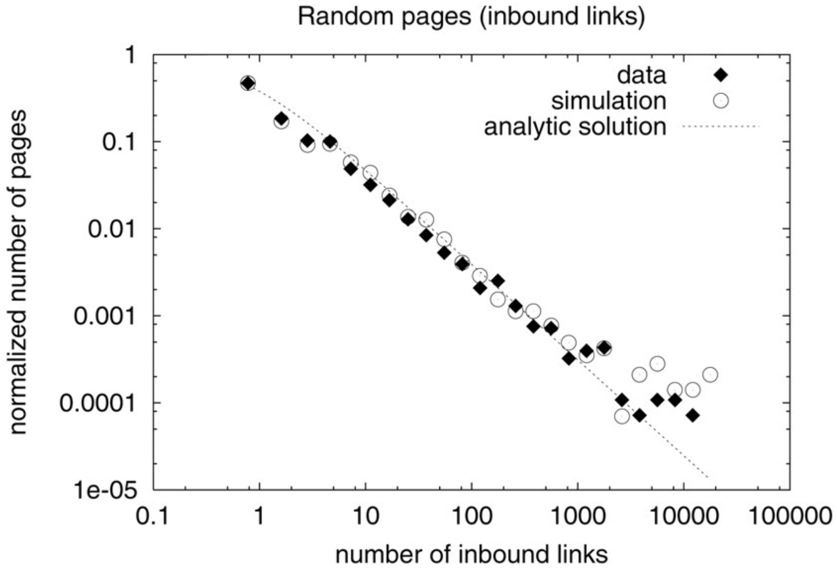
Networks in Web & IR

Outline

- the web graph
  - degree distribution
  - clustering
  - motif profile
  - communities
- ranking pages
  - PageRank
  - HITS
- more exotic applications
  - summarization LexRank
  - query connection subgraphs
  - machine translation

degree distribution

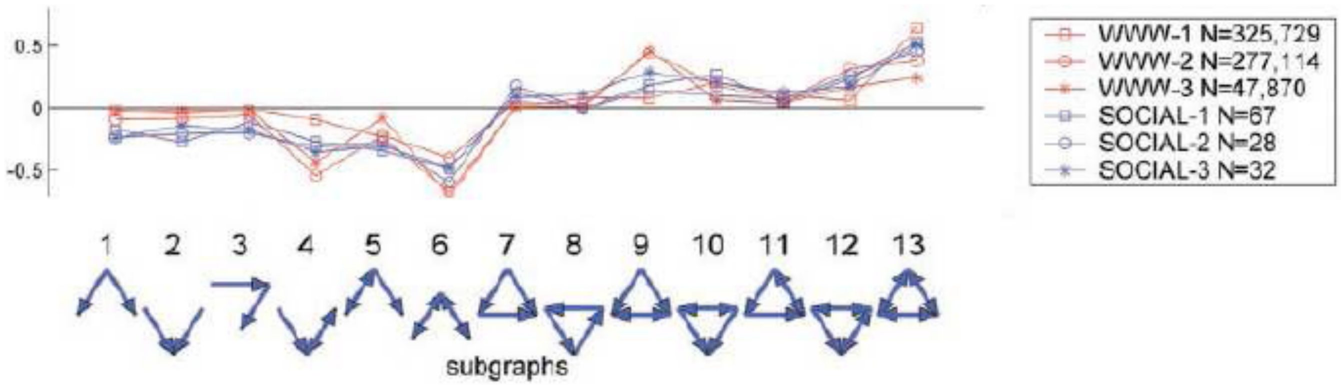
- indegree, α ~ 2.1
- outdegree, α ~ 2.4



source: Pennock et al.: Winners don't take all: Characterizing the competition for links on the web
PNAS April 16, 2002 vol. 99 no. 8 5207-5211

clustering & motifs

- clustering coefficient ~ 0.11 (at the site level)

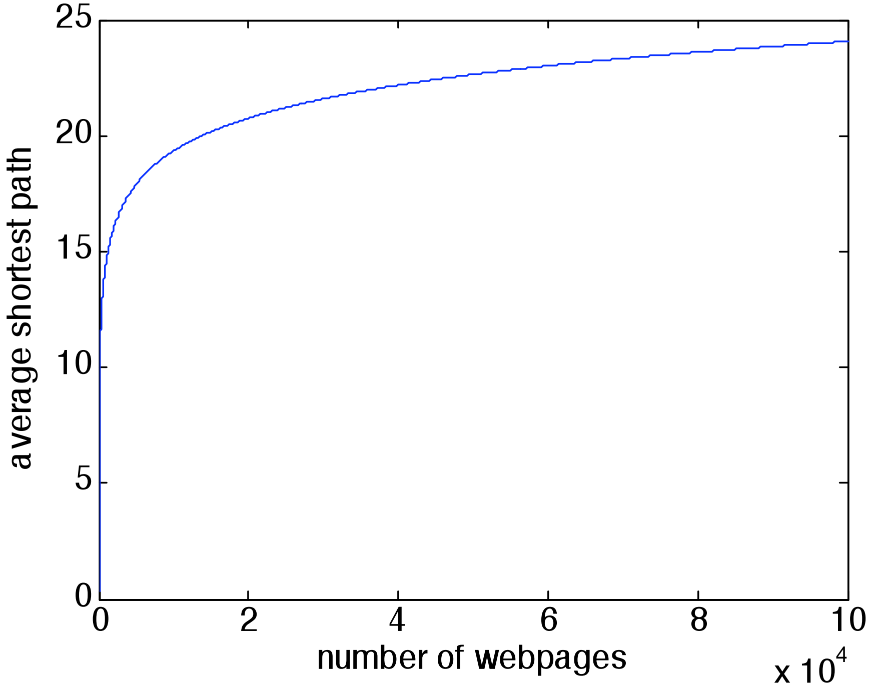


Source: Milo et al., "Superfamilies of evolved and designed networks", Science 303 (5663), p. 1538-1542, 2004.

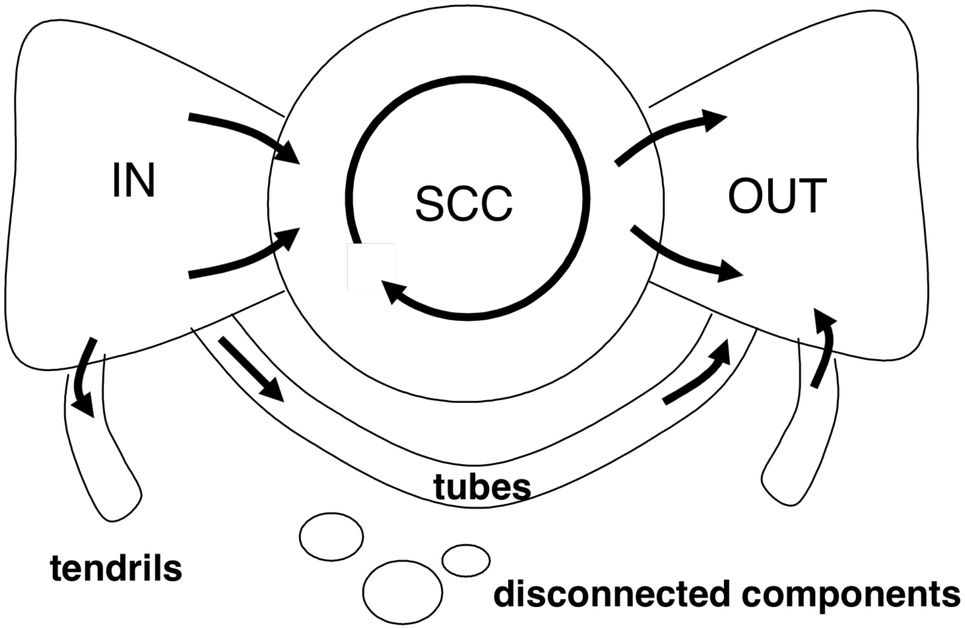
shortest paths

- <d> = 0.35 + 2.06 log(N)

- prediction: <d> = 17.5 for 200 million nodes
- actual: <d> = 16 for reachable pairs



bowtie



that was the web graph overall

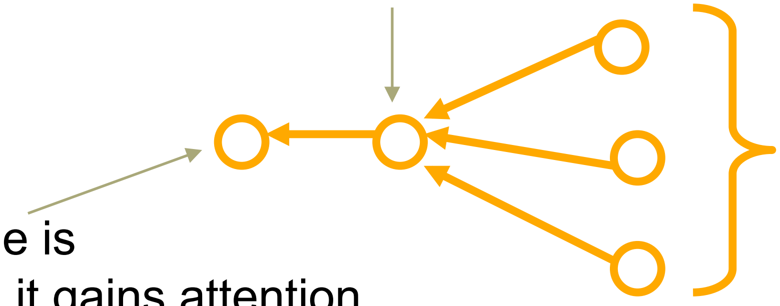
- How do we know which individual pages are important?

Outline

- the web graph
  - degree distribution
  - clustering
  - motif profile
  - communities
- ranking pages
  - PageRank
  - HITS
- more exotic applications
  - summarization LexRank
  - query connection subgraphs
  - machine translation

PageRank: bringing order to the web

- It’s in the links:
  - links to URLs can be interpreted as endorsements or recommendations
  - the more links a URL receives, the more likely it is to be a good/ entertaining/provocative/authoritative/interesting information source
  - but not all link sources are created equal
    - a link from a respected information source
    - a link from a page created by a spammer



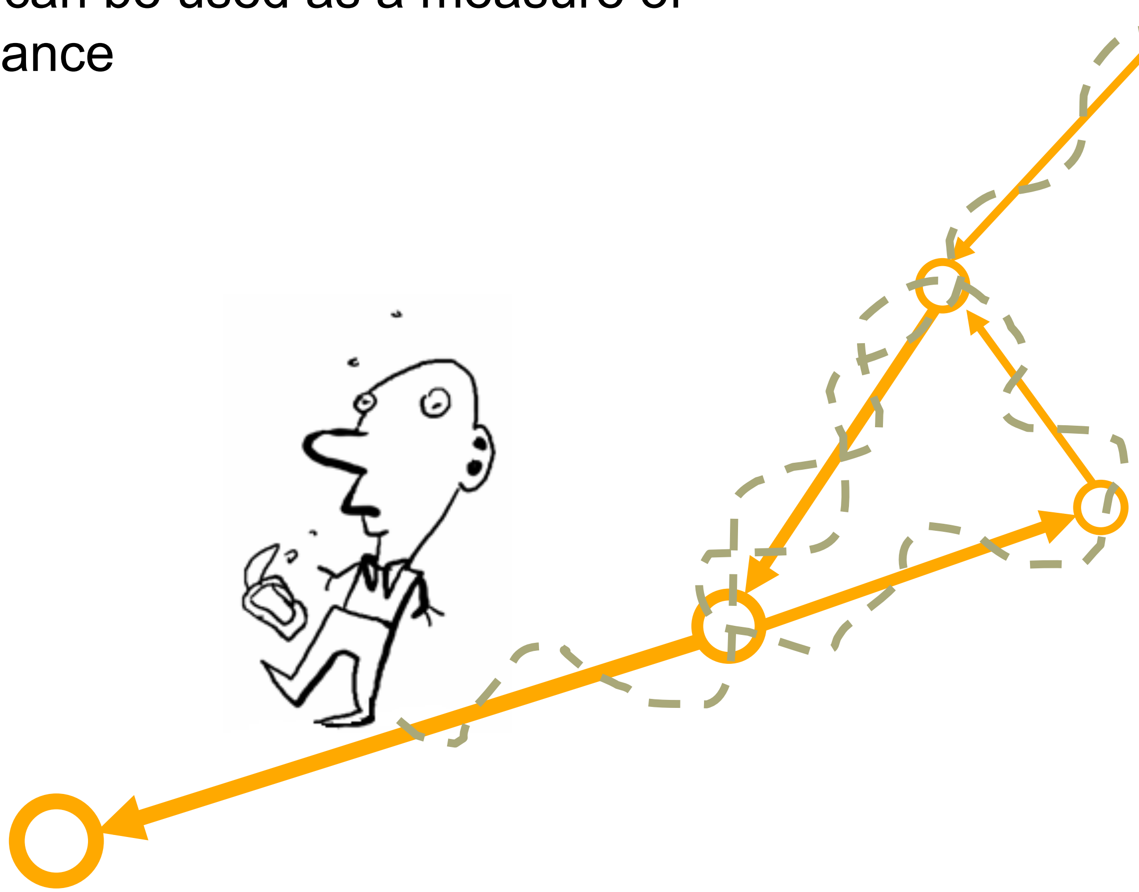
an important page, e.g. slashdot

Many webpages scattered across the web

if a web page is slashdotted, it gains attention

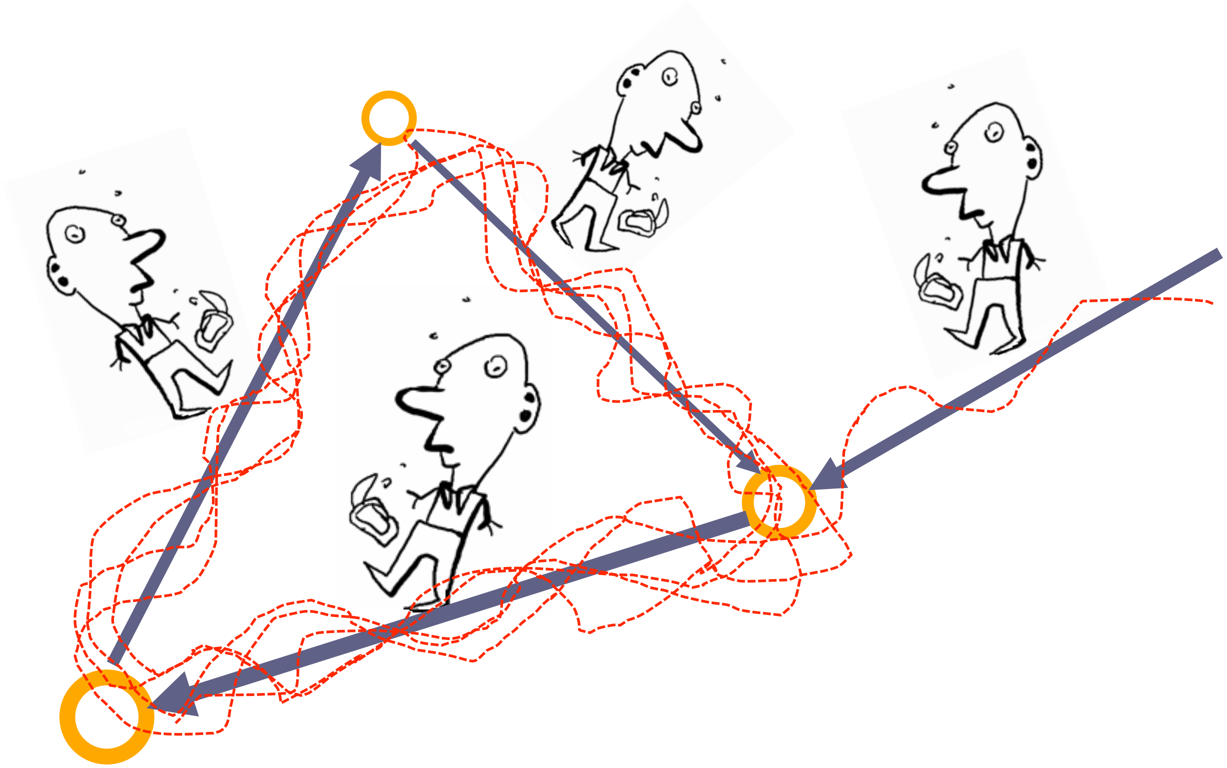
Ranking pages by tracking a drunk

- A random walker following edges in a network for a very long time will spend a proportion of time at each node which can be used as a measure of importance



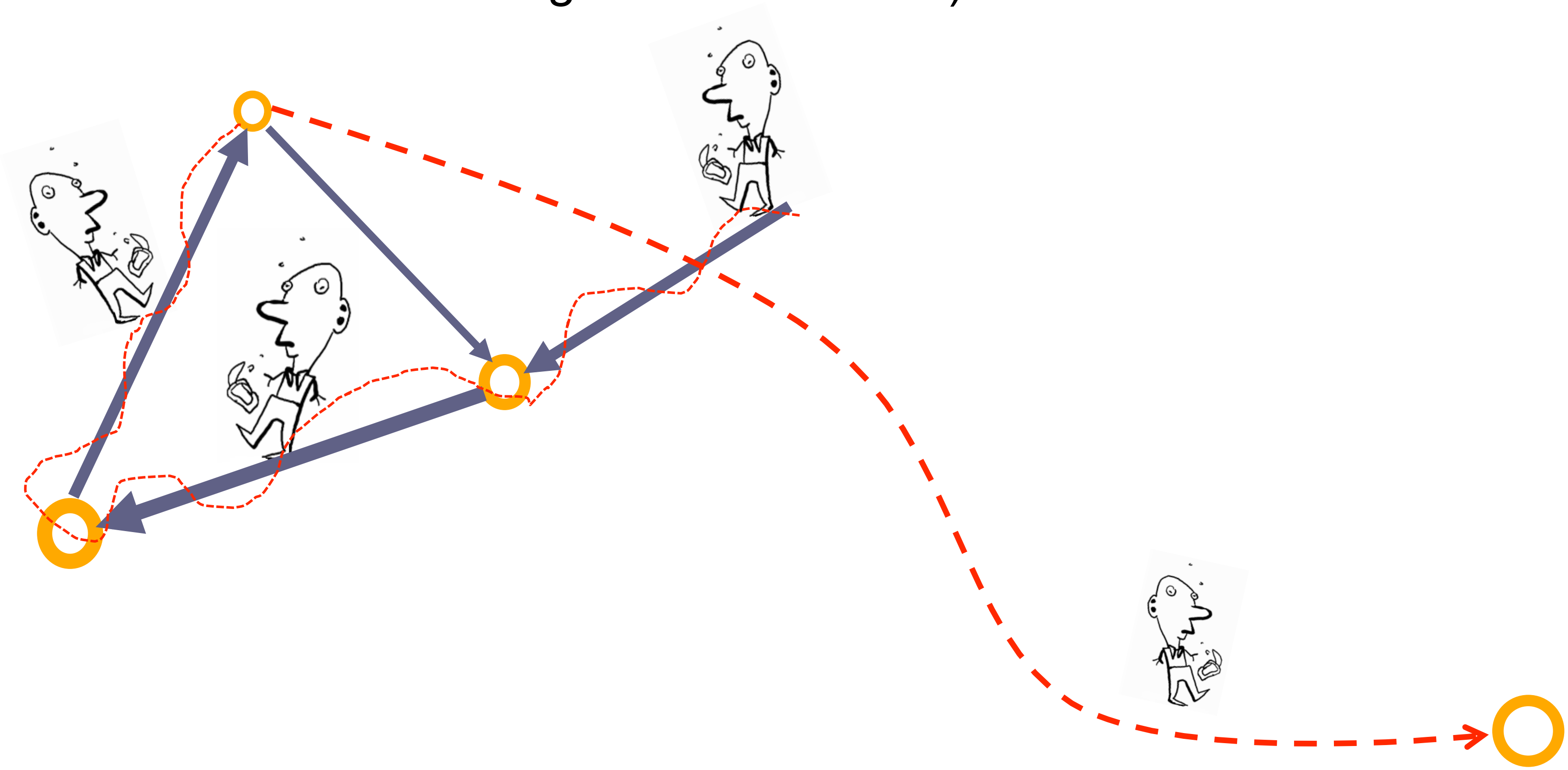
Trapping a drunk

- Problem with pure random walk metric:
  - Drunk can be “trapped” and end up going in circles

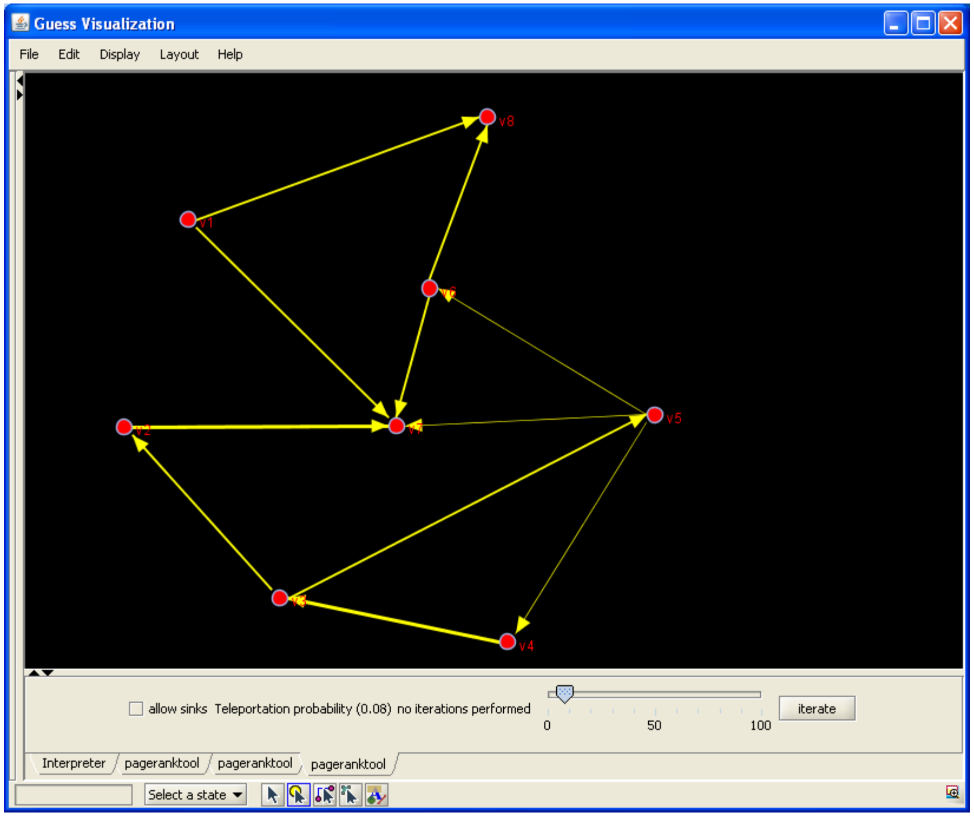


Ingenuity of the PageRank algorithm

- Allow drunk to teleport with some probability
  - e.g. random websurfer follows links for a while, but with some probability teleports to a “random” page (bookmarked page or uses a search engine to start anew)



lab exercise: PageRank



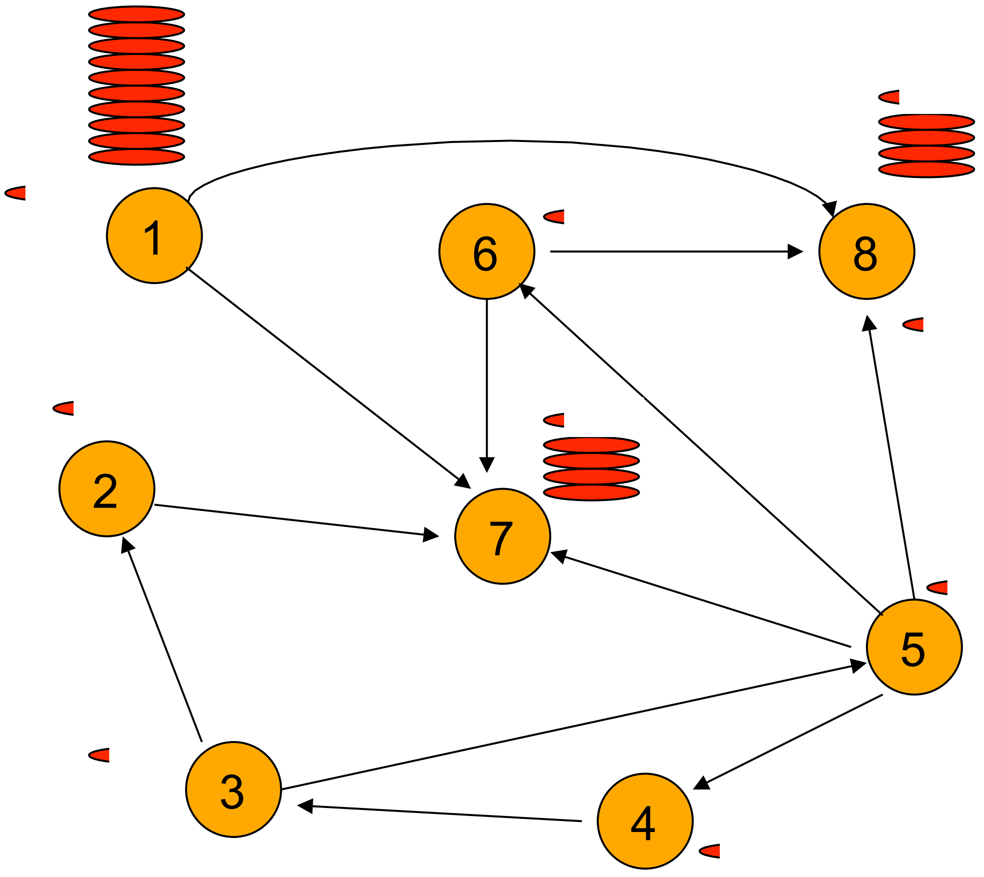
□ allow sinks Teleportation probability (0.08) no iterations performed

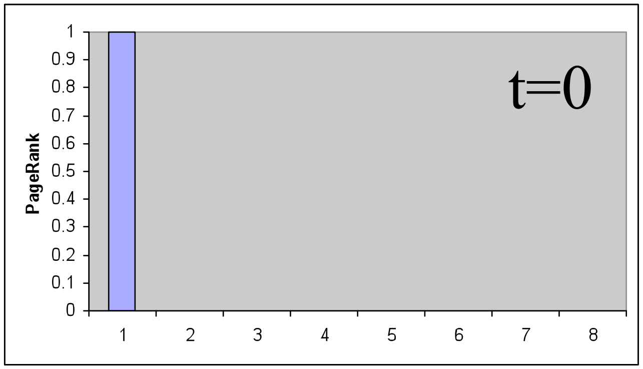
- What happens to the relative PageRank scores of the nodes as you increase the teleportation probability?
- Can you construct a network such that a node with low indegree has the highest PageRank?

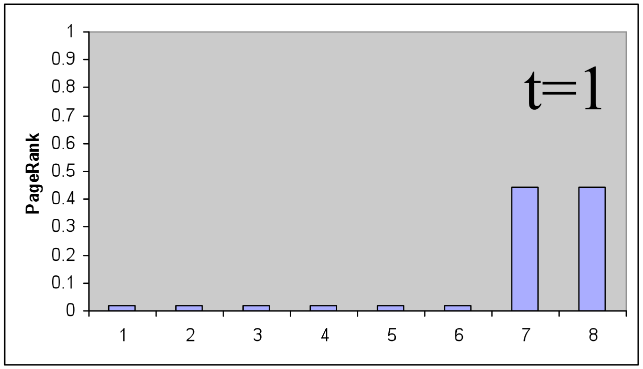
http://projects.si.umich.edu/netlearn/GUESS/pagerank.html

example: probable location of random walker after 1 step

20% teleportation probability

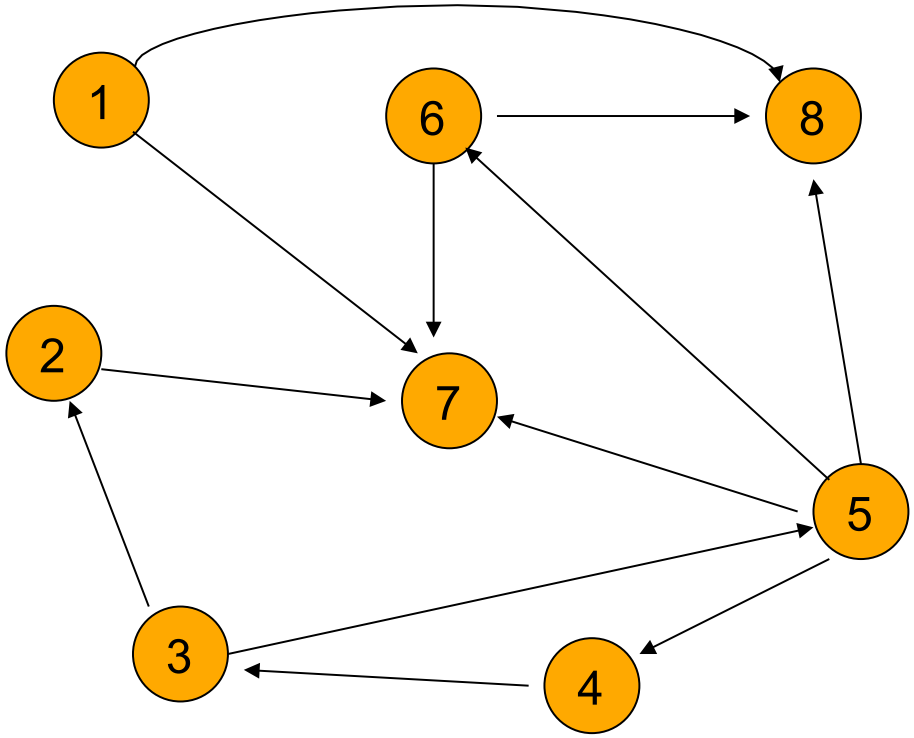


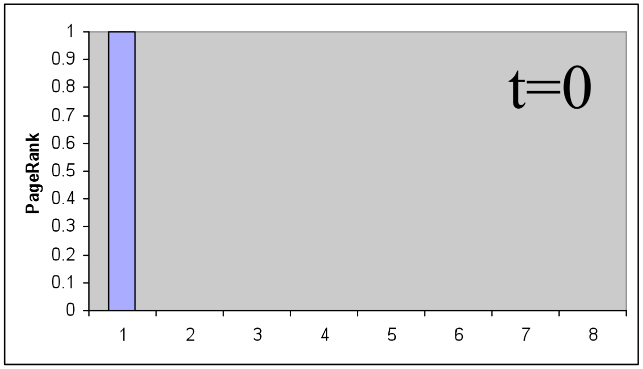


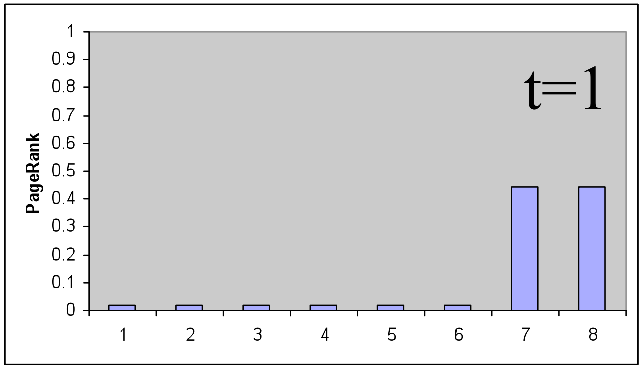


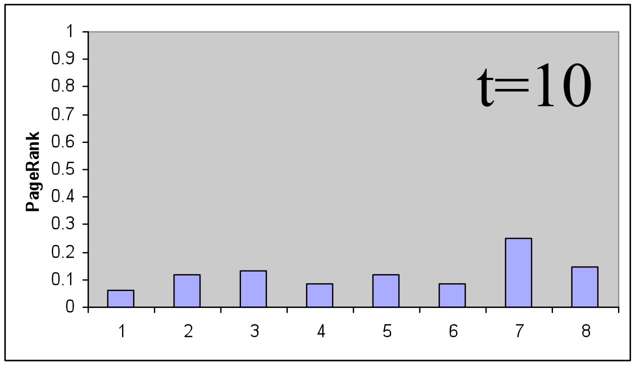
slide adapted from: Dragomir Radev

example: location probability after 10 steps







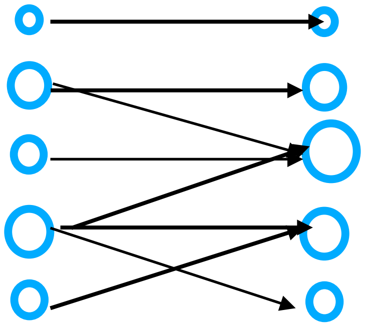


slide from: Dragomir Radev

alternate web page ranking algorithms: HITS

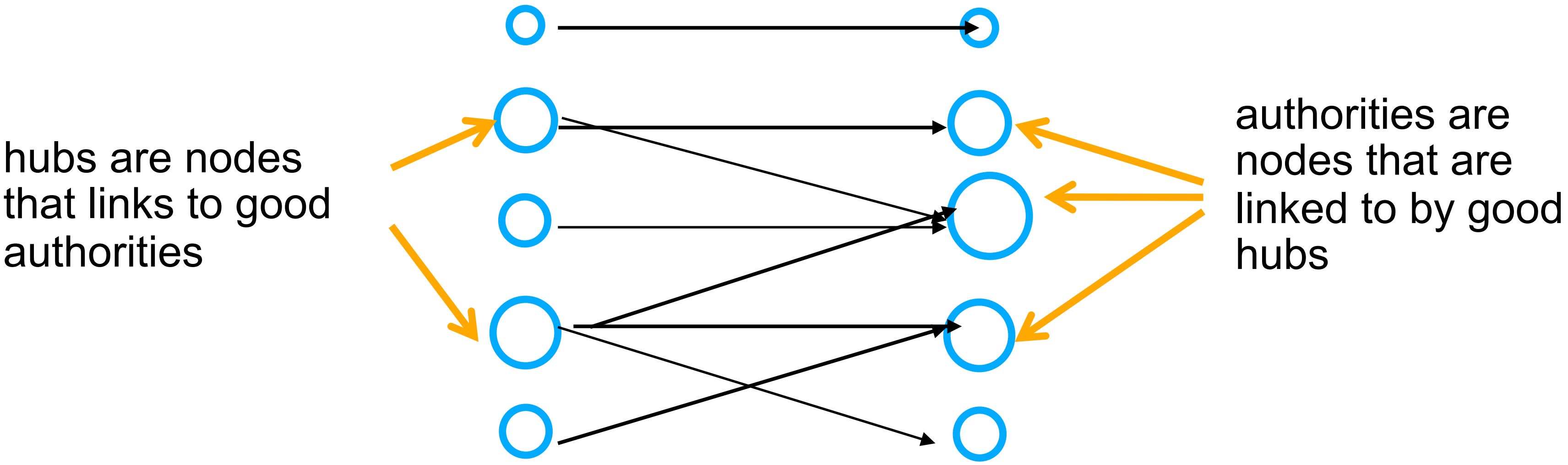
- HITS algorithm (developed by Jon Kleinberg, 1997):
  - start with a set of pages matching a query
  - expand the set by following forward and back links
  - take transition matrix E, where the i,j^th entry E_{ij} =1/n_i if i links to j, and n_i is the number of links *from i*
  - then one can compute the authority scores a, and hub scores h through an iterative approach:

$$\underline{a}' = E^T \underline{h} \qquad \underline{h}' = E\underline{a}$$



HITS: hubs and authorities

- recursive definition:



hubs are nodes that links to good authorities

authorities are nodes that are linked to by good hubs

Outline

- the web graph
  - degree distribution
  - clustering
  - motif profile
  - communities
- ranking pages
  - PageRank
  - HITS
- more exotic applications
  - summarization LexRank
  - query connection subgraphs
  - machine translation

Applications to IR beyond plain hyperlinks

- Can we use the notion of centrality to pick the best summary sentence?
- Can we use the subgraph of query results to infer something about the query
- Can we use a graph of word translations to expand dictionaries? disambiguate word meanings?

Centrality in summarization

- Extractive summarization (pick k sentences that are most representative of a collection of n sentences)
- Motivation: capture the most central words in a document or cluster
- Centroid score [Radev & al. 2000, 2004a]
- Alternative methods for computing centrality?

slide from: Dragomir Radev

Sample multidocument cluster

(DUC cluster d1003t)

1 (d1s1) Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.

2 (d2s1) Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.

3 (d2s2) Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.

4 (d2s3) Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.

5 (d3s1) The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.

6 (d3s2) Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, ``will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."

7 (d3s3) Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).

8 (d4s1) The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.

9 (d5s1) British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq ``did not end" and that Britain is still ``ready, prepared, and able to strike Iraq."

10 (d5s2) In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq ``will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.

11 (d5s3) A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

slide from: Dragomir Radev

Cosine between sentences

- Let s_1 and s_2 be two sentences.
- Let *x* and *y* be their representations in an *n*-dimensional vector space
- The cosine between is then computed based on the inner product of the two.

$$\cos(x, y) = \frac{\sum_{i=1,n} x_i y_i}{\|x\| \|y\|}$$

- The cosine ranges from 0 to 1.

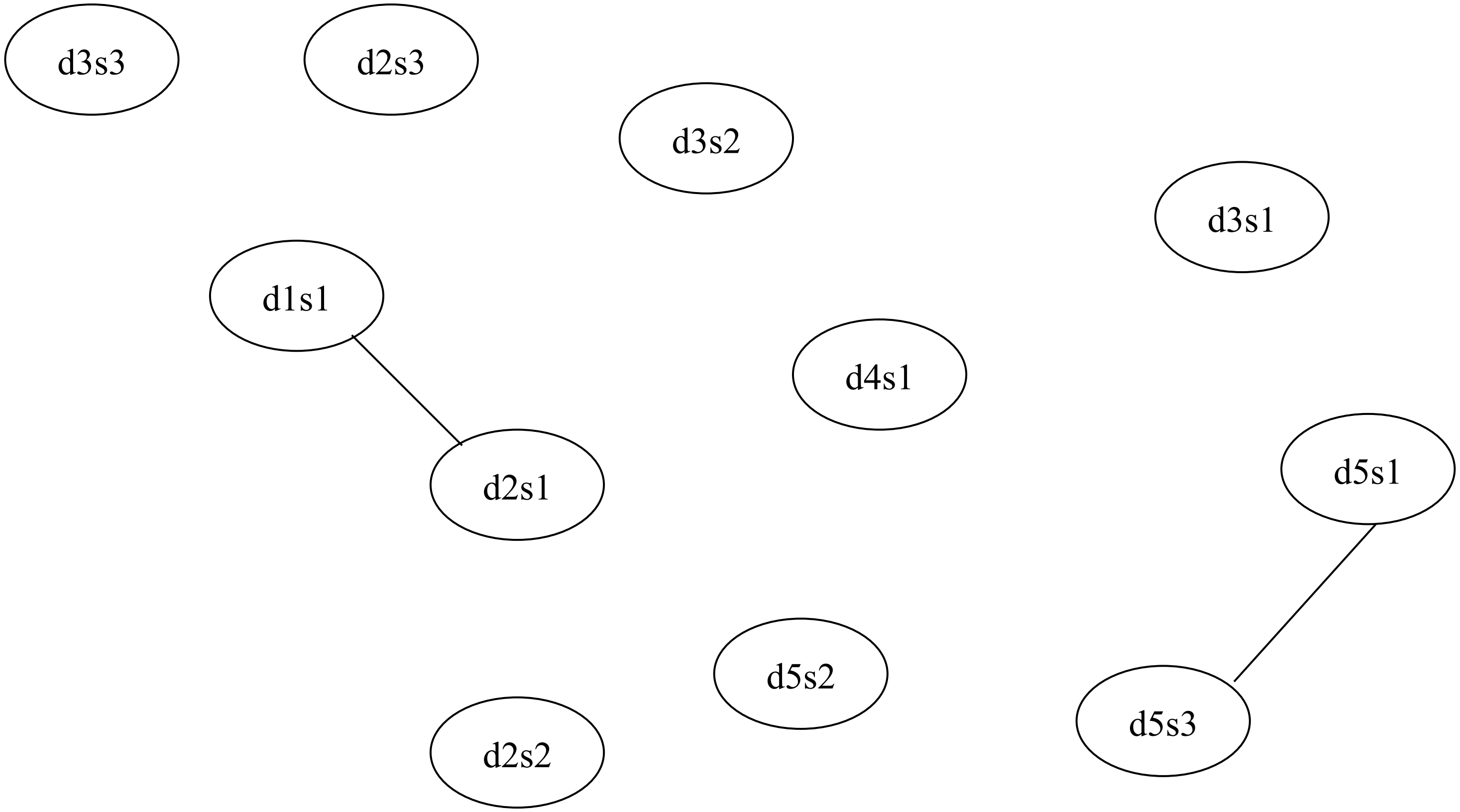
slide from: Dragomir Radev

LexRank (Cosine centrality)

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

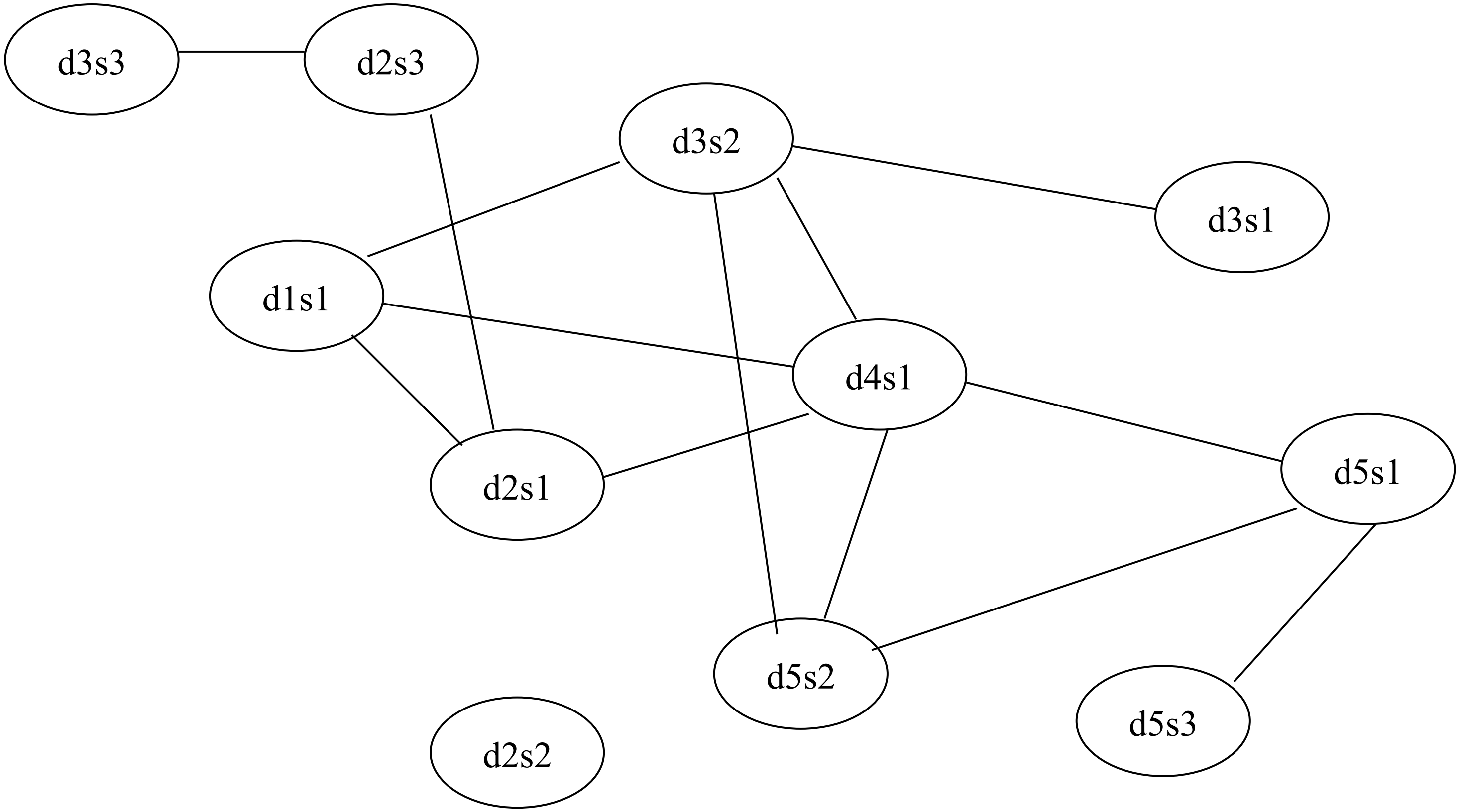
slide from: Dragomir Radev

Lexical centrality (t=0.3)



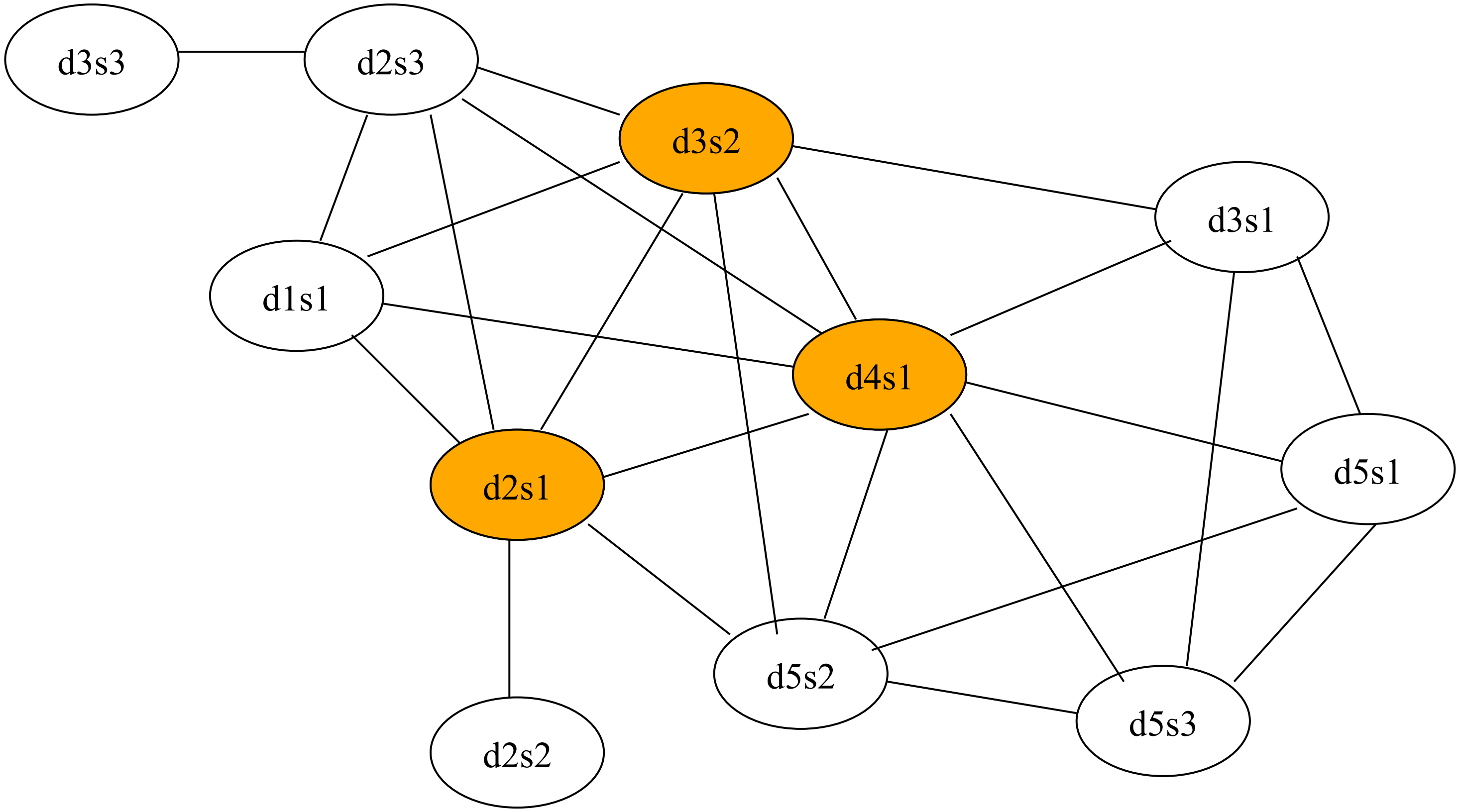
slide from: Dragomir Radev

Lexical centrality (t=0.2)



slide from: Dragomir Radev

Lexical centrality (t=0.1)



Sentences vote for the most central sentence…

slide from: Dragomir Radev

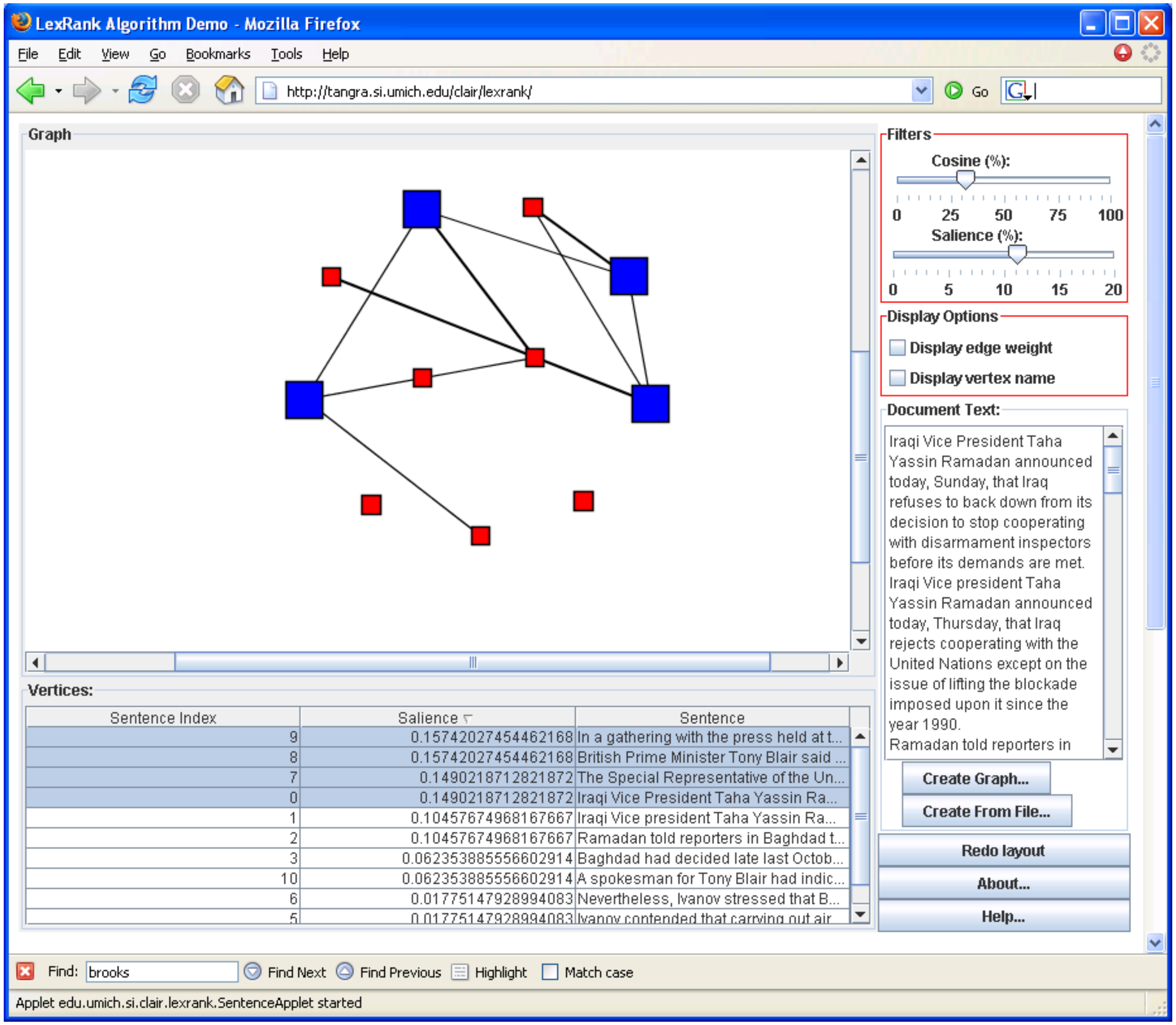
LexRank

$$p(Ti) = \frac{d}{c(T_1)} p(T_1)E(T_1, Ti) + \ldots + \frac{d}{c(T_n)} p(T_n)E(T_n, Ti) + \frac{1-d}{N}$$

- $T_1 \ldots T_n$ are pages that link to *A, c(*T_i*)* is the outdegree of pageT_i*, and N* is the total number of pages.
- *d* is the "damping factor", or the probability that we "jump" to a far-away node during the random walk. It accounts for disconnected components or periodic graphs.
- When *d* = 0, we have a strict uniform distribution.
  When *d* = 1, the method is not guaranteed to converge to a unique solution.
- Typical value for *d* is between [0.1,0.2] (Brin and Page, 1998).

slide from: Dragomir Radev

lab: Lexrank demo



□ Display edge weight

□ Display vertex name

Find: brooks

□ Match case

Applet edu.umich.si.clair.lexrank.SentenceApplet started

Vertices:

Sentence Index	Salience	Sentence
9	0.15742027454462168	In a gathering with the press held at t...
8	0.15742027454462168	British Prime Minister Tony Blair said ...
7	0.1490218712821872	The Special Representative of the Un...
0	0.1490218712821872	Iraqi Vice President Taha Yassin Ra...
1	0.10457674968167667	Iraqi Vice president Taha Yassin Ra...
2	0.10457674968167667	Ramadan told reporters in Baghdad t...
3	0.062353885556602914	Baghdad had decided late last Octob...
10	0.062353885556602914	A spokesman for Tony Blair had indic...
6	0.01775147928994083	Nevertheless, Ivanov stressed that B...
5	0.01775147928994083	Ivanov contended that carrying out air

Document Text:

Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
Ramadan told reporters in

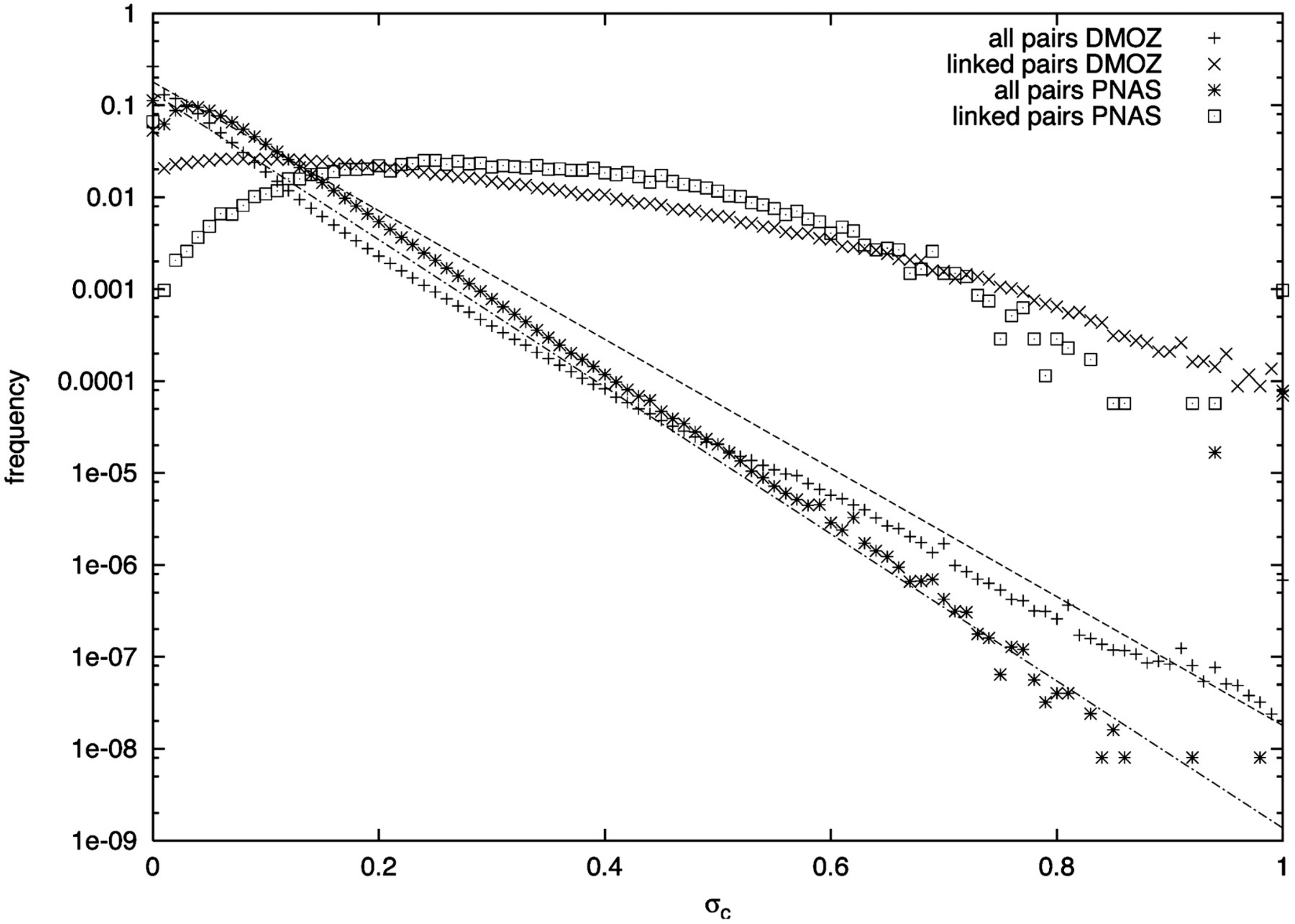
- how does the summary change as you:
- increase the cosine similarity threshold for an edge (how similar two sentences have to be)?
- increase the salience threshold (minimum degree of a node)

http://tangra.si.umich.edu/clair/lexrank/

How might one use the HITS algorithm for document summarization?

- Hint: consider a bipartite graph of sentences and words

Content similarity distributions for web pages (DMOZ) and scientific articles (PNAS)



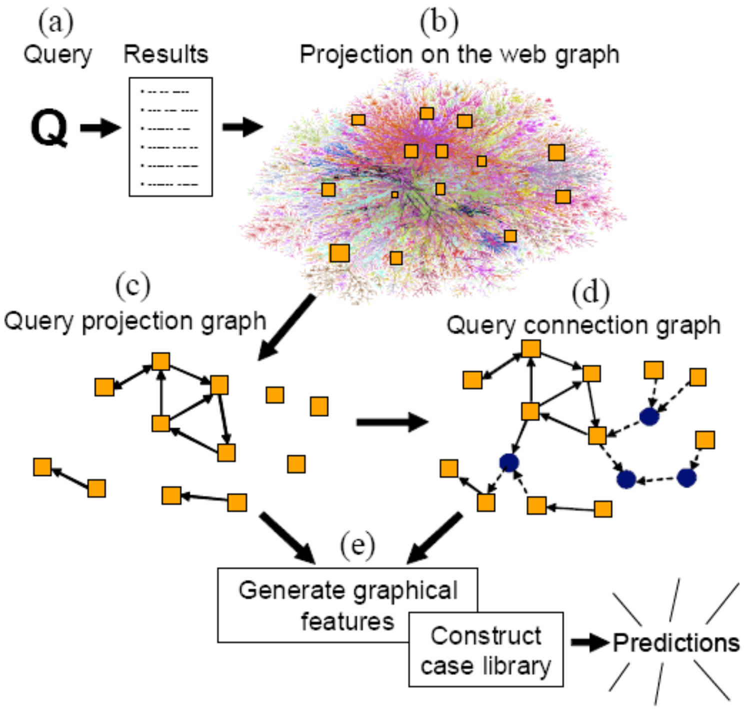
Menczer, Filippo (2004) Proc. Natl. Acad. Sci. USA 101, 5261-5265

PNAS

what is that good for?

- How could you take advantage of the fact that pages that are similar in content tend to link to one another?

What can networks of query results tell us about the query?



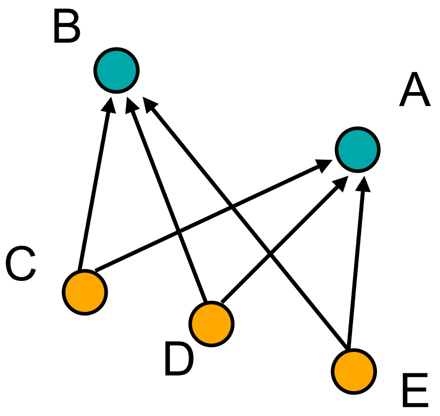
- If query results are highly interlinked, is this a narrow or broad query?

- How could you use query connection graphs to predict whether a query will be reformulated?

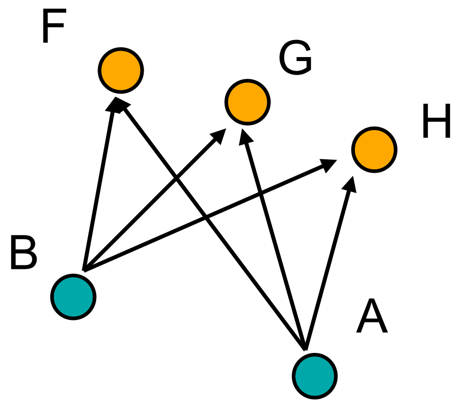
source: Web Projections: Learning from Contextual Subgraphs of the Web
Jure Leskovec, Susan Dumais, Eric Horvitz WWW2007

How can bipartite citation graphs be used to find related articles?

- co-citation: both A and B are cited by many other papers (C, D, E …)

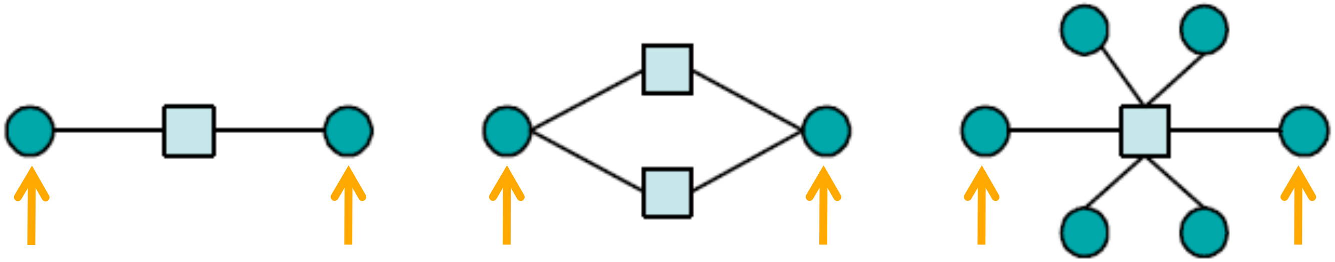


- bibliographic coupling: both A and B are cite many of the same articles (F,G,H …)



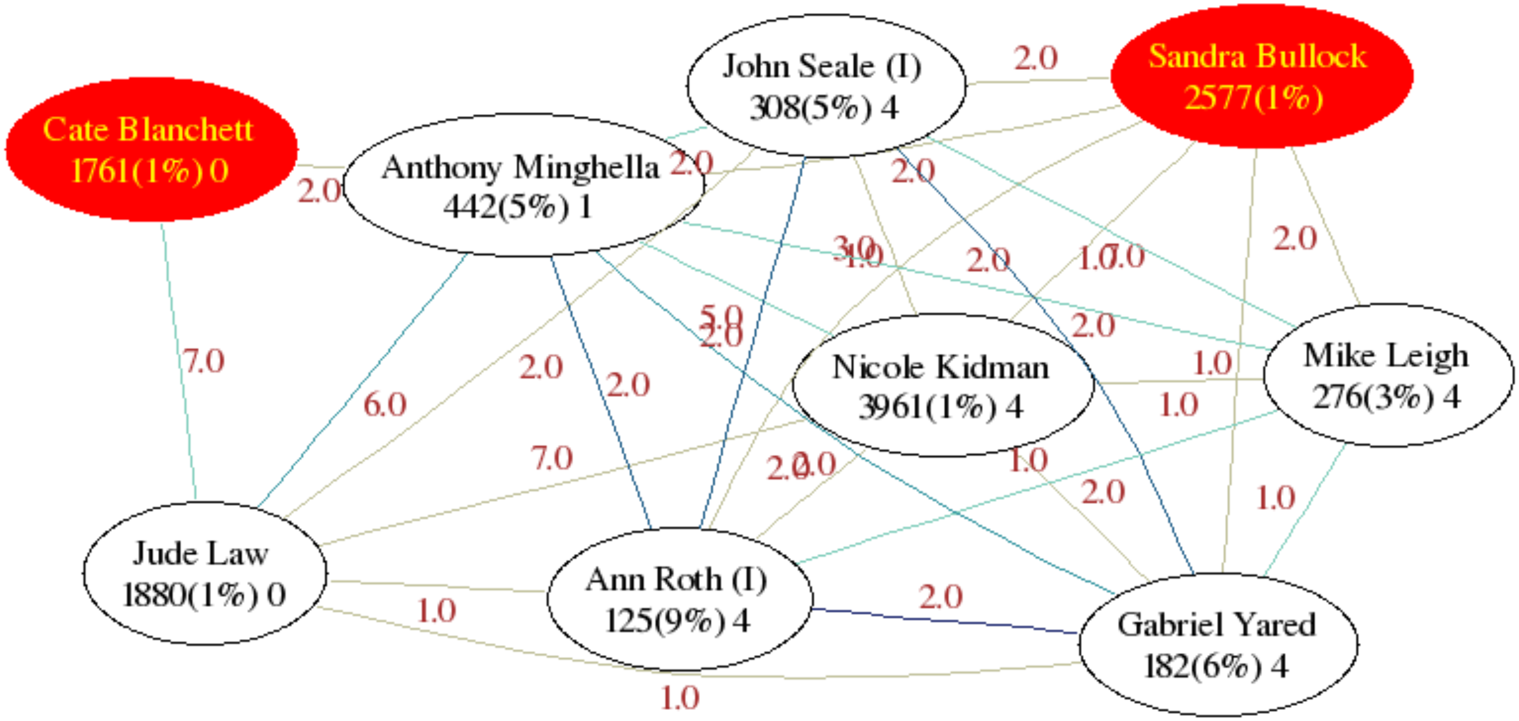
which of these pairs is more proximate

- according to cycle free effective conductance: the probability that you reach the other node before cycling back on yourself, while doing a random walk….

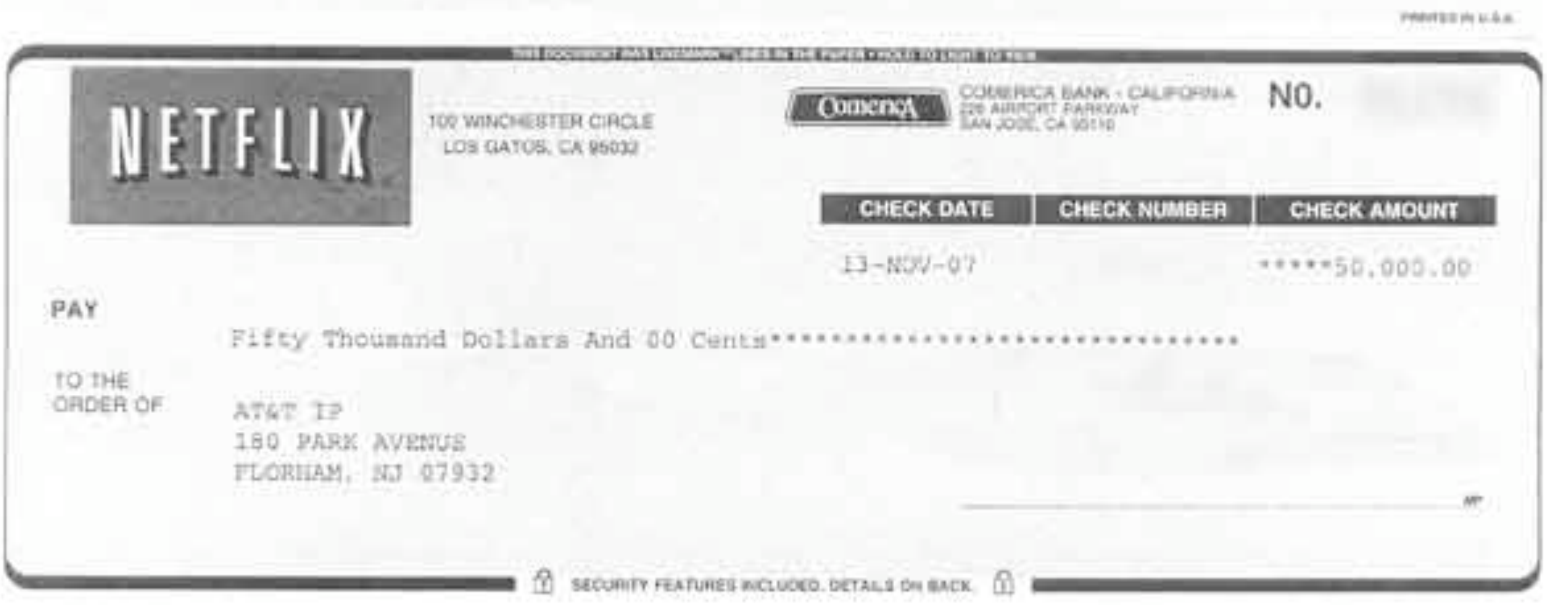


Proximity as cycle free effective conductance

- full slides: http://www.research.att.com/~volinsky/papers/KDD2006.ppt
- paper: Measuring and Extracting Proximity in Networks", by Yehuda Koren, Stephen C. North, Chris Volinsky, KDD 2006
- demo (with computer science authors or actors):http://public.research.att.com/~volinsky/cgi-bin/prox/prox.pl



Using network algorithms (specifically proximity) to improve movie recommendations can pay off



PRINTED IN U.S.A.

THIS DOCUMENT HAS LUMAMARK™ LINES IN THE PAPER • HOLD TO LIGHT TO VIEW

NETFLIX

100 WINCHESTER CIRCLE
LOS GATOS, CA 95032

Comerica

COMERICA BANK - CALIFORNIA
226 AIRPORT PARKWAY
SAN JOSE, CA 95110

NO. [redacted]

CHECK DATE	CHECK NUMBER	CHECK AMOUNT
13-NOV-07	[redacted]	*****50,000.00

PAY

Fifty Thousand Dollars And 00 Cents**********************************

TO THE
ORDER OF

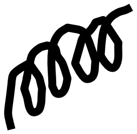
AT&T IP
180 PARK AVENUS
FLORHAM, NJ 07932

SECURITY FEATURES INCLUDED. DETAILS ON BACK.

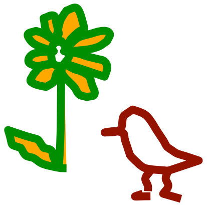
Source: undetermined

final IR application: machine translation

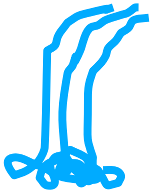
- not all pairwise translations are available (e.g. between rare languages)
- in some applications, e.g. image search, a word may have multiple meanings (“spring” is an example in english)



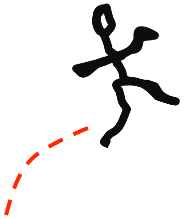
or



or



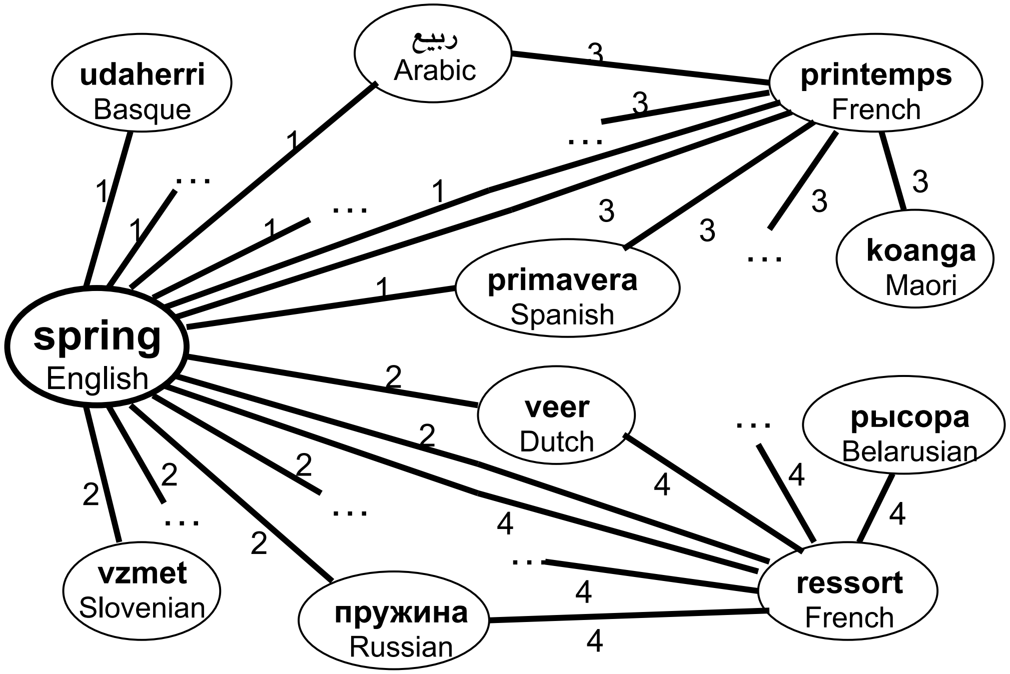
or



But in other languages, the word may be unambiguous.

- automated translation could be the key

final IR application: machine translation

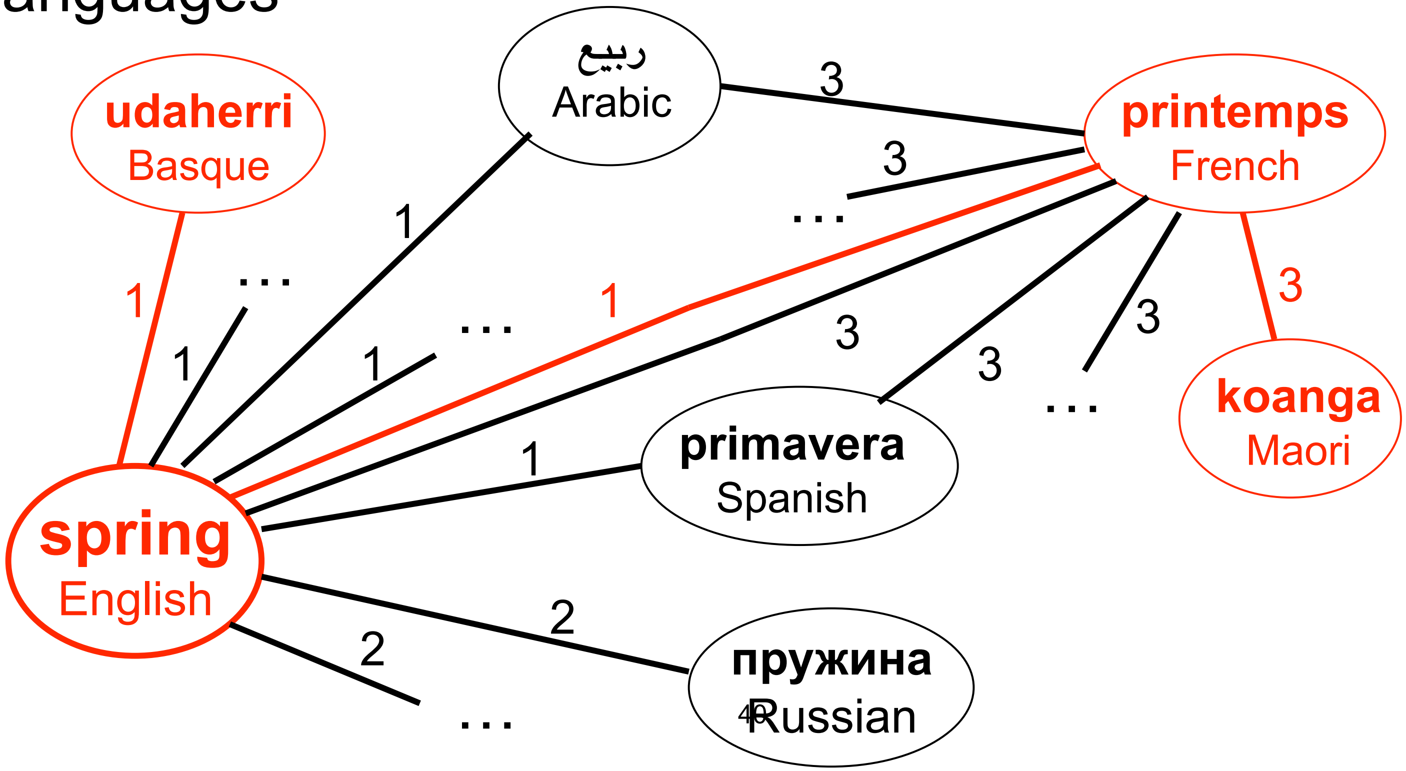


- if we combine all known word pairs, can we construct additional dictionaries between rare languages?

source: Reiter et al., ‘Lexical Translation with Application to Image Search on the Web ’, MT Summit 2007

Automatic translation & network structure

- Two words more likely to have same meaning if there are multiple indirect paths of length 2 through other languages



in lab: translating from English to Croatian (via German and French!)

summary

- the web can be studied as a network
- this is useful for retrieving relevant content
- network concepts can be used in other IR tasks
    - summarization
    - query prediction
    - machine translation