Link- based clustering and the power web

Prabhakar Raghavan

Yahoo! Research

4 lectures

- Link-based clustering
- Segmentation
- Vector spaces reborn
- VLSI
  - Variable Latent Semantic Indexing (Ravi Kumar)

Trawling the web graph

(w/ Ravi Kumar, Sridhar Rajagoplan
and Andrew Tomkins)

“Traditional” clustering

- Partition a set into cognate subsets
- Unclear whether all subsets are interesting

Trawling

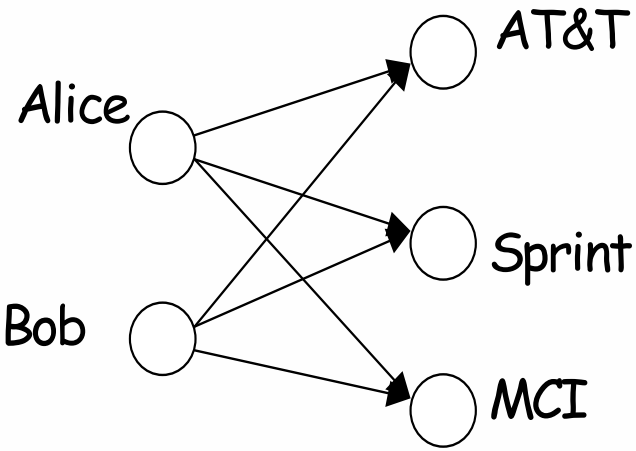
- Enumerate subsets
  - In our case, of web pages
- Subsets can overlap
  - In principle … though not in our expts
- Subsets should be “interesting”.
- Not a partition – most docs discarded.

What makes a subset interesting?

- Pages belong together.
- An audience segment cares about them
  - Has *utility*.
- A subset isn't too big
  - Fit for human consumption.

Trawling

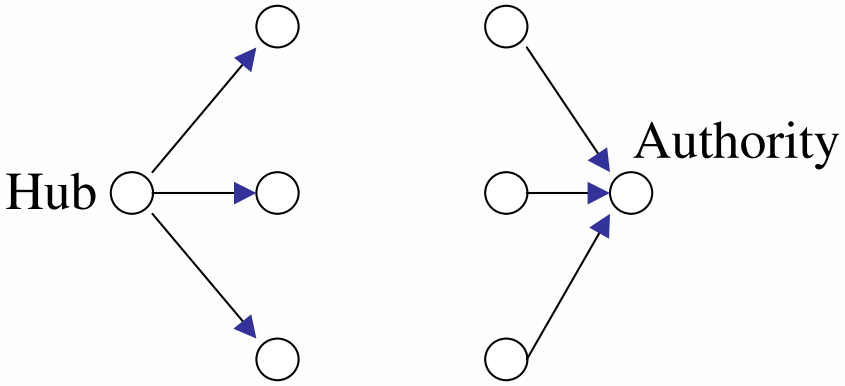
- Twist: will use purely link-based cues to decide whether docs are related
  - Look for all occurrences of a linkage pattern
- Recall from hubs/authorities link analysis:

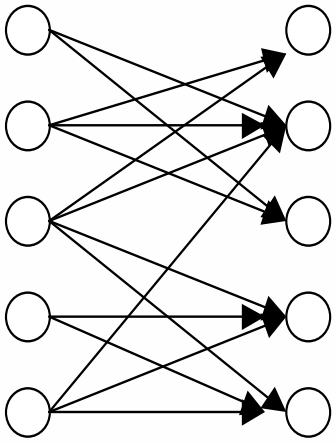


Not document clustering

- In clustering, we partition input docs into clusters.
- In *trawling*, we'll enumerate subsets of the corpus that "look related"
  - will discard lots of docs

Insights from hubs



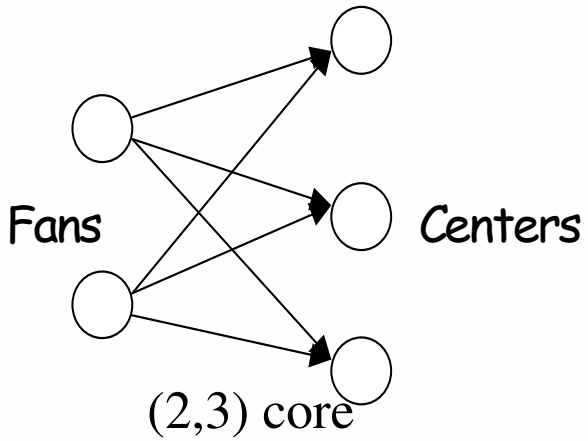


Link-based hypothesis:

Dense bipartite subgraph
⇔ Web community.

Communities from cores

- not easy, since web is huge
- what is a “dense subgraph”?
- define (*i*,*j*)-core: complete bipartite subgraph with *i* nodes all of which point to each of *j* others



Random graphs inspiration

Every “large” enough “dense” bipartite graph “almost surely” has a “non-trivial” core

e.g.,:

large = 3 by 10

dense = 50% edges

almost surely = 90% chance

non-trivial = 3 by 3

Approach

- Find all (*i*,*j*)-cores ($3 \leq i \leq 10$, $3 \leq j \leq 20$).
  - Why?
- Expand each core into its full community.

Finding cores

- "SQL" solution: find all triples of pages such that intersection of their outlinks is at least 3? Too expensive.
- Iterative pruning techniques actually work!

Initial data & preprocessing

- Crawl, then extract links
- Work with potential fans:

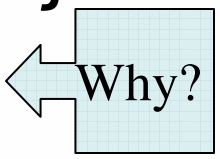
  nodes with ≥ *j* non-nepotistic links
- (Eliminate mirrors)
- Can sort URL's by either source or destination using disk-run sorting

Main requirements

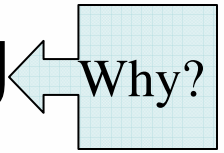
- Main memory conservation
- Few disk passes over data

Simple iterative pruning

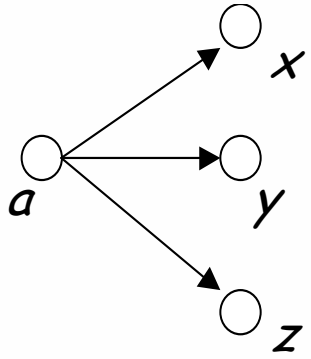
- Discard all pages of in-degree < *i* or out-degree < *j*.
- Repeat



- Reduces to a sequence of sorting operations on the edge list



Elimination/generation pruning



a is part of a (*3, 3*) core if and only if the intersection of inlinks of *x, y,* and *z* is at least *3*

- pick a node *a* of degree 3
- for each *a* output neighbors *x, y, z*
- use an index on centers to output in-links of *x, y, z*
- intersect to decide if *a* is a fan
- at each step, either eliminate a page (*a*) or generate a core

Exercise

- Work through the details of maintaining the index on centers to speed up elimination-generation pruning.

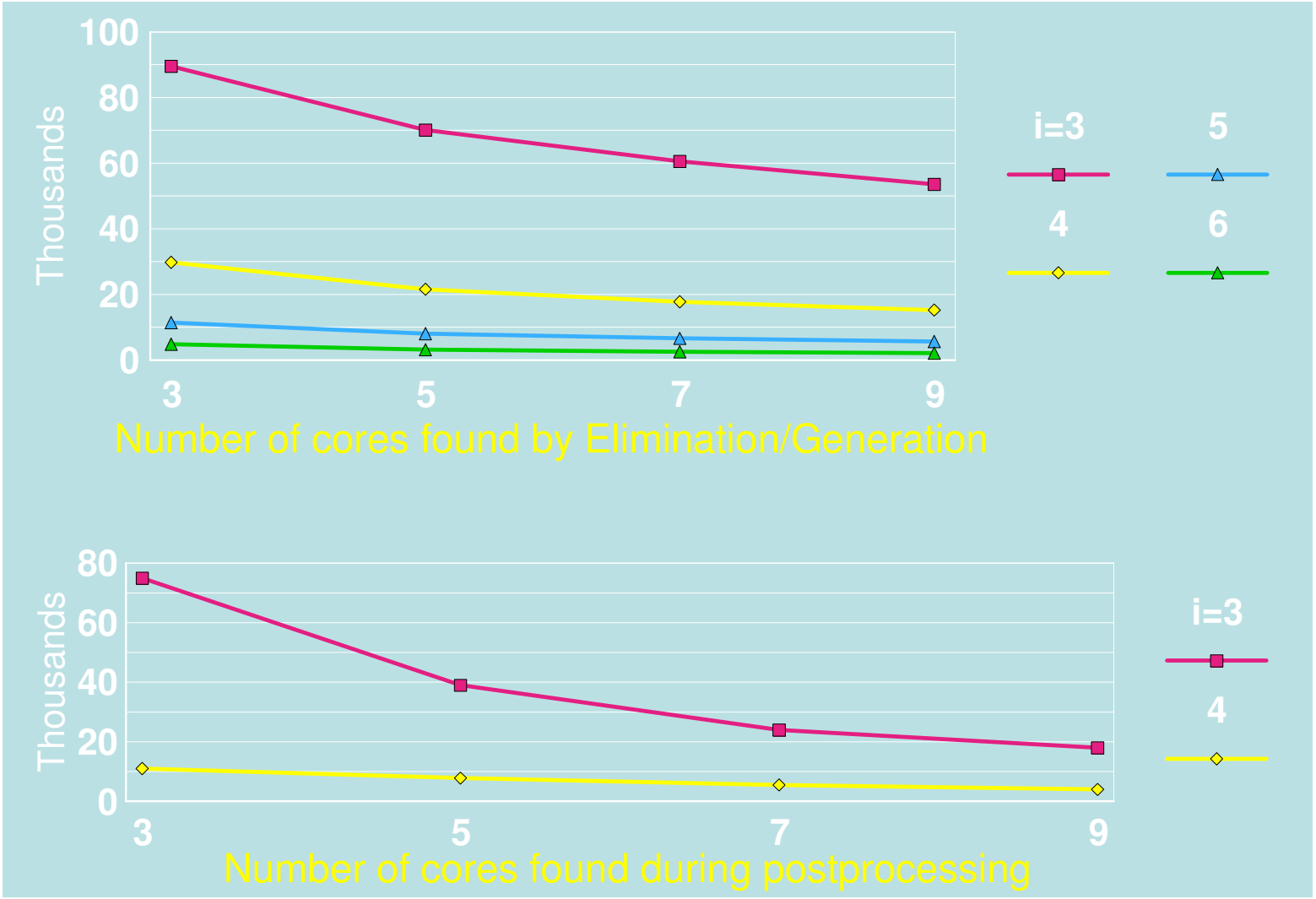
Begin crawl w/100M pages

- Elimination/generation pruning yields >100K non-overlapping cores for small *i,j.*
- 5M unpruned edges
  - small enough for postprocessing by *a priori*
  - build (*i+1, j*) cores from (*i, j*) cores
  - discard nodes, edges as you proceed

Exercise

- Detail the *a priori* algorithm to enumerating bipartite cores.

Some sense for core sizes



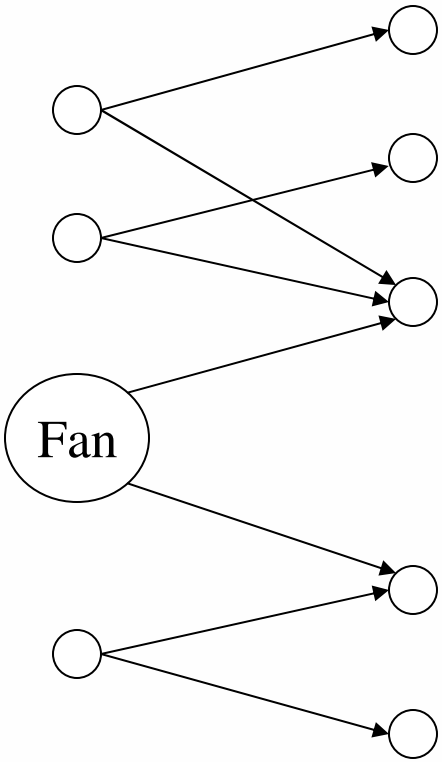
Sample cores

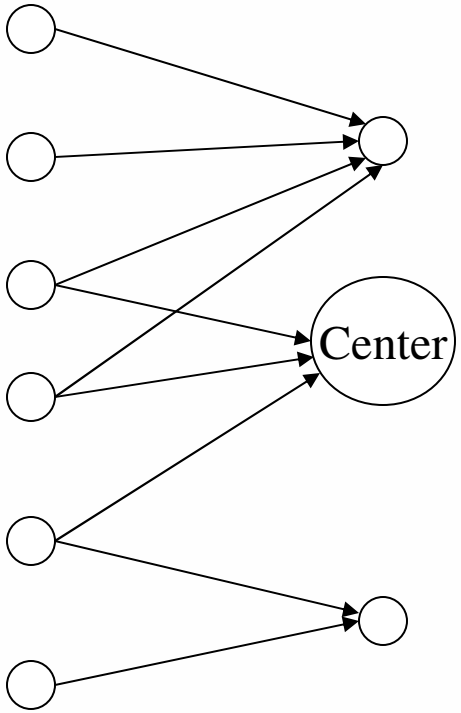
- hotels in Costa Rica
- clipart
- Turkish student associations
- oil spills off the coast of Japan
- Australian fire brigades
- aviation/aircraft vendors
- guitar manufacturers

From cores to communities

- Use hubs/authorities algorithm without text query - use fans/centers as samples
- Augment core with
  - all pages pointed to by any fan
    - all pages pointing into these
  - all pages pointing into any center
    - all pages pointed to by any of these

Using sample hubs/authorities





Finished communities

- On the set of resulting pages, run Kleinberg's HITS algorithm
- Output top-scoring hubs/authorities as the community

Costa Rican hotels and travel

- The Costa Rica Inte...ion on arts, busi...
- Informatica Interna...rvices in Costa Rica
- Cocos Island Research Center
- Aero Costa Rica
- Hotel Tilawa - Home Page
- COSTA RICA BY INTER@MERICA
- tamarindo.com
- Costa Rica
- New Page 5
- The Costa Rica Internet Directory.
- Costa Rica, Zarpe Travel and Casa Maria
- Si Como No Resort Hotels & Villas
- Apartotel El Sesteo... de San José, Cos...
- Spanish Abroad, Inc. Home Page
- Costa Rica's Pura V...ry - Reservation ...
- YELLOW\RESPALDO\HOTELES\Orquide1
- Costa Rica - Summary Profile
- COST RICA, MANUEL A...EPOS: VILLA
- Hotels and Travel in Costa Rica
- Nosara Hotels & Res...els & Restaurants...
- Costa Rica Travel, Tourism & Resorts
- Association Civica de Nosara
- Untitled: http://www...ca/hotels/mimos.html
- Costa Rica, Healthy...t Pura Vida
- Domestic & International Airline
- HOTELES / HOTELS - COSTA RICA
- tourgems
- Hotel Tilawa - Links
- Costa Rica Hotels T...On line Reservations
- Yellow pages Costa ...Rica Export
- INFOHUB Costa Rica Travel Guide
- Hotel Parador, Manuel Antonio, Costa Rica
- Destinations

Muslim student orgs.

- USC Muslim Students...ation Islamic Server
- The University of O...a Domain Name Change
- Caltech Muslim Students Home Page
- Islamic Society of Stanford University
- University of Texas...nformation Center...
- CSUN Muslim Students Association homepage
- HUDA
- Islamic Gateway
- Muslim Students' As...iversity of Michigan
- About Islam and Muslims
- Carnegie Mellon Uni...m Students Home Page
- Bookstore: The Onli...slamic Books, Isl...
- Islamic Texts and R... University at Bu...
- Kutkut - Islam
- Other MSAs and Organizations
- Other Resources rel...versity at Buffal.
- 777
- Huma's Mamalist of Islamic Links!
- Other MSAs
- ZUBAIR'S ISLAM PAGE
- MIDDLE EAST CONFLICTS
- Islamic Links at the Arabic Paper
- Middle East & Arab Hot Links
- MSA National: MSAs Home Page
- Islamic Page
- Info about Muslims (MSA @SUNY/Buffalo)
- Untitled: http://www...ev/mideast/islam.htm
- Aalim Fevens: Islam Home Page
- islam
- Links to MSAs

The power web

(incl ideas from Ashok Chandra)

The power web

- = All subsets of web pages
- So what?
- Some subsets are more interesting than others
- Assign a utility measure *U* on subsets of the web
  - Each user can have a different, time-varying utility function

Axioms?

- Utility small for very large sets
  - Human consumption
- Utility peaks on sets of “related” pages
  - Mirrors information needs
- Users assemble sets of high utility
  - Utility on other subsets an interpolation? How?
- Utility decays with age of pages in set?

Users assembling subsets

- What does it mean for users to assemble sets of pages?
  - Bookmarks
  - Hyperlinks
  - Annotation of collections cf. Yahoo!'s MyWeb

Where do these ideas lead?

- Microeconomics of the power web
  - Incentives for creating high-quality sets
- Concrete schemes for building systems
  - Trawling is only the beginning
  - How about trawling, but with time decay?


Questions?