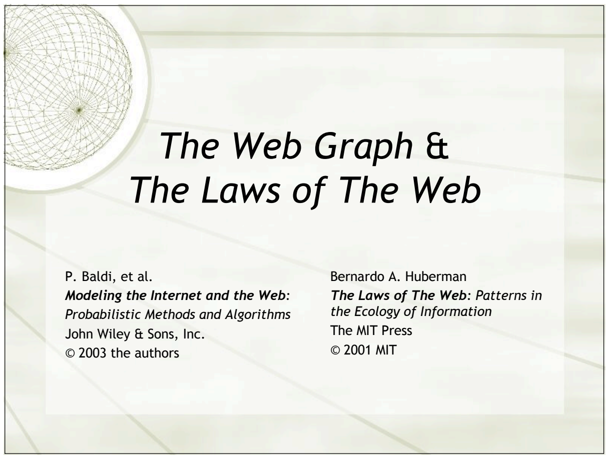
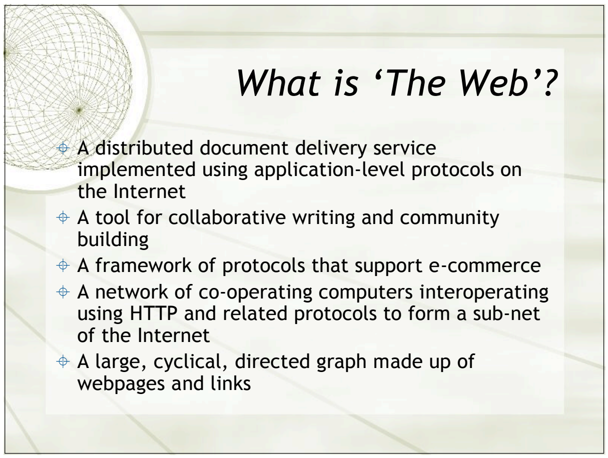


The Web Graph & The Laws of The Web

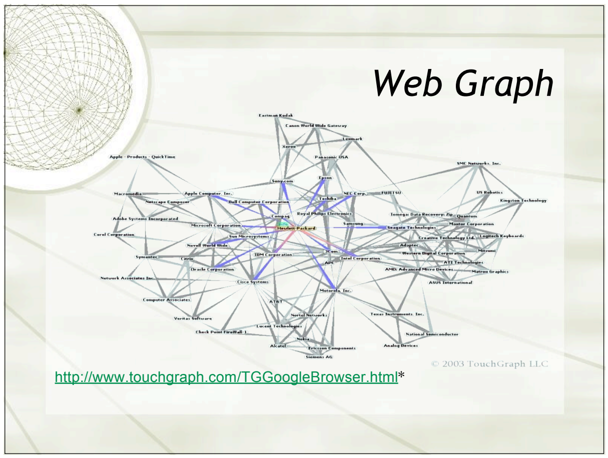
P. Baldi, et al.
Modeling the Internet and the Web:
Probabilistic Methods and Algorithms
John Wiley & Sons, Inc.
© 2003 the authors

Bernardo A. Huberman
***The Laws of The Web:** Patterns in the Ecology of Information*
The MIT Press
© 2001 MIT



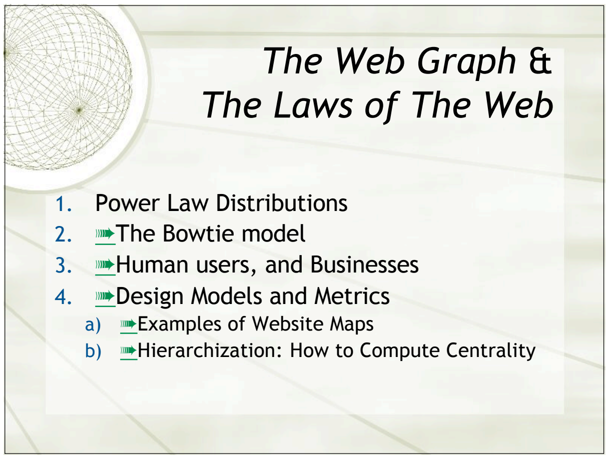
What is 'The Web'?

- A distributed document delivery service implemented using application-level protocols on the Internet
- A tool for collaborative writing and community building
- A framework of protocols that support e-commerce
- A network of co-operating computers interoperating using HTTP and related protocols to form a sub-net of the Internet
- A large, cyclical, directed graph made up of webpages and links



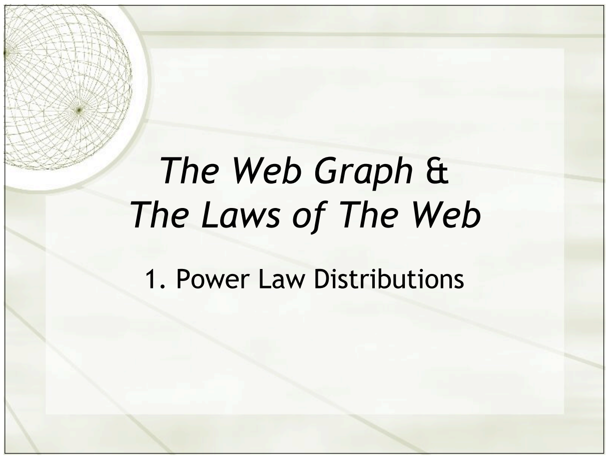
Web Graph

http://www.touchgraph.com/TGGoogleBrowser.html*



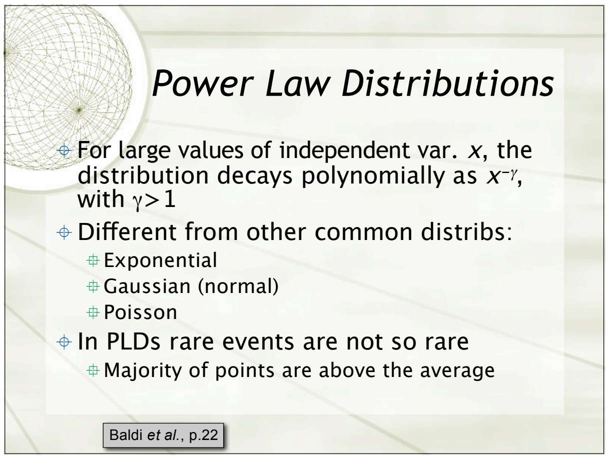
The Web Graph &
The Laws of The Web

1. Power Law Distributions
2. ➠The Bowtie model
3. ➠Human users, and Businesses
4. ➠Design Models and Metrics
   a) ➠Examples of Website Maps
   b) ➠Hierarchization: How to Compute Centrality



The Web Graph &
The Laws of The Web

1. Power Law Distributions

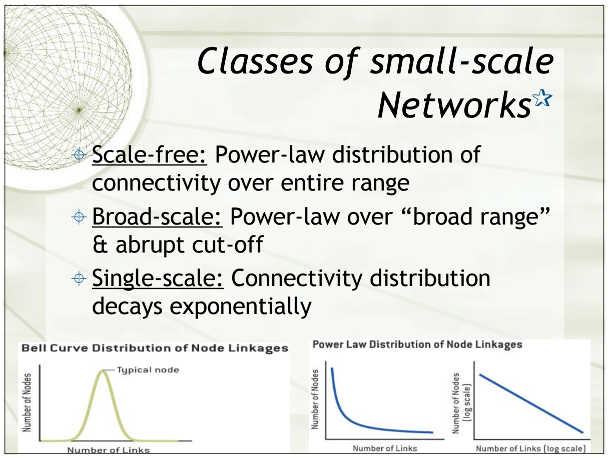


Power Law Distributions

- For large values of independent var. *x*, the distribution decays polynomially as $x^{-\gamma}$, with $\gamma>1$
- Different from other common distribs:
  - Exponential
  - Gaussian (normal)
  - Poisson
- In PLDs rare events are not so rare
  - Majority of points are above the average

Baldi *et al.*, p.22

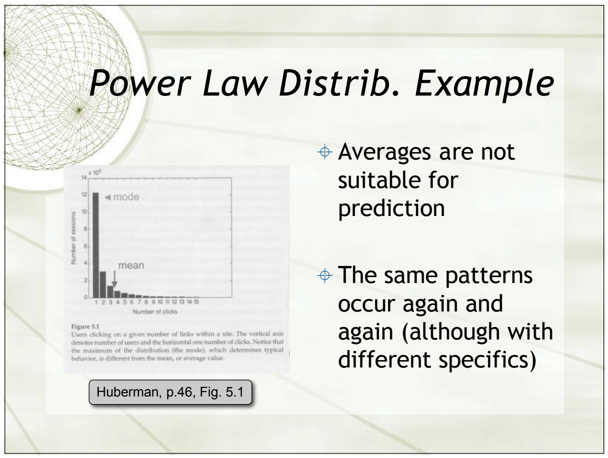
slide slightly modified from one by Baldi et al.



Classes of small-scale Networks☆

- Scale-free: Power-law distribution of connectivity over entire range
- Broad-scale: Power-law over “broad range” & abrupt cut-off
- Single-scale: Connectivity distribution decays exponentially

Bell Curve Distribution of Node Linkages
Typical node
Number of Nodes
Number of Links
Power Law Distribution of Node Linkages
Number of Nodes
Number of Links
Number of Nodes (log scale)
Number of Links (log scale)



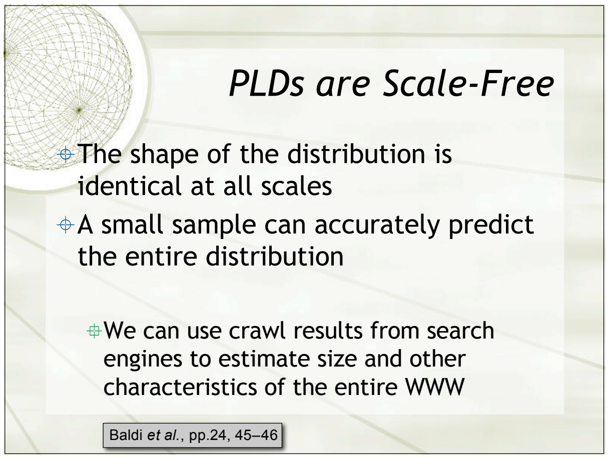
Power Law Distrib. Example

x 10⁵
14
12
10
8
6
4
2
0
Number of sessions
mode
mean
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Number of clicks

Figure 5.1
Users clicking on a given number of links within a site. The vertical axis denotes number of users and the horizontal one number of clicks. Notice that the maximum of the distribution (the mode), which determines typical behavior, is different from the mean, or average value.

Huberman, p.46, Fig. 5.1

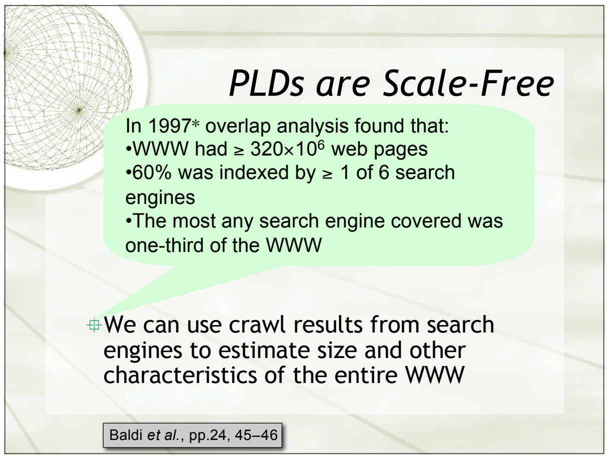
- Averages are not suitable for prediction
- The same patterns occur again and again (although with different specifics)



PLDs are Scale-Free

- The shape of the distribution is identical at all scales
- A small sample can accurately predict the entire distribution
  - We can use crawl results from search engines to estimate size and other characteristics of the entire WWW

Baldi *et al.*, pp.24, 45–46

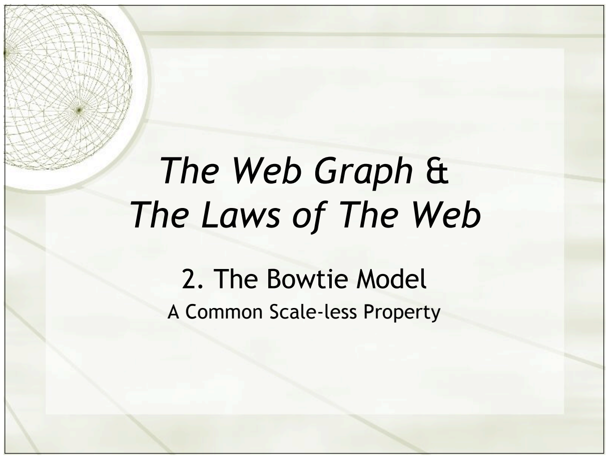


PLDs are Scale-Free

In 1997* overlap analysis found that:
- WWW had $\geq 320\times10^6$ web pages
- 60% was indexed by $\geq$ 1 of 6 search engines
- The most any search engine covered was one-third of the WWW

- We can use crawl results from search engines to estimate size and other characteristics of the entire WWW

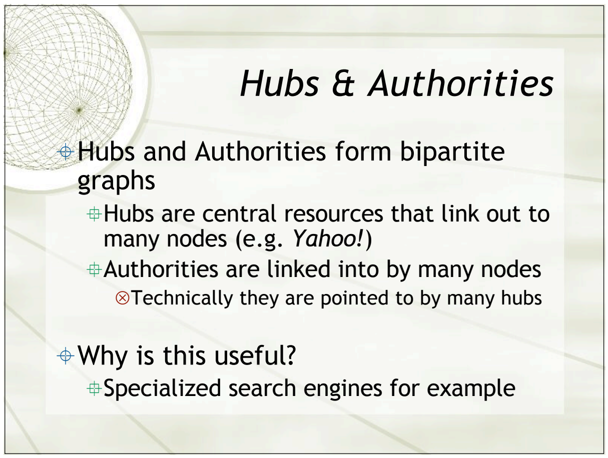
Baldi *et al.*, pp.24, 45–46



The Web Graph & The Laws of The Web

2. The Bowtie Model

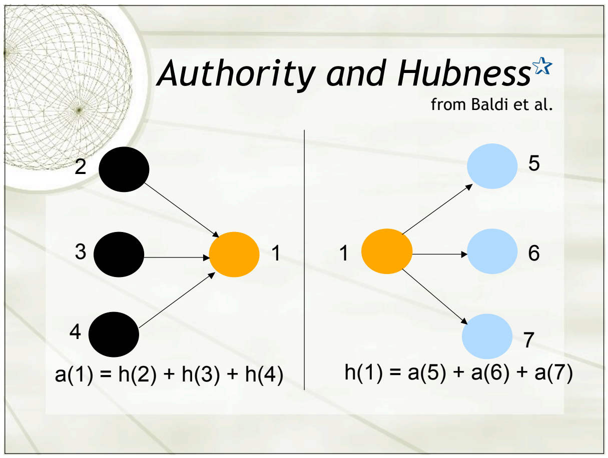
A Common Scale-less Property



Hubs & Authorities

- Hubs and Authorities form bipartite graphs
  - Hubs are central resources that link out to many nodes (e.g. *Yahoo!*)
  - Authorities are linked into by many nodes
    - ⊗Technically they are pointed to by many hubs
- Why is this useful?
  - Specialized search engines for example

This slide is slightly modified from one by Baldi et al. (<URL:http://ibook.ics.uci.edu/Slides/MIW%20Chapter%205.ppt>, slide 8 as of 2007-03-07)



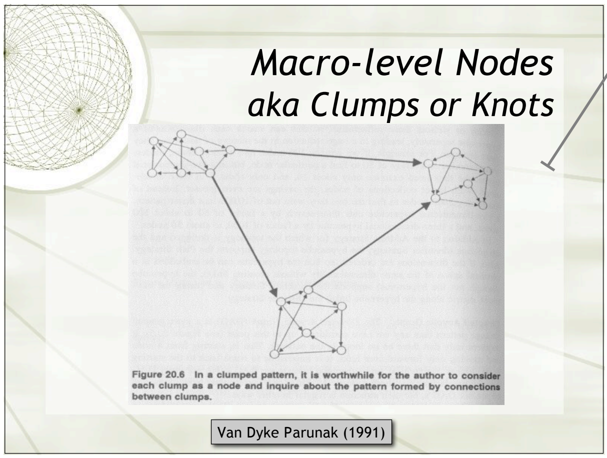
Authority and Hubness☆

from Baldi et al.

$a(1) = h(2) + h(3) + h(4)$

$h(1) = a(5) + a(6) + a(7)$

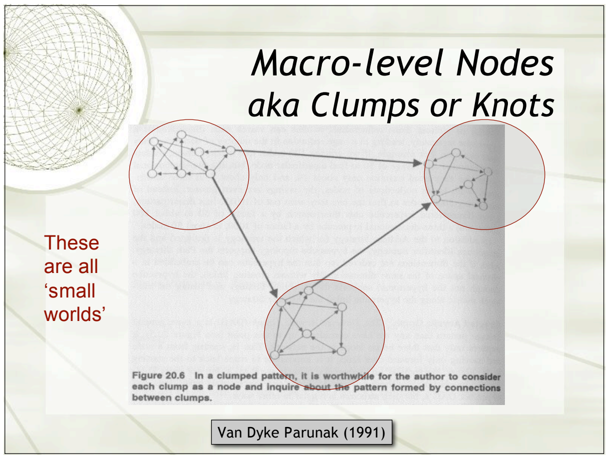
Figure 20.6 from page 310 of H. Van Dyke Parunak (1991). ‘Ordering the Information Graph’ Chapter 20 (pp.299–325) in *Hypertext/Hypermedia Handbook*, Emily Berk and Joseph Devlin(editors). Intertext Publications (New York, NY).



Macro-level Nodes aka Clumps or Knots

Figure 20.6 In a clumped pattern, it is worthwhile for the author to consider each clump as a node and inquire about the pattern formed by connections between clumps.

Van Dyke Parunak (1991)

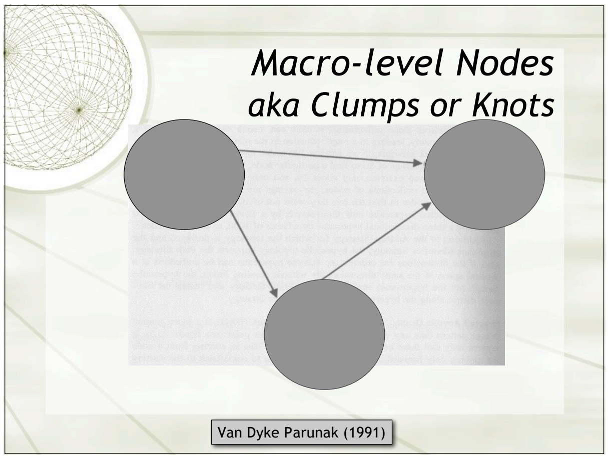


Macro-level Nodes aka Clumps or Knots

These are all ‘small worlds’

Figure 20.6 In a clumped pattern, it is worthwhile for the author to consider each clump as a node and inquire about the pattern formed by connections between clumps.

Van Dyke Parunak (1991)



Macro-level Nodes aka Clumps or Knots

Van Dyke Parunak (1991)

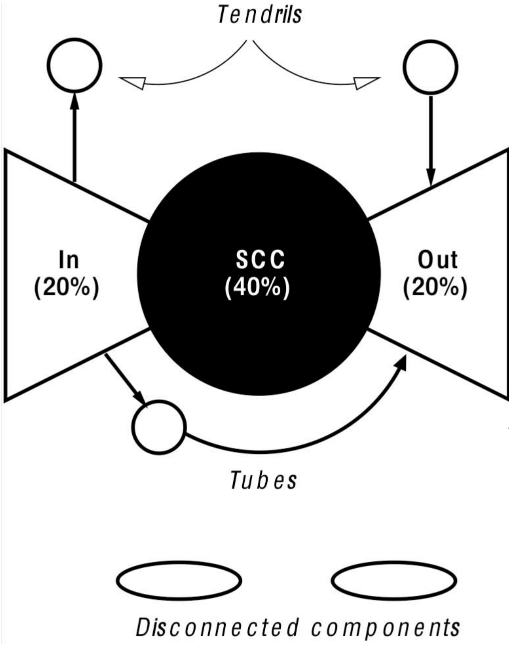
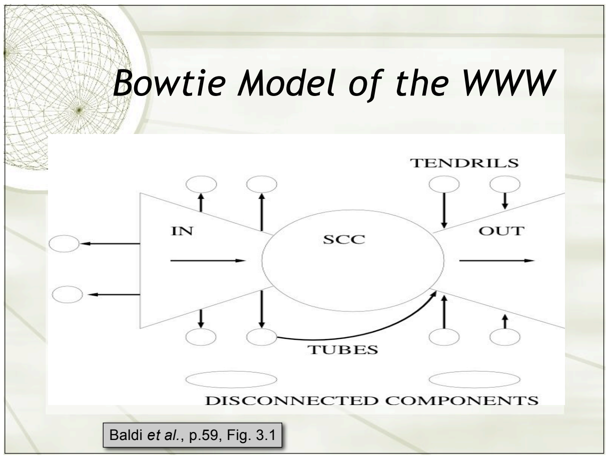
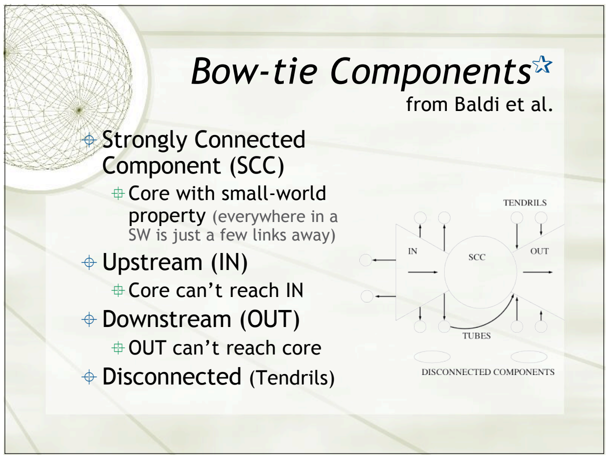


Image is a combination of Figure 3.1 from *Baldi et al.* and Business Wire Commercial Photo from Chris Sherman's article *New Web Map Reveals Previously Unseen 'Bow Tie' Organizational Structure* (22 May 2000) at Information Today, Inc. <URL: `http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=17813`> as it was on 23 Nov. 2009



Bowtie Model of the WWW

Baldi *et al.*, p.59, Fig. 3.1

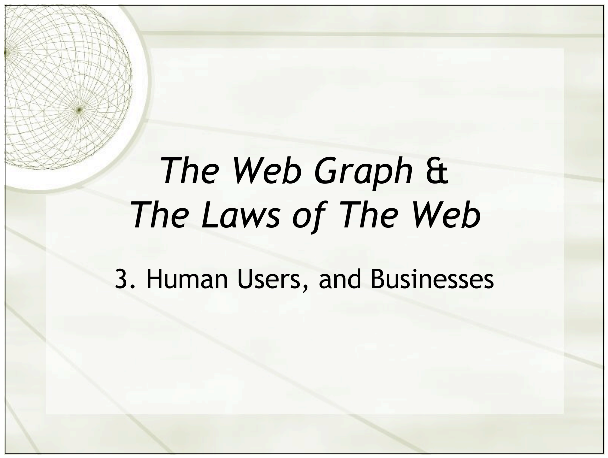


Bow-tie Components☆

from Baldi et al.

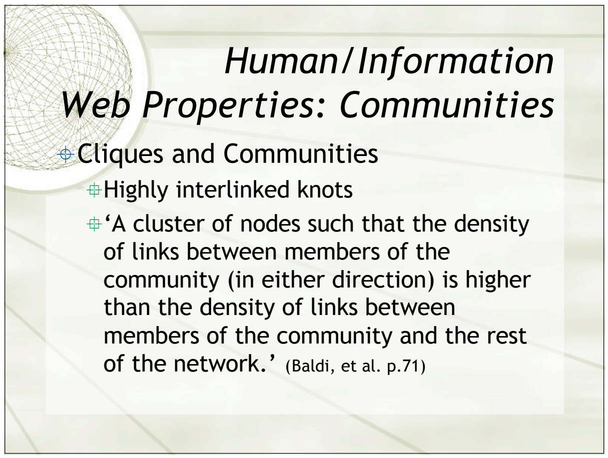
- Strongly Connected Component (SCC)
  - Core with small-world property (everywhere in a SW is just a few links away)
- Upstream (IN)
  - Core can't reach IN
- Downstream (OUT)
  - OUT can't reach core
- Disconnected (Tendrils)

TENDRILS
IN
SCC
OUT
TUBES
DISCONNECTED COMPONENTS



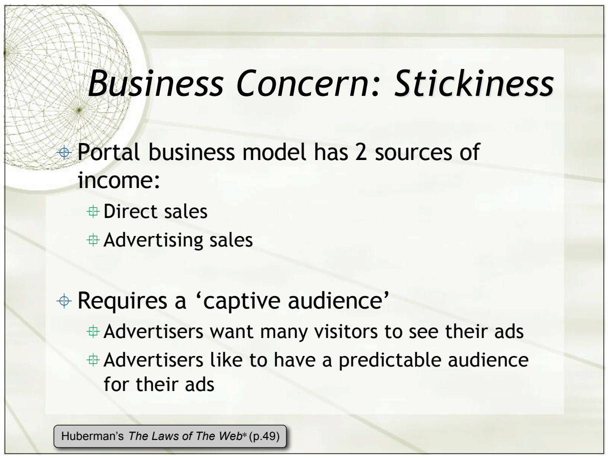
The Web Graph & The Laws of The Web

3. Human Users, and Businesses



Human/Information Web Properties: Communities

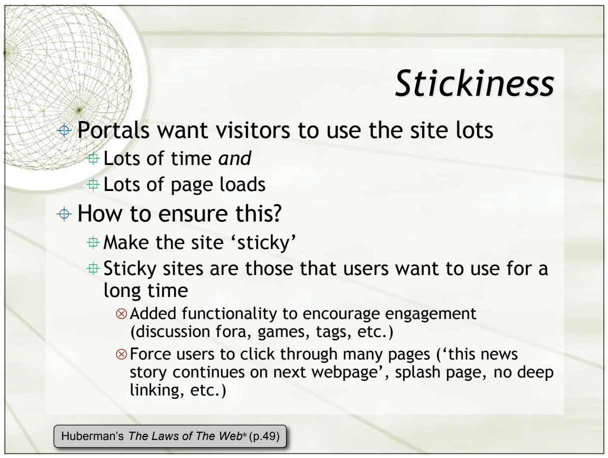
- Cliques and Communities
  - Highly interlinked knots
  - 'A cluster of nodes such that the density of links between members of the community (in either direction) is higher than the density of links between members of the community and the rest of the network.' (Baldi, et al. p.71)



Business Concern: Stickiness

- Portal business model has 2 sources of income:
  - Direct sales
  - Advertising sales
- Requires a 'captive audience'
  - Advertisers want many visitors to see their ads
  - Advertisers like to have a predictable audience for their ads

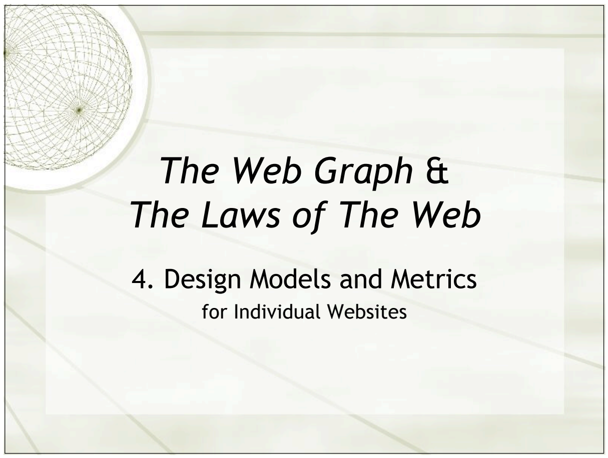
Huberman's *The Laws of The Web* (p.49)



Stickiness

- Portals want visitors to use the site lots
  - Lots of time *and*
  - Lots of page loads
- How to ensure this?
  - Make the site 'sticky'
  - Sticky sites are those that users want to use for a long time
    - ⊗ Added functionality to encourage engagement (discussion fora, games, tags, etc.)
    - ⊗ Force users to click through many pages ('this news story continues on next webpage', splash page, no deep linking, etc.)

Huberman's *The Laws of The Web*® (p.49)

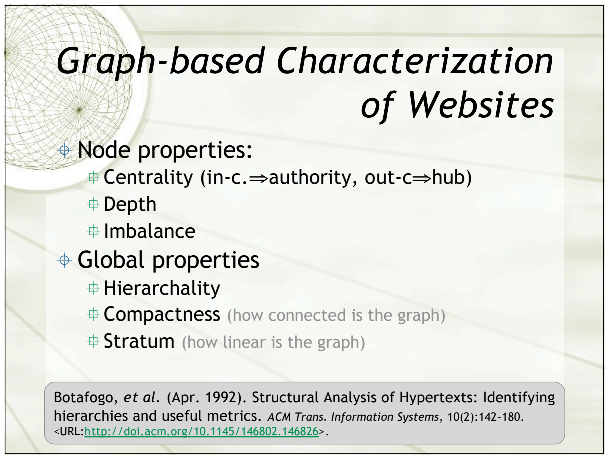


The Web Graph & The Laws of The Web

4. Design Models and Metrics

for Individual Websites

- *Depth* is distance from the root
- *Imbalance* refers to hierarchality: imbalanced nodes are at the root of trees that are not balanced (i.e. have more children in one branch than another)



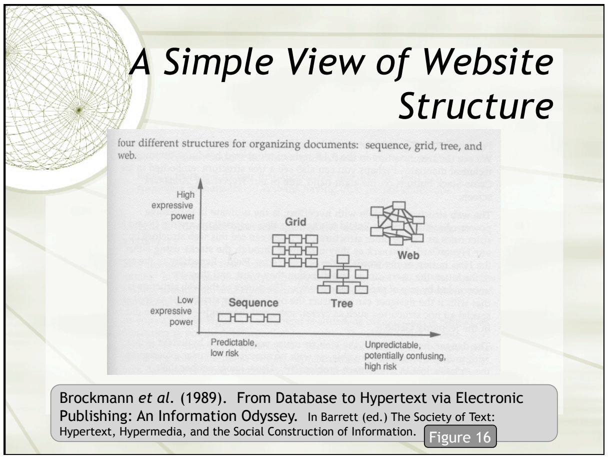
Graph-based Characterization of Websites

- Node properties:
  - Centrality (in-c.⇒authority, out-c⇒hub)
  - Depth
  - Imbalance
- Global properties
  - Hierarchality
  - Compactness (how connected is the graph)
  - Stratum (how linear is the graph)

Botafogo, *et al*. (Apr. 1992). Structural Analysis of Hypertexts: Identifying hierarchies and useful metrics. *ACM Trans. Information Systems*, 10(2):142–180. <URL:http://doi.acm.org/10.1145/146802.146826>.

Rodrigo A. Botafogo, Ehud Rivlin, and Ben Shneiderman, (Apr. 1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Trans. Information Systems*, 10(2):142–180. <URL:http://doi.acm.org/10.1145/146802.146826>.

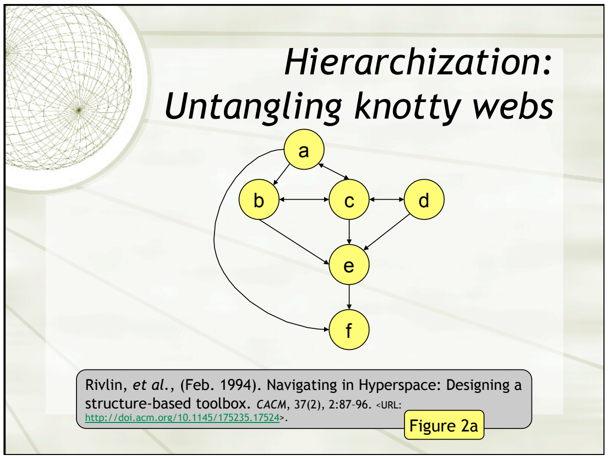
Ehud Rivlin, Rodrigo Botafogo, and Ben Shneiderman, (Feb. 1994). Navigating in hyperspace: designing a structure-based toolbox. *Communications of the ACM*, 37(2):87–96. <URL:http://doi.acm.org/10.1145/175235.175242>.



A Simple View of Website Structure

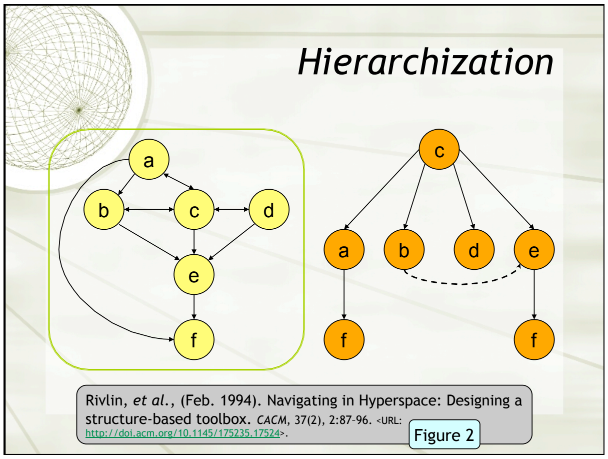
four different structures for organizing documents: sequence, grid, tree, and web.

Brockmann *et al.* (1989). From Database to Hypertext via Electronic Publishing: An Information Odyssey. In Barrett (ed.) The Society of Text: Hypertext, Hypermedia, and the Social Construction of Information. Figure 16



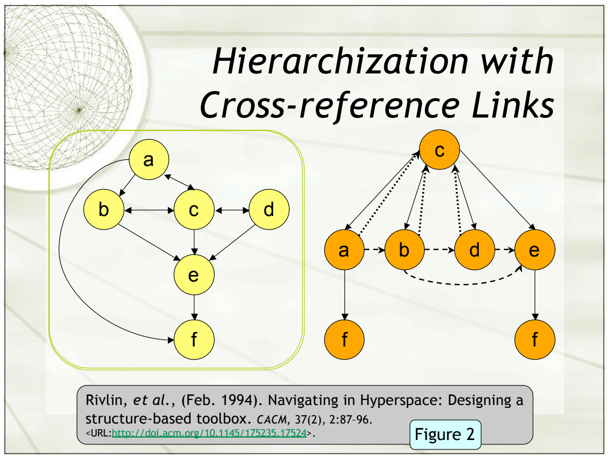
Hierarchization: Untangling knotty webs

Rivlin, *et al.*, (Feb. 1994). Navigating in Hyperspace: Designing a structure-based toolbox. *CACM*, 37(2), 2:87-96. <URL: http://doi.acm.org/10.1145/175235.17524>. Figure 2a



Hierarchization

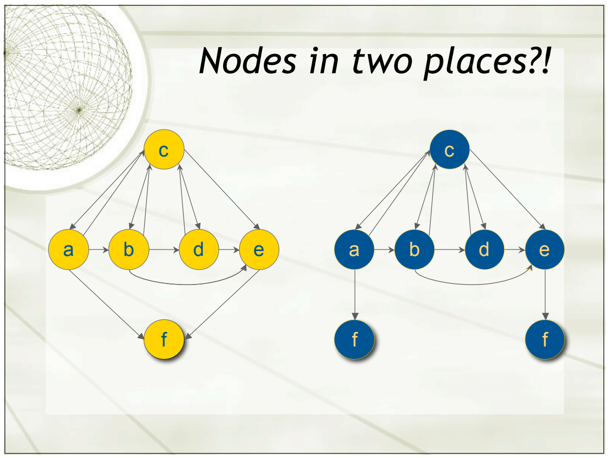
Rivlin, *et al.*, (Feb. 1994). Navigating in Hyperspace: Designing a structure-based toolbox. *CACM*, 37(2), 2:87-96. <URL: http://doi.acm.org/10.1145/175235.17524>. Figure 2



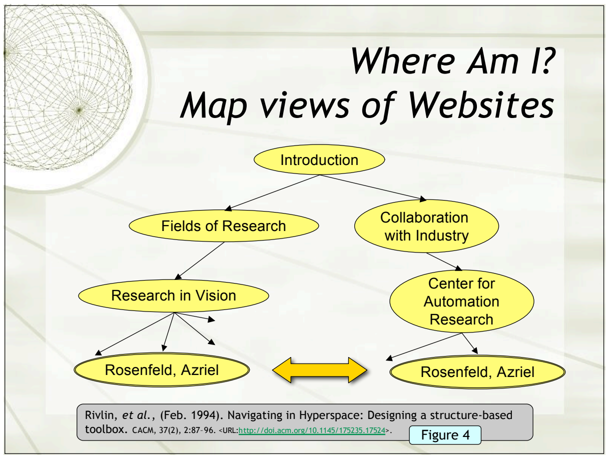
Hierarchization with Cross-reference Links

Rivlin, *et al.*, (Feb. 1994). Navigating in Hyperspace: Designing a structure-based toolbox. *CACM*, 37(2), 2:87-96. <URL:http://doi.acm.org/10.1145/175235.17524>.

Figure 2



Nodes in two places?!



Where Am I?
Map views of Websites

Rivlin, *et al.*, (Feb. 1994). Navigating in Hyperspace: Designing a structure-based toolbox. CACM, 37(2), 2:87-96. <URL:http://doi.acm.org/10.1145/175235.17524>.

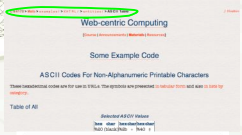
Figure 4

Types of Website Maps

- Breadcrumb lists
  - CS4173 examples
- Sitemap lists
  - CS4173 sitemap
- Sitemap pictures
  - CS4173 sitemap
- Multi-dimensional pictures
  - Colour, size, and position
  - Dynamic Diagrams, Inc.

Examples in picture form follow…

Some Sample Sitemaps



CS4173 > Mats > examples/ > XHTML/ > entities/ > ASCII Table

J. Blustein

Web-centric Computing

[Course | Announcements | Materials | Resources]

Some Example Code

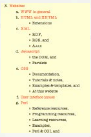
ASCII Codes For Non-Alphanumeric Printable Characters

These hexadecimal codes are for use in URLs. The symbols are presented in tabular form and also in lists by category.

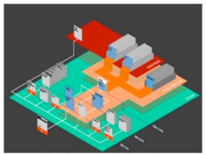
Table of All

Selected ASCII Values

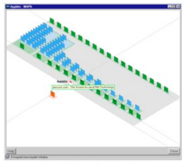
hex	char	hex	char	hex	char
%20	(blank)	%2b	+	%40	@



3. Websites
   a. WWW in general
   b. HTML and XHTML
      - Extensions
   c. XML
      - RDF,
      - RSS, and
      - Ajax
   d. Javascript
      - the DOM, and
      - Favelets
   e. CSS
      - Documentation,
      - Tutorials & notes,
      - Examples & templates, and
      - At this website
   f. User interface issues
   g. Perl
      - Reference resources,
      - Programming resources,
      - Learning resources,
      - Examples,
      - Perl & CGI, and

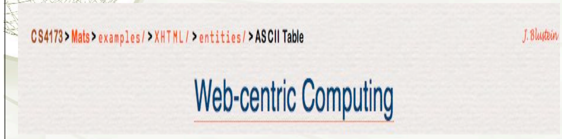


Map of the key design structures of the Nature Neuroscience Web site
© Dynamic Diagrams, Inc.


Applets - MAPA
Applets
java.sun.com - The Source for Java(TM) Technology
Help
Close
Unsigned Java Applet Window

A screen-shot of the MAPA map of part of the Javasoft Web site.
© Dynamic Diagrams, Inc.

Breadcrumb Detail

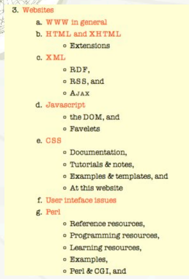


CS4173 > Mats > examples/ > XHTML/ > entities/ > ASCII Table

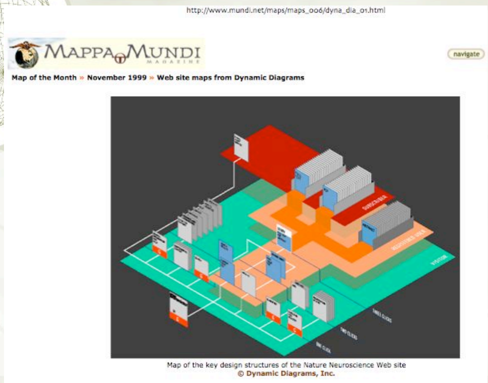
J. Blustein

Web-centric Computing

Sitemap List



3. Websites
   a. WWW in general
   b. HTML and XHTML
      - Extensions
   c. XML
      - RDF,
      - RSS, and
      - AJAX
   d. Javascript
      - the DOM, and
      - Favelets
   e. CSS
      - Documentation,
      - Tutorials & notes,
      - Examples & templates, and
      - At this website
   f. User inteface issues
   g. Perl
      - Reference resources,
      - Programming resources,
      - Learning resources,
      - Examples,
      - Perl & CGI, and

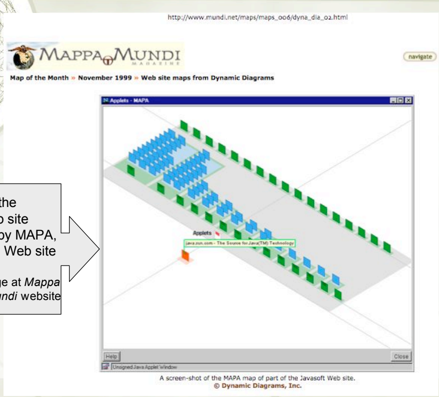


http://www.mundi.net/maps/maps_006/dyna_dia_01.html

MAPPA MUNDI MAGAZINE

Map of the Month » November 1999 » Web site maps from Dynamic Diagrams

Map of the key design structures of the Nature Neuroscience Web site
© Dynamic Diagrams, Inc.



http://www.mundi.net/maps/maps_006/dyna_dia_02.html

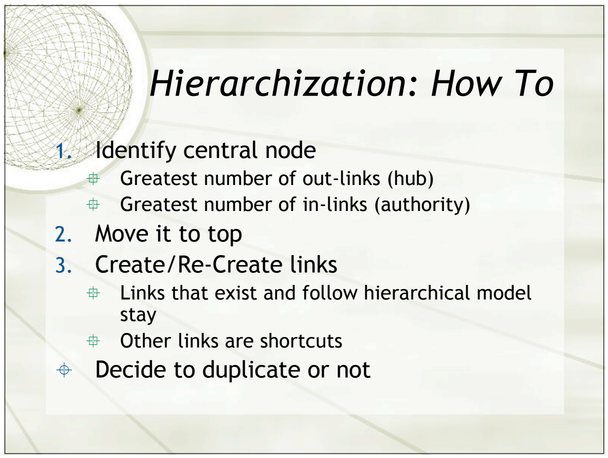
MAPPA MUNDI MAGAZINE

Map of the Month » November 1999 » Web site maps from Dynamic Diagrams

Applets - MAPA
Applets
java.sun.com - The Source for Java(TM) Technology
Help
Close
Unsigned Java Applet Window

A screen-shot of the MAPA map of part of the Javasoft Web site.
© Dynamic Diagrams, Inc.

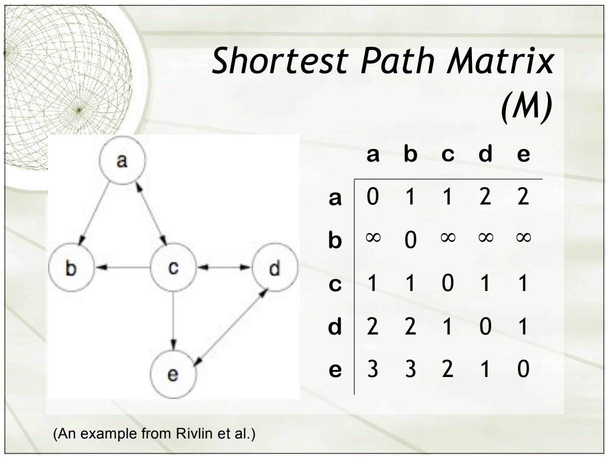
'A portion of the Javasoft Web site as rendered by MAPA, a data-driven Web site map system.'
— Martin Dodge at *Mappa Mundi* website



Hierarchization: How To

1. Identify central node
   - Greatest number of out-links (hub)
   - Greatest number of in-links (authority)
2. Move it to top
3. Create/Re-Create links
   - Links that exist and follow hierarchical model stay
   - Other links are shortcuts

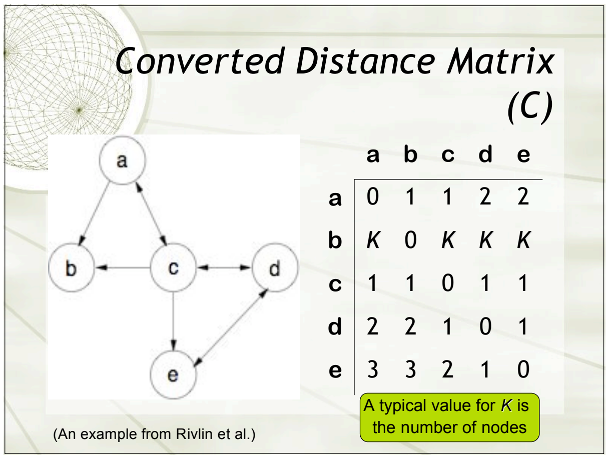
- Decide to duplicate or not



Shortest Path Matrix (M)

	a	b	c	d	e
a	0	1	1	2	2
b	∞	0	∞	∞	∞
c	1	1	0	1	1
d	2	2	1	0	1
e	3	3	2	1	0

(An example from Rivlin et al.)

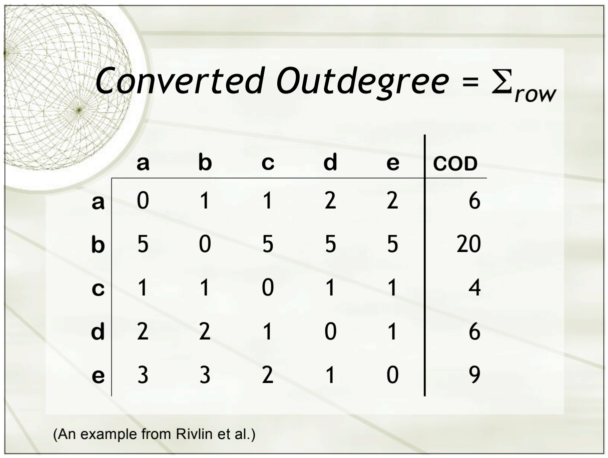


Converted Distance Matrix (C)

	a	b	c	d	e
a	0	1	1	2	2
b	*K*	0	*K*	*K*	*K*
c	1	1	0	1	1
d	2	2	1	0	1
e	3	3	2	1	0

A typical value for *K* is the number of nodes

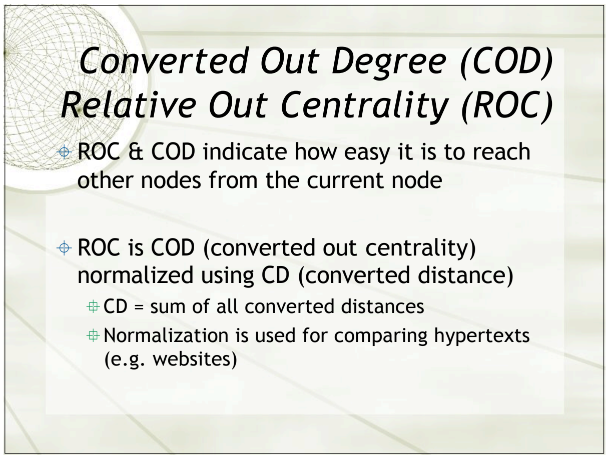
(An example from Rivlin et al.)



Converted Outdegree = Σ_{row}

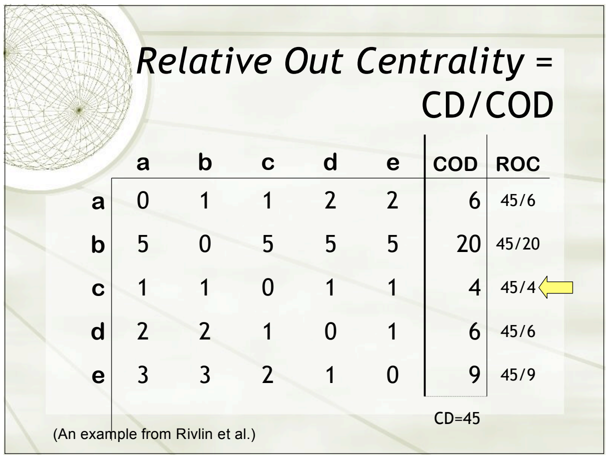
	a	b	c	d	e	COD
a	0	1	1	2	2	6
b	5	0	5	5	5	20
c	1	1	0	1	1	4
d	2	2	1	0	1	6
e	3	3	2	1	0	9

(An example from Rivlin et al.)



Converted Out Degree (COD) Relative Out Centrality (ROC)

- ROC & COD indicate how easy it is to reach other nodes from the current node
- ROC is COD (converted out centrality) normalized using CD (converted distance)
  - CD = sum of all converted distances
  - Normalization is used for comparing hypertexts (e.g. websites)



Relative Out Centrality = CD/COD

	a	b	c	d	e	COD	ROC
a	0	1	1	2	2	6	45/6
b	5	0	5	5	5	20	45/20
c	1	1	0	1	1	4	45/4 ⇦
d	2	2	1	0	1	6	45/6
e	3	3	2	1	0	9	45/9

CD=45

(An example from Rivlin et al.)