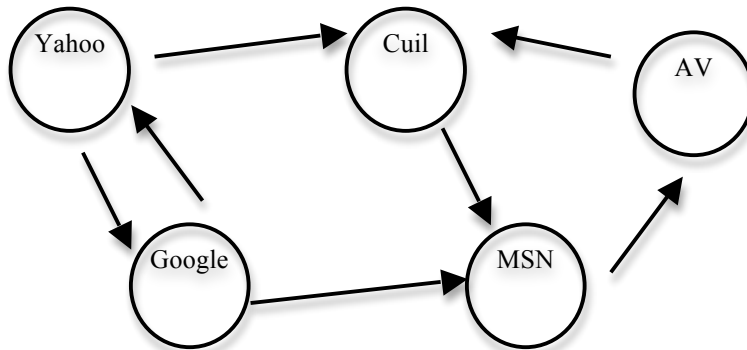


## Homework #5 (due in 3 weeks)

### Questions (30 points)

[1] Problem 21.10 (pg. 432 of IIR).

[2] Given the following directed graph of webpages (or sites), perform two iterations of PageRank computations. The arcs indicate outbound links between webpages. Initially give each page a PageRank score of 1. Use a 'teleport' (or transition) probability of 0.20. (Put differently, 80% of the time the random surfer visits a new page). Show the PageRank scores for each page after both iterations.



### Web Query Log Analysis (70 points)

On the course web site I have put a file containing queries that were submitted to an Internet search engine. (Note: the file is large, over 35 MB even compressed.) The data comes from the Excite search engine from 12/20/1999. This exercise asks you to work with this data. You can use any tools you want to perform your analysis. You can use: programs developed for earlier exercises; new programs; commercial or public domain tools (e.g., Excel, MySQL, Perl); or, Unix commands (e.g., grep, wc). Basically, use whatever tool(s) you see fit. The data in the log file is unfiltered and may contain objectionable content. The data are in four tab separated fields: timestamp; hashed user id; results-starting-point; and the query string (which may contain punctuation and spaces).

Analyze the data in the query log. I will leave the exact particulars to you, but I would expect you to include the easier items in the list below and some of the more interesting, but harder ones. I don't expect you to investigate all of the questions below.

- What is the mean number of queries per user id?
- Analyze the variability of query length (i.e., in words or in characters)
- What percentage of queries are mixed case? Upper case? Lower case?
- What percent of the time does a user request only the top 10 results? Top 20 results?
- Count the number of questions (look for patterns such as starting with Wh-words, or ending with a '?' symbol). What percentage of queries do questions make up? What is the most common type of question?
- What are the most common queries issued?
- What percent of queries contain stopwords like 'and', 'the', 'of', 'in', 'for'?
- How often is 'query' syntax used, like phrases in quotes, or '+' or '-' signs?
- What are the 10 most common words appearing in queries that contain the word *download*?
- What are the most common k words appearing in queries. (say for k=20)?
- What percentage of queries were asked by only one user?
- How often is a consecutive query a reformulation of the previous one? (Not the same query to greater depth.)
- What kind of spelling mistakes do users make?
- Which occurs more often "Al Gore" or "Johns Hopkins"? "Johns Hopkins" or "John Hopkins"?
- How often do URLs appear in queries?
- Is it likely that this web query log puts anyone's privacy at risk?
- Is query volume constant throughout the day?
- Other interesting questions you come up with.

Briefly describe the methods and tools you used, and summarize the conclusions you reach. In addition to your analysis, hand in any source code that you write for the assignment.

First 15 lines:

090000	B0A0F80A06A3AB6C	0	In what year did baseball become an official sport?
090000	95A33E619934A39B	0	wirehair pointing griffon
090000	E613C21C535BC636	30	ncic
090000	00CD4DE085A391DD	0	+ER +home +TV +Romano +picture
090000	5F48819400DB52D7	0	adolescent won't sleep in own bed
090000	D87CE5C149126B4B	0	where can i find free porno passwords
090000	47F6F715137F7C8D	0	play station codes . com
090000	40B1AACE633D9AC9	30	birth control and depression
090000	43D7E2332D3631DC	0	government
090000	87BE88FDCB1F7629	0	"WorkAbility I"+conference
090000	687340036669C45D	0	kitchen appliances
090000	E43DD6D82BFBD0B8	0	where can I find a chine rosewood
090000	CA52ECD1524E737D	0	jennifer love hewitt naked
090000	2B4FAF545C0E6EF0	40	pageant trim
090000	6456584F5B316AAE	100	tiger electronics

Last 15 lines:

165959	5F083C02AF42D762	10	"master boot record, fdisk"
165959	9F57839B10170414	0	beastiality
165959	2302407F00A4D6F9	870	www. lynn white.com
165959	4AD556DDCA079EB8	0	Where can i find crafts over the internet?
165959	FD647F92E62C1999	0	Where can I find a child's lilac sweater?
165959	3DF4E9B0AFF6B808	0	windows fix irq
165959	590F4121ABC62C02	0	what are the longterm physical effects of methamphetamine use in women?
165959	E43DD6D82BFBD0B8	0	chugach mountain
165959	4EB35F3114240AEE	110	alphabet hawiiian
165959	8BA362CFA3B96117	0	"free internet access"
165959	C81EA097CF872C51	0	yahoo
165959	3365AF9F3B5CB4D9	0	alta vista
165959	32E290F942064B3A	0	body surface area drug dosage
165959	4D71604181DC294A	10	SLSA AND Australia
165959	302D2CA498C522F4	0	start up win95