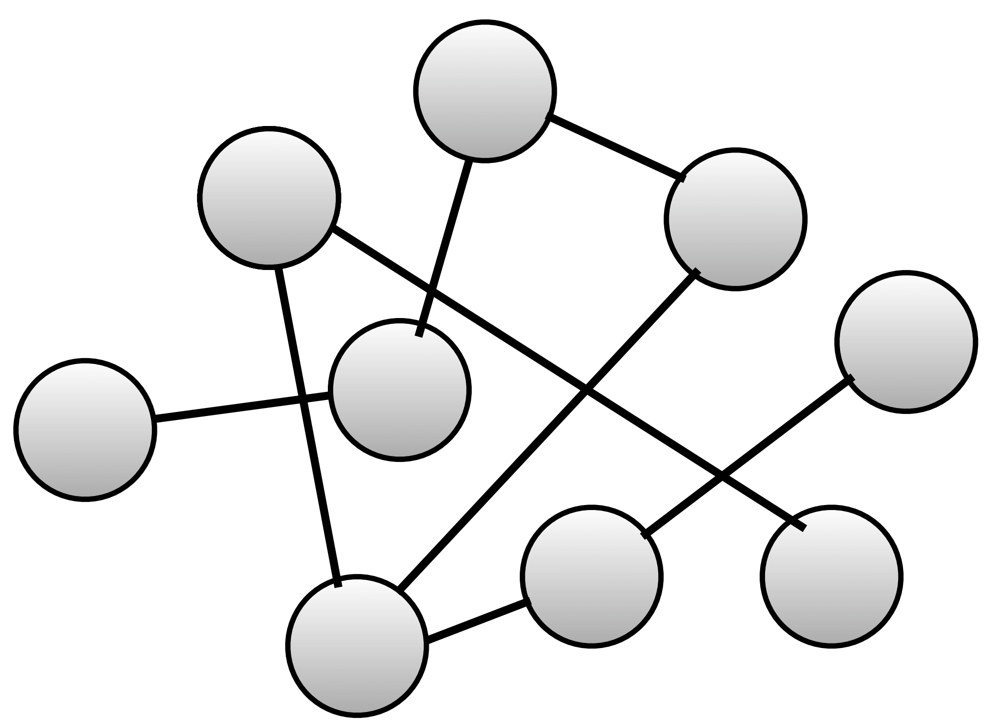
# Mining di Dati Web

Lezione 2 - Webgraph & its Models

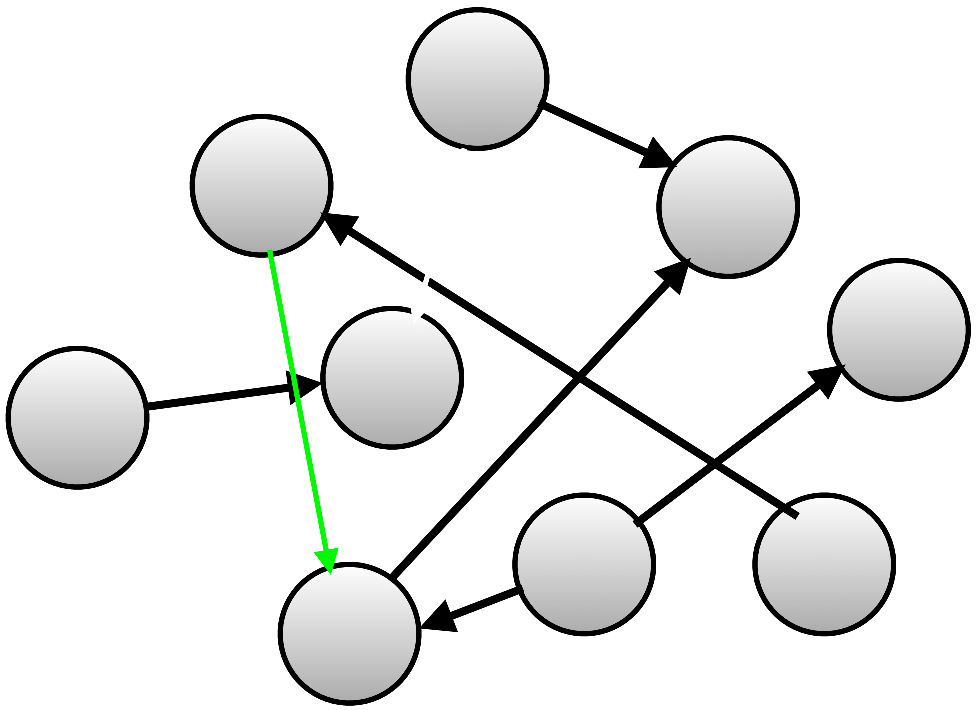
# Introduction

- A graph G=(V,E) is characterized by a set of nodes (vertexes) V and a set of Edges E whose elements are pairs $(v_1,v_2)$ where $v_1,v_2$ are vertexes in V.



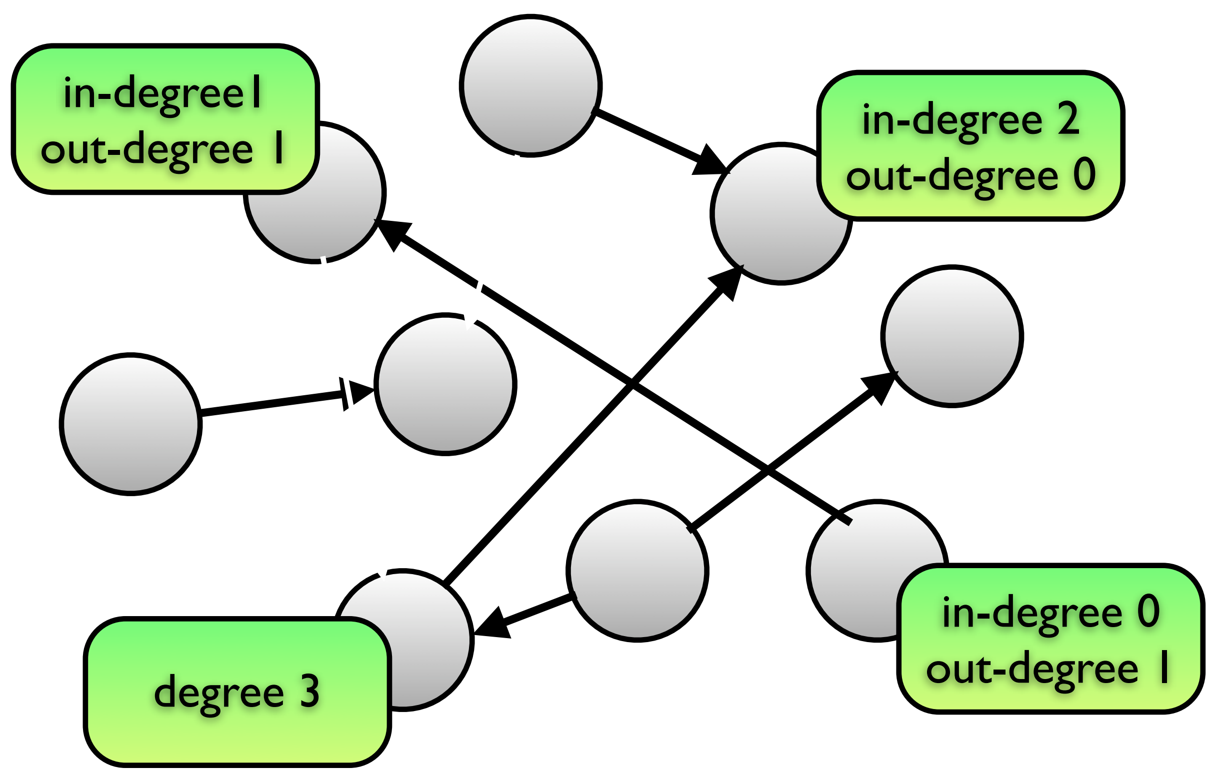
# Directed Graph

- A graph G=(V,E) is directed (a.k.a. digraph) if edges in E are ordered pairs of vertexes.



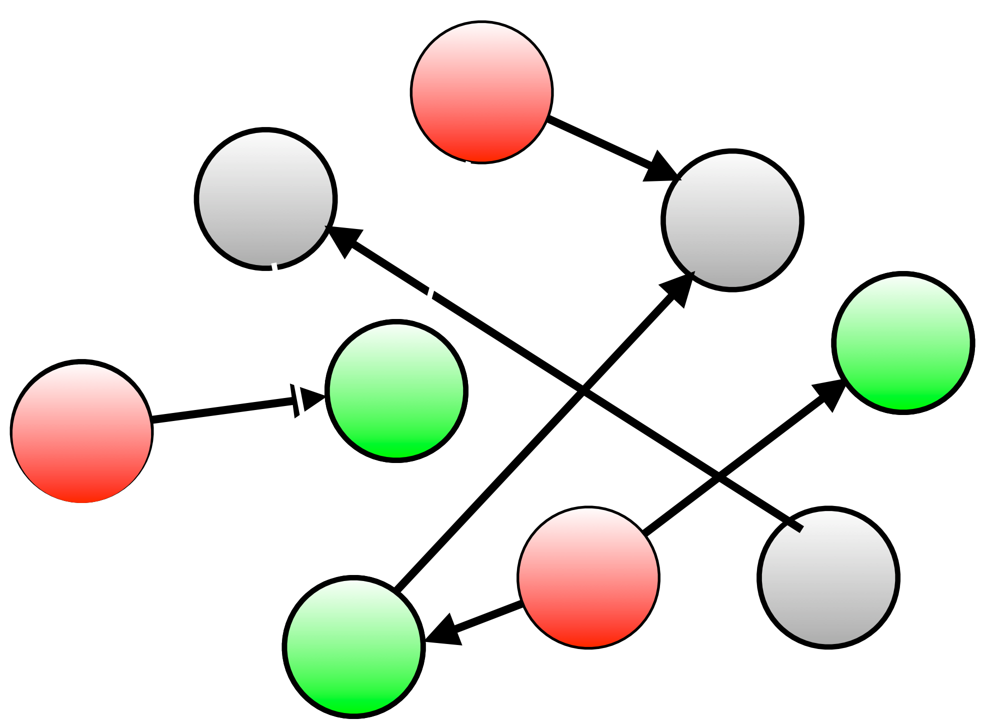
# Features of a (Di)Graph

- The degree of a vertex is the number of edges incident to it
- The in-degree (out-degree) of a node in a digraph is the number of incoming (outgoing) edges.



# Successor and Predecessor

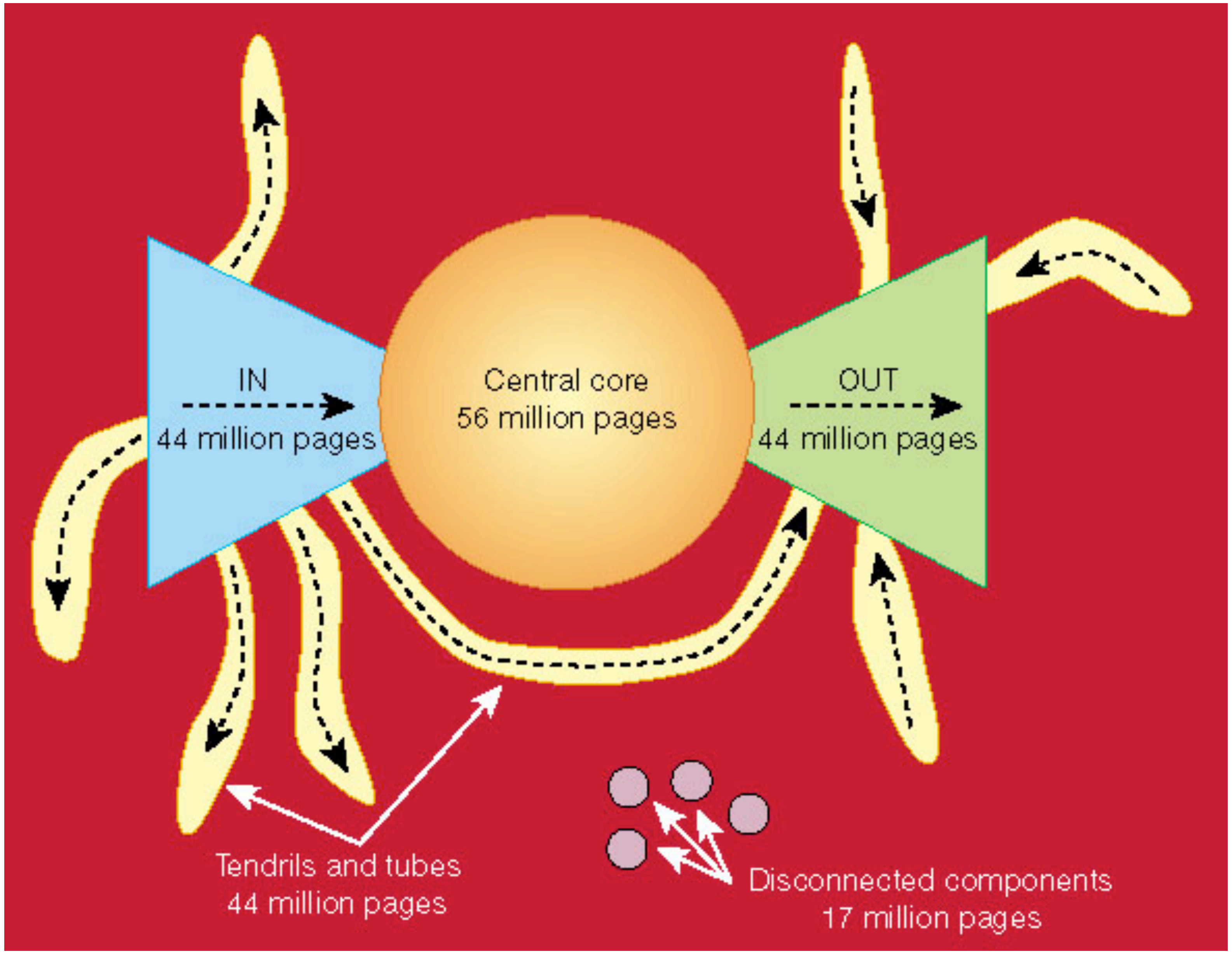
- We call successors of a node v, all the nodes pointed by v
- We call predecessors of a node v, all the nodes that point to v



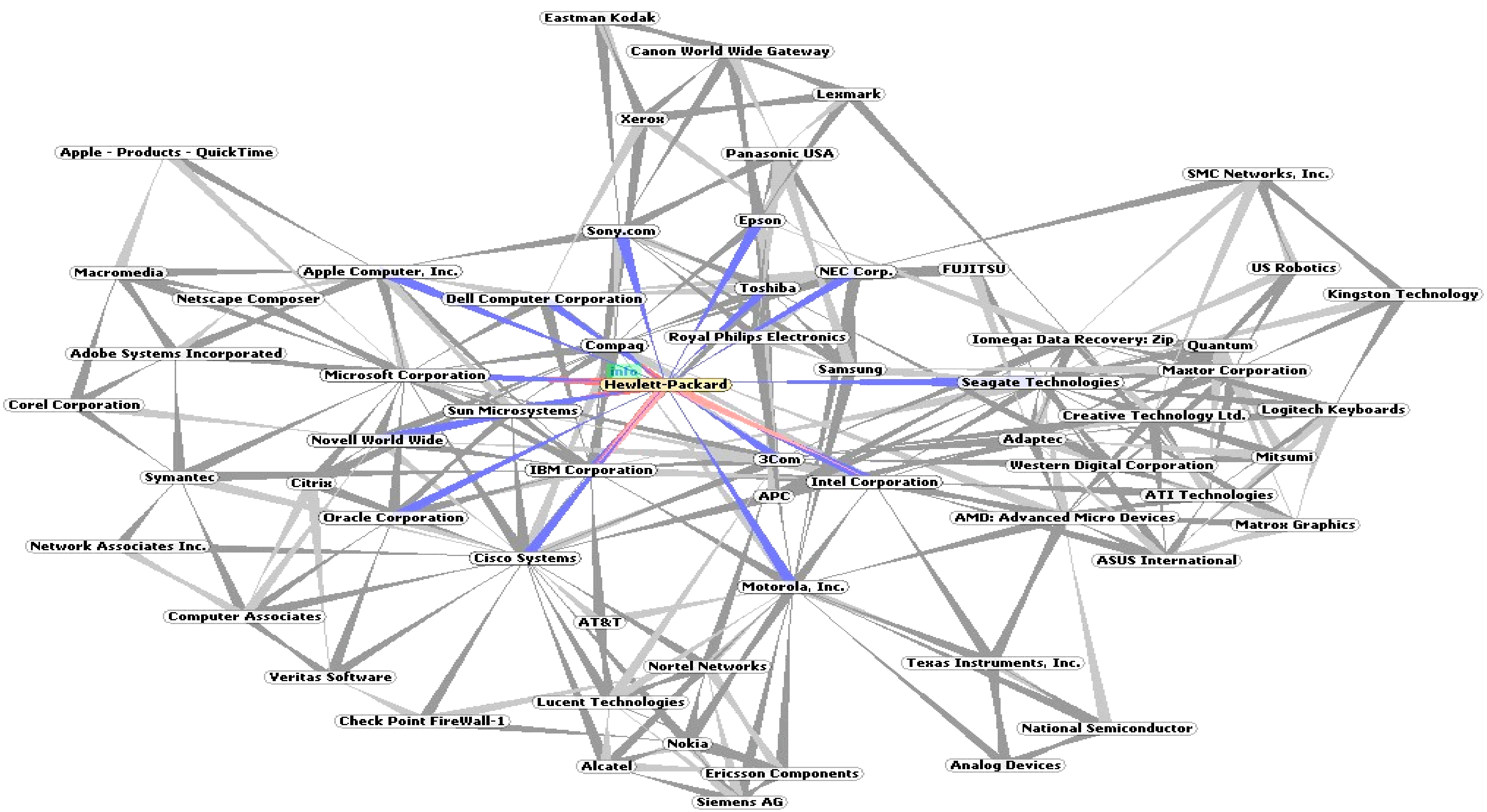
# Subset of Nodes

- A subset of nodes S of V is a connected component *iff* for every pair o vertices u,v in S, u is *reachable* from v.
- A graph is connected iff for every pair of vertices u,v in V, u is *reachable* from v.
- A set of nodes S is a strongly connected component (SCC) of a *digraph* iff, for every pair of nodes A,B in S, there exists a directed path from A to B <u>and</u> from B to A, <u>and</u> the set is *maximal*.

# The Webgraph



# A “sort of” Webgraph



# Well...

# The Size of Webgraph

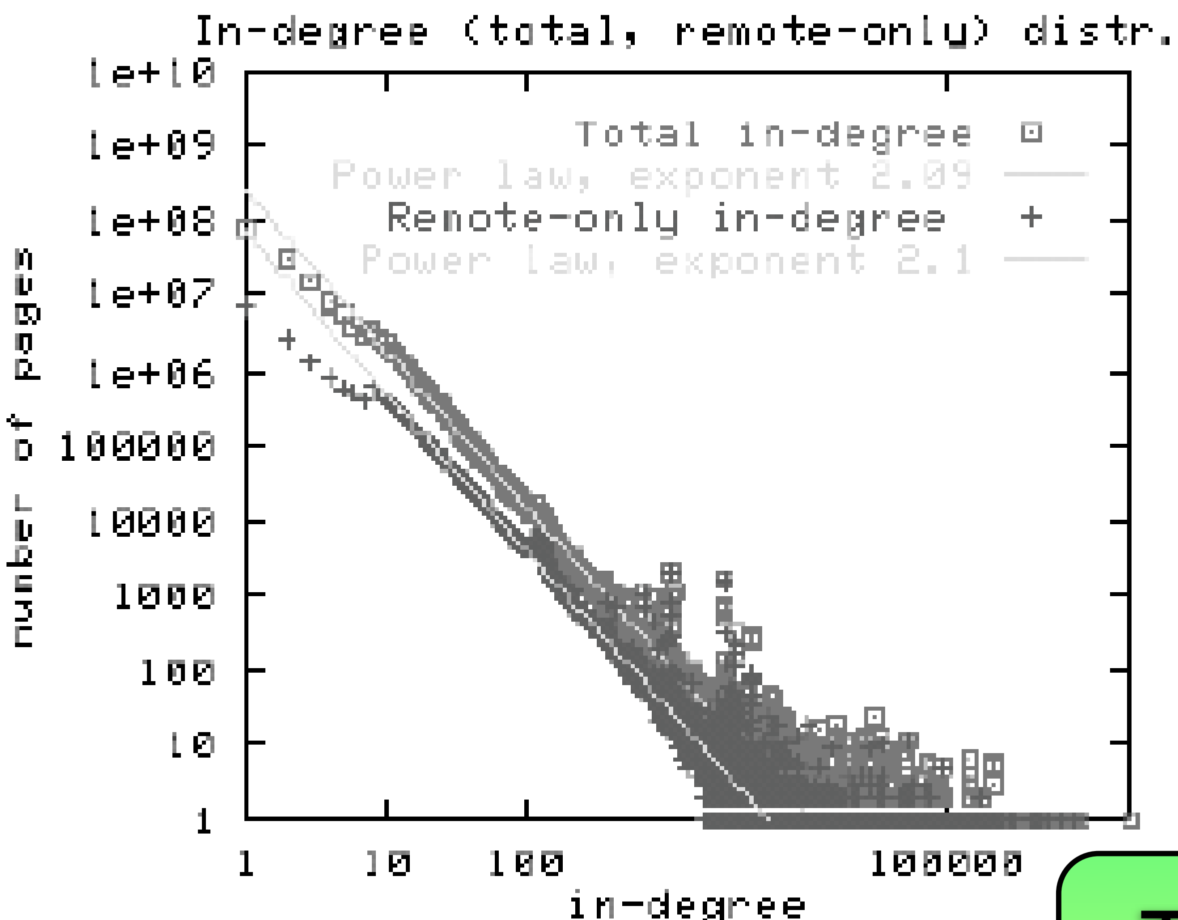
- The web is really infinite
  - Dynamic content, e.g. calendars, online organizers, etc.
  - http://www.raingod.com/raingod/resources/Programming/JavaScript/Software/RandomStrings/index.html
- Static web contains syntactic duplication, mostly due to mirroring (~ 20-30%)
- Some servers are seldom connected.

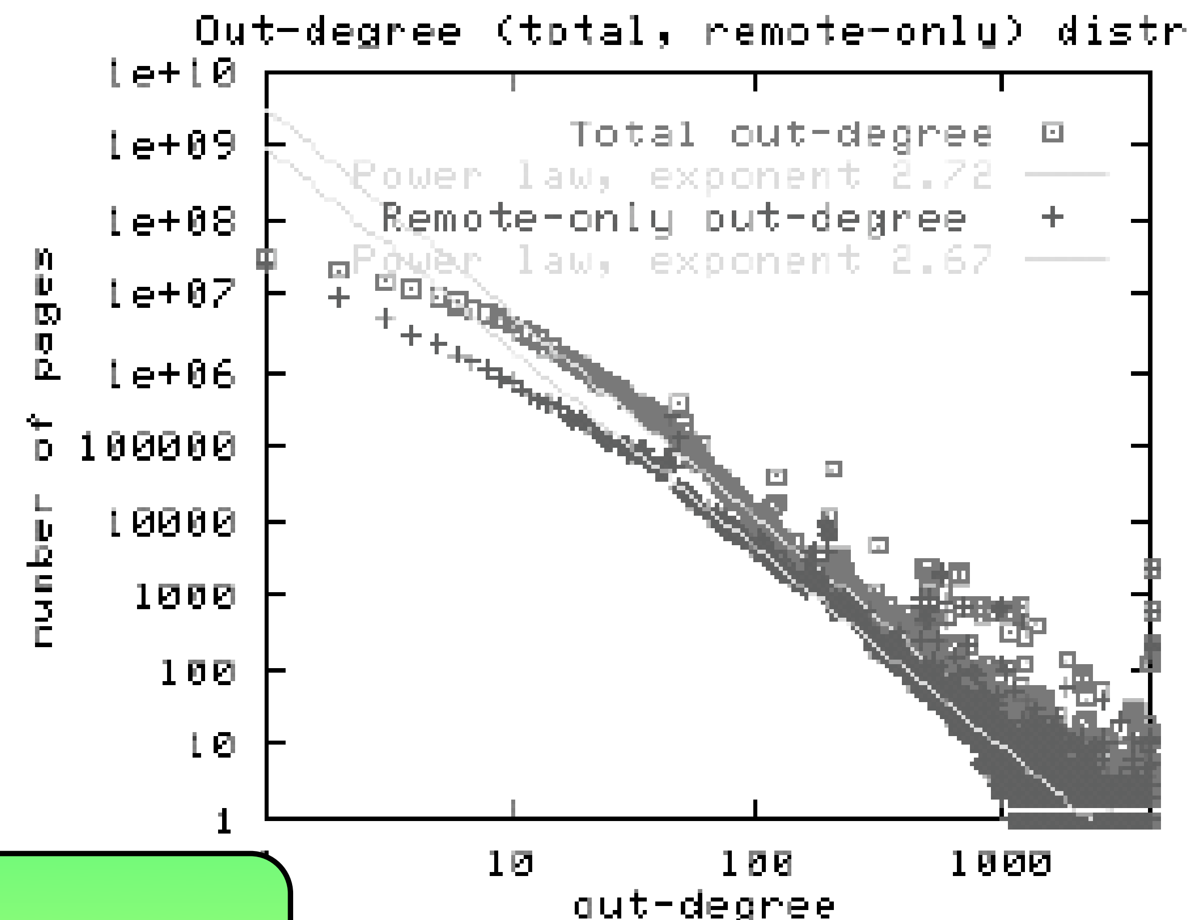
# Recent Measurement

- A. Gullì and A. Signorini. The Indexable Web is More than 11.5 Billion Pages. WWW2005.
- 2.3B the pages unknown to popular Search Engines.
- 35-120B of pages are within the hidden web.
- The index intersection between the largest available search engines - namely Google, Yahoo!, MSN, Ask/Teoma - is estimated to be

We'll dedicate a lesson on this at the end.

# Let's Characterize it Better




These are power-law distributions!

# Power-laws (an Informal Definition)

- Power law trends arise in many different natural contexts:
  - Telephone call networks.
  - Java program networks.
  - E-mail networks.
  - Scientific citations.
  - Protein-protein interactions in a cell.
  - http://wordcount.org/main.php (Zipf's law)
  - ...

# Power-laws (an Informal Definition)

- Sometimes called heavy-tail or long-tail distributions.
- In a power law network many nodes have degree equal to 1 and very few of them have higher degrees.

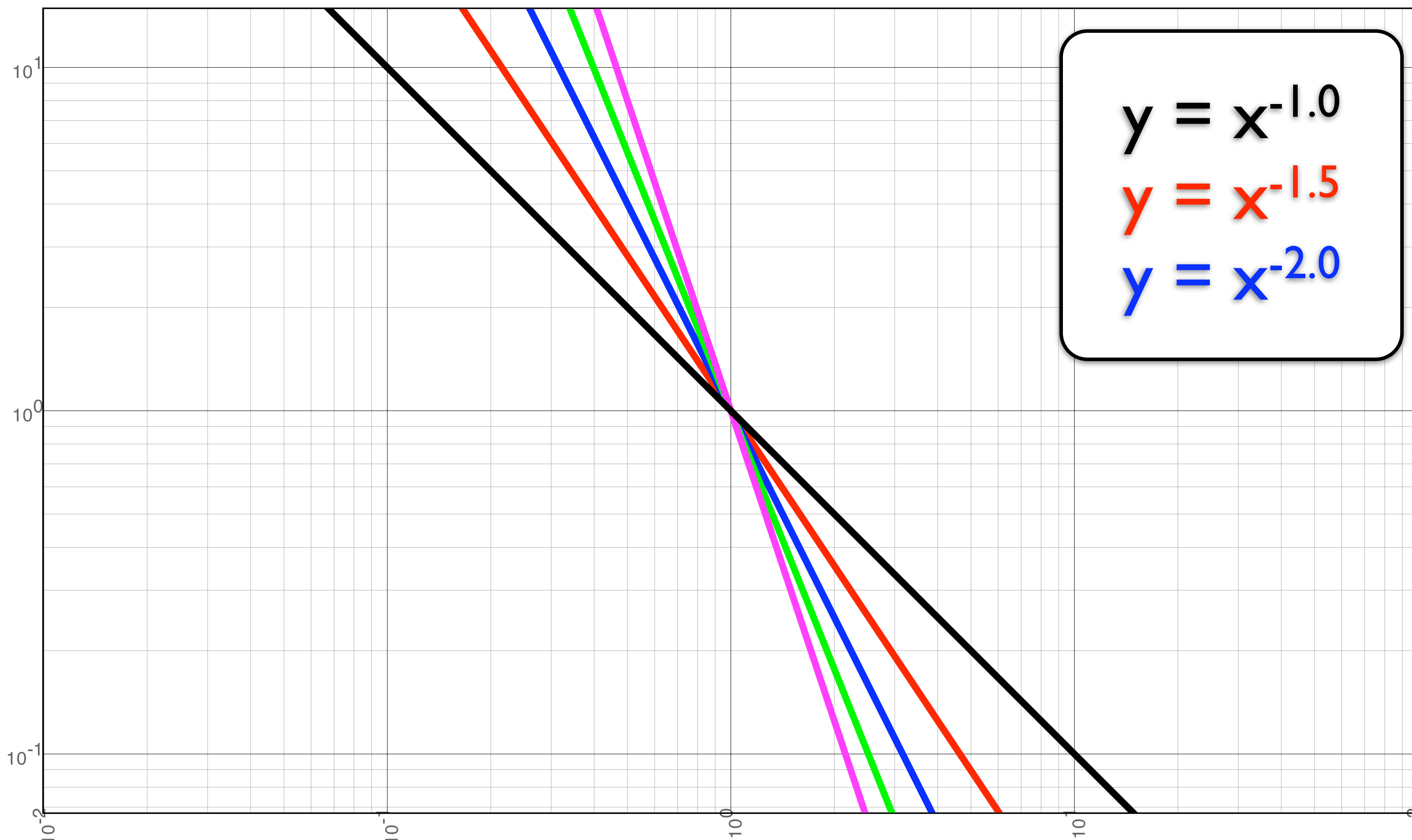
# Power-law

- Two discrete random variables x and y are related by a power-law when:
  - $y(x) = Kx^{-a}$
- where K and a are positive constants
- The constant a is often called the power law exponent.

# Power-law Distribution

- A discrete random variable is distributed according to a power-law when the probability density function (pdf) is given by:
  - $p(x)=Kx^{-a}$

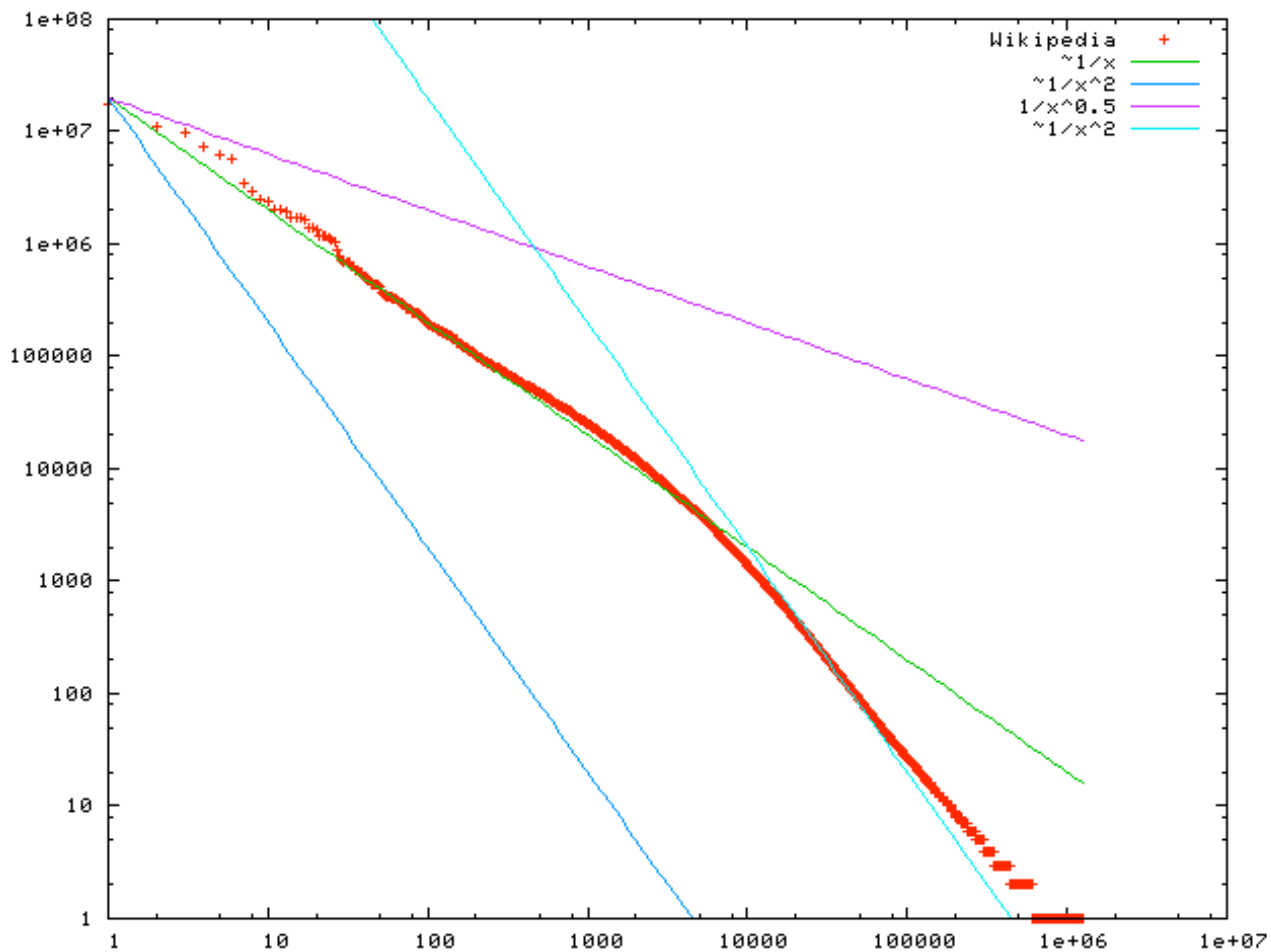
# Examples of Power-laws



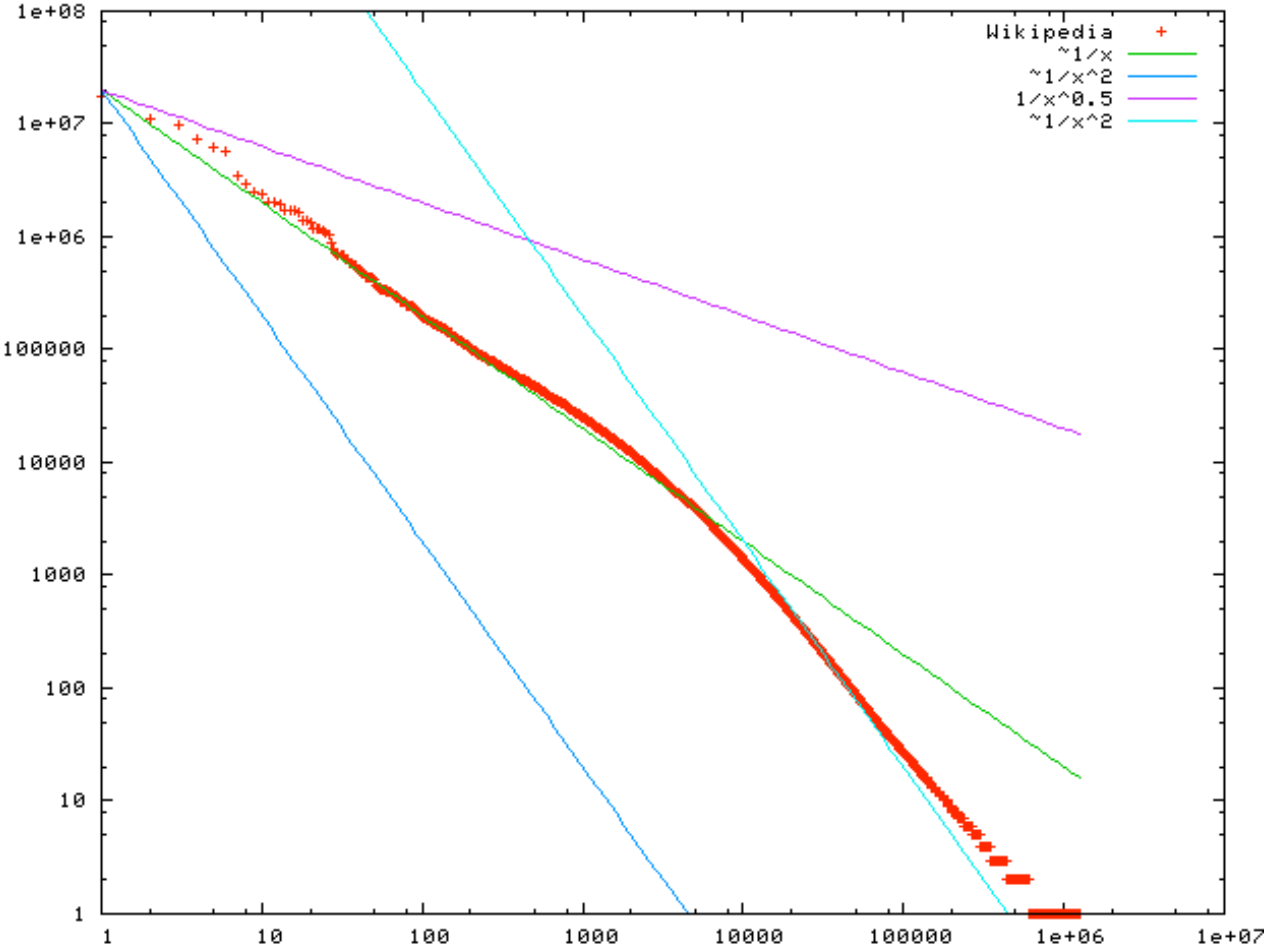
# Semantic of Power-law Distributions

- Roughly speaking a variable is distributed according to a power-law when there are few values having a very high probability of occurring, whereas the majority of the values occurs very rarely.
- For instance: words in english texts are distributed according a power-law of parameter a=1 (*Zipf's Law*)

# Wikipedia's Word Distribution



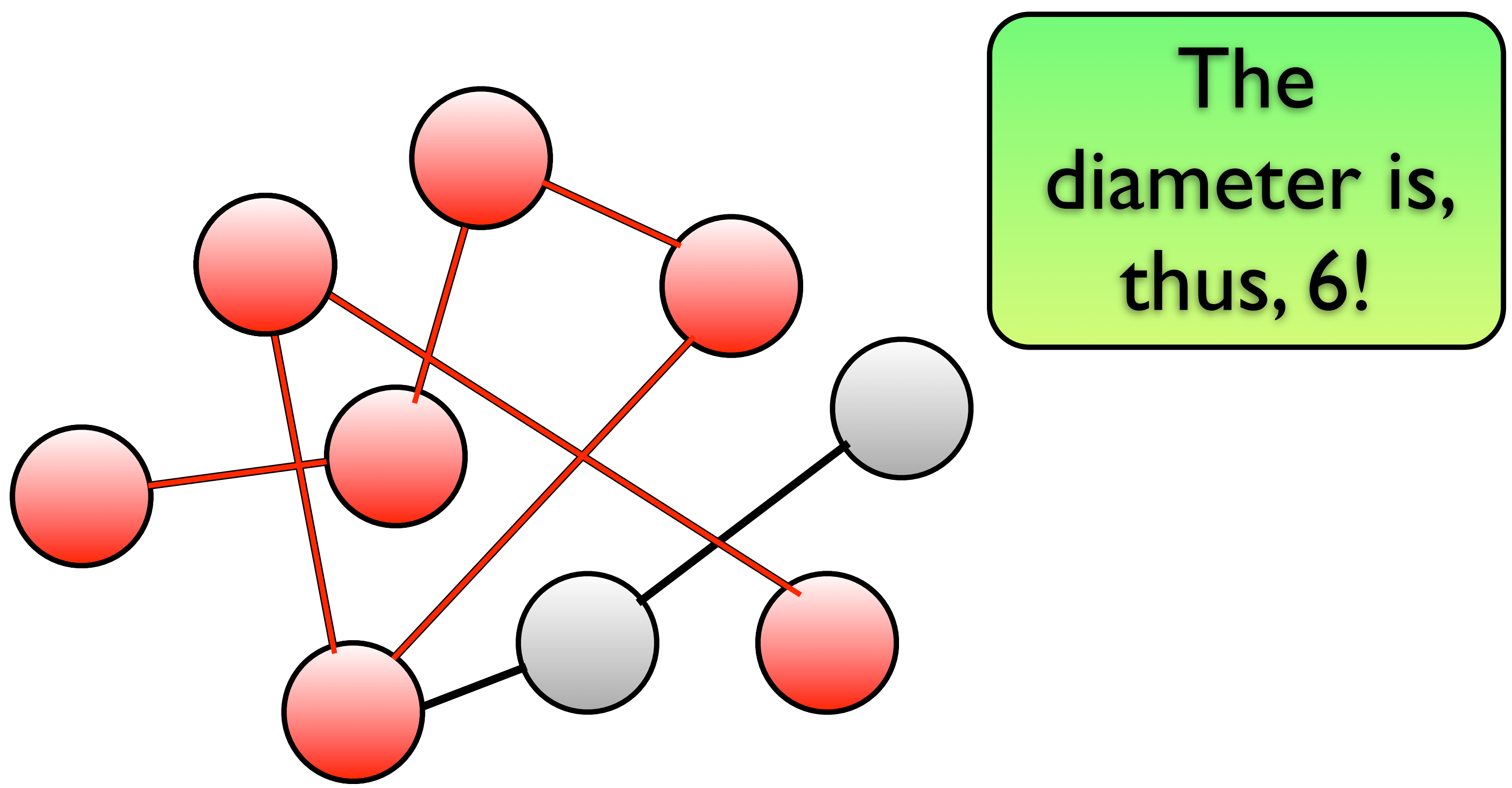

From http://en.wikipedia.org/wiki/Zipf's_law



Wikipedia
~1/x
~1/x^2
1/x^0.5
~1/x^2
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
1e+06
1e+07

# Diameter of a Graph

- Informally it is the "longest shortest path"



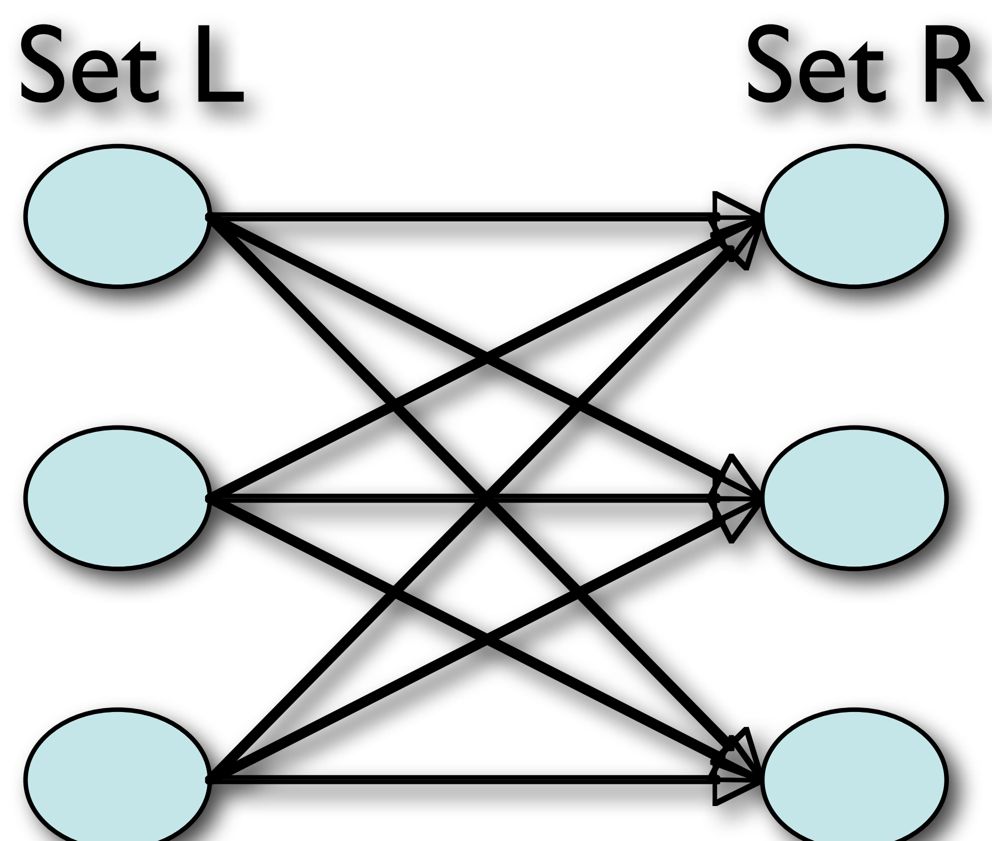
# Diameter of Webgraphs

- In Webgraphs the diameter should be "as small as possible"
- If N is the number of nodes of the graph, Webgraphs exhibit logarithmic diameters - i.e. O(log N)
- This property is also known as:
  - Scale-free: because doubling the nodes increase the diameter by only 1
  - Small World: because every two nodes are linked by very few vertexes

Typically diameter in a Webgraph is 19

# Bipartite Cores

- Informally a bipartite core in a graph consists of two sets of nodes L and R such that every node in L links to every node in R.



# Models of the Webgraph

- On-line property.
  - The number of nodes and edges changes with time.
- Power law degree distribution.
- Small world property.
- Many bipartite substructures.

# Random Graphs

- RGs are structures introduced by Paul Erdos and Alfred Reny.
- There are several models of RGs. We are concerned with the model $G_{n,p}$.
- A graph G = (V,E) $G_{n,p}$ is such that |V|=n and an edge (u,v) is selected uniformly at random with probability p.

# Why Webgraph Cannot be a Random Graph?

- Suppose $X_v$ is the degree of node v.
- Suppose $X_{v,w}$ be a r.v. equal to 1 if there is an edge joining v and w ($v \neq w$), 0 otherwise.

$$X_v = \sum_w X_{v,w}$$

$$E\left[X_v\right] = \sum_w E\left[X_{v,w}\right]$$

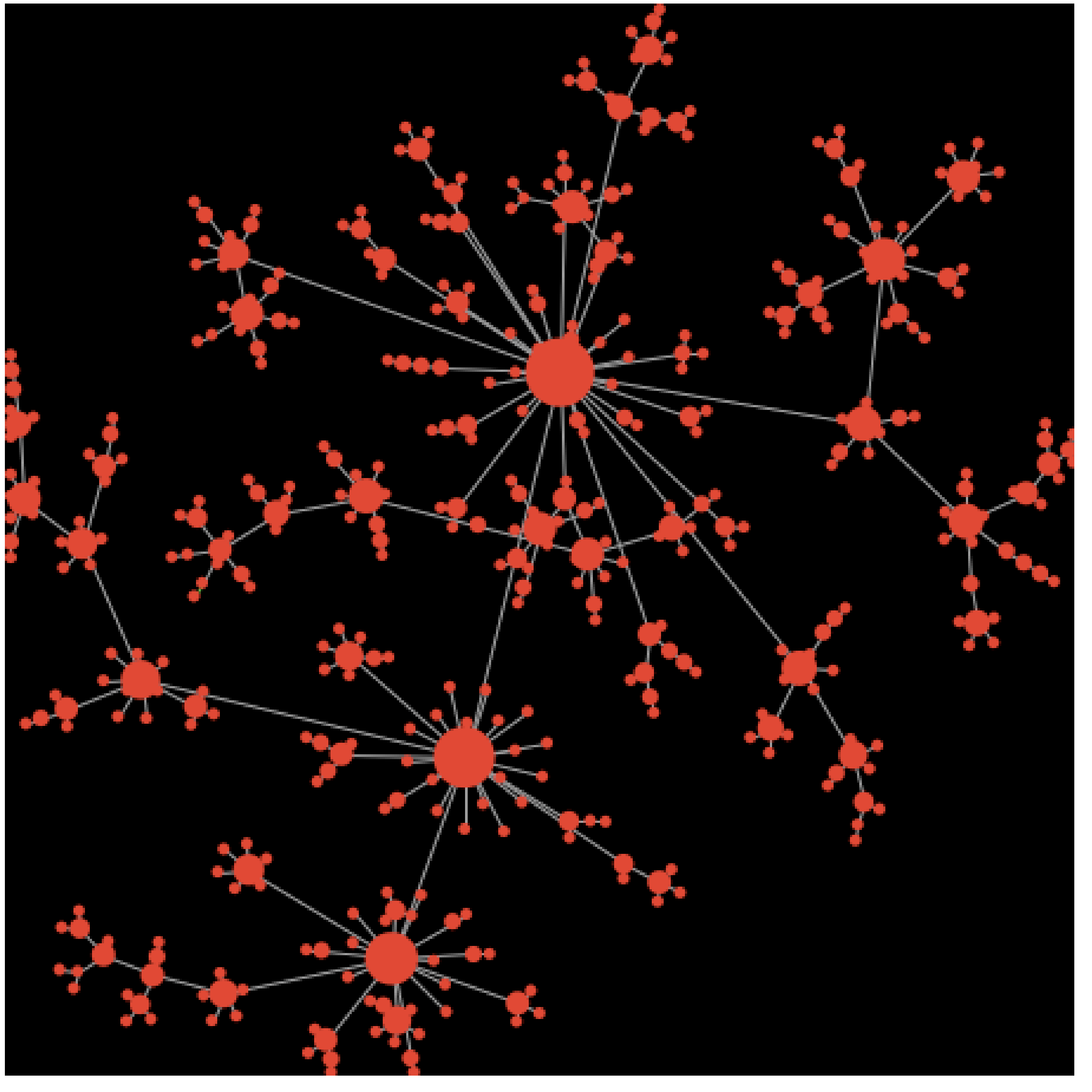
$$= \sum_w p = (n-1)\,p$$

- Thus $X_v$ is distributed as a Binomial(n-1,k) not a power-law.

# Preferential Attachment (PA)

- Parameter: m a positive integer
- At time 0, add a single edge
- At time t+1, add m edges from a new node $v_{t+1}$ to existing nodes
  - the edge $(v_{t+1}, v_s)$ is added with probability $\text{degree}(v_s)/2t$.

# An example



Generated with
http://ccl.northwestern.edu/netlogo/models/PreferentialAttachment

# PA in-degree

- Fix m a positive integer, fix an epsilon > 0. For k a non-negative integer, define

$$\alpha_{m,k} = \frac{2m(m+1)}{(k+m)(k+m+1)(k+m+2)}$$

Then with probability tending to 1 as t goes to infinity, for all k satisfying 0≤k≤t1/5

$$(1-\epsilon)\,\alpha_{m,k} \le p(k) \le (1+\epsilon)\,\alpha_{m,k}$$

# PA Diameter

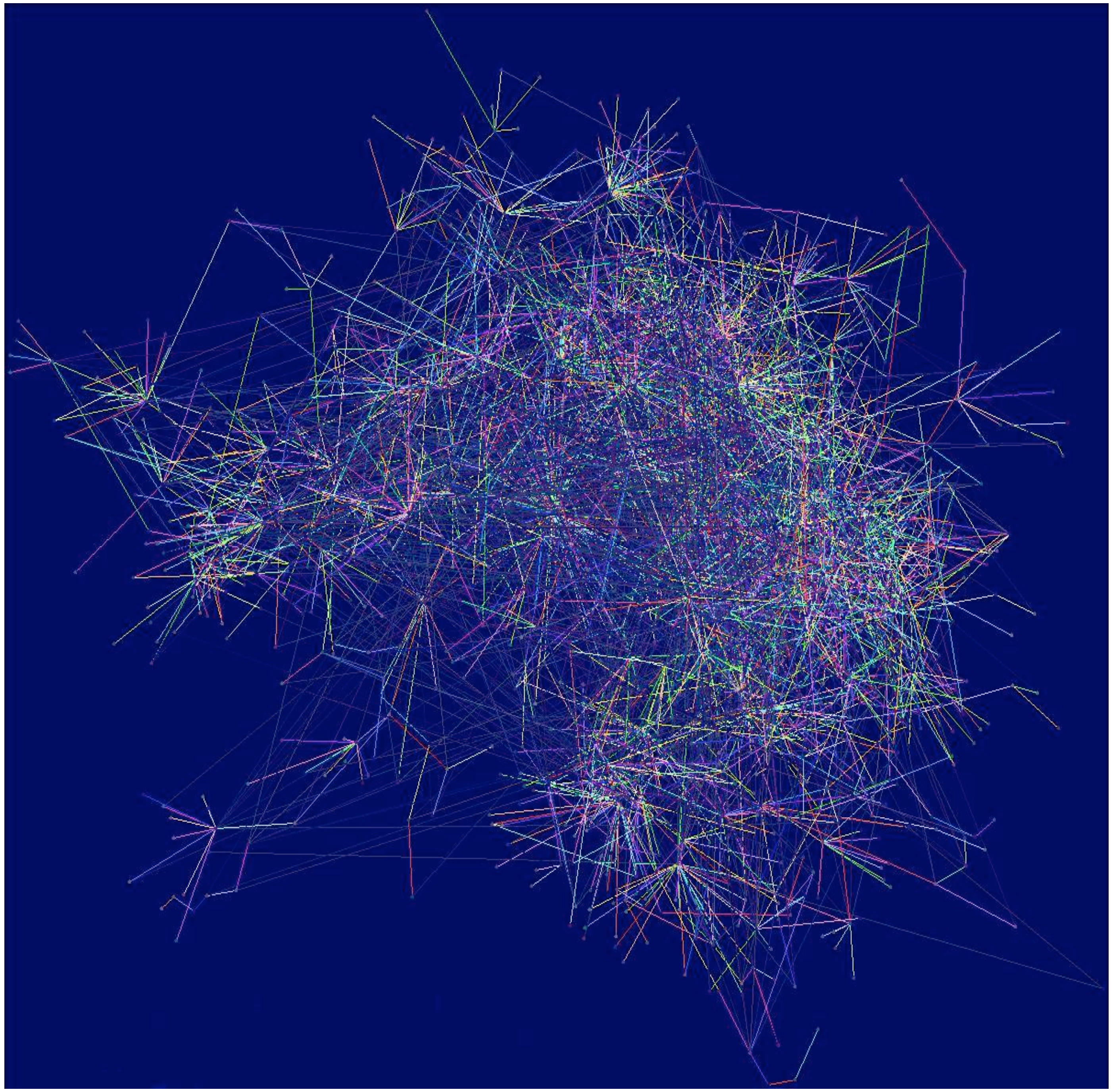
- Fix an integer $m \geq 2$ and a positive real number epsilon. With probability 1 as t goes to infinity, $G_m(t)$ is connected and

$$(1-\epsilon)\frac{\log t}{\log\log t} \leq \mathrm{diam}\left(G_m\left(t\right)\right) \leq (1+\epsilon)\frac{\log t}{\log\log t}$$

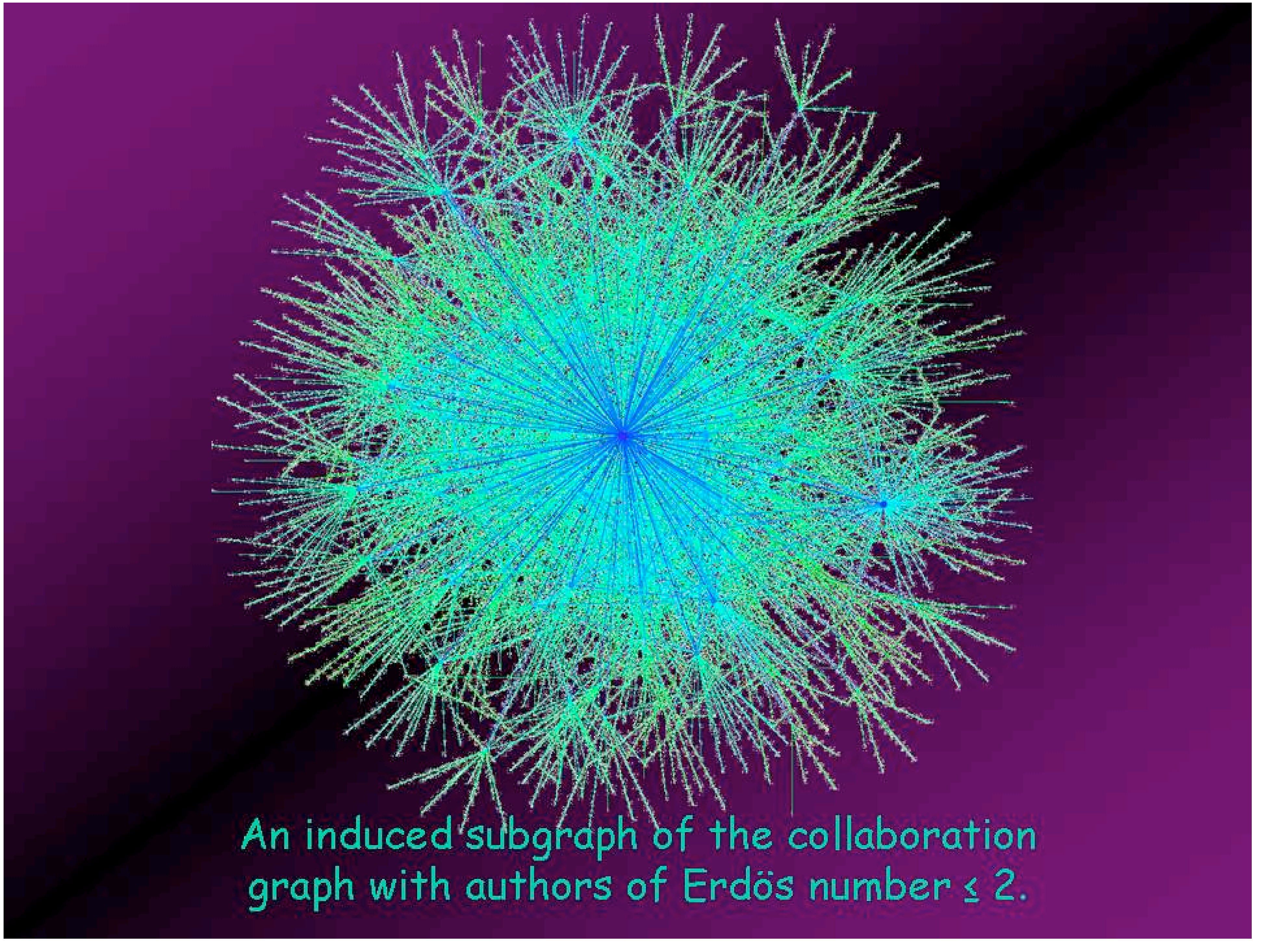
# Scale-Free Networks

- Network analysis is in its infancy
- Many different examples of networks exists.

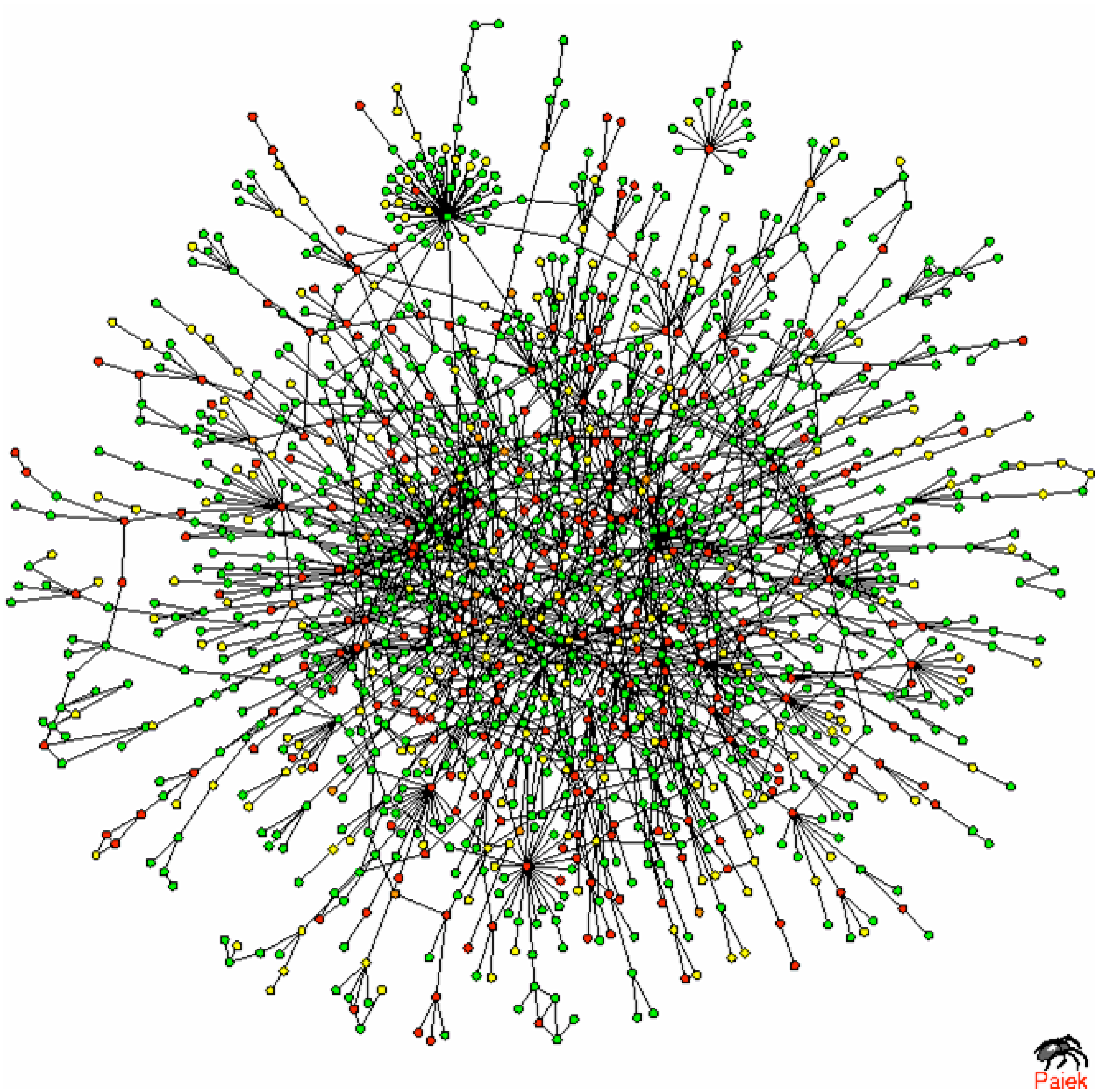
# co-authors Network



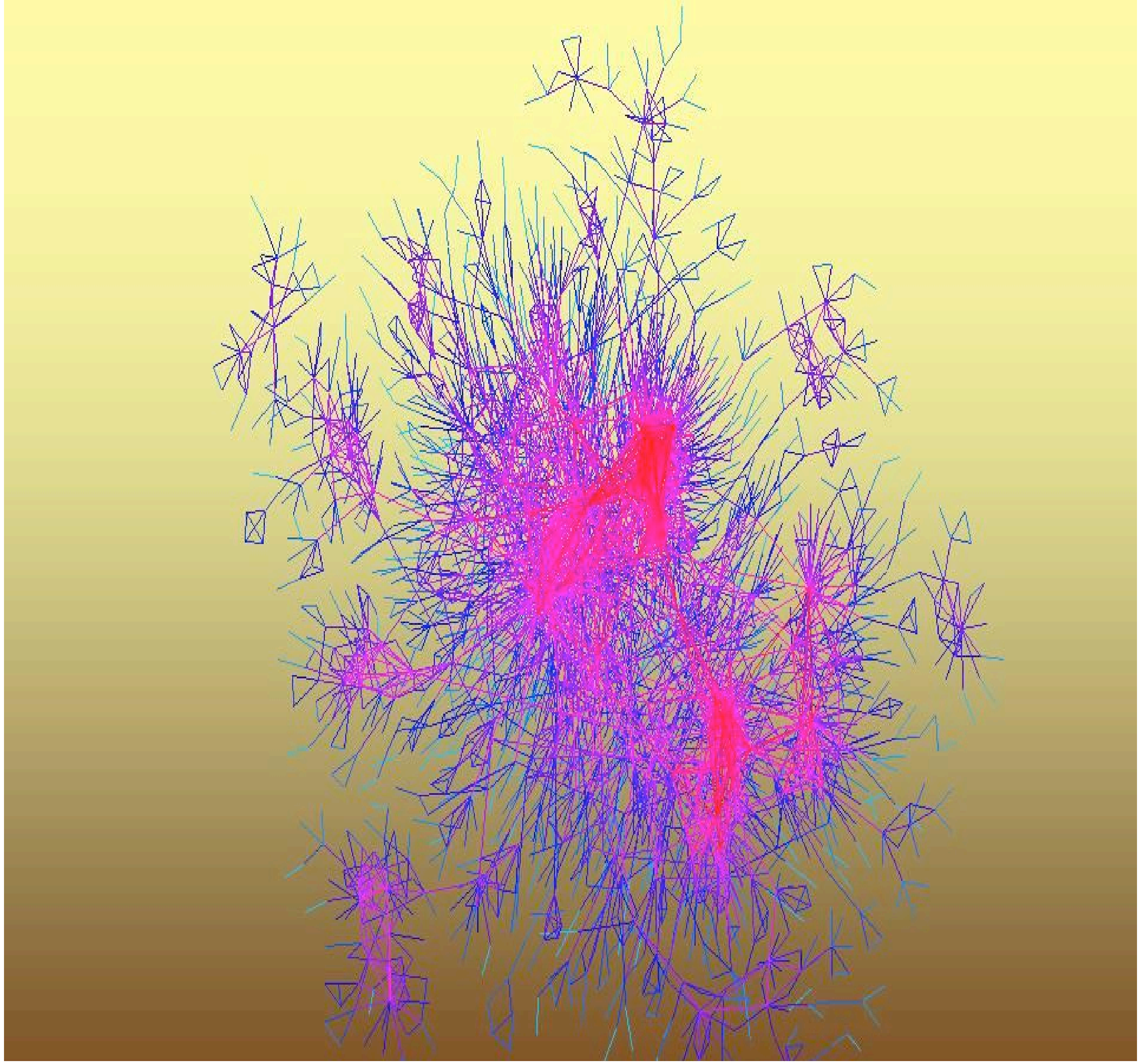
# Those with Erdos number ≤ 2



# Protein-Protein Interactions



# Hollywood Network



# The Lesson is Over