IMA Tutorial (part II): Measurement and modeling of the web and related data sets

Andrew Tomkins
IBM Almaden Research Center
May 5, 2003
Setup

- This hour: data analysis on the web
- Next hour: probabilistic generative models, particularly focused on models that generate distributions that are power laws in the limit
Context

- Data Analysis *on the web*…
- …as a hyperlinked corpus
- Note: Many areas of document analysis are highly relevant to the web, and should not be ignored (but will be):
  - Supervised/unsupervised classification (Jon – combinatorial side)
  - Machine learning (Jon – a little)
  - Information retrieval (Jon – dimensionality reduction)
  - Information extraction
  - NLP
  - Discourse analysis
  - Relationship induction
  - etc
Focus Areas

- Web Measurement
- Self similarity on the web
- Extraction of information from large graphs
- A word on evolution
One view of the Internet: Inter-Domain Connectivity

- **Core**: maximal clique of high-degree nodes
- **Shells**: nodes in 1-neighborhood of core, or of previous shell, with degree > 1
- **Legs**: 1-degree nodes

[Tauro, Palmer, Siganos, Faloutsos, 2001 Global Internet]
Another view of the web: the hyperlink graph

- Each static html page = a node
- Each hyperlink = a directed edge
- Currently $\sim 10^{10}$ nodes (mostly junk), $10^{11}$ edges
Getting started – structure at the hyperlink level

- Measure properties of the link structure of the web.
- Study a sample of the web that contains a reasonable fraction of the entire web.
- Apply tools from graph theory to understand the structure.

[Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, Wiener, 2001]
Terminology

- SCC – strongly connected component
- WCC – “weakly connected component” – connected component in the underlying *undirected* graph
Data

- Altavista crawls, up to 500M pages
- Ran strong and weak connected component algorithms
- Ran random directed breadth-first searches from 1000 starting nodes, both forwards and backwards along links
Breadth-first search from random starts

- How many vertices are reachable from a random vertex?
A Picture of (~200M) pages.
Some distance measurements

- \( \Pr[u \text{ reachable from } v] \sim \frac{1}{4} \)
- Max distance between 2 SCC nodes: 28
- Max distance between 2 nodes (if there is a path) > 900
- Avg distance between 2 SCC nodes: 16
Facts (about the crawl).

- Indegree and Outdegree distributions satisfy the power law. Consistent over time and scale.

The distribution of indegrees on the web is given by a Power Law --- Heavy-tailed distribution, with many high-indegree pages (eg, Yahoo)
Analysis of power law

\[ \text{Pr [ page has } k \text{ inlinks ]} \approx k^{-2.1} \]

\[ \text{Pr [ page has } k \text{ outlinks ]} \approx k^{-2.7} \]

Corollary:

\[ \text{Pr [ page has } > k \text{ inlinks ]} \approx 1/k \]
Component sizes.

- Component sizes are distributed by the power law.

![WCC distribution](image1)

![SCC distribution](image2)
Other observed power laws in the web

- Depths of URLs
- Sizes of sites
- Eigenvalues of adjacency matrix of hyperlink graph [Mihail and Papadimitriou shed some light here]
- Many different traffic measures
- Linkage between hosts and domains
- Many of the above measures on particular subsets of the graph

... [Faloutsos, Faloutsos, Faloutsos 99]

[Bharat, Chang, Henzinger, Ruhl 02]
More Characterization: Self-Similarity
Ways to Slice the Web

- Domain (*.it)
- Host (www.ibm.com)
- Geography (pages with a geographical reference in the Western US)
- Content
  - Keyword: Math, subdivided by Math Geometry
  - Keyword: MP3, subdivided by MP3 Napster

We call these slices “Thematically Unified Communities”, or TUCs
Self-Similarity on the Web

- Pervasive: holds for all reasonable characteristics
- Robust: holds for all reasonable slices
- “Theorem:”
  - TUCs share properties with the web at large
  - TUCs are linked by a “navigational backbone”
In particular...

- All TUCs have:
  - Power laws for degree, SCC, and WCC distributions
  - Similar exponents for power laws
  - Similar “bow tie” structure
  - Large number of dense subgraphs
Is this surprising?

- YES (for downsampling general graphs). Example:
  - This graph has 1 SCC containing all nodes
  - Remove any nonzero fraction of edges – graph has \( n \) components of size 1
  - Generally: random subset of size \( n^{1/2} \) in a graph with \( O(n) \) edges will have only constant number of edges
A structural explanation

- Each TUC has a “bow tie” – how do they relate?
Each TUC contains a large SCC that is well-connected to the SCCs of other TUCs
[Dill, Kumar, McCurley, Rajagopalan, Sivakumar, Tomkins 2002]
Information Extraction from Large Graphs
Overview

Goal: Create higher-level "knowledge bases" of web information for further processing.

[Kumar, Raghavan, Rajagopalan, Tomkins 1999]
Many approaches to this problem

- Databases over the web:
  - Web SQL, Lore, ParaSite, etc
- Data mining
  - A priori, Query flocks, etc
- Information foraging
- Community extraction
  - [Lawrence et al]
- Authority-based search
  - HITS, and variants
General approach

- It’s hard (though getting easier) to analyze the content of all pages on the web
- It’s easier (though still hard) to analyze the graph
- How successfully can we extract useful semantic knowledge (i.e., community structure) from links alone?
Different communities appear to have very different structure.
Web Communities

Fishing

Linux

But both contain a common “footprint”: two pages (🔴) that both Point to three other pages in common (🟢)
Communities and cores

Definition: A "core" $K_{ij}$ consists of $i$ left nodes, $j$ right nodes, and all left->right edges.

Example $K_{2,3}$

Critical facts:
1. Almost all communities contain a core [expected]
2. Almost all cores betoken a community [unexpected]
Other footprint structures

- Newsgroup thread
- Web ring
- Corporate partnership
- Intranet fragment
Subgraph enumeration

- Goal: Given a graph-theoretic "footprint" for structures of interest, find ALL occurrences of these footprints.
Enumerating cores

Clean data by removing:
- mirrors (true and approximate)
- empty pages, too-popular pages,
  nepotistic pages

Preprocessing

When no more pruning is possible, finish using database techniques

Postprocessing

a belongs to a $K_{2,3}$ if and only if some node points to b1, b2, b3.

Inclusion/Exclusion Pruning
Results for cores

Number of cores found by Elimination/Generation

Number of cores found during postprocessing
The cores are interesting

Explicit communities.
- Yahoo!, Excite, Infoseek
- webrings
- news groups
- mailing lists

Implicit communities
- japanese elementary schools
- turkish student associations
- oil spills off the coast of japan
- australian fire brigades

(1) Implicit communities are defined by cores.
(2) There are an order of magnitude more of these. ($10^{5+}$)
(3) Can grow the core to the community using further processing.
Elementary Schools in Japan

- The American School in Japan
- The Link Page
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- Kids’ Space
- The Link Page-13
- “ù–¼¬Šw Z
- Fuzoku Home Page
- welcome to Miasa E&J school
- "½¬Šw Z
- fuzoku-es.fukui-u.ac.jp
- Torisu primary school
- Welcome to Yakumo Elementary, Hokkaido, Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

- schools
- LINK Page-13
- "ù–¼¬Šw Z
- a‰„, ¬Šw Z
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...net and Education
- http://www...iglobe.ne.jp/~IKESAN
- Torisu primary school
- TOYODA HOMEPAGE
- Education
- Cay’s Homepage (Japanese)
- UNIVERSITY
- %Ø—³¬Šw Z DRAGON97-TOP
- Yakumo Elementary, Hokkaido, Japan
- FuZOKU Home Page
- Kamishibun Elementary School...
So...

- Possible to extract order-of-magnitude more communities than currently known.
- Few (4%) of these appear coincidental.
- Entirely automatic extraction.
- Open question: how to use implicit communities?
A word on evolution
A word on evolution

- Phenomenon to characterize: A topic in a temporal stream occurs in a “burst of activity”
- Model source as multi-state
- Each state has certain emission properties
- Traversal between states is controlled by a Markov Model
- Determine most likely underlying state sequence over time, given observable output

[Kleinberg02]
Example

State 1: Output rate: very low
1

0.01

0.005

2

State 2: Output rate: very high

I've been thinking about your idea with the asparagus...
Uh huh
I think I see...
Uh huh
Yeah, that's what I'm saying...
So then I said "Hey, let's give it a try"
And anyway she said maybe, okay?

Time

Pr[1] ~ 1
Pr[1] ~ 10
Pr[1] ~ 5
Pr[1] ~ 10
Pr[1] ~ 2
Pr[1] ~ 1
Pr[1] ~ 2
Pr[2] ~ 5
Pr[2] ~ 2
Pr[2] ~ 5
Pr[2] ~ 2
Pr[2] ~ 7
Pr[2] ~ 10
Pr[2] ~ 10

Most likely “hidden” sequence:

1
1
1
1
1
2
2
2
More bursts

- Infinite chain of increasingly high-output states
- Allows hierarchical bursts
- Example 1: email messages
- Example 2: conference titles
Integrating bursts and graph analysis

Number of blog pages that belong to a community

Number of blog communities

Number of communities identified automatically as exhibiting “bursty” behavior – measure of cohesiveness of the blogspace

Wired magazine publishes an article on weblogs that impacts the tech community

Newsweek magazine publishes an article that reaches the population at large, responding to emergence, and triggering mainstream adoption

[KNRT03]