Mining social networks for knowledge management

Prabhakar Raghavan
Verity, Inc.

Overview of this talk

- What are social networks?
  - Connection to knowledge management
- Web mining experiments
  - Bowtie structures
  - Community mining
  - Web behavior models
- The knowledge management challenge
  - Enterprise complications
  - First steps - a demo
- A research agenda

Milgram's experiments

- Began with volunteers from Omaha, NE.
- Asked to get a letter to a physician near Boston.
- Could only send to first-name acquaintance, to be forwarded etc.
- Median path length of successful deliveries was 6.
- Led to famous "6 degrees of separation" folklore.

eMail cliques: Schwartz / Wood

- Studied eMail (sub)graph.
- Proposed metrics for groups of people to share interests; cluster analysis.
- Qualitatively “good” results.
- Raised issues of ethical use of data and privacy.

Various other projects

- PHOAKS
  - Extracting heavily cited resources in newsgroups, etc.
- Call graphs
  - Discerning home, business and fax lines
  - Calling circles.
- Recommendation systems
  - Input: users' product endorsements.
  - Output: product recommendations to each user.

Yenta: Forman

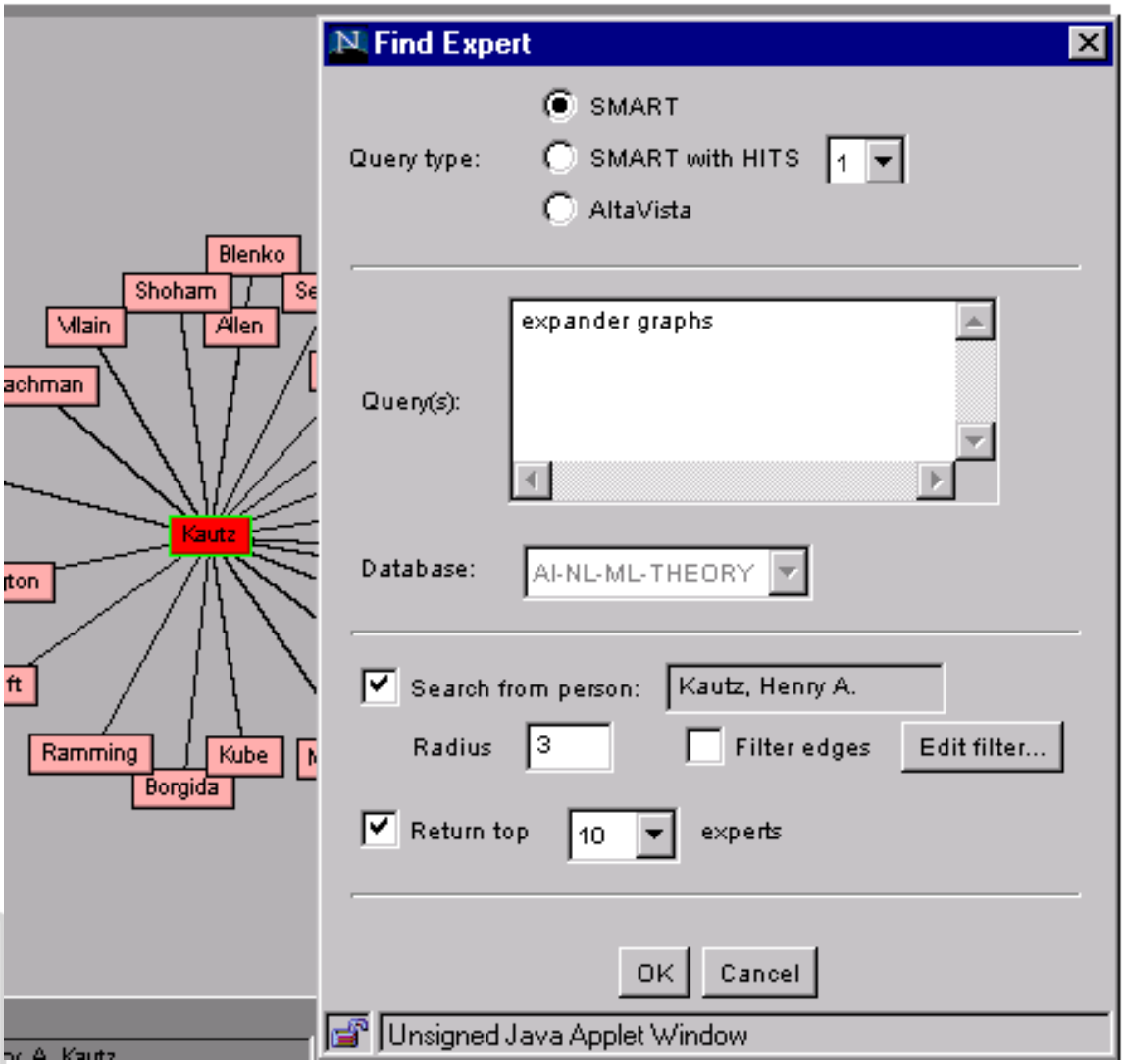
- Analyzes documents "associated" with each user.
- Distils significant "interests" for each.
- Matches/clusters groups of users with overlapping interests.
- Decentralized; aims for privacy protection.
- Elements of peer-to-peer operation.

ReferralWeb: Kautz/Selman

- Establishes links between people, e.g.,
  - co-authorship
  - colleagues in an organization
- Allows search through this social network, e.g.,
  - find me someone within distance 2 who will referee a paper on xyz ...

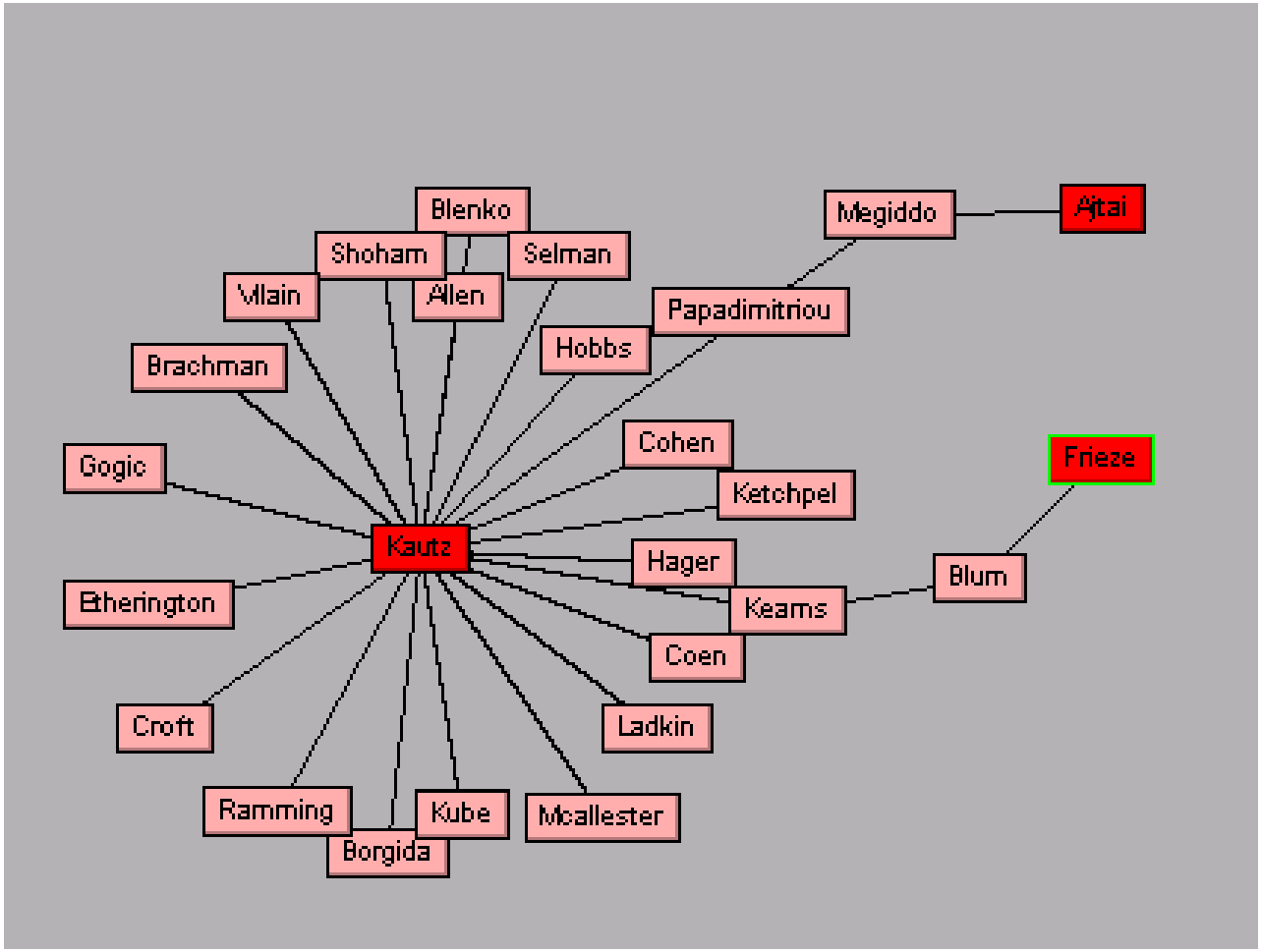
Who can I ask to review a paper on "expander graphs"?

Source: H. Kautz & B. Selman



Paths to Experts

Source: H. Kautz & B. Selman



Observations

Source: H. Kautz & B. Selman

- Official company hierarchy only a sparse subset of the corporate social network
- Shortest (and often best) paths involve a combination of official and unofficial links
  - Conditions for trust and evaluation may greatly differ
  - Global social network is the union of many different kinds of sub-networks

Studying web graphs

- The web is a good test bed for social networks in knowledge management
  - Large scale
  - Plenty of data
  - Large fraction of humanity participates
- Lacks enterprise challenges
  - will return to these later
- Used in knowledge management
  - google and other services

Graphs on the web

- The Web (di)graph
  - Each static html page = a node
  - Each hyperlink = a directed edge
  - Currently ~10^9 nodes, 10^{10} edges

Questions about the web graph

- How big is the graph?
  - How many links on a page (outdegree)?
  - How many links to a page (indegree)?
- Can we browse from any page to any other?
  - How many clicks?
- Pick a random page on the web.
  - Search engine measurement.

Questions about the web graph

- Can we exploit the structure of the web?
  - for searching and mining?
- What does the web graph reveal about social dynamics?
- How different is browsing from a "random walk"?

Why?

- Exploit structure for Web algorithms
  - Crawl strategies
  - Search
  - Mining communities
- Classification/organization
- Web anthropology
  - Prediction, discovery of structures
  - Sociological understanding

"The web is a small world"

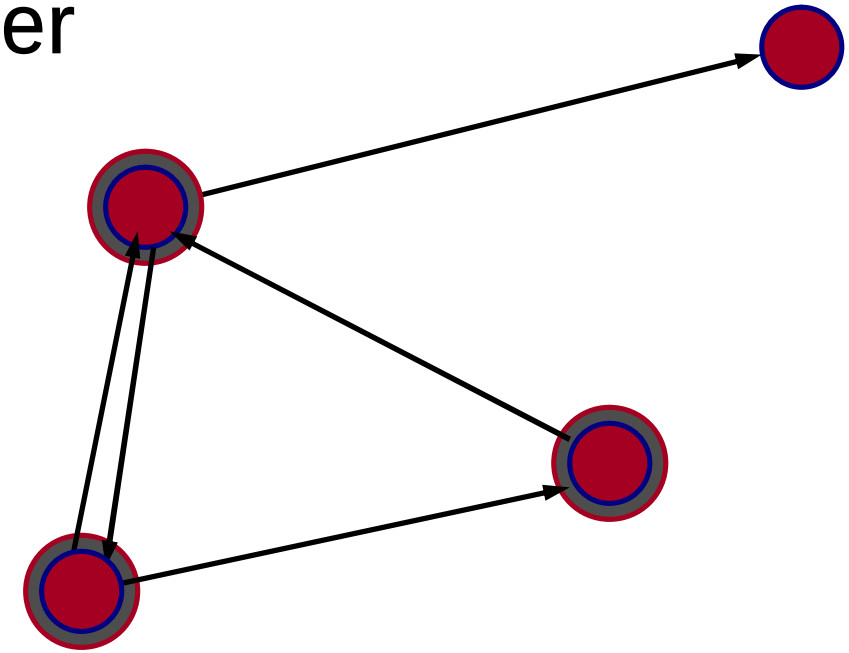
- Analogy to Milgram's six degrees.
- [Barabasi and Albert 99, Albert-Jeong-Barabasi 99].
  - Crawl nd.edu domain, build simple model for web graphs.
  - Predict that most pages are "within 19 links" of each other $\rightarrow$ 19 degrees of separation.

Web snapshots

- Altavista crawls (May 99/Oct 99/Feb 00)
- 220/317/500M pages
- 1.5/2.1B/5B hyperlinks
- Compaq CS2 connectivity server
  - back-link information
  - 10bytes/url, 3.4bytes/link, 0.15μs/access
  - given pages, return their in/out neighborhood

Algorithms

- Weakly connected components (WCC)
- Strongly connected components (SCC)
- Breadth-first search (BFS)
- Diameter



Challenges from scale

- Typical diameter algorithm:
  - number of steps ~ pages × links.
  - For 500 million pages, 5 billion links, even at a *very* optimistic 0.15μs/step, we need

    ~4 billion seconds.

  Hopeless.
- Will estimate diameter/distance metrics.

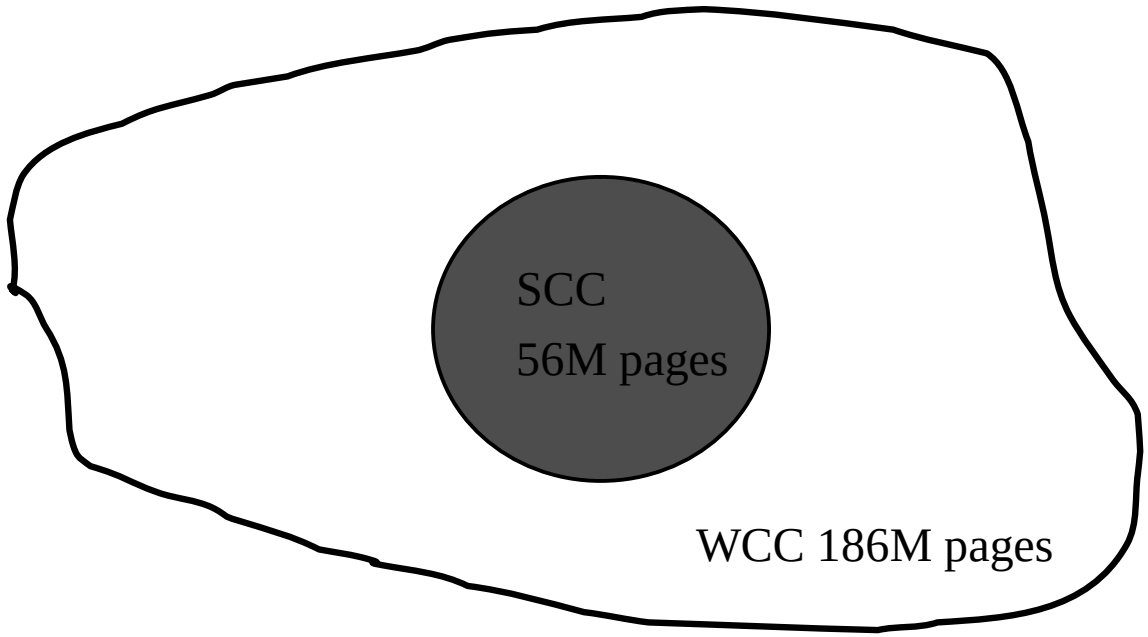
Scale

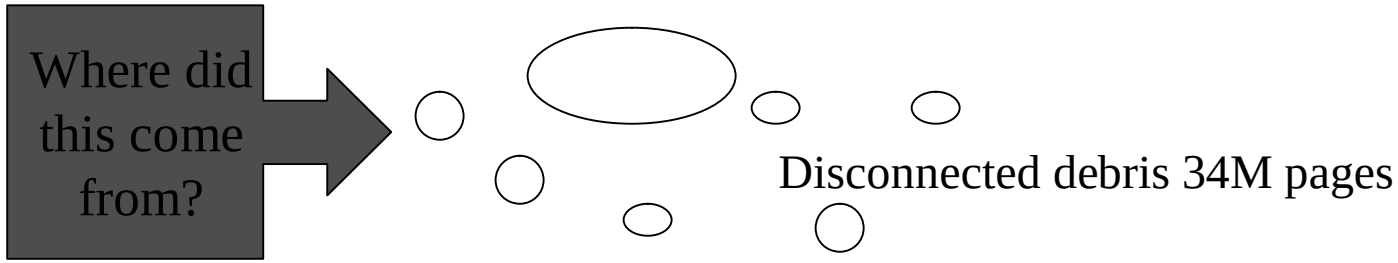
- On the other hand, can handle tasks linear in the links (5 billion) at ~1 μs/step.
  - E.g., breadth-first search
- First eliminate duplicate pages/mirrors.
- Linear-time implementations for WCC and SCC.

May 1999 crawl

- 220 million pages after duplicate elimination.
- Giant WCC has ~186 million pages.
- Giant SCC has ~56 million pages.
  - Cannot browse your way from any page to any other
  - Next biggest SCC ~150K pages
- Fractions roughly the same in other crawls.

Tentative picture





Breadth-first search (BFS)

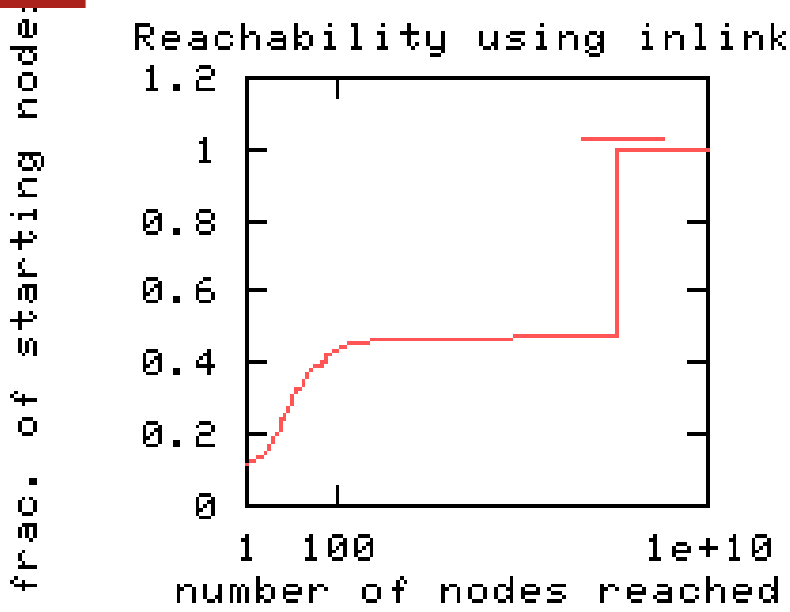
- Start at a page *p*
  - get its neighbors;
  - their neighbors, etc.
- Get profile of the number of pages reached by crawling out of *p*, as a function of distance *d*
- Can do this following links forwards as well as backwards

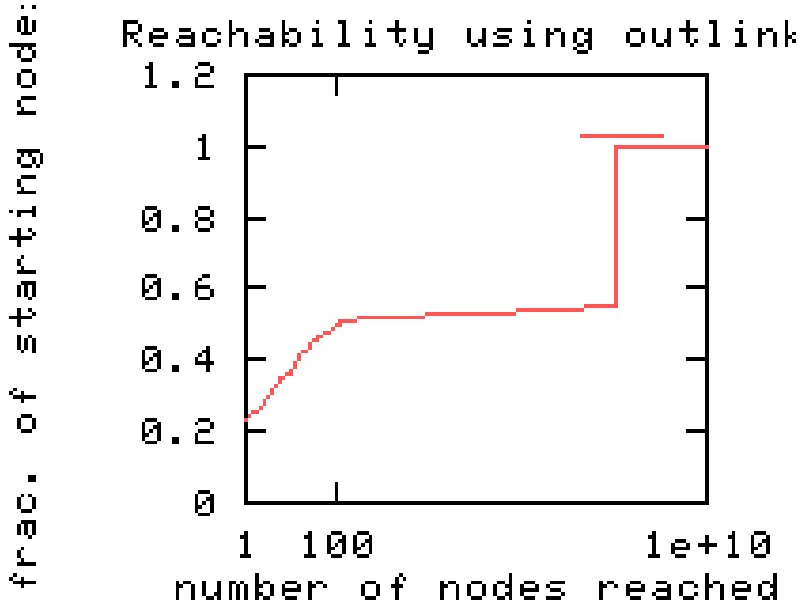
BFS experiment

- Start at 1000+ random pages
- For each start page, build BFS (reachability vs. distance) profiles going forwards, and backwards

Reachability

How many pages are reachable from a random page?





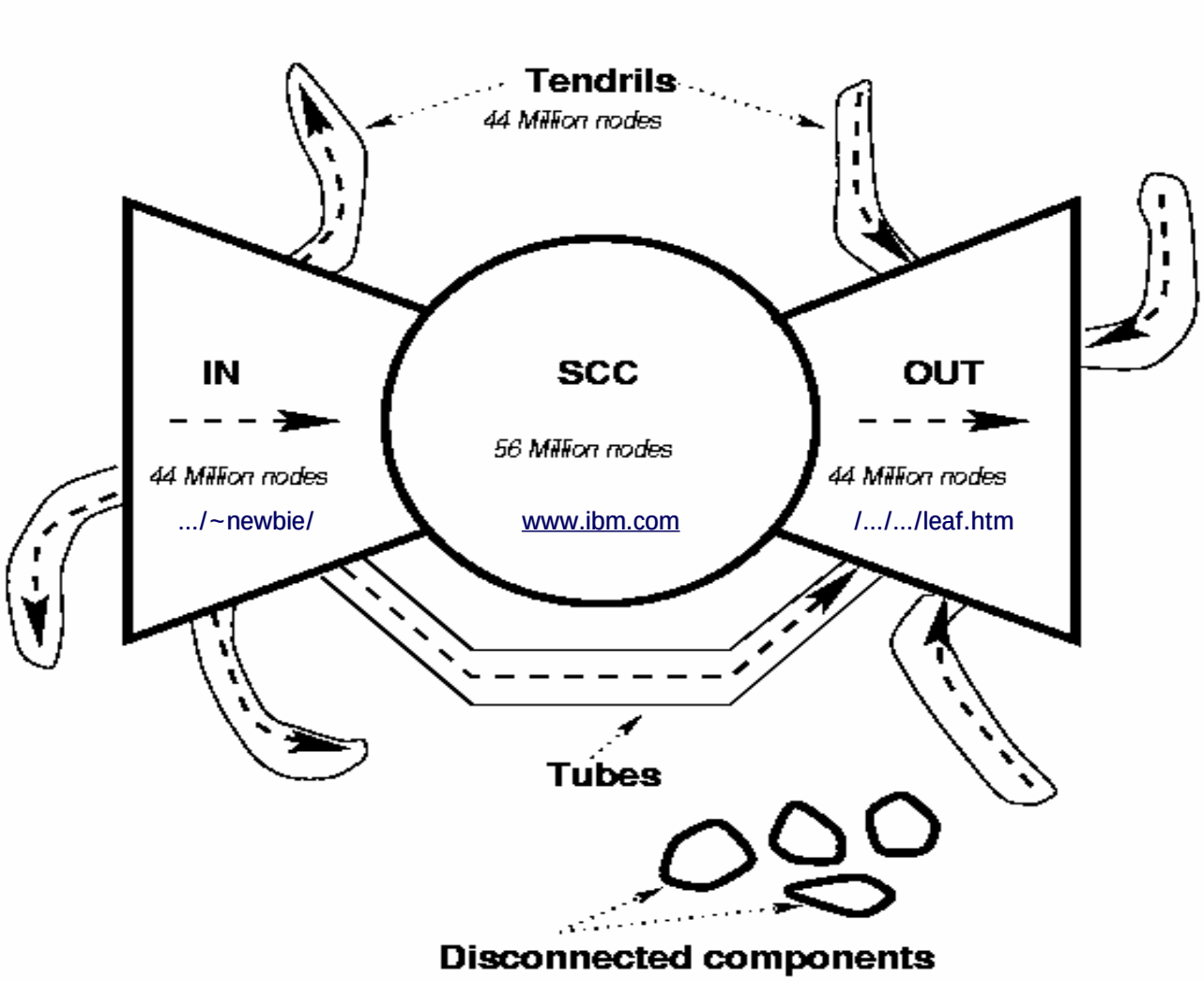
Net of BFS experiments

- BFS out of a page
  - either dies quickly (~100 pages reached)
  - "explodes" and reaches ~100 million pages
    - somewhat over 50% of starting pages
  - SCC pages ~25% of total, reach >56M pages
- Qualitatively the same following in- or out-links

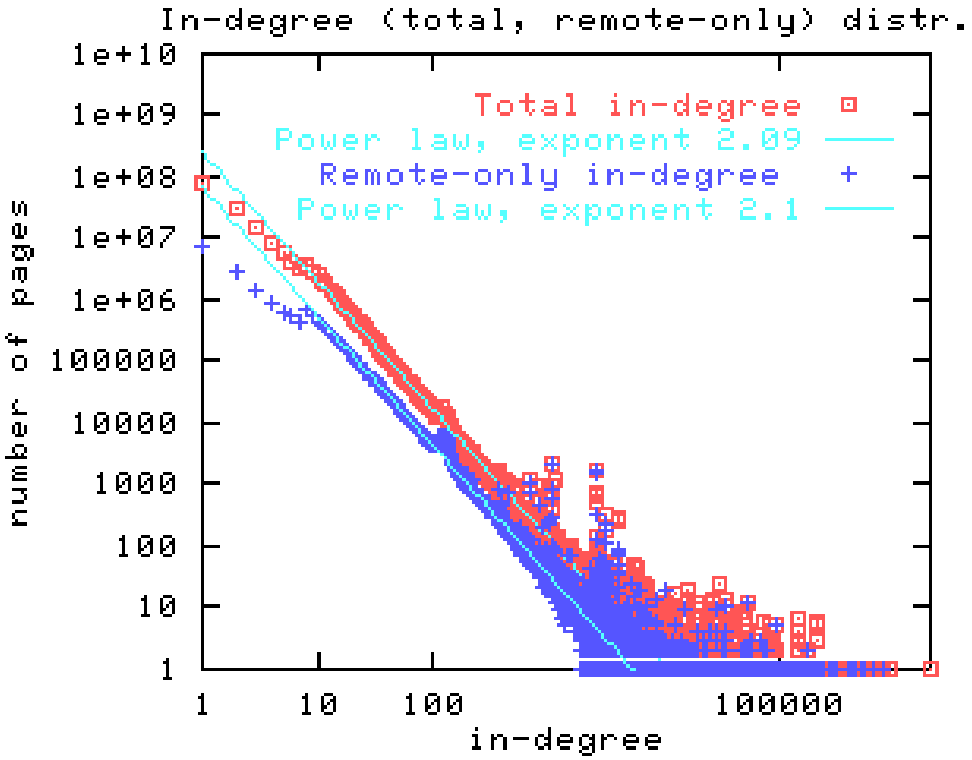
Interpreting BFS expts

- Need another 100-56 = 44M pages reachable from SCC
  - gives us 100M pages reachable from SCC
- Likewise, need another ~44M pages reachable from SCC going backwards
- These together don't account for all 186M pages in giant WCC.

The shape of the Web



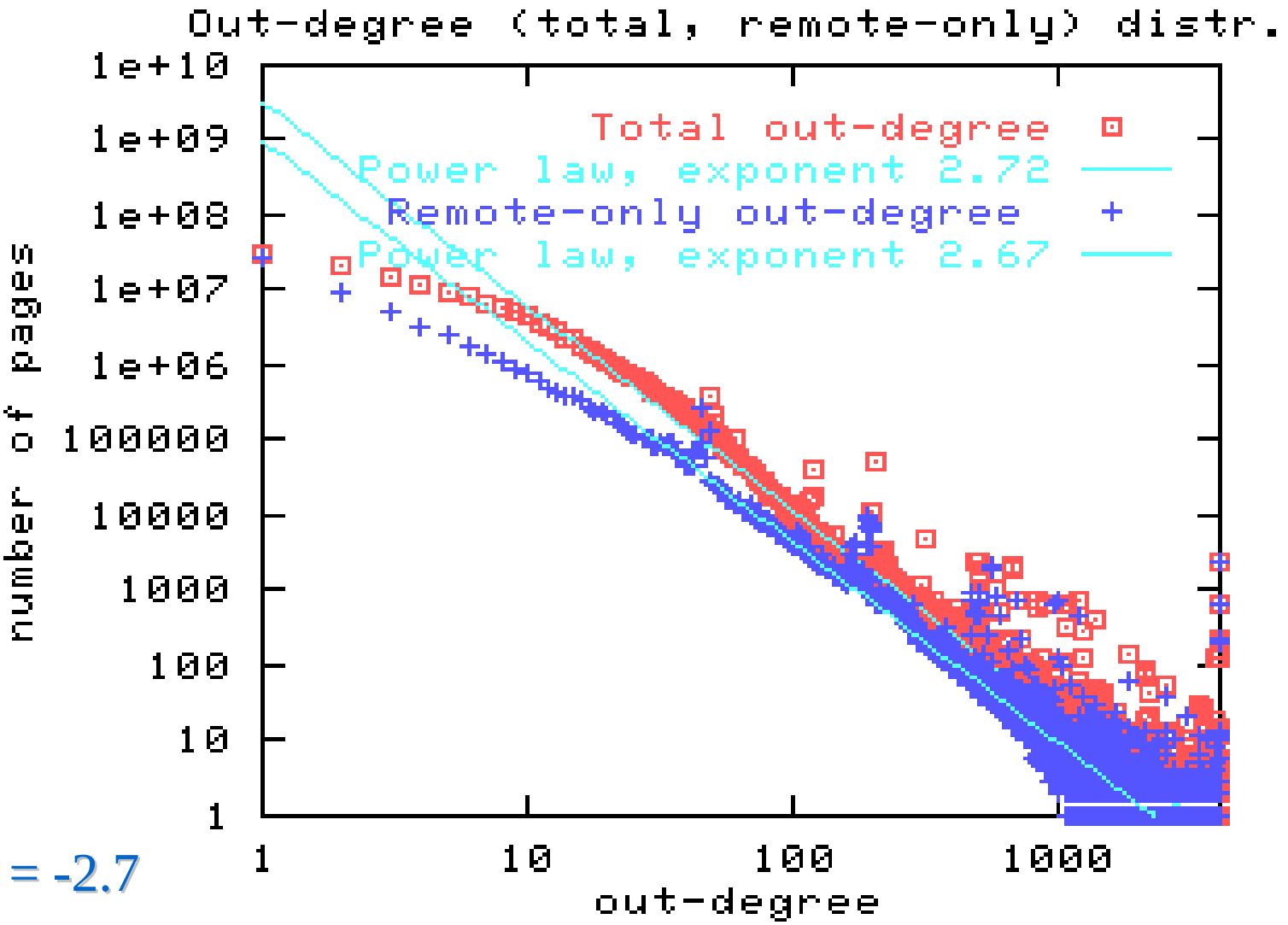
Distribution of in-degrees



$$I(u) = |\{v \to u\}|$$

$$\Pr_u(I(u) = k) \cong k^{-2.1}$$

Distribution of out-degrees



Slope = -2.7

Recurring phenomena

- Many interesting distributions
  - term frequencies in a corpus
  - citations
  - in-links to web pages
  - populations of US cities
  - degrees of internet nodes
  - document access frequencies ...

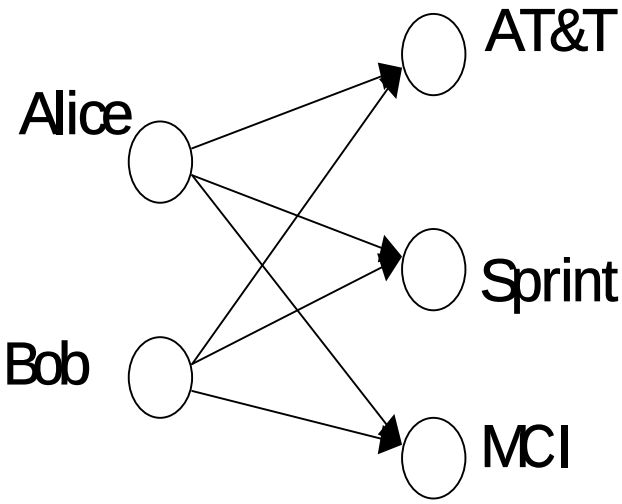
  follow an inverse polynomial function.

What leads to power laws?

- "Scale free" growth.
- "Highly optimized tolerance".
- Behavioral models.
  - Model behavior of individuals in social network.
- See Christos Papadimitriou's class notes
  http://www.cs.berkeley.edu/~christos/games/powerlaw.ps

Communities from cores

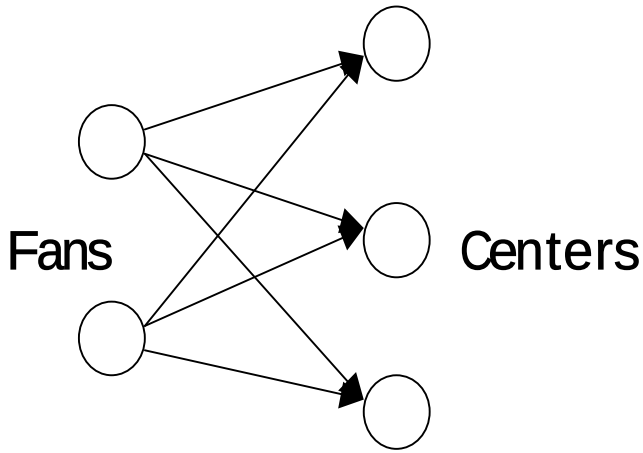
- A dense bipartite graph implies a community.
- What is a "dense bipartite subgraph"?
- Define (*i*,*j*)-core: complete bipartite subgraph with *i* nodes all pointing to each of *j* nodes.
- Enumerate (*i*,*j*)-cores for various small *i*,*j*.



(2,3)-core

Trawling bipartite cliques

- Take the (directed) Web link graph.
- Enumerate all (small) bipartite cliques.



(2,3)-core

R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins (1998).

Enumeration by iterative pruning

- Cannot try all 6-tuples for (3,3)-cores
  - (all $i+j$ tuples for (i,j)-cores).
- Iterative pruning:
  - throw away all nodes with < 3 neighbors
  - this can be iterated
- Eventually have small, dense graph in memory
  - additional pruning and data mining tricks.

The cores are interesting

- We find ~200,000 of them.
- Examples:
  - Japanese elementary schools
  - Turkish student associations
  - oil spills off the coast of Japan
  - Australian fire brigades

Japanese Elementary Schools

Fans

- schools
- LINK Page-13
- “ú–{‚ÌŠw Z
- a‰„ ¬Šw Zƒz [ƒ ƒy [ƒW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www...iglobe.ne.jp/~IKESAN
- ‚l‚f‚j ¬Šw Z‚U”N‚P‘g• ¨ Œê
- ÒŠ—’¬— § ÒŠ—“Œ ¬Šw Z
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- –y“ì ¬Šw Z‚Ìƒz [ƒ ƒy [ƒW
- UNIVERSITY
- ‰J—³ ¬Šw Z DRAGON97-TOP
- Â‰ª ¬Šw Z‚T”N‚P‘gƒz [ƒ ƒy [ƒW
- ¶µ° é¼ÂÁ© ¥á¥Ë¥å¡¼ ¥á¥Ë¥å¡¼

Centers

- The American School in Japan
- The Link Page
- ‰ª è s— § ˆä“c ¬Šw Zƒz [ƒ ƒy [ƒW
- Kids' Space
- ˆÀ é s— § ˆÀ é ¼•” ¬Šw Z
- ‹{ é‹³ˆç‘åŠw• ‘® ¬Šw Z
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- _“Þ ìŒ § E‰¡•l s— § ’† ì ¼ ¬Šw Z‚Ìƒy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

Knowledge management

- The big challenge:

  Increase productivity in knowledge workers by getting them the expertise they need at all times

  > the right information (documents?)

  > the right experts.

- Enterprise: group of people engaged in a collective endeavour, typically with proprietary content.

Challenges in enterprises

- Information resides in heterogeneous
  - formats (email, pdf, word, ...)
  - repositories (Lotus, MS Exchange, Documentum, databases ...)
  - applications (HR, ERP, Siebel, ...)
- Data security: documents have different access classes.
  - compound documents have pieces, each with its own access lists (ACL's).
  - My search should hit the doc only if it hits the pieces I can see.

Need a general formulation

- How do we combine different sources of content and context?
  - terms in docs
  - links between docs
  - users' access patterns
  - users' profile information.

General formulation

- Every item of interest - each term, query, doc, person, treated as a node.
- Impose similarity metric between pairs of nodes.
- Need to be able to measure proximity from sets of nodes (a person+a doc they're viewing+a query they've issued) to nodes of a target type (a person).

Issues in formulation

- If a user is close to two docs d1 and d2, are the docs d1 and d2 close to each other?
- How do you measure proximity from a set of nodes?
- How do you capture collaborative (as opposed to content and context-based filtering).
- How do you succinctly represent and manipulate similarities?

Upshot

- Verity social networks project Screenshot.
- Security/privacy issues remain difficult.
- What aspects of social behavior can we exploit in the algorithms?
  - Power laws?

A research agenda

- Need new ways of combining content, context and collaboration.
- Analyze and model structures in social networks.
  - Tune algorithms on models.
  - Build on "standard" mining paradigms: associations, clustering …
- Incorporate enterprise constraints:
  - Roles and profile information from apps
  - Security and Privacy!

Domo arigatou gozaimashita