Assignment Project Exam Help

The Structure of the Web

https://powcoder.com

Add WeChat powcoder

Mark Levene

(Follow the links to learn more!)

Questions

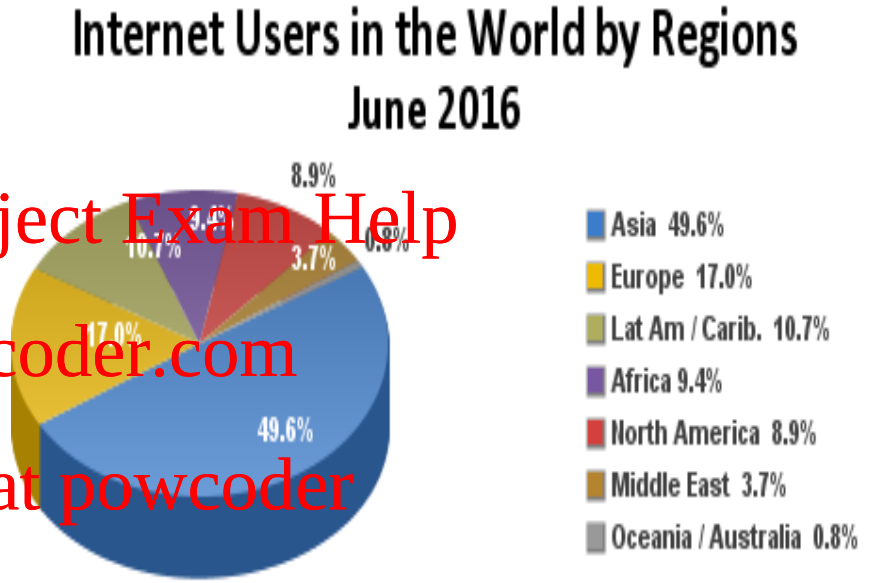
- How many people use the web?
- What is the size of the web?
- How many web sites are there?
- How many searches per day?
- How do web pages change?
- What is the graph structure of the web?
- How could the structure arise?
- What can we do with link analysis?

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Global Internet Statistics

- 49.5% of world population is online
- 91.6% online in the UK
- 96.3% online is Norway
- 73.5% online in Europe
- 89.0% online in USA

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

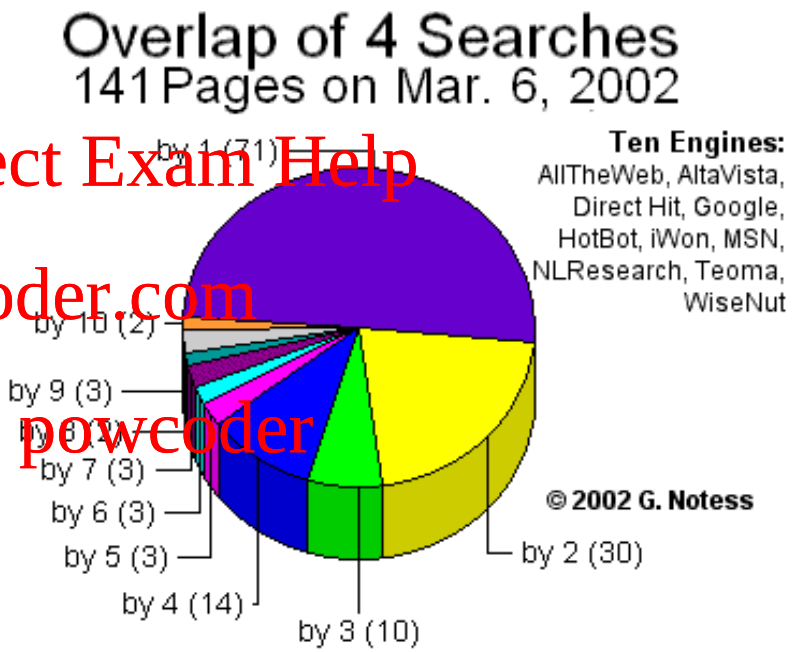


Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 3,611,375,813 Internet users on June 30, 2016
Copyright © 2016, Miniwatts Marketing Group

See more stats by following the link.

The Size of the Web

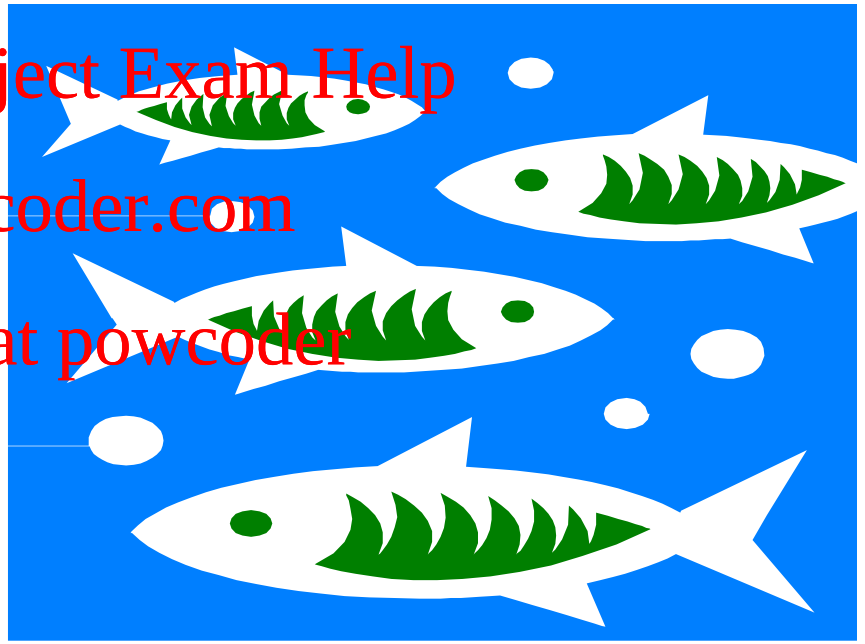
- Lawrence and Giles 1999 – initial estimate of 800 million web pages
- Over 10 Billion in 2005.
- Trillions in 2016
- Coverage – about 40% in 1999
- Overlap - low
- The *deep* (or *hidden* or *invisible)* web contains 400-550 times more information.



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Capture Recapture

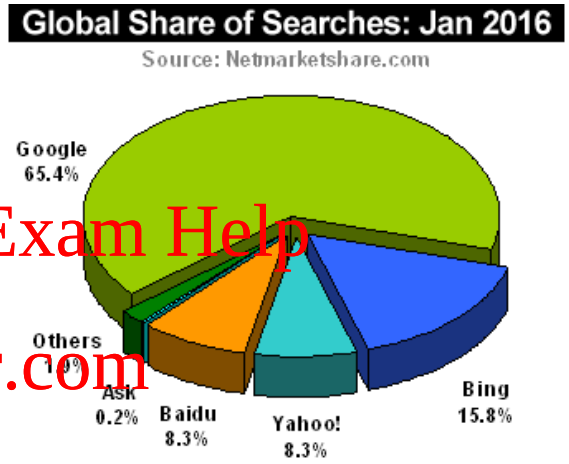
- **SE1** – reported size of search engine 1.
- **Q** – set of queries.
- **QSE1** and **QSE2** – pages returned for **Q** from two engines.
- **OVR** – overlap of **QSE1** and **QSE2.**
- Estimate of **Web** size: (**QSE2** x **SE1) / OVR**



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Search Engine Statistics

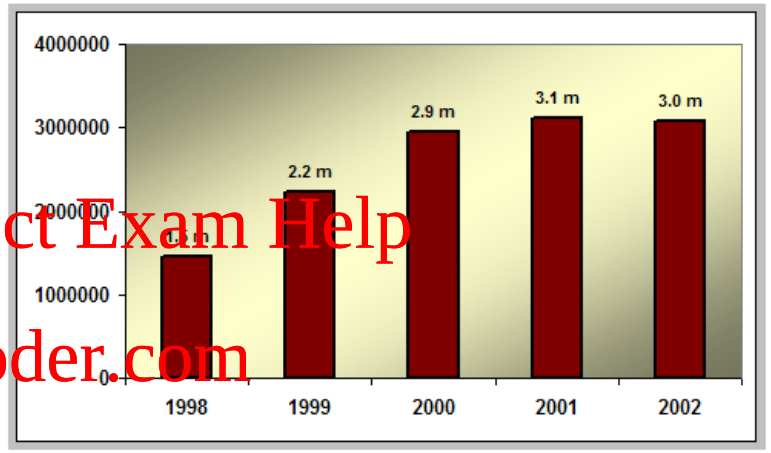
- Google has over 3.5 billion searches a day
- January 2016- Google has 65.4% searches.
- How much time do users spend on the internet per day? About 3 hours (Google it yourself.)



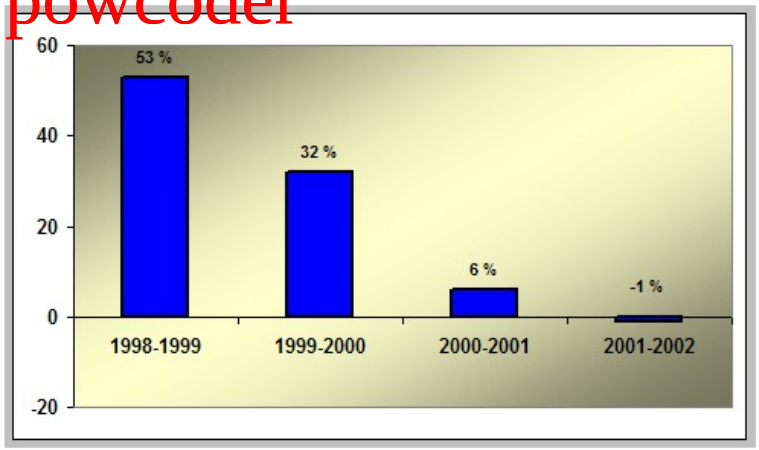
Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Growth in number of Public Sites

- Number of web sites identified by capture-recapture method by sampling random IPs.
- Average size of web site 441 pages.
- Decrease in 2002 – no rush to get online, economic factors.



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder



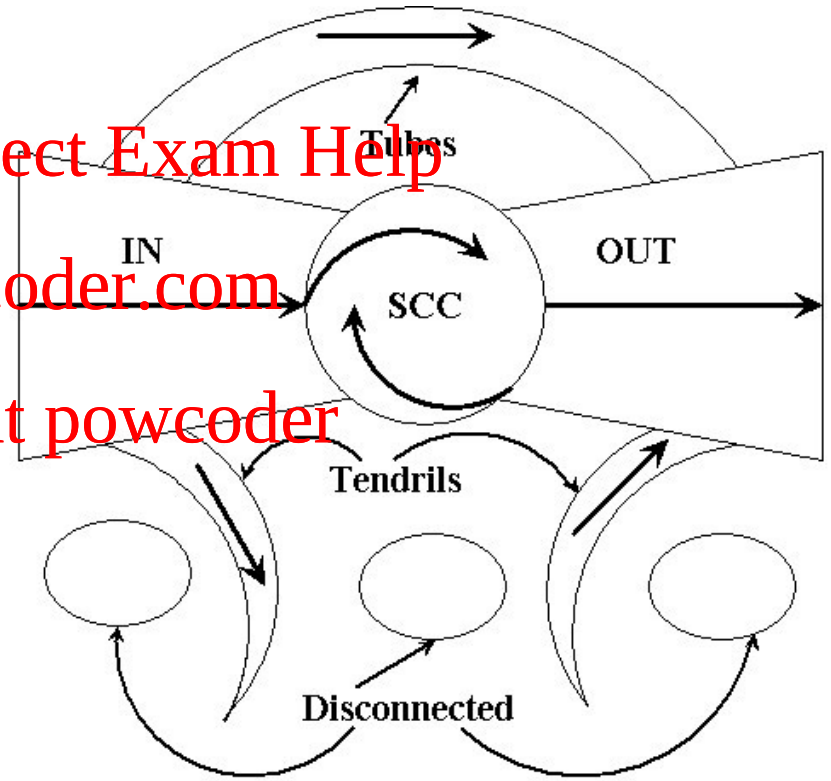
How do Web Pages Change

- Most pages do not change much.
- Larger pages change more often.
- Commercial pages change more often.
- Past change to a web page is a good indicator of future change.
- About 30% of pages are very similar to other pages, and being a near-duplicate is fairly stable.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Bowtie Model of the Web

- Broder et al. 1999 – crawl of over 200 million pages and 1.5 billion links.
- SCC – 27.5%
- IN and OUT – 21.5%
- Tendrils and tubes – 21.5%
- Disconnected – 8%



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Diameter of the Web

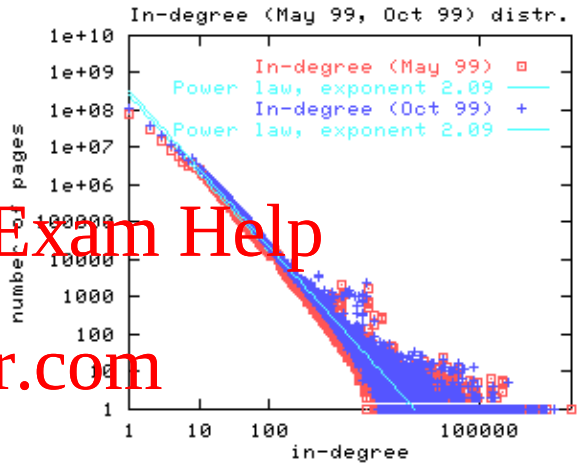
- Compute Average shortest path between pairs of pages that have a path from one to the other.
- Broder 99 – directed 16.2, undirected 6.8
- Barabasi 99 – directed for nd.edu 19
- Small diameter is a characteristic of a *small world* network
- Choose random source and destination – 75% of the time **no** directed path between them.

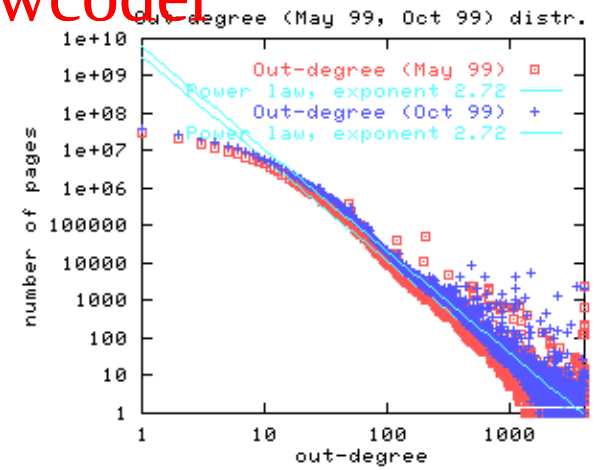
Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Web Structure Distributions

- Average out-degree between 7 and 8
- Degree distributions – how many page have n=1,2,… links:
  - indegree : $\frac{1}{n^{2.1}}$
  - outdegree : $\frac{1}{n^{2.72}}$
- Log-log plots

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder





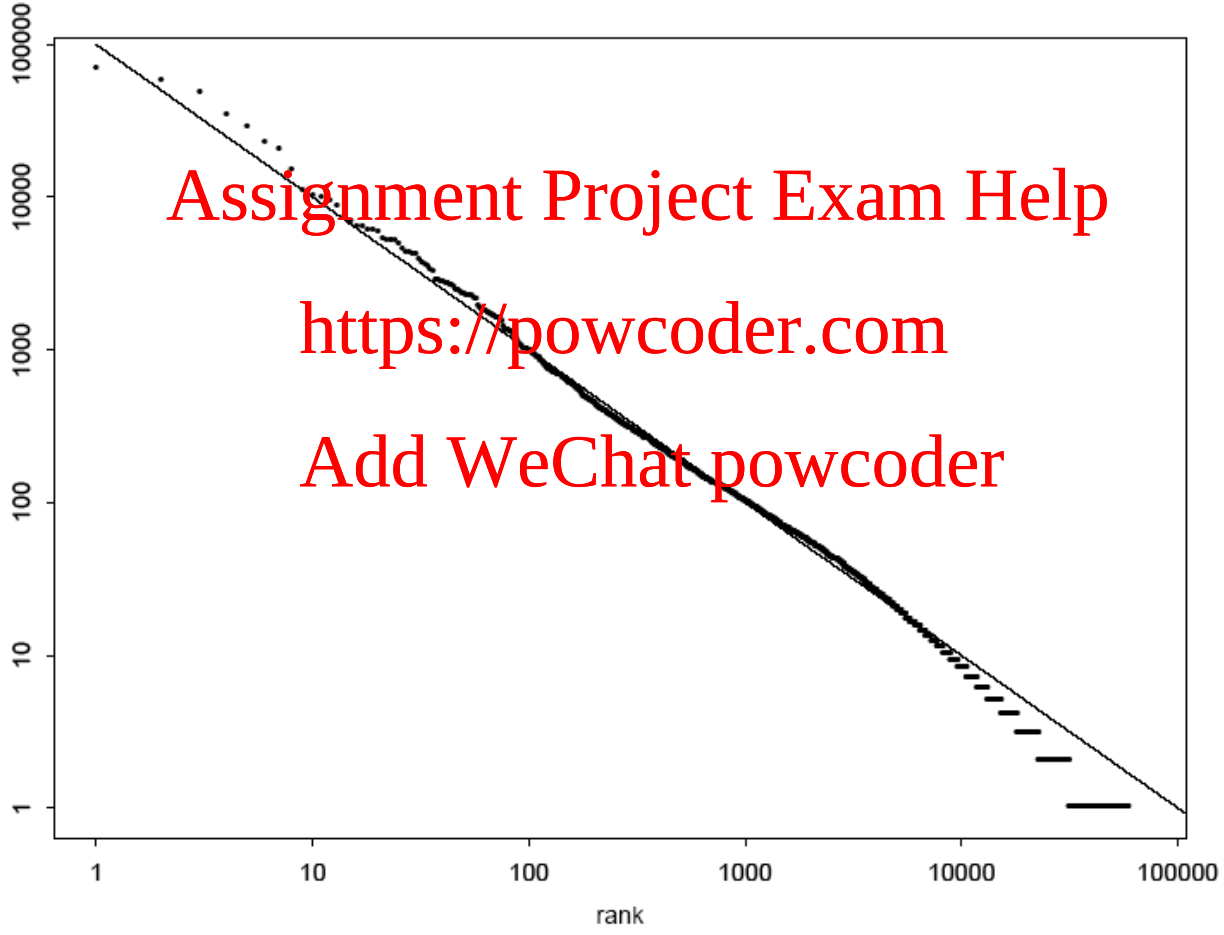
What is a Power Law

$$f(i) = \frac{C}{i^{\tau}}$$

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- *f(i)* is the proportion of objects having property *i*
- *E.g. f(i) = # pages, i = # inlinks*
- *E.g. f(i) = # sites, i = # pages*
- *E.g. f(i) = # sites i = # users*
- *E.g. f(i) = frequency of word , i = rank of word, from most freqeunt to least frequent*
- *Log-log plot - linear relationship (straight line)*

Zipf's Distribution for Brown Corpus (1 million words – f(r) approx. C/r)



Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

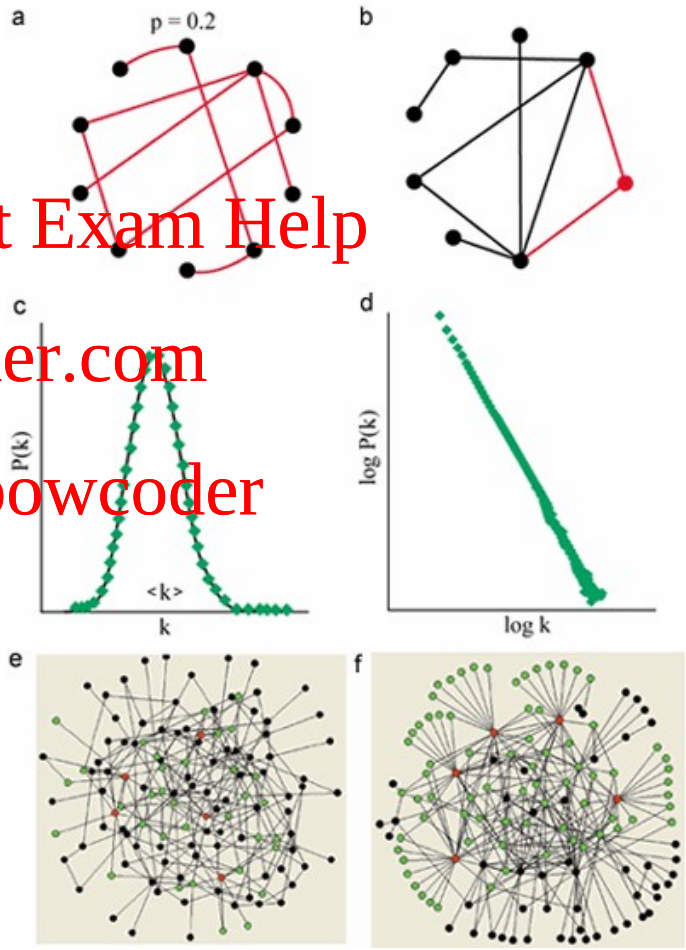
Word Frequency for Brown Corpus

	Word	Instances	% Frequency
1.	The	69970	6.8872
2.	of	36410	3.5839
3.	and	28854	2.8401
4.	to	26154	2.5744
5.	a	23363	2.2996
6.	in	21345	2.1010
7.	that	10594	1.0428
8.	is	10102	0.9943
9.	was	9815	0.9661
10.	He	9542	0.9392

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

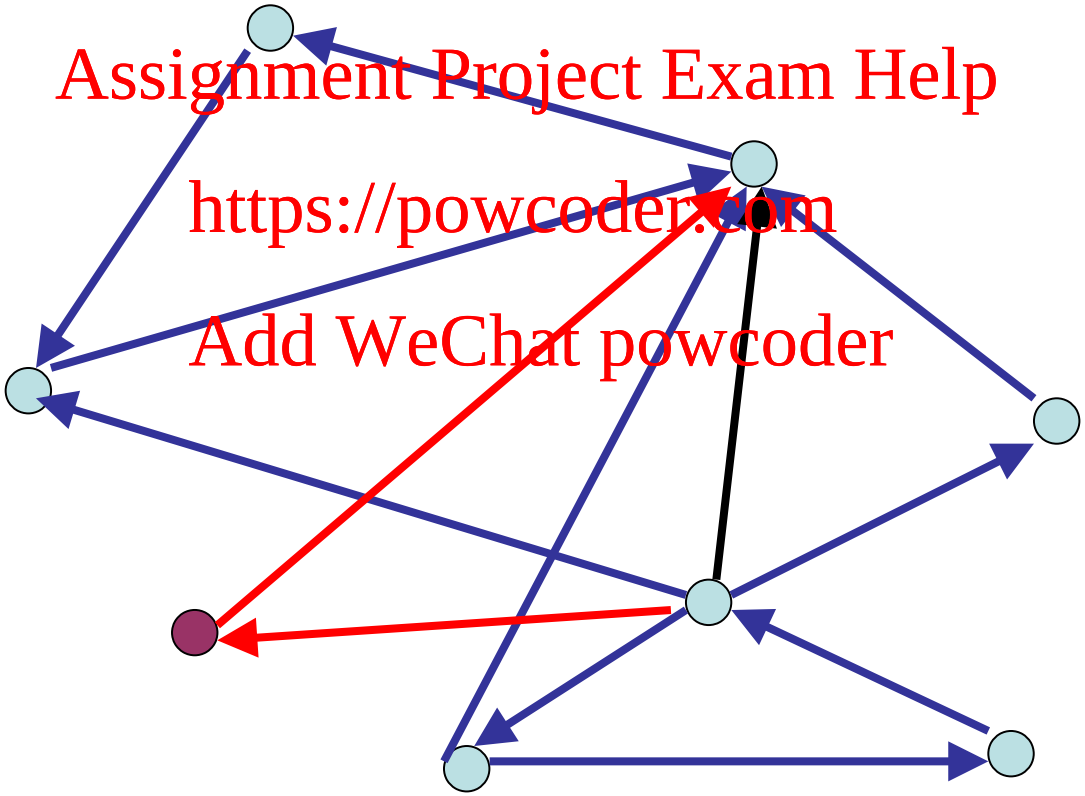
Evolving Random Networks

- Classical random graphs – all links have the same probability p – degree distribution is poisson
- Evolving networks – log-log degree distribution is linear
- Model – add new node and randomly link to it with probability p, or with probability 1-p choose an existing node with proportion to its inllinks.



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

How Power Laws Arise - Preferential Attachment or The Rich Get Richer



Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Power Laws on the Web

- inlinks (2.1)
- outlinks (2.72)
- Strongly connected components (2.54)
- No. of web pages in a site (2.2)
- No. of visitors to a site during a day (2.07)
- No. links clicked by web surfers (1.5)
- PageRank (2.1)

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Robustness and Vulnerability of Power Law Networks

- The web is extremely robust against attacks targeted at random web sites.
- The web is vunerable against an attack targeted at well-connected nodes.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- Has implications, e.g. on the spread of viruses on the Internet.