Power Laws and Preferential Attachment
Web Science (VU) (707.000)

Denis Helic
KTI, TU Graz
April 15, 2020
Outline

1. Popularity
2. A Simple Hypothesis
3. Log-normal Distributions
4. Power Laws
5. Rich-Get-Richer Models
6. Preferential Attachment
7. Multiplicative Random Processes
Popularity
Popularity

- Popularity is a phenomenon characterized by extreme imbalances
- Almost everyone is known only to people in their immediate social circles
- A few people achieve wider visibility
- A very few attain global name recognition
- Analogy with books, movies, scientific papers
- Everything that requires an audience
Popularity: questions

- How can we quantify imbalances?
Popularity: questions

- How can we quantify imbalances?
- Analyze distributions
- Why do these imbalances arise?
- What are the mechanisms and processes that cause them?
- Are they intrinsic (generalizable, universal) to popularity?
Web as an example

- To begin the analysis we take the Web as an example
- On the Web it is easy to measure popularity very accurately
- E.g. it is difficult to estimate the number of people worldwide who have heard of Bill Gates
- How can we achieve this on the Web?
Web as an example

- To begin the analysis we take the Web as an example
- On the Web it is easy to measure popularity very accurately
- E.g. it is difficult to estimate the number of people worldwide who have heard of Bill Gates
- How can we achieve this on the Web?
- Take a snapshot of the Web and count the number of *in-links* to Bill Gates homepage
- Calculate the authority score of Bill Gates homepage
- Calculate the PageRank of Bill Gates homepage
- We will learn how to calculate these quantities later in the course
The popularity question: a basic version

- As a function of $k$, what fraction of pages on the Web have $k$ in-links
- Larger values of $k$ indicate greater popularity
- Technically, what is the question about?
The popularity question: a basic version

- As a function of $k$, what fraction of pages on the Web have $k$ in-links
- Larger values of $k$ indicate greater popularity
- Technically, what is the question about?
- Distribution of the number of in-links (in-degree distribution) over a set of Web pages
- What is the interpretation of this question/answer?
The popularity question: a basic version

- As a function of $k$, what fraction of pages on the Web have $k$ in-links
- Larger values of $k$ indicate greater popularity
- Technically, what is the question about?
- Distribution of the number of in-links (in-degree distribution) over a set of Web pages
- What is the interpretation of this question/answer?
- Distribution of popularity over a set of Web pages
A Simple Hypothesis
A simple hypothesis

- Before trying to resolve the question
- What do we expect the answer to be?
- What distribution do we expect?
- What was the degree distribution in the random graph $G(n, p)$?
A simple hypothesis

- Before trying to resolve the question
- What do we expect the answer to be?
- What distribution do we expect?
- What was the degree distribution in the random graph $G(n, p)$?
- Binomial and approximation was Poisson
A simple hypothesis

\[ P(k) = \binom{n-1}{k} p^k (1 - p)^{n-1-k} \]

\[ P(k) = \frac{\lambda^k}{k!} e^{-\lambda} \]
Degree distribution (Binomial)

Degree distribution random graph ($n = 21$); differing $p$ values

- $p = 0.10$
- $p = 0.40$
A Simple Hypothesis

Degree distribution (Poisson)

Degree distribution random graph \((n = 21)\); differing \(\lambda\) values

\(\lambda = 2.0\)

\(\lambda = 8.0\)
A Simple Hypothesis

Degree distribution (Poisson approximation)

Degree distribution random graph ($n = 21$); Poisson approx. binomial

$p = 0.40$

$\lambda = 8.0$
A simple hypothesis

- From our experience how are some typical quantities distributed in our world?
- People’s height, weight, and strength
- In engineering and natural sciences
- Errors of measurement, position and velocities of particles in various physical processes, etc.
- Continuous approximation of Binomial and Poisson: **Normal Distribution**
A Simple Hypothesis

Normal (Gaussian) distribution

- It occurs so often in nature, engineering and society: Normal
- Characterized by a mean value $\mu$ and a standard deviation around the mean $\sigma$

**PDF**

$$f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**CDF**

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x'^2}{2}} dx'$$
Normal (Gaussian) distribution

PDF of a Normal random variable; differing $\mu$ and $\sigma$ values

- $\mu = 0.0, \sigma = 1.0$
- $\mu = -2.0, \sigma = 2.0$
Standard normal distribution

- If $\mu = 0$ and $\sigma = 1$ we talk about standard normal distribution

PDF

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- Please note, that you can always standardize a random variable $X$ with:

Standardizing

$$Z = \frac{X - \mu}{\sigma}$$
A Simple Hypothesis

Normal (Gaussian) distribution

- The basic fact: the density for a value that exceed mean by more than $c$ times the standard deviation decreases exponentially in $c$

\[
\begin{align*}
    r(1) &= \frac{f(1)}{f(0)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-1/2}}{\frac{1}{\sqrt{2\pi}}} \\
    &= \frac{1}{\sqrt{e}} \approx 0.6
\end{align*}
\]

\[
\begin{align*}
    r(c\sigma) &= r(c) = \frac{f(c)}{f(0)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-c^2/2}}{\frac{1}{\sqrt{2\pi}}} \\
    &= e^{-c^2/2} = O(e^{-c^2})
\end{align*}
\]
A Simple Hypothesis

Normal (Gaussian) distribution

Prob. of value more than $x$ times exceeding the mean: $N(0, 1)$

$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$P(|X| > x)$
Normal (Gaussian) distribution

- Why is normal distribution so ubiquitous
- Theoretical result: Central Limit Theorem provides an explanation
- Informally, we take any sequence of small independent and identically distributed (i.i.d) random quantities
- In the limit of infinitely long sequences their sum (or their average) are distributed normally
A Simple Hypothesis

Central Limit Theorem

**Theorem**

Suppose $X_1, \ldots, X_n$ are independent and identical r.v. with the expectation $\mu$ and variance $\sigma^2$. Let $S_n$ be the $n$-th partial sum of $X_i$:

$$S_n = \sum_{i=1}^{n} X_i.$$

Let $Z_n$ be a r.v. defined as (standardized $S_n$):

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

The CDF $F_n(z)$ tends to CDF of a standard normal r.v. for $n \to \infty$:

$$\lim_{n \to \infty} F_n(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}$$

Denis Helic (KTI, TU Graz)
Central limit theorem with unif. dist. and $Z_{30}$: $\mu = -0.007, \sigma^2 = 1.00474$
Central Limit Theorem

Central limit theorem with unif. dist. and $Z_{100}: \mu = 0.001, \sigma^2 = 0.99159$
Central Limit Theorem: Proof

- Now we present a proof sketch (to better understand the assumptions that CLT makes)
- For the proof we need some preliminaries

**Definition**

Characteristic function of a real valued r.v. $X$ is defined as expectation of the complex function $e^{itX}$:

$$\varphi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} f(x) \, dx,$$

where $t$ is the parameter and $f(x)$ is PDF of r.v. $X$.

- A characteristic function completely defines PDF of a r.v.
Central Limit Theorem: Proof

- To calculate characteristic function we typically apply Taylor expansion:

\[ e^{itX} = \sum_{n=0}^{\infty} \frac{(itx)^n}{n!} = 1 + itx - \frac{(tx)^2}{2} + O(t^3) \]
Central Limit Theorem: Proof

- Substituting the expansion into the integral:

\[
\varphi_X(t) = \int_{-\infty}^{\infty} f(x) \, dx + \int_{-\infty}^{\infty} itxf(x) \, dx - \int_{-\infty}^{\infty} \frac{(tx)^2}{2} f(x) \, dx + O(t^3)
\]

\[
= 1 + itE[X] - \frac{t^2}{2} E[X^2] + O(t^3)
\]

- Now suppose that we have a r.v. $X$ with 0 mean and variance 1 (which can be always achieved by standardizing a r.v. with finite mean and variance):

\[
\varphi_X(t) = 1 - \frac{t^2}{2} + O(t^3)
\]
Another important fact of the characteristic functions
Suppose $X$ and $Y$ are two independent r.v.
We want to calculate the characteristic function of r.v. $Z = X + Y$:

$$\varphi_{X+Y}(t) = E[e^{it(X+Y)}] = E[e^{itX}e^{itY}] = E[e^{itX}]E[e^{itY}]$$

$$= \varphi_X(t)\varphi_Y(t)$$

Last equality in the first row follows from the independence
The last fact that we need: if $Z \sim N(0, 1)$ then $\varphi_Z(t) = e^{-t^2/2}$
Central Limit Theorem: Proof

- Suppose now we have a set of random variables with individual 
  \( X_i \sim (\mu, \sigma^2) \) which are all independent and identically distributed (i.i.d.)

- Note that we do not make assumptions on the distribution of \( X_i \) just that they have finite \( \mu \) and \( \sigma^2 \)

- We build a new r.v. \( S_n = \sum_{i=1}^{n} X_i \) as the \( n \)-th partial sum

\[
E[S_n] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n \mu
\]

\[
Var(S_n) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma^2 = n \sigma^2
\]
Central Limit Theorem: Proof

- Now we standardize $S_n$ to obtain $Z_n$:

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma}} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n\sigma}}$$

- By introducing $Y_i = \frac{X_i - \mu}{\sigma}$ (please note that $Y_i$ is standardization of $X_i$, i.e. $Y_i \sim (0, 1)$):

$$Z_n = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}$$
Central Limit Theorem: Proof

- Now let us calculate $\varphi_{Z_n}(t)$ (where we use the fact that characteristic function of the sum equals to the product of characteristic functions if r.v. are independent and we scale the parameter $t$ with $1/\sqrt{n}$):

$$
\varphi_{Z_n} = \prod_{i=1}^{n} \varphi_{Y_i}(t/\sqrt{n}) = [\varphi_Y(t/\sqrt{n})]^n
$$

$$
= [1 - \frac{t^2}{2n} + O((t/\sqrt{n})^3)]^n
$$
Central Limit Theorem: Proof

- Now we are interested what happens when \( n \to \infty \)
- Obviously \( O((t/\sqrt{n})^3) \to 0 \)
- Thus, we have:

\[
\lim_{n \to \infty} \varphi_{Z_n} = \lim_{n \to \infty} \left[ 1 - \frac{t^2}{2n} \right]^n = e^{-t^2/2}
\]

- We obtain the characteristic function of standard normal and thus

\[
\lim_{n \to \infty} Z_n \sim N(0, 1)
\]
Central Limit Theorem

- How can we interpret this result?

Any quantity that can be viewed as a sum of many small independent random effects will have a normal distribution. For example, we take a lot of measurements of a fixed physical quantity. Variations in the measurements across trials are cumulative results of many independent sources of errors, such as errors in the equipment, human errors, and changes in external factors. Therefore, the distribution of measured values is normally distributed.
Central Limit Theorem

- How can we interpret this result?
- Any quantity that can be viewed as a sum of many small independent random effects will have a normal distribution
- E.g. we take a lot of measurements of a fixed physical quantity
- Variations in the measurements across trials are cumulative results of many independent sources of errors
- E.g. errors in the equipment, human errors, changes in external factors
- Then the distribution of measured values is normally distributed
Central Limit Theorem

- Can you explain why examination grades tend to be normally distributed?
Can you explain why examination grades tend to be normally distributed?

- Each student is a small “random factor”
- The points for each question are a random variable, which are i.i.d
- Then the sum (average) of the points will be according to CLT normally distributed
- If the distribution of exam grades for a course is not normal what can be going on?
Can you explain why examination grades tend to be normally distributed?

Each student is a small “random factor”

The points for each question are a random variable, which are i.i.d

Then the sum (average) of the points will be according to CLT normally distributed

If the distribution of exam grades for a course is not normal what can be going on?

Too strict, too loose, discrimination, independence is broken, not identically distributed, etc.
A Simple Hypothesis

How to apply this on the Web?

- If we model the link structure by assuming that each page decides \textit{independently} at random to which page to link to
- Then the number of in-links for any given page is the sum of many i.i.d quantities
- Hence, we expect it to be normally distributed
- If we believe that this model is correct:
- Then the number of pages with $k$ in-links should decrease exponentially in $k$ as $k$ grows
Log-Normal Distribution
Log-Normal Distribution

- If $X$ is log-normally distributed $\iff Y = \ln(X)$ is normally distributed
- If $Y$ is normally distributed $\iff X = e^Y$ is log-normally distributed
- Characterized by a mean value $\mu$ and a standard deviation around the mean $\sigma$

PDF

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$
Log-Normal Distribution

PDF of a Log-Normal random variable; differing $\mu$ and $\sigma$ values

$\mu = 0.00, \sigma = 0.25$

$\mu = 0.00, \sigma = 1.00$
Multiplicative random processes lead to log-normal distributions

Suppose we have a set of random variables with individual $X_i \sim (\mu, \sigma^2)$ which are all independent and identically distributed (i.i.d.)

Note that we do not make assumptions on the distribution of $X_i$ just that they have finite $\mu$ and $\sigma^2$

We build a new r.v. $P_n = \prod_{i=1}^{n} X_i$ as the $n$-th partial product

We claim that $\lim_{n \to \infty} P_n$ is log-normally distributed
Multiplicative random processes

\[ P_n = \prod_{i=1}^{n} X_i \]

\[ \ln(P_n) = \sum_{i=1}^{n} \ln(X_i) \]

- From the CLT we now that \( \ln(P_n) \) tends to standard normal
- Thus, \( P_n \) tends to log-normal distribution
Power Laws
When people measured the distribution of links on the Web they found something very different to Normal distribution.

In all studies over many different Web snapshots:

- The fraction of Web pages that have $k$ in-links is approximately proportional to $1/k^2$.
- More precisely the exponent on $k$ is slightly larger than 2.
What is the difference to the normal distribution?

- $1/k^2$ decreases much more slowly as $k$ increases.
- Pages with large number of in-links are much more common than we would expect with a normal distribution.
- E.g. $1/k^2$ for $k = 1000$ is one in million.
- One page in million will have 1000 in-links.
- For a function like $e^{-k}$ or $2^{-k}$ this is unimaginably small.
- No page will have 1000 in-links.
A function that decreases as $k$ to some fixed power $1/k^c$, e.g. $1/k^2$ is called **power law**

- The basic property: it is possible to see very large values of $k$
- This is a quantitative explanation of popularity imbalance
- It accords to our intuition for the Web: there is a reasonable large number of extremely popular Web pages
- We observe similar power laws in many other domains
- The fraction of books that are bought by $k$ people: $1/k^3$
- The fraction of scientific papers that receive $k$ citations: $1/k^3$, etc.
The normal distribution is widespread in natural sciences and engineering.

Power laws seem to dominate whenever popularity is involved, i.e. (informally) in social sciences and/or e.g. psychology.

Conclusion: if you analyze the user data of any kind

E.g. the number of downloads, the number of emails, the number of tweets.

**Expect to see a power law**

Test for power law: histogram + test if $1/k^c$ for some $c$

If yes estimate $c$
Power Law Histogram

PMF of a power law r. v.; differing $c$ values

$k$

$p(k)$

$c = 2.0$

$c = 3.0$
A simple visual method

Let $f(k)$ be the fraction of items that have value $k$

We want to know if $f(k) = a/k^c$ approximately holds for some exponent $c$ and some proportion constant $a$

Let us take the logarithms of both sides

$$\ln(f(k)) = \ln(a) - c \cdot \ln(k)$$
Power Law Log-Log Plot

PMF of a power law r. v.; differing $c$ values

- $c = 2.0$
- $c = 3.0$
Power Law check: a simple method

- If we plot $f(k)$ on a log-log scale we expect to see a straight line
- $-c$ is the slope and $\ln(a)$ will be the $y$-intercept
- This is only a simple check to see if there is an apparent power law behavior
- **Do not use this method to estimate the parameters!**
- There are statistically sound methods to that
- We discuss them in some other courses e.g. Network Science
Power Law check: a simple method

Figure: From Broder et al. (Graph Structure in the Web)
Why Power Law?

- We need a simple explanation for what causes Power Laws?
- Central Limit Theorem gives us a basic reason to expect the normal distribution
- Technically, we also need to find out why CLT does not apply in this case
- Which of its assumptions are broken?
- Sum of independent random effects
- What is broken?
Why Power Law?

- We need a simple explanation for what causes Power Laws?
- Central Limit Theorem gives us a basic reason to expect the normal distribution
- Technically, we also need to find out why CLT does not apply in this case
- Which of its assumptions are broken?
- Sum of independent random effects
- What is broken?
- Independence assumption
Power Laws arise from the feedback introduced by **correlated decisions** across a population.

- In networks person’s decisions depend on the choices of other people.
- E.g. peer influence/pressure.
- E.g. success, activity, but also examples of bad influence.
Why Power Law?

- In an information network you are exposed to the information by the others, not necessarily only peers.
- E.g. reply, retweet, post, etc.
- An assumption: people tend to copy the decisions of people who act before them.
- E.g. people tend to copy their friends when they buy books, go to movies, etc.
Why Power Law?

Many different possibilities to generate power laws such as:

1. Rich-get-richer models, aka preferential attachment, aka correlated models
2. Multiplicative random processes
Rich-Get-Richer Models
Simple copying model

- Creation of links among Web pages
  1. Pages are created in order and named 1, 2, 3, ..., \( N \)
  2. When page \( j \) is created it produces a link to an earlier Web page \( (i < j) \) with \( p \) being a number between 0 and 1:
     a. With probability \( p \), page \( j \) chooses a page \( i \) uniformly at random and links to \( i \)
     b. With probability \( 1 - p \), page \( j \) chooses a page \( i \) uniformly at random and creates a link to the page that \( i \) points to
     c. The step number 2 may be repeated multiple times to create multiple links
Simple copying model

- Part 2(b) is the key
- After finding a random page \( i \) in the population the author of page \( j \) does not link to \( i \)
- Instead the author copies the decision made by the author of \( i \)
- The main result about this model is that if you run it for many pages
- The fraction of pages with \( k \) in-links will be distributed approximately as a \( 1/k^c \)
- The exponent \( c \) depends on the choice of \( p \)
- Intuition: if \( p \) gets smaller what do you expect
Simple copying model

- Part 2(b) is the key
- After finding a random page \( i \) in the population the author of page \( j \) does not link to \( i \)
- Instead the author copies the decision made by the author of \( i \)
- The main result about this model is that if you run it for many pages
- The fraction of pages with \( k \) in-links will be distributed approximately as a \( 1/k^c \)
- The exponent \( c \) depends on the choice of \( p \)
- Intuition: if \( p \) gets smaller what do you expect
- More copying makes seeing extremely popular pages more likely
Rich-get-richer dynamics

- The copying mechanism in 2(b) is an implementation of the following “rich-get-richer” mechanism.
- When you copy the decision of a random earlier page, what is the probability of linking to a page $\ell$?
Rich-get-richer dynamics

- The copying mechanism in 2(b) is an implementation of the following “rich-get-richer” mechanism
- When you copy the decision of a random earlier page what is the probability of linking to a page $\ell$
- It is proportional to the total number of pages that currently link to $\ell$

(a) ...
(b) With probability $1 - p$, page $j$ chooses a page $\ell$ with probability proportional to $\ell$’s current number of in-links and links to $\ell$
(c) ...
Preferential Attachment
Preferential attachment

- Why do we call this “rich-get-richer” rule?
- The probability that page $\ell$ increases its popularity is directly proportional to $\ell$’s current popularity
- This phenomenon is also known as preferential attachment
- E.g. the more well known someone is, the more likely you are to hear their name in conversations
- A page that gets a small lead over others tends to extend that lead
- On contrary, the idea behind CLT is that small *independent* random values tend to cancel each other out
Arguments for simple models

- The goal of simple models is not to capture all the reasons why people create links on the Web
- The goal is to show that a simple principle leads directly to observable properties, e.g. Power Laws
- Thus, they are not as surprising as they might first appear
- “Rich-get-richer” models suggest also a basis for Power Laws in other areas as well
- E.g. the populations of cities
Analytic handling of simple models

- Simple models can be sometimes handled analytically
- This allows also for *prediction* of how networks may evolve
- We can also easily cover extensions of the model
- Predict consequences of these extensions
Simple “rich-get-richer” model

- Creation of links among Web pages
  1. Pages are created in order and named $1, 2, 3, \ldots, N$
  2. When page $j$ is created it produces a link to an earlier Web page ($i < j$) with $p$ being a number between 0 and 1:
     a. With probability $p$, page $j$ chooses a page $i$ uniformly at random and links to $i$
     b. With probability $1 - p$, page $j$ chooses a page $\ell$ with probability proportional to $\ell$’s current number of in-links and links to $\ell$
     c. The step number 2 may be repeated multiple times to create multiple links
Analysis of the simple “rich-get-richer” model

- We have specified a randomized process that runs for $N$ steps
- We want to determine the *expected* number of pages with $k$ in-links at the end of the process
- In other words, we want to analyze the distribution of the in-degree
- Many possibilities to approach this
- We will make a continuous approximation to be able to use introductory calculus
Properties of the original model

- The number of in-links to a node $j$ at time $t \geq j$ is a random variable $X_j(t)$.

- Two facts that we know about $X_j(t)$:
  1. The initial condition: node $j$ starts with no in-links when it is created, i.e. $X_j(j) = 0$.
  2. The expected change to $X_j(t)$ over time, i.e. probability that node $j$ gains an in-link at time $t + 1$:
     - With probability $p$ the new node links to a random node – probability to choose $j$ is $1/t$, i.e. altogether $p/t$.
     - With probability $1 - p$ the new node links proportionally to the current number of in-links – probability to choose $j$ is $X_j(t)/t$, i.e. altogether $(1 - p)X_j(t)/t$.
  3. The overall probability that node $t + 1$ links to $j$: $\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$.
Approximation

- We have now an equation which tells us how the expected number of in-links evolves in *discrete* time.
- We will approximate this function by a *continuous* function of time $x_j(t)$ (to be able to use calculus).
- The two properties of $X_j(t)$ now translate into:
  1. The initial condition: $x_j(j) = 0$ since $X_j(j) = 0$.
  2. The expected gain in the number of in-links now becomes the *growth equation* (which is a differential equation):

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1 - p)x_j}{t}$$

- Now by solving the differential equation we can explore the consequences.
Solution

- For notational simplicity, let $q = 1 - p$
- The differential equation becomes:

$$\frac{dx_j}{dt} = \frac{p + qx_j}{t}$$

- Separate variables ($x$ on the left side, $t$ on the right side):

$$\frac{dx_j}{p + qx_j} = \frac{dt}{t}$$
Solution

- Integrate both sides:

\[
\int \frac{dx_j}{p + qx_j} = \int \frac{dt}{t}
\]

- We obtain:

\[
\ln(p + qx_j) = q\ln(t) + c
\]
Solution

- Exponentiating both sides (and writing $C = e^c$):

$$p + qx_j = C t^q$$

- Rearranging:

$$x_j(t) = \frac{1}{q} (C t^q - p)$$
Solution

- We can determine $C$ from the initial condition ($x_j(j) = 0$):

  \[
  0 = \frac{1}{q} (C j^q - p) \\
  C = \frac{p}{j^q}
  \]

- Final solution:

  \[
  x_j(t) = \frac{1}{q} \left( \frac{p}{j^q t^q} - p \right) = \frac{p}{q} \left[ \left( \frac{t}{j} \right)^q - 1 \right]
  \]
Identifying a power law

- Now we know how $x_j$ evolves in time
- We want to answer question: for a given value of $k$ and a time $t$ what fraction of nodes have at least $k$ in-links at time $t$
- In other words what fraction of functions $x_j(t)$ satisfies: $x_j(t) \geq k$

$$x_j(t) = \frac{p}{q} \left[ \left( \frac{t}{j} \right)^q - 1 \right] \geq k$$

- Rewriting in terms of $j$:

$$j \leq t \left[ \frac{q}{p} k + 1 \right]^{-1/q}$$
The fraction of values $j$ that satisfy the condition is simply:

$$\frac{1}{t} \left[ \frac{q}{p} k + 1 \right]^{-1/q} = \left[ \frac{q}{p} k + 1 \right]^{-1/q}$$

This is the fraction of nodes that have at least $k$ in-links.

In probability this is complementary cumulative distribution function (CCDF) $F(k)$.

The probability density $f(k)$ (the fraction of nodes that has exactly $k$ in-links) is then $f(k) = -\frac{dF(k)}{dk}$.
Identifying a power law

Differentiating:

\[ f(k) = -\frac{dF(k)}{dk} = \frac{1}{q} \frac{q}{pp} \left[ \frac{q}{p}k + 1 \right]^{-1-1/q} = \frac{1}{p} \left[ \frac{q}{p}k + 1 \right]^{-1-1/q} \]

The fraction of nodes with \( k \) in-links is proportional to \( k^{-(1+1/q)} \)

It is a power law with exponent:

\[ 1 + \frac{1}{q} = 1 + \frac{1}{1 - p} \]
Discussion of the results

- What happens with the exponent when we vary $p$
  - When $p$ is close to 1 the links creation is mainly random
  - The power law exponent tends to infinity and nodes with large number of in-links are increasingly rare
  - When $p$ is close to 0 the growth of the network is strongly governed by “rich-get-richer” behavior
  - The exponent decreases towards 2 allowing for many nodes with large number of in-links
  - 2 is natural limit for the exponent and this fits very well in what has been observed on the Web (exponents are slightly over 2)
- Simple model but extensions are possible
Multiplicative Random Processes
Multiplicative random processes

- Multiplicative random processes lead to log-normal distributions.
- With a small modification of the process we can also obtain power law distributions.
- Suppose we have a set of random variables with individual $X_i \sim (\mu, \sigma^2)$ which are all independent and identically distributed (i.i.d.)
- Note that we do not make assumptions on the distribution of $X_i$ just that they have finite $\mu$ and $\sigma^2$.
- We build a new r.v. $P_n = \prod_{i=1}^{n} X_i$ as the $n$-th partial product.
- We also introduce a threshold that defines a minimal value for the product.
- If the product falls below the threshold we reset it to the threshold.
- This results in a power law distribution.
Summary

We have learned about:

- Popularity as a network phenomenon
- CLT and sums of independent random quantities
- Power Laws
- “Rich-get-richer” and preferential attachment
- Multiplicative random processes
Some Practical Examples

- The long tail in the media industry
- Selling “blockbusters” vs. selling “niche products”
- Various strategies in recommender systems
- E.g. recommend “niche products” to make money from the long tail
- We can either reduce or amplify “rich-get-richer” effects
Thanks for your attention - Questions?

Slides use figures from Chapter 18, Crowds and Markets by Easley and Kleinberg (2010)