveals **clusters**

Example:

“Abortion”

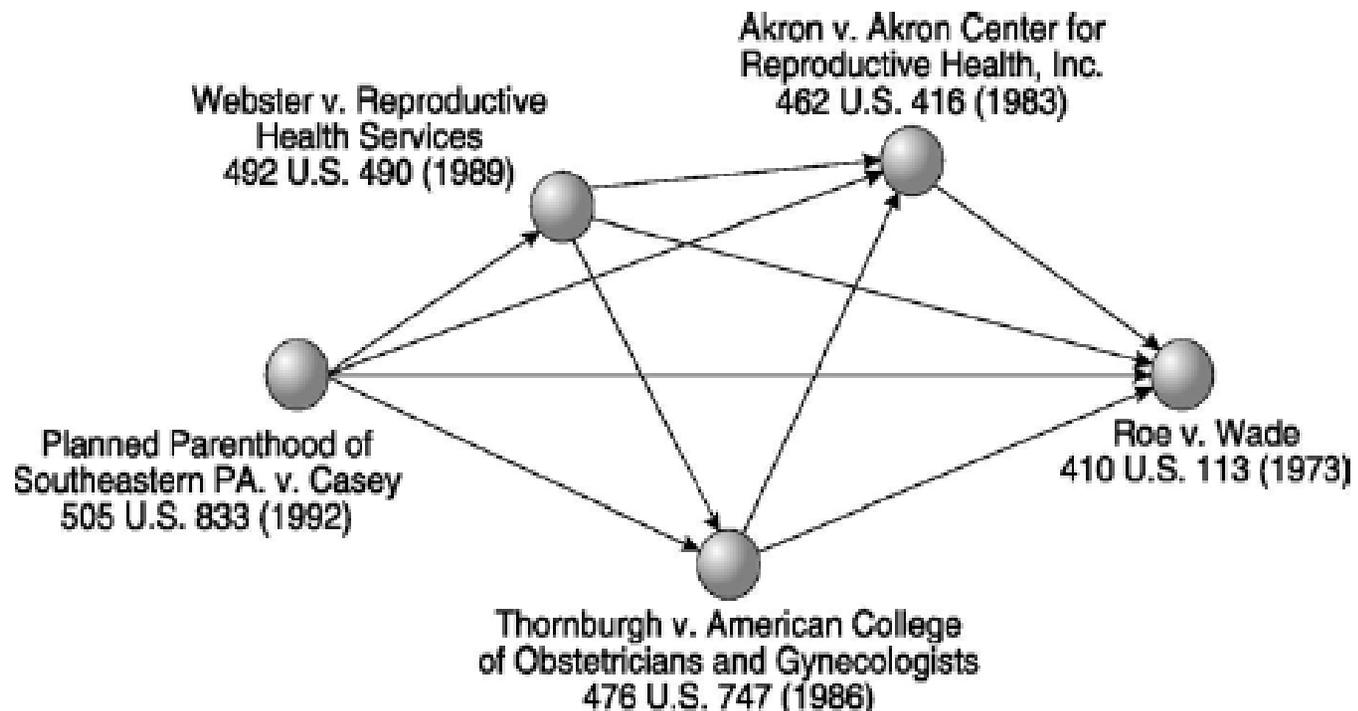


- **Data Mining** – Process of finding answers that are not in the data and must be inferred.

Example: “How is a person who shops at Whole Foods & REI likely to vote?”

[Fowler and Jeon, '05]

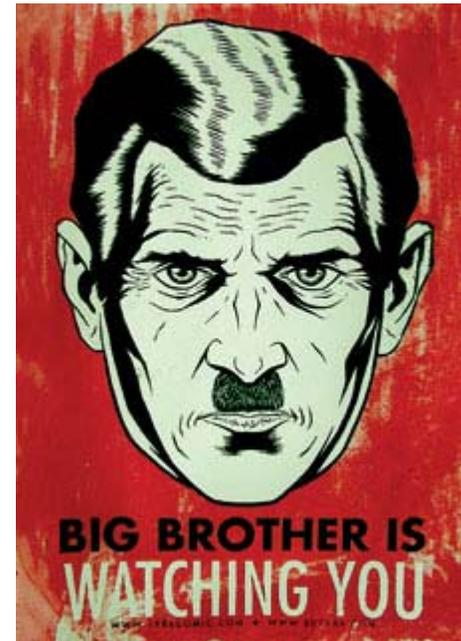
FIGURE 1. Network of Selected Landmark Abortion Decisions



Concerns

From **users**:

- Privacy
- Privacy
- Privacy



From **Computer scientists**:

- Formalize privacy
- How to safeguard privacy while allowing legitimate computations



Qs for next time:

What is computation?

What can computers not do?

Also, 10-min discussion of readings for today's lecture.