Social and Information Networks

University of Toronto CSC303
Winter/Spring 2024

Week 6: Feb 12-16

Mon. Feb 12th: Announcements & Corrections

- Mid-course evaluation now available on Quercus
  - Anonymous, 5-10 minutes, open until Feb 23
  - I really appreciate knowing how things are going, and how to do better :)

- Assignment 1 is due this Thursday :(
  - ... but it's reading week after that :)

This week's high-level learning goals

- Define the power law distribution and Zipf's law, and explain the similarities and differences
    - Recall the dynamics that often give rise to power laws, give examples, and debate whether they apply to a given scenario
- Define the Kumar et al. rich-get-richer model
    - Explain the connection between the rich-get-richer model and dynamics that give rise to power laws
    - Recall the expected distribution, and explain the relevant parameters of the model
- Summarize the Salganik et al. music popularity experiment
- Explain the problem of ranking web results
    - Explain the hubs and authorities algorithm, and execute on examples
    - Explain the (scaled) Page rank algorithm, and execute on examples
    - Describe high-level proof-sketches of their convergence

Roadmap: where we have been and whats next

Chapter 20 started off with a discussion of the small worlds phenomena and an insightful understanding of how decentralized search can work.

Previously, we were led to the observation that geographical distance (or social distance) correlates with friendship such that Prob[v is a friend of u] $\approx [rank_u(v)]^{-1}$.

Furthermore, there is a sense that long distance friendships are "rare".

Roadmap: where we have been and whats next

Chapter 20 started off with a discussion of the small worlds phenomena and an insightful understanding of how decentralized search can work.

Previously, we were led to the observation that geographical distance (or social distance) correlates with friendship such that Prob[v is a friend of u] $\approx [rank_u(v)]^{-1}$.

Furthermore, there is a sense that long distance friendships are "rare".

We even saw a claim (by Oscar Sandberg) that decentralized search might implicitly be a partial explanation of network dynamics and structure

Roadmap: where we have been and whats next

Chapter 20 started off with a discussion of the small worlds phenomena and an insightful understanding of how decentralized search can work.

Previously, we were led to the observation that geographical distance (or social distance) correlates with friendship such that Prob[v is a friend of u] $\approx [rank_u(v)]^{-1}$.

Furthermore, there is a sense that long distance friendships are "rare".

We even saw a claim (by Oscar Sandberg) that decentralized search might implicitly be a partial explanation of network dynamics and structure

We have seen earlier (Chapters 3 and 4) how selection (i.e. homophily in the sense of "birds of a feather flock together") causes friendship links. Chapter 5 also relates to how links can form or change to achieve structural balance.

Roadmap: where we have been and whats next

Chapter 20 started off with a discussion of the small worlds phenomena and an insightful understanding of how decentralized search can work.

Previously, we were led to the observation that geographical distance (or social distance) correlates with friendship such that Prob[v is a friend of u] $\approx [rank_u(v)]^{-1}$.

Furthermore, there is a sense that long distance friendships are "rare".

We even saw a claim (by Oscar Sandberg) that decentralized search might implicitly be a partial explanation of network dynamics and structure

We have seen earlier (Chapters 3 and 4) how selection (i.e. homophily in the sense of "birds of a feather flock together") causes friendship links. Chapter 5 also relates to how links can form or change to achieve structural balance.

This week we will be building on these ideas.

Power law distributions

A power law distribution for discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c. (We often just focus on on the exponent c.)

Power law distributions

A power law distribution for discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c. (We often just focus on on the exponent c.)

Closely related (and sometimes used interchangeably) is Zipf's Law, which relates the frequency, f (i.e. count) of something with it's rank, r.

- $r = 1$ being the most frequent, $r = 2$ being the second most frequent, and so on.

Power law distributions

A power law distribution for discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c. (We often just focus on on the exponent c.)

Closely related (and sometimes used interchangeably) is Zipf's Law, which relates the frequency, f (i.e. count) of something with it's rank, r.

- $r = 1$ being the most frequent, $r = 2$ being the second most frequent, and so on.

A phenomena satisfies Zipf's Law when:

$$f \approx \frac{a}{r^c}$$

for some constants a and c

Why care about power laws?

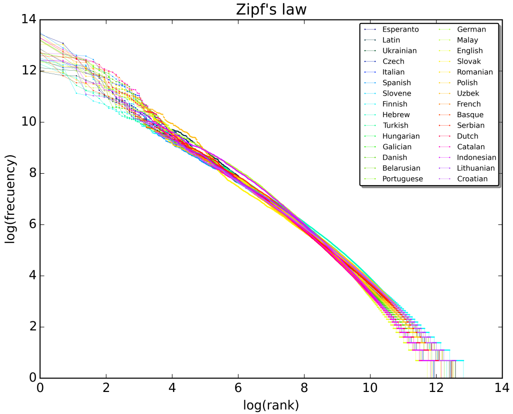
- Chapter 18 calls attention to the fact that power law distributions often occur in network and natural phenomena.
- We've already seen power laws emerge in the probability of friendship forming with respect to both distances, and ranks

Why care about power laws?

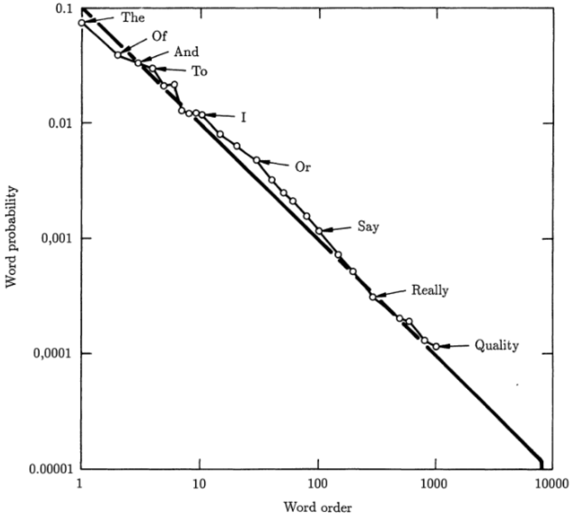
- Chapter 18 calls attention to the fact that power law distributions often occur in network and natural phenomena.
- We've already seen power laws emerge in the probability of friendship forming with respect to both distances, and ranks
- Where else do they appear?

Zipf's law in text

- In any book, let w be a word within. If we calculate it's frequency f_w (i.e. number of occurrences in the text) and it's rank r_w (i.e. is the word the first, second, ... nth most common word in the book) then we find that $f_w \propto 1/r_w$ (or equivalently, $\log f_w \approx -\log r_w + C$)



Power law in text



[Image from Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise by Manfred Schroeder]

Zipf's law in population centres

- Similarly, if we consider cities t and let f_t be the population of a city, and r_t be the city's rank by population, we have $f_t \approx a/r_t$

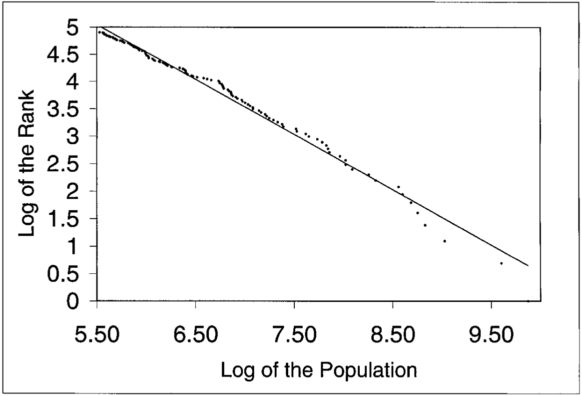


FIGURE I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

[Image from Gabaix, 1999]

Power laws in websites and products

- Power laws also arise in the popularity of websites and commercial products

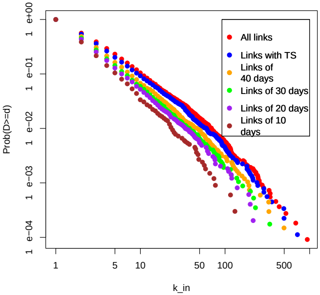


Fig. 6: *Temporal changes in the in-degree distributions in TREC.*

[Image from Shi et al.]

- Empirically, in the web network (i.e. an *information network*), the probability that a site will have k in-links is proportional to k^{-2}. (More precisely, proportional to $k^{-(2+\epsilon)}$ for some $\epsilon > 0$.

Power law distributions

- Why are power laws so prevalent?

Power law distributions

- Why are power laws so prevalent?
- Power law distributions in social and information networks often arise from coupled or correlated individual decisions

Power law distributions

- Why are power laws so prevalent?
- Power law distributions in social and information networks often arise from coupled or correlated individual decisions
  - e.g., the *popularity* of certain books or cities, occurrences of specific words in a natural language, etc...
  - We'll be considering the frequency of in-links to web sites in more detail

Power law distributions

- Why are power laws so prevalent?
- Power law distributions in social and information networks often arise from coupled or correlated individual decisions
  - e.g., the *popularity* of certain books or cities, occurrences of specific words in a natural language, etc...
  - We'll be considering the frequency of in-links to web sites in more detail
- Consequence: Events may be less rare then they appear at first glance

Power law distributions

- Why are power laws so prevalent?
- Power law distributions in social and information networks often arise from coupled or correlated individual decisions
  - e.g., the *popularity* of certain books or cities, occurrences of specific words in a natural language, etc...
  - We'll be considering the frequency of in-links to web sites in more detail
- Consequence: Events may be less rare then they appear at first glance

- Key takeaway: extreme events (e.g., for a site to have very many in-links) is **not so rare** when compared with what would be predicted by independent decisions.

How rare is rare when compared with averages over independent actions?

How rare is rare when compared with averages over independent actions?

- What if people chose where to live independent of the city? What would be (the distribution of) the population of cities?
- What if we all independently chose to read books not dependent on current events or what friends (or an online system) recommended? How rare would it be to have a huge best seller?
- What, if each web site chose their out-links independently and without some underlying dynamics to guide the process?

How rare is rare when compared with averages over independent actions?

- What if people chose where to live independent of the city? What would be (the distribution of) the population of cities?
- What if we all independently chose to read books not dependent on current events or what friends (or an online system) recommended? How rare would it be to have a huge best seller?
- What, if each web site chose their out-links independently and without some underlying dynamics to guide the process?

As is well understood, the **Central Limit Theorem** tells us that "a quantity that can be viewed as the sum (or average) of many small independent random effects will be well-approximated" by a *normal distribution*.

The normal distribution

The normal or Gaussian distribution has the following probability density function:

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

As we know, normal distributions have a *bell shaped curve*.

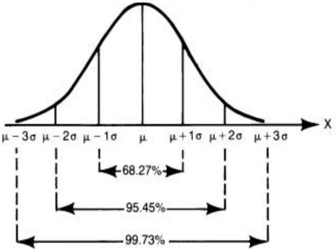


Figure 2

Percent	99.73%	99%	95.45%	95%	90%	80%	68.27%
No. of ± σ's	3.00	2.58	2.00	1.96	1.645	1.28	1.00

From: http://www.answers.com/topic/normal-distribution

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large “outliers” would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

- $k = 2$:
  - $(2)^{-2} = 1/4$
  - $2^{-2} = 1/4$

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

- $k = 2$:
  - $(2)^{-2} = 1/4$
  - $2^{-2} = 1/4$
- $k = 30$

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

- $k = 2$:
  - $(2)^{-2} = 1/4$
  - $2^{-2} = 1/4$
- $k = 30$
  - $(30)^{-2} = 1/900$

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

- $k = 2$:
  - $(2)^{-2} = 1/4$
  - $2^{-2} = 1/4$
- $k = 30$
  - $(30)^{-2} = 1/900$
  - $2^{-30} \approx 1/10^9$

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

- $k = 2$:
  - $(2)^{-2} = 1/4$
  - $2^{-2} = 1/4$
- $k = 30$
  - $(30)^{-2} = 1/900$
  - $2^{-30} \approx 1/10^9$
- One in a billion vs better than 1 in a 1000!

So how rare is rare?

- In a normal distribution, the probability of an outlier (i.e. an exceptional event) decreases exponentially with distance from the mean
  - If in-links followed a normal distribution, then the probability that a given site would have k links would decrease *exponentially* in k
  - Very small or large "outliers" would be highly improbable

How improbable? Let's compare Prob[k in-links] $\approx k^{-2}$ (power law) with Prob[k in-links] $\approx 2^{-k}$ (exponential decay):

- $k = 2$:
  - $(2)^{-2} = 1/4$
  - $2^{-2} = 1/4$
- $k = 30$
  - $(30)^{-2} = 1/900$
  - $2^{-30} \approx 1/10^9$
- One in a billion vs better than 1 in a 1000!
  - Note: These are not probabilities (we're missing the normalizing constants)
  - However, they illustrate the difference in the rate of decay

So where are we going?

As we have mentioned before, one of the most fundamental questions for social networks concerns how they evolve. What is the interplay between selection and influence?

So where are we going?

As we have mentioned before, one of the most fundamental questions for social networks concerns how they evolve. What is the interplay between selection and influence?

This is a difficult question! Perhaps the dynamics of information networks created by individuals can be better understood than the dynamics of friendships, political affiliations, opinion formation, etc...

We will see a network dynamic that leads to a power law distribution.

Recap

- Power Laws
  - Definition
  - Zipf's law
  - Dynamics that often give rise to power laws

A power law distribution and network dynamics

- *A power law distribution* for a discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c

A power law distribution and network dynamics

- *A power law distribution* for a discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c
    - We often just focus on on the exponent c and say that $Prob[X = k] \propto k^{-c}$

A power law distribution and network dynamics

- *A power law distribution* for a discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c
  - We often just focus on on the exponent c and say that $Prob[X = k] \propto k^{-c}$
- Having observed power law distributions emerge from events in social and information networks (e.g., website in-links), the big question is how this happens
  - We saw that it could *not* evolve from independent decisions that have averaged out; therefore it must arise from correlated decisions

A power law distribution and network dynamics

- *A power law distribution* for a discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c
  - We often just focus on on the exponent c and say that $Prob[X = k] \propto k^{-c}$
- Having observed power law distributions emerge from events in social and information networks (e.g., website in-links), the big question is how this happens
  - We saw that it could *not* evolve from independent decisions that have averaged out; therefore it must arise from correlated decisions
- Kumar et al [2000] proposed a preferential attachment model that can explain the power law distribution

A power law distribution and network dynamics

- *A power law distribution* for a discrete random variable X satisfies $Prob[X = k] \approx \frac{a}{k^c}$ for some constants a and c
  - We often just focus on on the exponent c and say that $Prob[X = k] \propto k^{-c}$
- Having observed power law distributions emerge from events in social and information networks (e.g., website in-links), the big question is how this happens
  - We saw that it could *not* evolve from independent decisions that have averaged out; therefore it must arise from correlated decisions
- Kumar et al [2000] proposed a preferential attachment model that can explain the power law distribution
  - Recall, the observed distribution of in-links is: Prob[a site has k in-links] $\propto k^{-(2+\epsilon)}$ for a small $\epsilon > 0$

A "rich get richer model" for in-links on the Web

Here is the model proposed in Kumar et al article:

1. Web pages are created sequentially, and named $1, 2, \ldots N$.
   (Of course, N keeps growing but we are looking at the web at some point in time)

A “rich get richer model” for in-links on the Web

Here is the model proposed in Kumar et al article:

1. Web pages are created sequentially, and named $1, 2, \ldots N$.
   (Of course, N keeps growing but we are looking at the web at some point in time)
2. When the j^{th} page is created, we choose a page $i < j$ uniformly at random.

A "rich get richer model" for in-links on the Web

Here is the model proposed in Kumar et al article:

1. Web pages are created sequentially, and named $1, 2, \dots N$.
   (Of course, N keeps growing but we are looking at the web at some point in time)
2. When the j^{th} page is created, we choose a page $i < j$ uniformly at random. Next:

A "rich get richer model" for in-links on the Web

Here is the model proposed in Kumar et al article:

1. Web pages are created sequentially, and named $1, 2, \ldots N$.
   (Of course, N keeps growing but we are looking at the web at some point in time)
2. When the j^{th} page is created, we choose a page $i < j$ uniformly at random. Next:
   - With probability p, page j links to page i

A “rich get richer model” for in-links on the Web

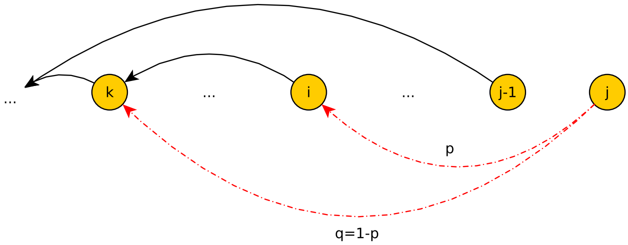
Here is the model proposed in Kumar et al article:

1. Web pages are created sequentially, and named $1, 2, \dots N$.
   (Of course, N keeps growing but we are looking at the web at some point in time)
2. When the j^{th} page is created, we choose a page $i < j$ uniformly at random. Next:
   - With probability p, page j links to page i
   - Otherwise, page j links to the page (say $k < i$) to which i has a link

A “rich get richer model” for in-links on the Web

Here is the model proposed in Kumar et al article:

1. Web pages are created sequentially, and named $1, 2, \ldots N$.
   (Of course, N keeps growing but we are looking at the web at some point in time)
2. When the j^{th} page is created, we choose a page $i < j$ uniformly at random. Next:
   - With probability p, page j links to page i
   - Otherwise, page j links to the page (say $k < i$) to which i has a link



Aside

- The full model by Kumar et al. is more general: specifically, multiple links from page j are created in the full model
- Our simplified model only creates one link

Aside

- The full model by Kumar et al. is more general: specifically, multiple links from page j are created in the full model
- Our simplified model only creates one link
- Despite this, **the power law exponent does not change**
    - The key parameter is p – whether or not we're linking to a page selected uniformly at random

The linking model continued

There is an equivalent way to frame the indirect linking that highlights the "rich get richer" preferential attachment phenomena.

The linking model continued

There is an equivalent way to frame the indirect linking that highlights the "rich get richer" preferential attachment phenomena.

- Original rule: Select a prior page, i, uniformly at random. With probability p link to i, otherwise link to the node pointed to by i's outgoing link

The linking model continued

There is an equivalent way to frame the indirect linking that highlights the "rich get richer" preferential attachment phenomena.

- Original rule: Select a prior page, i, uniformly at random. With probability p link to i, otherwise link to the node pointed to by i's outgoing link
- Equivalent rule:
  - With probability p, page j links to a page $\ell < j$, chosen uniformly at random
  - Else, page j links to the page $\ell < j$ with probability proportional to ℓ's current number of in-links

The linking model continued

There is an equivalent way to frame the indirect linking that highlights the "rich get richer" preferential attachment phenomena.

- Original rule: Select a prior page, i, uniformly at random. With probability p link to i, otherwise link to the node pointed to by i's outgoing link
- Equivalent rule:
  - With probability p, page j links to a page $\ell < j$, chosen uniformly at random
  - Else, page j links to the page $\ell < j$ with probability proportional to ℓ's current number of in-links

This is, of course, the idea behind *popularity*!

- e.g., the more people that are reading a current novel, the more likely that you might want to read it
- For various social and economic reasons why some large cities continue to grow.

The linking model continued

- Equivalent rule:
  - With probability p, page j links to a page $\ell < j$, chosen uniformly at random
  - Else, link to the page $\ell < j$ with probability proportional to ℓ's current number of in-links

Note: For $q := 1 - p$, the as $p \to 0$ (and $q \to 1$), pages are more likely to copy the same previous pages and the more likely that the process is creating some popular pages

The linking model continued

- Equivalent rule:
  - With probability p, page j links to a page $\ell < j$, chosen uniformly at random
  - Else, link to the page $\ell < j$ with probability proportional to ℓ's current number of in-links

Note: For $q := 1 - p$, the as $p \to 0$ (and $q \to 1$), pages are more likely to copy the same previous pages and the more likely that the process is creating some popular pages

Hedge: As the text states clearly, the goal of this model is not to capture all the reasons why people create links on the Web (or links in other networks) but rather to explain why it is reasonable to expect power laws to arise from such popularity phenomena.

An informal analysis for the simplified preferential attachment model proposed for Web in-links

A precise analysis of even the simple one link per page preferential attachment model is technical.

There is a heuristic argument that shows how the power law exponent is determined by the probability p (of the j^{th} page linking uniformly at random)

An informal analysis for the simplified preferential attachment model proposed for Web in-links

A precise analysis of even the simple one link per page preferential attachment model is technical.
There is a heuristic argument that shows how the power law exponent is determined by the probability p (of the j^{th} page linking uniformly at random)

While we often discretize continuous processes, it is often advantageous to model a sequence of discrete events as a continuous process

An informal analysis for the simplified preferential attachment model proposed for Web in-links

A precise analysis of even the simple one link per page preferential attachment model is technical.
There is a heuristic argument that shows how the power law exponent is determined by the probability p (of the j^{th} page linking uniformly at random)

While we often discretize continuous processes, it is often advantageous to model a sequence of discrete events as a continuous process

Specifically, we'll consider a continuous deterministic variable $x_\ell(t)$, that approximates the discrete random variable $X_\ell(t)$, the number of in-links to a page ℓ at time $t \geq 0$.

The deterministic continuous model

- Let $X_\ell(t)$ is the number of in-links to a page ℓ at time $t \geq 0$
  - $X_\ell(t)$ is a discrete random variable

The deterministic continuous model

- Let $X_\ell(t)$ is the number of in-links to a page ℓ at time $t \geq 0$
  - $X_\ell(t)$ is a discrete random variable
- We will create a continuous deterministic variable $x_\ell(t)$, that approximates $X_\ell(t)$

The deterministic continuous model

- Let $X_\ell(t)$ is the number of in-links to a page ℓ at time $t \geq 0$
  - $X_\ell(t)$ is a discrete random variable
- We will create a continuous deterministic variable $x_\ell(t)$, that approximates $X_\ell(t)$
- Assuming that page ℓ is added at time ℓ, then $X_\ell(\ell) = x_\ell(\ell) = 0$.
  - the soonest a link to ℓ can be added is at $\ell + 1$

The deterministic continuous model

- Let $X_\ell(t)$ is the number of in-links to a page ℓ at time $t \geq 0$
  - $X_\ell(t)$ is a discrete random variable
- We will create a continuous deterministic variable $x_\ell(t)$, that approximates $X_\ell(t)$
- Assuming that page ℓ is added at time ℓ, then $X_\ell(\ell) = x_\ell(\ell) = 0$.
  - the soonest a link to ℓ can be added is at $\ell + 1$
- In the discrete model, for $t \geq \ell$, the probability that the number of links to a page ℓ increases at time $t + 1$ is:

$$p\frac{1}{t} + q\frac{X_\ell(t)}{t}$$

The deterministic continuous model

- Let $X_\ell(t)$ is the number of in-links to a page ℓ at time $t \geq 0$
  - $X_\ell(t)$ is a discrete random variable
- We will create a continuous deterministic variable $x_\ell(t)$, that approximates $X_\ell(t)$
- Assuming that page ℓ is added at time ℓ, then $X_\ell(\ell) = x_\ell(\ell) = 0$.
  - the soonest a link to ℓ can be added is at $\ell + 1$
- In the discrete model, for $t \geq \ell$, the probability that the number of links to a page ℓ increases at time $t+1$ is:

$$p\frac{1}{t} + q\frac{X_\ell(t)}{t}$$

- For $t \geq \ell$, and the corresponding continuous model obeys the differential equation:

$$\frac{dx_\ell}{dt} = p\frac{1}{t} + q\frac{x_\ell(t)}{t}$$

The deterministic continuous model

- Let $X_\ell(t)$ is the number of in-links to a page ℓ at time $t \geq 0$
  - $X_\ell(t)$ is a discrete random variable
- We will create a continuous deterministic variable $x_\ell(t)$, that approximates $X_\ell(t)$
- Assuming that page ℓ is added at time ℓ, then $X_\ell(\ell) = x_\ell(\ell) = 0$.
  - the soonest a link to ℓ can be added is at $\ell + 1$
- In the discrete model, for $t \geq \ell$, the probability that the number of links to a page ℓ increases at time $t + 1$ is:

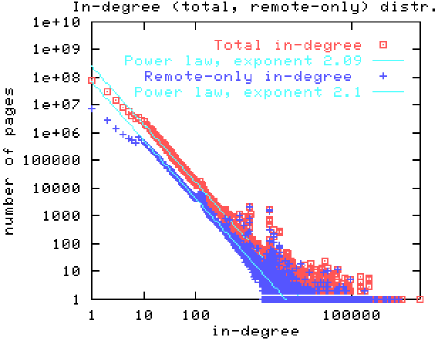
$$p\frac{1}{t} + q\frac{X_\ell(t)}{t}$$

- For $t \geq \ell$, and the corresponding continuous model obeys the differential equation:

$$\frac{dx_\ell}{dt} = p\frac{1}{t} + q\frac{x_\ell(t)}{t}$$

- From some basic calculus (see Ch 18.7) this leads to a power law distribution proportional to k^{-c} with $c = 1 + 1/q$

The deterministic continuous model



[Fig 18-2 in E&K]

- As $p \to 0$, the exponent $c = 1 + 1/q$ limits to the observed exponent $c = 2 + \epsilon$ for the observed in-link power law distribution
- As $p \to 1$, the exponent limits to ∞ making a large number of in-links very unlikely.

Wed. Feb 14th: Announcements & Corrections

- A0 marks are out
- A1 due tomorrow
- Please do fill out the mid-course evaluation :)
- Critical review groups & paper choices due by Fri Mar 1 (email me)
- Midterm will be Fri Mar 8 (make up test on Fri Mar 15)
  - Covers up to the end of this week's slides (we might finish covering PageRank next lecture)
- Reading week next week
  - Try and get some rest! I've tried to avoid making things due right after reading week :)
  - No regularly scheduled office hours in reading week; but I'm available by appointment – email me :)
- Happy Valentines day!

Aside: Open Questions

- It's worth noting that although the preferential attachment model suggests that popularity phenomena lead to power laws, it still cannot explain all the examples we saw
- $c = 1 + 1/q \in [2, \infty)$, yet in text and cities the observed exponent is 1

Aside: Open Questions

- It's worth noting that although the preferential attachment model suggests that popularity phenomena lead to power laws, it still cannot explain all the examples we saw
- $c = 1 + 1/q \in [2, \infty)$, yet in text and cities the observed exponent is 1

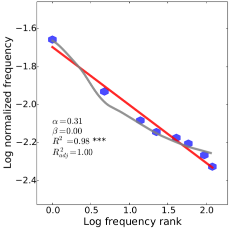


Figure 8: An approximate power law distribution of novel alien names used by subjects in making up a story.

[Figure from Piantadosi (2014)]

- Piantadosi (2014)'s prompt was "An alien space ship crashes in the Nevada desert. Eight creatures emerge, a Wug, a Plit, a Blicket, a Flark, a Warit, a Jupe, a Ralex, and a Timon. In at least 2000 words, describe what happens next"; observed is $c = 0.31$

Sensitivity to unpredictable initial stages in network dynamics

- It is never clear why say some “pop” singers become so popular while other (perhaps of equal talent) never “make it”
  - Clearly, the initial stages of a dynamic process are critical and that is why advertising, promotions, etc. are so important

Sensitivity to unpredictable initial stages in network dynamics

- It is never clear why say some “pop” singers become so popular while other (perhaps of equal talent) never “make it”
  - Clearly, the initial stages of a dynamic process are critical and that is why advertising, promotions, etc. are so important
- How can we better understand the impact of the randomness in the initial stages of a dynamic process?
  - if we could replay history many times, we would expect the resulting distribution to be the same
  - But would the same books, the same movies, the same pop stars, the same web pages, etc continue to be the most popular?

Sensitivity to unpredictable initial stages in network dynamics

- It is never clear why say some "pop" singers become so popular while other (perhaps of equal talent) never "make it"
  - Clearly, the initial stages of a dynamic process are critical and that is why advertising, promotions, etc. are so important

- How can we better understand the impact of the randomness in the initial stages of a dynamic process?
  - if we could replay history many times, we would expect the resulting distribution to be the same
  - But would the same books, the same movies, the same pop stars, the same web pages, etc continue to be the most popular?

- Intuition suggests there is considerable "luck" in exactly who or what becomes popular; yet we also believe that "quality" is also important

- How can we rewind history", to try and find out?

An experiment to “rewind history”

Although we can’t rewind history, Salganik et al perform an interesting experiment (in fact, two experiments at different times with different participants) to observe the impact of the initial random stages in a dynamic process. (the article is available on the course website).

The Salganik et al experiment

Here is their experiment:

- They created 9 copies of a music streaming site with 48 “obscure” (as determined by some experts) songs of varying “quality”
- Approximately 7200 young participants were recruited to listen to the music, knowing only the name of the band and the song.
- In each of the copies, participants sequentially listened to some music selections, rated the music and then were given the opportunity to download copies of songs they liked.

The Salganik et al experiment

Here is their experiment:

- They created 9 copies of a music streaming site with 48 “obscure” (as determined by some experts) songs of varying “quality”
- Approximately 7200 young participants were recruited to listen to the music, knowing only the name of the band and the song.
- In each of the copies, participants sequentially listened to some music selections, rated the music and then were given the opportunity to download copies of songs they liked.
- In 8 copies of the site (each with 10% of the participants), they were also given the number of times each song had been previously downloaded.

The Salganik et al experiment

Here is their experiment:

- They created 9 copies of a music streaming site with 48 "obscure" (as determined by some experts) songs of varying "quality"
- Approximately 7200 young participants were recruited to listen to the music, knowing only the name of the band and the song.
- In each of the copies, participants sequentially listened to some music selections, rated the music and then were given the opportunity to download copies of songs they liked.
- In 8 copies of the site (each with 10% of the participants), they were also given the number of times each song had been previously downloaded.
- In the 9th version, this previous history of downloads was not provided to the remaining 20% of the participants. The average of the ratings (from 1 = "I hated it" to 5 = "I loved it") in this "no influence" version determined the "true" song "quality".

The Salganik et al experiment

Here is their experiment:

- They created 9 copies of a music streaming site with 48 “obscure” (as determined by some experts) songs of varying “quality”
- Approximately 7200 young participants were recruited to listen to the music, knowing only the name of the band and the song.
- In each of the copies, participants sequentially listened to some music selections, rated the music and then were given the opportunity to download copies of songs they liked.
- In 8 copies of the site (each with 10% of the participants), they were also given the number of times each song had been previously downloaded.
- In the 9th version, this previous history of downloads was not provided to the remaining 20% of the participants. The average of the ratings (from 1 = “I hated it” to 5 = “I loved it”) in this “no influence” version determined the “true” song “quality”.
- The experiment was then repeated, with the 8 site copies displaying songs sorted by downloads instead of randomly

The findings in the Salganik experiment

The experiment was designed to measure the extent that social influence leads to different outcomes in the “success” (i.e. the number of downloads) of a particular song

The findings in the Salganik experiment

The experiment was designed to measure the extent that social influence leads to different outcomes in the "success" (i.e. the number of downloads) of a particular song

Simply stated, the results show that:

- Increasing the strength of social influence (by sorting songs by downloads) increased both the inequality (i.e. degrees of popularity) and unpredictability (i.e., relation to quality) of success

The findings in the Salganik experiment

The experiment was designed to measure the extent that social influence leads to different outcomes in the "success" (i.e. the number of downloads) of a particular song

Simply stated, the results show that:

- Increasing the strength of social influence (by sorting songs by downloads) increased both the inequality (i.e. degrees of popularity) and unpredictability (i.e., relation to quality) of success
- However, quality was also a factor: the best rated songs rarely did poorly and the the worst songs rarely did well.

The findings in the Salganik experiment

The experiment was designed to measure the extent that social influence leads to different outcomes in the "success" (i.e. the number of downloads) of a particular song

Simply stated, the results show that:

- Increasing the strength of social influence (by sorting songs by downloads) increased both the inequality (i.e. degrees of popularity) and unpredictability (i.e., relation to quality) of success
- However, quality was also a factor: the best rated songs rarely did poorly and the the worst songs rarely did well.

As I said, this is an interesting study and one where the authors carefully try to eliminate sources of bias. The article is worth reading.

The findings in the Salganik experiment

The experiment was designed to measure the extent that social influence leads to different outcomes in the "success" (i.e. the number of downloads) of a particular song

Simply stated, the results show that:

- Increasing the strength of social influence (by sorting songs by downloads) increased both the inequality (i.e. degrees of popularity) and unpredictability (i.e., relation to quality) of success
- However, quality was also a factor: the best rated songs rarely did poorly and the the worst songs rarely did well.

As I said, this is an interesting study and one where the authors carefully try to eliminate sources of bias. The article is worth reading.

As the text points out in section 18.6, how recommendation systems are designed can impact how people make choices, leading to increased "rich get richer" phenomena, or alternatively exposing people to less popular items.

Visualizing the long tail of a power law distribution

- Once we accept a power law nature of popularity, it is instructive to consider the consequences for a given industry
  - Consider that the nature of the sales curve that would be dictated by a power law distribution

Visualizing the long tail of a power law distribution

- Once we accept a power law nature of popularity, it is instructive to consider the consequences for a given industry
  - Consider that the nature of the sales curve that would be dictated by a power law distribution

The shape of the long tail in a power distribution raises the question as to how many sales can be obtained from less popular (e.g. niche items).

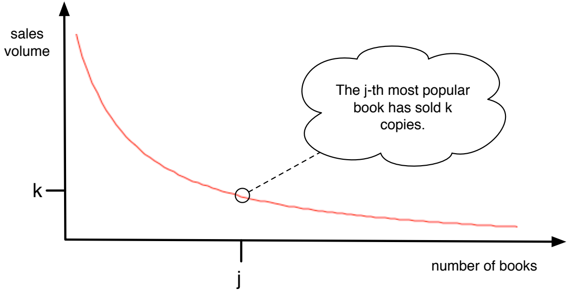


Figure: [Fig 18-4 in E&K] text; how many copies of the j^{th} most popular items have been sold.

Recap

- Power Laws
  - Definition
  - Zipf's law
  - Dynamics that often give rise to power laws
- Rich-get-richer model
- Salganik et al. music popularity experiment

Search and ranking on the Web

Our next topic is to understand how the popularity of a web page is determined and how that impacts its rank in the responses to a query.

But first, how do search engines find and rank responses to a query?

Search and ranking on the Web

Our next topic is to understand how the popularity of a web page is determined and how that impacts its rank in the responses to a query.

But first, how do search engines find and rank responses to a query?

The specific algorithms used by search engines such as Bing and Google is a trade secret. To some extent this has to be kept secret as there is always a "war" between a search engine and companies that create web sites to enhance the ranking of a site.

However, we do have a basic idea as to how these search engines rank sites given a query. In fact, at the most elementary level, the main idea is an old one.

Search and ranking on the Web

Our next topic is to understand how the popularity of a web page is determined and how that impacts its rank in the responses to a query.

But first, how do search engines find and rank responses to a query?

The specific algorithms used by search engines such as Bing and Google is a trade secret. To some extent this has to be kept secret as there is always a "war" between a search engine and companies that create web sites to enhance the ranking of a site.

However, we do have a basic idea as to how these search engines rank sites given a query. In fact, at the most elementary level, the main idea is an old one.

Aside: In the 1960s and 70s, there was a basic argument as to whether online search and ranking was a more or less normal algorithmic search and optimization problem or one that required "intelligence" (i.e. the ability to understand natural language). Who won this argument?

Search and ranking of Web documents; the role of link popularity

The most basic approach is to treat a document as a bag of words and then use “normalized” word counts (and pairs,triplets of words) to identify and rank documents relating to the query. This became enhanced by more sophisticated contextual aspects of word occurrences, etc and today machine learning algorithms are also used in classifying a search query.

Search and ranking of Web documents; the role of link popularity

The most basic approach is to treat a document as a bag of words and then use “normalized” word counts (and pairs,triplets of words) to identify and rank documents relating to the query. This became enhanced by more sophisticated contextual aspects of word occurrences, etc and today machine learning algorithms are also used in classifying a search query.

But early in the development of popular search engines, a popularity aspect was added where the ranking of a document also depended on the link structure and the popularity of a Web page in the Web network (or at least in that part that seems relevant to the query).

Search and ranking of Web documents; the role of link popularity

The most basic approach is to treat a document as a bag of words and then use “normalized” word counts (and pairs,triplets of words) to identify and rank documents relating to the query. This became enhanced by more sophisticated contextual aspects of word occurrences, etc and today machine learning algorithms are also used in classifying a search query.

But early in the development of popular search engines, a popularity aspect was added where the ranking of a document also depended on the link structure and the popularity of a Web page in the Web network (or at least in that part that seems relevant to the query).

Two algorithms were independently proposed for determining the popularity of a Web page, namely Hubs and Authorities developed at IBM, and Page Rank, developed and integrated into Google’s search engine.

Link analysis and page popularity

Neither Hubs and Authorities nor Page Rank use link in-degree as the popularity measure but link analysis is (or at least was) used to determine page popularity. Currently, it seems clear that popularity also depends on recent behaviour of users to related queries.

Link analysis and page popularity

Neither Hubs and Authorities nor Page Rank use link in-degree as the popularity measure but link analysis is (or at least was) used to determine page popularity. Currently, it seems clear that popularity also depends on recent behaviour of users to related queries.

We will not try to infer more precisely how say Google (or any search engine) precisely determines the ranking of a document in response to a query. In particular, we do not know how much page ranking depends on content vs link analysis. But we do know that this ranking is essential in determining how often a page will be downloaded. The quality of the ranking algorithm leads to user activity and thus the resulting advertising.

Link analysis and page popularity

Neither Hubs and Authorities nor Page Rank use link in-degree as the popularity measure but link analysis is (or at least was) used to determine page popularity. Currently, it seems clear that popularity also depends on recent behaviour of users to related queries.

We will not try to infer more precisely how say Google (or any search engine) precisely determines the ranking of a document in response to a query. In particular, we do not know how much page ranking depends on content vs link analysis. But we do know that this ranking is essential in determining how often a page will be downloaded. The quality of the ranking algorithm leads to user activity and thus the resulting advertising.

We will begin with the Hubs and Authorities ranking algorithm and then the Page Rank algorithm.

Hubs and Authorities

- A simple way to utilize links to rank web pages would be to think of each link from A to B as an endorsement or vote by A for B.

- Question: Assuming it's tractable, then what's wrong with just counting the number of in-links?

Hubs and Authorities

- A simple way to utilize links to rank web pages would be to think of each link from A to B as an endorsement or vote by A for B.
- Question: Assuming it's tractable, then what's wrong with just counting the number of in-links?
- If we use the number (or weight) of endorsements to determine rank, then one would have to adjust such scores coming from say the same domain name.

Hubs and Authorities

- A simple way to utilize links to rank web pages would be to think of each link from A to B as an endorsement or vote by A for B.
- Question: Assuming it's tractable, then what's wrong with just counting the number of in-links?
- If we use the number (or weight) of endorsements to determine rank, then one would have to adjust such scores coming from say the same domain name.
- Even after adjusting for such "vote fixing", there are problems
  - if Wayne Gretzky has a web site with a link suggesting where to buy hockey equipment, that should be more meaningful than my recommendation about hockey equipment

Hubs and Authorities

- A simple way to utilize links to rank web pages would be to think of each link from A to B as an endorsement or vote by A for B.
- Question: Assuming it's tractable, then what's wrong with just counting the number of in-links?
- If we use the number (or weight) of endorsements to determine rank, then one would have to adjust such scores coming from say the same domain name.
- Even after adjusting for such "vote fixing", there are problems
  - if Wayne Gretzky has a web site with a link suggesting where to buy hockey equipment, that should be more meaningful than my recommendation about hockey equipment

  Spoiler alert: I don't play or watch hockey

Reinforcement of Hubs and Authorities.

- This then becomes the motivation (and seemingly circular reasoning) behind hubs and authorities:
  - The best "authorities" on a subject (places to buy equipment) are being endorsed by the best "hubs" (people who know where to buy equipment)
  - Similarly, the best hubs are those sites that recommend the best authorities

Reinforcement of Hubs and Authorities.

- This then becomes the motivation (and seemingly circular reasoning) behind hubs and authorities:
    - The best "authorities" on a subject (places to buy equipment) are being endorsed by the best "hubs" (people who know where to buy equipment)
    - Similarly, the best hubs are those sites that recommend the best authorities
    - Conceptually the link structure induces a bipartite graph, however the same web page can be both a hub and an authority

Reinforcement of Hubs and Authorities.

- This then becomes the motivation (and seemingly circular reasoning) behind hubs and authorities:
  - The best "authorities" on a subject (places to buy equipment) are being endorsed by the best "hubs" (people who know where to buy equipment)
  - Similarly, the best hubs are those sites that recommend the best authorities
  - Conceptually the link structure induces a bipartite graph, however the same web page can be both a hub and an authority

- Comment: The word "authority" is not generally an accurate way to describe high ranking documents. These might better be referred to (barring other information) as the most relied upon sites. This is also different from "the most popular" sites which might better be measured in terms of the number of clicks being received. Hubs then are the most reliable endorsers.

Hubs & Authorities procedure

- The hubs and authorities procedure is as follows:
  - Initialize each node' hub value to some positive number (perhaps depending on usage or content)

Hubs & Authorities procedure

- The hubs and authorities procedure is as follows:
  - Initialize each node' hub value to some positive number (perhaps depending on usage or content)
  - For sufficiently large k, repeat the following k times

Hubs & Authorities procedure

- The hubs and authorities procedure is as follows:
  - Initialize each node' hub value to some positive number (perhaps depending on usage or content)
  - For sufficiently large k, repeat the following k times
    1. Apply authority update rule to each page, p (i.e., set the authority value to the sum of the hub values of the nodes endorsing the page p)

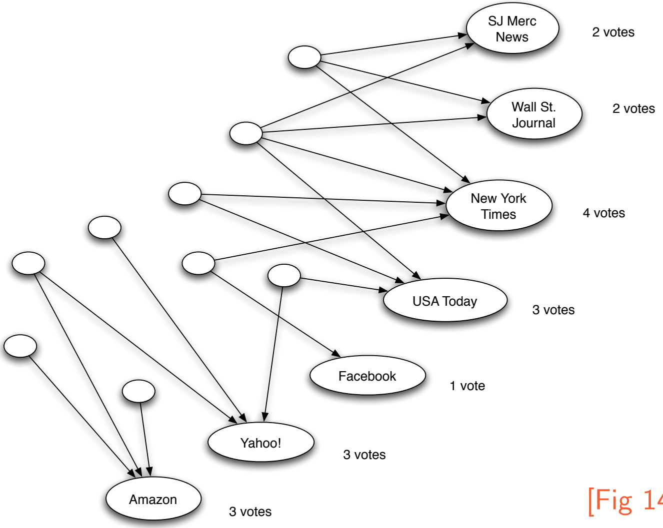
Hubs & Authorities procedure

- The hubs and authorities procedure is as follows:
  - Initialize each node' hub value to some positive number (perhaps depending on usage or content)
  - For sufficiently large k, repeat the following k times
    1. Apply authority update rule to each page, p (i.e., set the authority value to the sum of the hub values of the nodes endorsing the page p)
    2. Apply hub update rule to each page, p (i.e., set the hub value to the sum of the authority values of the nodes endorsed by the page p)

Hubs & Authorities procedure

- The hubs and authorities procedure is as follows:
  - Initialize each node' hub value to some positive number (perhaps depending on usage or content)
  - For sufficiently large k, repeat the following k times
    1. Apply authority update rule to each page, p (i.e., set the authority value to the sum of the hub values of the nodes endorsing the page p)
    2. Apply hub update rule to each page, p (i.e., set the hub value to the sum of the authority values of the nodes endorsed by the page p)
    3. Normalize so that sum of A weights is 1 and sum of H weights is 1.

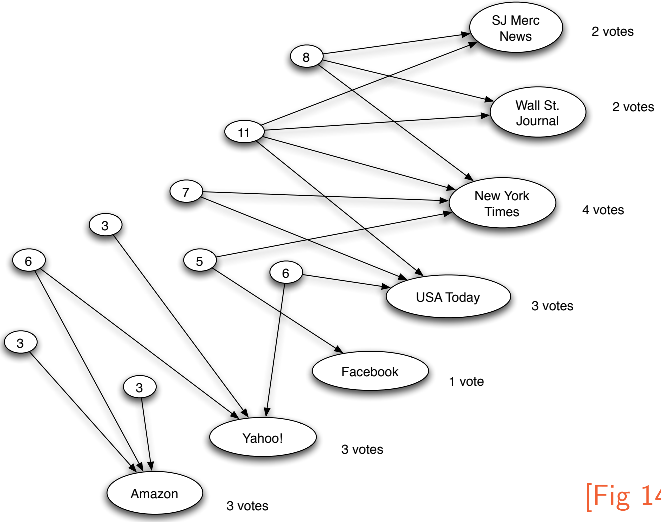
- The result of applying the authority update rule with all hub values initially 1: for each page p, auth(p) is the sum of hub values (initially just the number) of hubs pointing to p.



[Fig 14.1, E&K]

Figure: Counting in-links to pages for the query "newspapers."

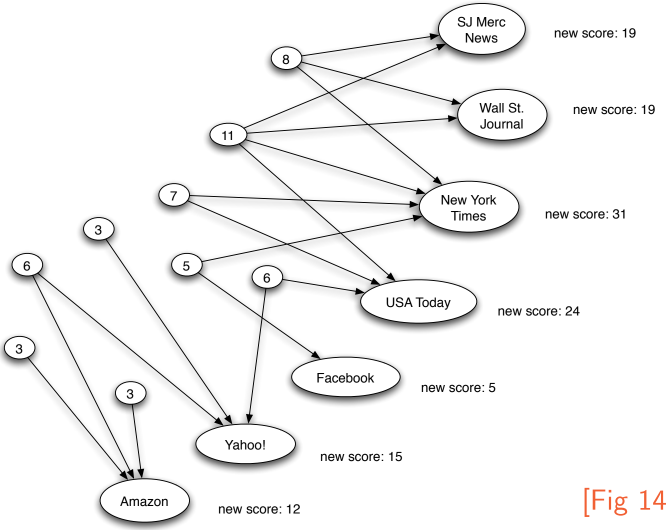
- Then to recalibrate hub values, we use the hub update rule: for each page p, hub(p) is the sum of values of all authorities that p points to.



[Fig 14.2, E&K]

Figure: Finding good lists for the query "newspapers": each page's value as a list is written as a number inside it.

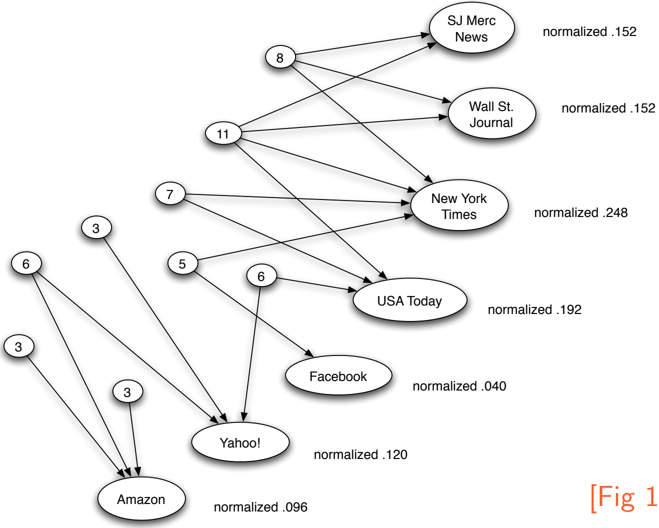
- Applying the authority update rule again we get figure 14.3.



[Fig 14.3, E&K]

Figure: Re-weighting votes for the query "newspapers": each of the labeled page's new score is equal to the sum of the values of all lists that point to it.

- Since we only care about the relative values of these numbers, both authority and hub scores can be normalized to sum to 1 (to allow convergence and avoid dealing with large numbers).



[Fig 14.4, E&K]

Figure: Re-weighting votes after normalizing for the query "newspapers".

Keep repeating a good idea

- Now having recalibrated and normalized both the authority and hub scores, we can continue this process to continue to refine these scores.

- That is, the hubs and authorities procedure is as follows:
  - Initialize all hub values to some positive vector (perhaps depending on usage or content)
  - For sufficiently large k, perform the following k times
    1. Apply authority update rule to each page
    2. Apply hub update rule to each page
    3. Normalize so that sum of A weights and sum of H weights are both 1.

Keep repeating a good idea

- Now having recalibrated and normalized both the authority and hub scores, we can continue this process to continue to refine these scores.
- That is, the hubs and authorities procedure is as follows:
  - Initialize all hub values to some positive vector (perhaps depending on usage or content)
  - For sufficiently large k, perform the following k times
    1. Apply authority update rule to each page
    2. Apply hub update rule to each page
    3. Normalize so that sum of A weights and sum of H weights are both 1.
- Using linear algebra, it can be shown (in Section 14.6) that these A and H normalized values will converge to a limit as $k \to \infty$ (which can be approximated by some sufficiently large k)! The limiting value is an equilibrium.

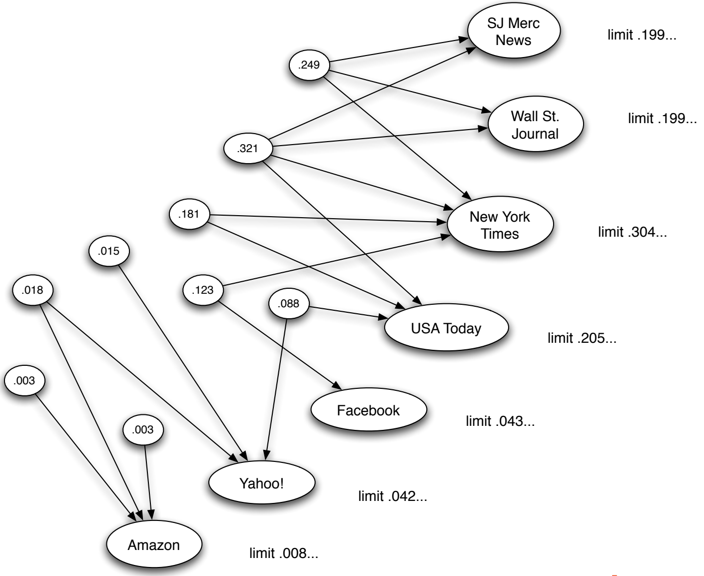
Keep repeating a good idea

- Now having recalibrated and normalized both the authority and hub scores, we can continue this process to continue to refine these scores.

- That is, the hubs and authorities procedure is as follows:
  - Initialize all hub values to some positive vector (perhaps depending on usage or content)
  - For sufficiently large k, perform the following k times
    1. Apply authority update rule to each page
    2. Apply hub update rule to each page
    3. Normalize so that sum of A weights and sum of H weights are both 1.

- Using linear algebra, it can be shown (in Section 14.6) that these A and H normalized values will converge to a limit as $k \to \infty$ (which can be approximated by some sufficiently large k)! The limiting value is an equilibrium.

- Hubs and Authorities can be extended to work for weighted edges (e,g. weighting links in anchor text, or near a section heading, etc.)



[Fig 14.5, E&K]

Figure: Limiting hub and authority values for the query "newspapers".

Page Rank

- The motivation behind page rank is a somewhat different view of how authority is conferred.
  - Endorsement of authority is conveyed by other authorities
  - That is, no hub concept
  - This is how peer review works in the academic and scholarly world.

Page Rank

- The motivation behind page rank is a somewhat different view of how authority is conferred.
  - Endorsement of authority is conveyed by other authorities
  - That is, no hub concept
  - This is how peer review works in the academic and scholarly world.

- Authorities themselves convey authority on those they link to. This naturally leads to a formulation in terms of two equivalent views of page rank:

Page Rank

- The motivation behind page rank is a somewhat different view of how authority is conferred.
  - Endorsement of authority is conveyed by other authorities
  - That is, no hub concept
  - This is how peer review works in the academic and scholarly world.

- Authorities themselves convey authority on those they link to. This naturally leads to a formulation in terms of two equivalent views of page rank:
  1. Authorities directly conveying authority (without hubs)

Page Rank

- The motivation behind page rank is a somewhat different view of how authority is conferred.
  - Endorsement of authority is conveyed by other authorities
  - That is, no hub concept
  - This is how peer review works in the academic and scholarly world.

- Authorities themselves convey authority on those they link to. This naturally leads to a formulation in terms of two equivalent views of page rank:
  1. Authorities directly conveying authority (without hubs)
  2. Authority values resulting from long term behaviour of a random walk on a graph

How does Page rank spread authority?

- Suppose at any point of time we have relevant authority scores
  - A page spreads its authority equally amongst all of its out links
  - If a page has no outlinks then all authority stays there
- This redistributes the authority scores. (We are not creating or losing any authority, only moving it)

How does Page rank spread authority?

- Suppose at any point of time we have relevant authority scores
  - A page spreads its authority equally amongst all of its out links
  - If a page has no outlinks then all authority stays there
- This redistributes the authority scores. (We are not creating or losing any authority, only moving it)
- We can initially start with every relevant page having authority $1/n$ where there are n pages. Then we repeat this process k times for some sufficiently large k

How does Page rank spread authority?

- Suppose at any point of time we have relevant authority scores
  - A page spreads its authority equally amongst all of its out links
  - If a page has no outlinks then all authority stays there
- This redistributes the authority scores. (We are not creating or losing any authority, only moving it)
- We can initially start with every relevant page having authority $1/n$ where there are n pages. Then we repeat this process k times for some sufficiently large k
- With the exception of some "degenerate cases" (e.g. the process is periodic) it can be proven (again using linear algebra) that this process has a limiting behaviour as $k \to \infty$
- The resulting limit values will form an equilibrium
- If the network is strongly connected then there is a unique equilibrium.

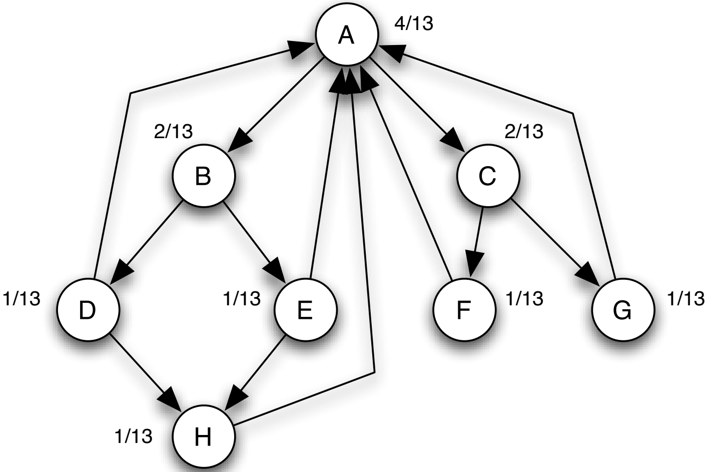
How does Page rank spread authority?

- Suppose at any point of time we have relevant authority scores
  - A page spreads its authority equally amongst all of its out links
  - If a page has no outlinks then all authority stays there
- This redistributes the authority scores. (We are not creating or losing any authority, only moving it)
- We can initially start with every relevant page having authority $1/n$ where there are n pages. Then we repeat this process k times for some sufficiently large k
- With the exception of some "degenerate cases" (e.g. the process is periodic) it can be proven (again using linear algebra) that this process has a limiting behaviour as $k \to \infty$
- The resulting limit values will form an equilibrium
- If the network is strongly connected then there is a unique equilibrium.

Remark

In many cases this won't reflect the desired authority. Namely, if the network has any sinks which it will surely have, then all of the authority will pass to such sinks.

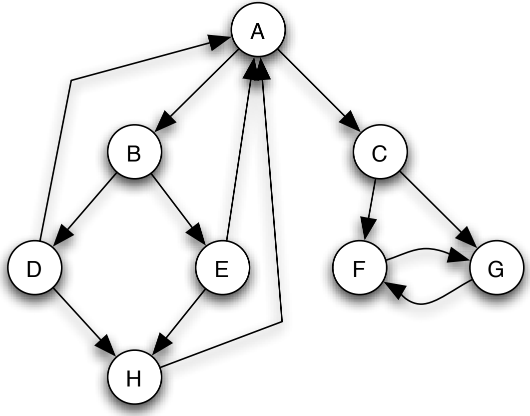
Page rank equilibrium for a network



[Fig 14.7, E&K]

Figure: Equilibrium PageRank values for the network of eight Web page.

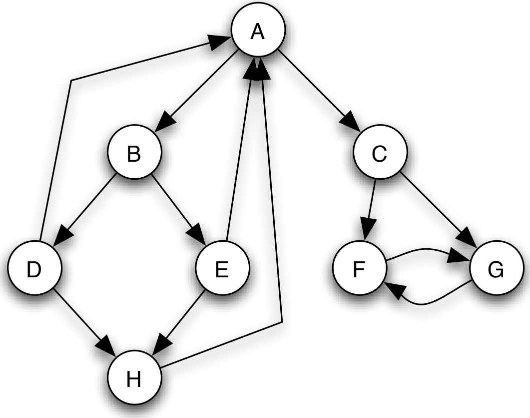
Where has all the authority gone when we redirect (F, A) and (G, A) edges?



[Fig 14.8, E&K]

The same collection of eight pages, but F and G have changed their links to point to each other instead of to A.

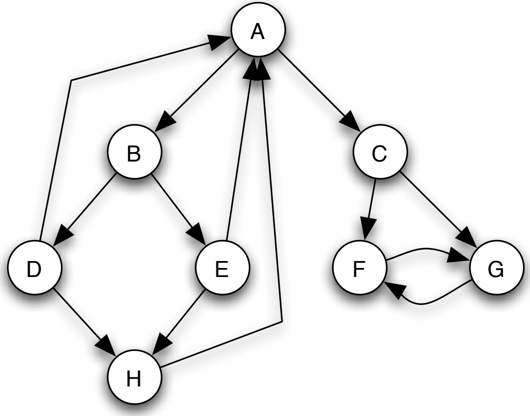
Where has all the authority gone when we redirect (F, A) and (G, A) edges?



[Fig 14.8, E&K]

The same collection of eight pages, but F and G have changed their links to point to each other instead of to A. Without "scaling", all the PageRank would go to F and G.

Where has all the authority gone when we redirect (F, A) and (G, A) edges?



[Fig 14.8, E&K]

The same collection of eight pages, but F and G have changed their links to point to each other instead of to A. Without "scaling", all the PageRank would go to F and G. NetLogo demo time!

How does PageRank spread authority?

- Even if the network converges to a unique equilibrium, the equilibrium may be undesirable

Definition (Sink)

A (typically small) strict subset of the nodes with no outgoing edges that are reachable from all nodes in the network

Remark

In many cases PageRank won't reflect the desired authority. Namely, if the network has any sinks which it will surely have, then all of the authority will pass to such sinks.

Scaled page rank

- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed according to PageRank
  - the remaining $(1 - s)$ fraction is distributed equally amongst **all** pages (including this page!)

Scaled page rank

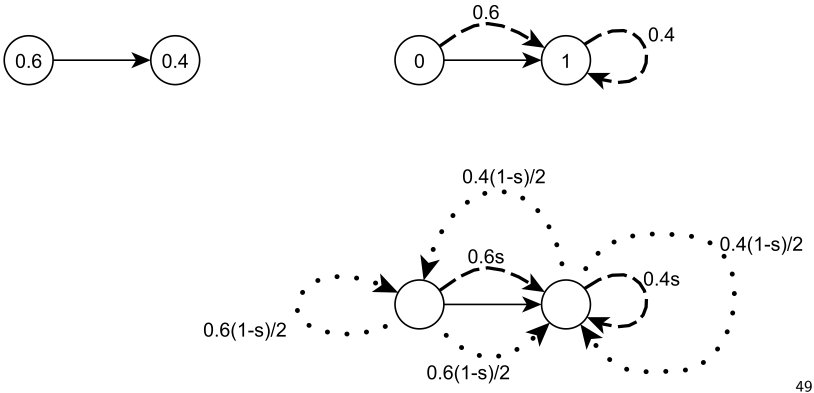
- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed according to PageRank
  - the remaining $(1 - s)$ fraction is distributed equally amongst **all** pages (including this page!)

Unscaled Page Rank:

Scaled page rank

- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed according to PageRank
  - the remaining $(1 - s)$ fraction is distributed equally amongst **all** pages (including this page!)

Unscaled Page Rank (top) vs. Scaled Page Rank (bottom):



Scaled page rank

- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed according to PageRank
  - the remaining $(1 - s)$ fraction is distributed equally amongst **all** pages (including this page!)

- For any value of $s < 1$ (which effectively makes the graph strongly connected), we get convergence to a unique set of scores for each page and that is its page rank (for that particular value of s). It is reported that Google used $0.8 \leq s \leq 0.9$

Scaled page rank

- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed according to PageRank
  - the remaining $(1 - s)$ fraction is distributed equally amongst **all** pages (including this page!)
- For any value of $s < 1$ (which effectively makes the graph strongly connected), we get convergence to a unique set of scores for each page and that is its page rank (for that particular value of s). It is reported that Google used $0.8 \leq s \leq 0.9$
- (See the footnote on page 410 of E&K as to why in the previous example, nodes F and G will still get most of the authority but that for realistically large networks, the process works well.
  Hint: "bow-tie" structure)

Scaled page rank

- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed according to PageRank
  - the remaining $(1 - s)$ fraction is distributed equally amongst **all** pages (including this page!)

- For any value of $s < 1$ (which effectively makes the graph strongly connected), we get convergence to a unique set of scores for each page and that is its page rank (for that particular value of s). It is reported that Google used $0.8 \leq s \leq 0.9$

- (See the footnote on page 410 of E&K as to why in the previous example, nodes F and G will still get most of the authority but that for realistically large networks, the process works well.
  Hint: "bow-tie" structure)

- NetLogo demo!

Mon. Feb 26th: Announcements & Corrections

- Welcome back from reading week! I hope you got some rest :)
- Consider being a volunteer notetaker! It's a chance to help a fellow student, it may help study, and you get recognition for doing so – see the Quercus announcement for details :)
- Critical review project groups & paper choices are due this Friday!
- Assignment 1 is being graded, and we hope to get the marks out before the midterm next week

Mon. Feb 26th: Announcements & Corrections

- Mid-term eval is in: A big thank you to the 4 respondents, and for the rest hopefully no news is good news :)
  - In case you were busy but still feedback, the anonymous feedback survey is always open!
  - One suggestion: "Making the practice quizzes worth a small percentage of grade."
    - We used to have things organized that way, ultimately it was changed for pedagogical & equity reasons (The quizzes are meant to be formative feedback so making mistakes should be encouraged instead of punished, and quizzes with a completion grade risk grading students on their spare time instead of understanding)
  - Zoom access is helpful, having slides early is good, & enthusiasm is appreciated!

Some additional remarks

- The limiting scores for both the authority and hubs approach and the page rank approach are equilibrium points for an appropriate algebraic process
- That is, if we actually were in the limiting state, we would be in the equilibrium state. In practice, we can stop the process when the change in each iteration is sufficiently small

Some additional remarks

- The limiting scores for both the authority and hubs approach and the page rank approach are equilibrium points for an appropriate algebraic process
- That is, if we actually were in the limiting state, we would be in the equilibrium state. In practice, we can stop the process when the change in each iteration is sufficiently small
- We can weight the network edges (say according to some concept of link importance) and apply the same authority and hubs or page rank approach distributing authority in proportion to these weights

Advanced material (section 14.6): Handwaving argument why these processes converge

Both the page rank and hubs and authorities processes can be understood as a linear transformation

Advanced material (section 14.6): Handwaving argument why these processes converge

Both the page rank and hubs and authorities processes can be understood as a linear transformation

- Consider a web network of n pages
  - We can represent the hub, authority or page rank values at any time k of the process by an n-dimensional (column) vector, denoted (respectively) by $\mathbf{h}^{(k)}, \mathbf{a}^{(k)}, \mathbf{r}^{(k)}$.

Advanced material (section 14.6): Handwaving argument why these processes converge

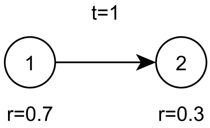
Both the page rank and hubs and authorities processes can be understood as a linear transformation

- Consider a web network of n pages
  - We can represent the hub, authority or page rank values at any time k of the process by an n-dimensional (column) vector, denoted (respectively) by $\mathbf{h}^{(k)}, \mathbf{a}^{(k)}, \mathbf{r}^{(k)}$.
- Boldface $\mathbf{v} =< v_1, \ldots, v_n >^T$ represents a vector whose components are the v_j
  - e.g., $r_j^{(k)}$ represents the page rank of the j^{th} web page after k steps of the page rank process

Advanced material (section 14.6): Handwaving argument why these processes converge

Both the page rank and hubs and authorities processes can be understood as a linear transformation

- Consider a web network of n pages
  - We can represent the hub, authority or page rank values at any time k of the process by an n-dimensional (column) vector, denoted (respectively) by $\mathbf{h}^{(k)}, \mathbf{a}^{(k)}, \mathbf{r}^{(k)}$.
- Boldface $\mathbf{v} =< v_1, \ldots, v_n >^T$ represents a vector whose components are the v_j
  - e.g., $r_j^{(k)}$ represents the page rank of the j^{th} web page after k steps of the page rank process

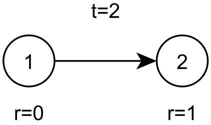


$$\mathbf{r}^{(1)} =< 0.7, 0.3 >^T$$

Advanced material (section 14.6): Handwaving argument why these processes converge

Both the page rank and hubs and authorities processes can be understood as a linear transformation

- Consider a web network of n pages
  - We can represent the hub, authority or page rank values at any time k of the process by an n-dimensional (column) vector, denoted (respectively) by $\mathbf{h}^{(k)}, \mathbf{a}^{(k)}, \mathbf{r}^{(k)}$.
- Boldface $\mathbf{v} =< v_1, \ldots, v_n >^T$ represents a vector whose components are the v_j
  - e.g., $r_j^{(k)}$ represents the page rank of the j^{th} web page after k steps of the page rank process



$$\mathbf{r}^{(2)} =< 0, 1 >^T$$

Advanced material (section 14.6): Handwaving argument why these processes converge

Let **v** be any of the hub, authority or page rank vectors

- PageRank & Hubs and Authorities can both be viewed as a linear transformation $\mathbf{v}^{(k+1)} = M\mathbf{v}^{(k)}$
- M is an appropriate $n \times n$ matrix, whose entries are non negative real numbers
  - Why $n \times n$?

Advanced material (section 14.6): Handwaving argument why these processes converge

Let **v** be any of the hub, authority or page rank vectors

- PageRank & Hubs and Authorities can both be viewed as a linear transformation $\mathbf{v}^{(k+1)} = M\mathbf{v}^{(k)}$
- M is an appropriate $n \times n$ matrix, whose entries are non negative real numbers
  - Why $n \times n$? We have n webpages, each has a value before & after the update

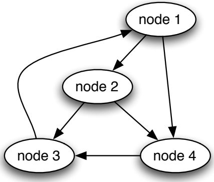
Advanced material (section 14.6): Handwaving argument why these processes converge

Let $\mathbf{v}$ be any of the hub, authority or page rank vectors

- PageRank & Hubs and Authorities can both be viewed as a linear transformation $\mathbf{v}^{(k+1)} = M\mathbf{v}^{(k)}$
- M is an appropriate $n \times n$ matrix, whose entries are non negative real numbers
  - Why $n \times n$? We have n webpages, each has a value before & after the update
- Furthermore, there are conditions that will guarantee the convergence of the process
  - i.e., when there exists $\mathbf{v}^{(*)} = \lim_{k\to\infty} \mathbf{v}^{(k)}$ and when this limit vector $\mathbf{v}^{(*)}$ is unique and independent of the starting vector $\mathbf{v}^{(0)}$

Advanced material continued: page rank convergence

- Figure 14.3 of the text illustrates a simple directed graph and the matrix N that defines the unscaled page rank update process. That is, $\mathbf{r}^{(\mathbf{k+1})} = N^T \mathbf{r}^{(\mathbf{k})}$ where N^T is the transpose of matrix N.



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

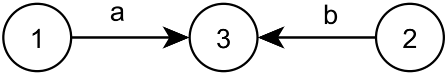
[Fig 14.13, E&K]

Figure: A toy web graph and the associated matrix N describing the unscaled update process.

- N is a weighted adjacency matrix
  - Weights are the proportion of authority that's transferred along the edge

Advanced material continued: page rank convergence

- Note: N_{ij} is the proportion of node i's rank, that should go to node j in the next update
  - Therefore, N^T_{ij} is the proportion of j's rank that i should receive

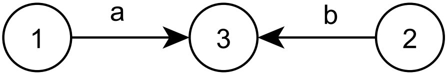


- Let's focus on node 3
- We can see that $r_3^{k+1} = ar_1^k + br_2^k + r_3^k$

$$N = \begin{bmatrix} * & * & a \\ * & * & b \\ * & * & 1 \end{bmatrix}$$

Advanced material continued: page rank convergence

- Note: N_{ij} is the proportion of node i's rank, that should go to node j in the next update
  - Therefore, N^T_{ij} is the proportion of j's rank that i should receive

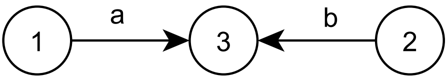


- Let's focus on node 3
- We can see that $r_3^{k+1} = ar_1^k + br_2^k + r_3^k$

$$N = \begin{bmatrix} * & * & a \\ * & * & b \\ * & * & 1 \end{bmatrix} \quad N^T = \begin{bmatrix} * & * & * \\ * & * & * \\ a & b & 1 \end{bmatrix}$$

Advanced material continued: page rank convergence

- Note: N_{ij} is the proportion of node i's rank, that should go to node j in the next update
  - Therefore, N^T_{ij} is the proportion of j's rank that i should receive

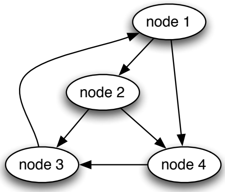


- Let's focus on node 3
- We can see that $r_3^{k+1} = ar_1^k + br_2^k + r_3^k$

$$N = \begin{bmatrix} * & * & a \\ * & * & b \\ * & * & 1 \end{bmatrix} \quad N^T\mathbf{r^k} = \begin{bmatrix} * & * & * \\ * & * & * \\ a & b & 1 \end{bmatrix} \begin{bmatrix} r_1^k \\ r_2^k \\ r_3^k \end{bmatrix} = \begin{bmatrix} * \\ * \\ ar_1^k + br_2^k + r_3^k \end{bmatrix}$$

Page rank analysis for the scaled update

Similarly Figure 14.4 illustrates the same graph and the matrix $\tilde{N}$ that defines the scaled page rank update process with scaling factor $s = .8$.



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

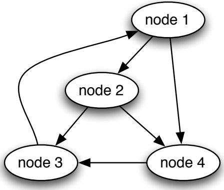
[Fig 14.14, E&K]

Figure: The same toy web graph and the associated matrix $\tilde{N}$ describing the scaled update process with $s = 0.8$.

- $\tilde{N}^T = sN^T + (1-s)\frac{\mathbf{1}(\mathbf{1}^T)}{n}$

Page rank analysis for the scaled update

Similarly Figure 14.4 illustrates the same graph and the matrix $\tilde{N}$ that defines the scaled page rank update process with scaling factor $s = .8$.



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

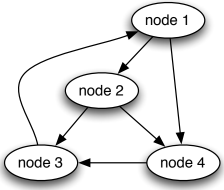
[Fig 14.14, E&K]

Figure: The same toy web graph and the associated matrix $\tilde{N}$ describing the scaled update process with $s = 0.8$.

- $\tilde{N}^T = sN^T + (1-s)\frac{\mathbf{1}(\mathbf{1}^T)}{n}$
- It follows that $\mathbf{r^{(k)}} = (\tilde{N}^T)^k \mathbf{r^{(0)}}$

Page rank analysis for the scaled update

Similarly Figure 14.4 illustrates the same graph and the matrix $\tilde{N}$ that defines the scaled page rank update process with scaling factor $s = .8$.



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

[Fig 14.14, E&K]

Figure: The same toy web graph and the associated matrix $\tilde{N}$ describing the scaled update process with $s = 0.8$.

- $\tilde{N}^T = sN^T + (1-s)\frac{\mathbf{1}(\mathbf{1}^T)}{n}$
- It follows that $\mathbf{r}^{(\mathbf{k})} = (\tilde{N}^T)^k \mathbf{r}^{(\mathbf{0})}$
- If the process is converging then it would be converging to some $\mathbf{r}^*$ satisfying $\mathbf{r}^* = \tilde{N}^T \mathbf{r}^*$

Now comes the necessary linear algebra

Matrices are more than a convenient representation of the process; linear algebra review time!

- Let $M_{n \times n}$ be a full rank matrix. Recall that the matrix-vector multiplication $M\mathbf{v}$ can rotate and expand/shrink the vector $\mathbf{v}$
  - Simpler to visualize the meaning of such a linear transformation in 2-space or 3-space

Now comes the necessary linear algebra

Matrices are more than a convenient representation of the process; linear algebra review time!

- Let $M_{n \times n}$ be a full rank matrix. Recall that the matrix-vector multiplication $M\mathbf{v}$ can rotate and expand/shrink the vector $\mathbf{v}$
  - Simpler to visualize the meaning of such a linear transformation in 2-space or 3-space
- A vector $\mathbf{v}$ is an eigenvector of M with associated eigenvalue λ if $M\mathbf{v} = \lambda\mathbf{v}$
  - Note that $\mathbf{v}$ is also an eigenvector of M^k with eigenvalue λ^k

Now comes the necessary linear algebra

Matrices are more than a convenient representation of the process; linear algebra review time!

- Let $M_{n \times n}$ be a full rank matrix. Recall that the matrix-vector multiplication $M\mathbf{v}$ can rotate and expand/shrink the vector $\mathbf{v}$
  - Simpler to visualize the meaning of such a linear transformation in 2-space or 3-space
- A vector $\mathbf{v}$ is an eigenvector of M with associated eigenvalue λ if $M\mathbf{v} = \lambda\mathbf{v}$
  - Note that $\mathbf{v}$ is also an eigenvector of M^k with eigenvalue λ^k
- Question: What does this tell us about the equilibrium of the process (i.e., $\mathbf{v}$ s.t. $M\mathbf{v} = \mathbf{v}$)?

Now comes the necessary linear algebra

Matrices are more than a convenient representation of the process; linear algebra review time!

- Let $M_{n \times n}$ be a full rank matrix. Recall that the matrix-vector multiplication $M\mathbf{v}$ can rotate and expand/shrink the vector $\mathbf{v}$
  - Simpler to visualize the meaning of such a linear transformation in 2-space or 3-space
- A vector $\mathbf{v}$ is an eigenvector of M with associated eigenvalue λ if $M\mathbf{v} = \lambda\mathbf{v}$
  - Note that $\mathbf{v}$ is also an eigenvector of M^k with eigenvalue λ^k
- Question: What does this tell us about the equilibrium of the process (i.e., $\mathbf{v}$ s.t. $M\mathbf{v} = \mathbf{v}$)?
- When $\lambda = 1$, the eigenvector is an equilibrium of the process!

More linear algebra

- For some matrices there is a set of n eigenvectors with (not necessarily distinct) associated eigenvalues $\lambda_1, \ldots, \lambda_n$; these eigenvectors span the n-dimensional Euclidean space so that any vector can be expressed as a linear combination of the eigenvectors

More linear algebra

- For some matrices there is a set of *n* eigenvectors with (not necessarily distinct) associated eigenvalues $\lambda_1, \ldots, \lambda_n$; these eigenvectors span the *n*-dimensional Euclidean space so that any vector can be expressed as a linear combination of the eigenvectors
- Perron's Theorem states that for any matrix which has all positive entries there is:
  - A unique positive eigenvector **y** that corresponds to the largest positive eigenvalue, λ_1

More linear algebra

- For some matrices there is a set of n eigenvectors with (not necessarily distinct) associated eigenvalues $\lambda_1, \ldots, \lambda_n$; these eigenvectors span the n-dimensional Euclidean space so that any vector can be expressed as a linear combination of the eigenvectors
- Perron's Theorem states that for any matrix which has all positive entries there is:
  - A unique positive eigenvector $\mathbf{y}$ that corresponds to the largest positive eigenvalue, λ_1
  - Furthermore $\lambda_1 > |\lambda_i|$ for $i > 1$.

More linear algebra

- For some matrices there is a set of n eigenvectors with (not necessarily distinct) associated eigenvalues $\lambda_1, \ldots, \lambda_n$; these eigenvectors span the n-dimensional Euclidean space so that any vector can be expressed as a linear combination of the eigenvectors
- Perron's Theorem states that for any matrix which has all positive entries there is:
  - A unique positive eigenvector $\mathbf{y}$ that corresponds to the largest positive eigenvalue, λ_1
  - Furthermore $\lambda_1 > |\lambda_i|$ for $i > 1$.
- Since $\lambda_1 > |\lambda_i|$ for $i > 1$, and since every vector is a linear combination of the eigenvectors, it follows that as $k \to \infty$, the transformation M^k is being dominated by the largest eigenvalue acting on its associated eigenvector

More linear algebra

- For some matrices there is a set of n eigenvectors with (not necessarily distinct) associated eigenvalues $\lambda_1, \ldots, \lambda_n$; these eigenvectors span the n-dimensional Euclidean space so that any vector can be expressed as a linear combination of the eigenvectors
- Perron's Theorem states that for any matrix which has all positive entries there is:
  - A unique positive eigenvector **y** that corresponds to the largest positive eigenvalue, λ_1
  - Furthermore $\lambda_1 > |\lambda_i|$ for $i > 1$.
- Since $\lambda_1 > |\lambda_i|$ for $i > 1$, and since every vector is a linear combination of the eigenvectors, it follows that as $k \to \infty$, the transformation M^k is being dominated by the largest eigenvalue acting on its associated eigenvector
  - Time for a demo!

More linear algebra

- For some matrices there is a set of n eigenvectors with (not necessarily distinct) associated eigenvalues $\lambda_1, \ldots, \lambda_n$; these eigenvectors span the n-dimensional Euclidean space so that any vector can be expressed as a linear combination of the eigenvectors
- Perron's Theorem states that for any matrix which has all positive entries there is:
  - A unique positive eigenvector $\mathbf{y}$ that corresponds to the largest positive eigenvalue, λ_1
  - Furthermore $\lambda_1 > |\lambda_i|$ for $i > 1$.
- Since $\lambda_1 > |\lambda_i|$ for $i > 1$, and since every vector is a linear combination of the eigenvectors, it follows that as $k \to \infty$, the transformation M^k is being dominated by the largest eigenvalue acting on its associated eigenvector
  - Time for a demo!
- For the scaled matrix $\tilde{N}^T$, all entries are positive and the largest eigenvalue is 1
  - Therefore as $k \to \infty$, $(\tilde{N}^T)^k\mathbf{v}$ will converge to the eigenvector $\mathbf{y}$ associated with the largest eigenvalue 1

Aside: Random Walk

- Remember that we said that PageRank was equivalent to a random walk on the graph?

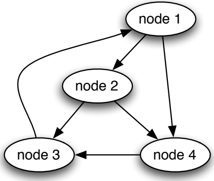
Aside: Random Walk

- Remember that we said that PageRank was equivalent to a random walk on the graph?
  - We can view this as a Markov chain
  - i.e., a type of probabilistic finite state machine that's represented as a graph, where each timestep we follow an edge based on the corresponding probabilities

Aside: Random Walk

- Remember that we said that PageRank was equivalent to a random walk on the graph?
  - We can view this as a Markov chain
  - i.e., a type of probabilistic finite state machine that's represented as a graph, where each timestep we follow an edge based on the corresponding probabilities
- Similarly, scaled page rank can be viewed as the same Markov chain but with added low probability edges between every pair of nodes

Aside: Random Walk

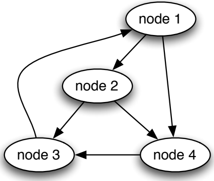


$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

[Fig 14.13, E&K]

- Specifically, the *transition matrices* of these Markov chains are N and $\tilde{N}$ respectively (N_{ij} is the probability of transitioning from state i to state j)

Aside: Random Walk

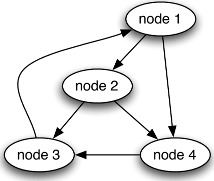


$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

[Fig 14.13, E&K]

- Specifically, the *transition matrices* of these Markov chains are N and $\tilde{N}$ respectively (N_{ij} is the probability of transitioning from state i to state j)
- If the chain is *irreducible* (we can reach state j from all states i and vice versa), and *aperiodic* (there is no state i such that if you leave i, you can only return on timesteps that are multiples of some $p > 1$) then there is a unique *stationary distribution* π that the chain converges to
- This is the distribution that we found!

Aside: Random Walk

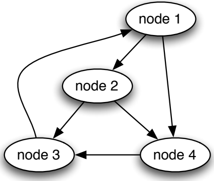


$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

[Fig 14.13, E&K]

- Specifically, the *transition matrices* of these Markov chains are N and $\tilde{N}$ respectively (N_{ij} is the probability of transitioning from state i to state j)
- If the chain is *irreducible* (we can reach state j from all states i and vice versa), and *aperiodic* (there is no state i such that if you leave i, you can only return on timesteps that are multiples of some $p > 1$) then there is a unique *stationary distribution* π that the chain converges to
- This is the distribution that we found!
- When π exists, then all initial probability distributions over the nodes converges to π

Aside: Random Walk



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

[Fig 14.13, E&K]

- Specifically, the *transition matrices* of these Markov chains are N and $\tilde{N}$ respectively (N_{ij} is the probability of transitioning from state i to state j)
- If the chain is *irreducible* (we can reach state j from all states i and vice versa), and *aperiodic* (there is no state i such that if you leave i, you can only return on timesteps that are multiples of some $p > 1$) then there is a unique *stationary distribution* π that the chain converges to
- This is the distribution that we found!
- When π exists, then all initial probability distributions over the nodes converges to π
- NetLogo time!

Similar analysis for hubs and authorities

- If M is the adjacency matrix of the web graph, then the (unnormalized) process can be described by $\mathbf{h} = M\mathbf{a}$ and $\mathbf{a} = M^T\mathbf{h}$.

- Exercise: Convince yourself this is true!

Similar analysis for hubs and authorities

- If M is the adjacency matrix of the web graph, then the (unnormalized) process can be described by $\mathbf{h} = M\mathbf{a}$ and $\mathbf{a} = M^T\mathbf{h}$.

- Exercise: Convince yourself this is true!

- Then
  1. $\mathbf{a}^{(1)} = M^T\mathbf{h}^{(0)}$
  2. $\mathbf{h}^{(1)} = M\mathbf{a}^{(1)} = MM^T\mathbf{h}^{(0)}$

- It follows that
  1. $\mathbf{a}^{(k)} = (M^TM)^{k-1}M^T\mathbf{h}^{(0)}$
  2. $\mathbf{h}^{(k)} = (MM^T)^k\mathbf{h}^{(0)}$

Hubs and authorities analysis continued

- The matrices (MM^T) and (M^TM) are symmetric and have non-negative entries
- Any $n \times n$ symmetric matrix S with non negative entries has an orthonormal set of n eigenvectors all of whose associated eigenvalues are real
    - By normalizing the scores, we assume that the largest eigenvalue $\lambda_1 = 1$

Hubs and authorities analysis continued

- The matrices (MM^T) and (M^TM) are symmetric and have non-negative entries
- Any $n \times n$ symmetric matrix S with non negative entries has an orthonormal set of n eigenvectors all of whose associated eigenvalues are real
  - By normalizing the scores, we assume that the largest eigenvalue $\lambda_1 = 1$
- If the largest eigenvalue is unique (which is what would happen in a real web graph), then the same analysis for page rank applies (assuming that the starting hub scores are all positive).

Returning to the issue of influence

In some sense or another we are often talking about social influence in this course. Even in web ranking (Ch 14), we can view hubs as influencing which Web pages will be ranked highly.

In chapter 18, we observed two sequential processes where previous individual decisions had a significant impact:
1) The evolution of links on the Web, and
2) The evolution of opinions in evaluating music.

The music evaluation experiment is closer to reality in the sense that it explicitly integrates a measure of quality (a simplification of selection?) into the decision making process.

Recap

With practice & review, you'll be able to:

- Define the power law distribution and Zipf's law, and explain the similarities and differences
  - Recall the dynamics that often give rise to power laws, give examples, and debate whether they apply to a given scenario
- Define the Kumar et al. rich-get-richer model
  - Explain the connection between the rich-get-richer model and dynamics that give rise to power laws
  - Recall the expected distribution, and explain the relevant parameters of the model
- Summarize the Salganik et al. music popularity experiment
- Explain the problem of ranking web results
  - Explain the hubs and authorities algorithm, and execute on examples
  - Explain the (scaled) Page rank algorithm, and execute on examples
  - Describe high-level proof-sketches of their convergence