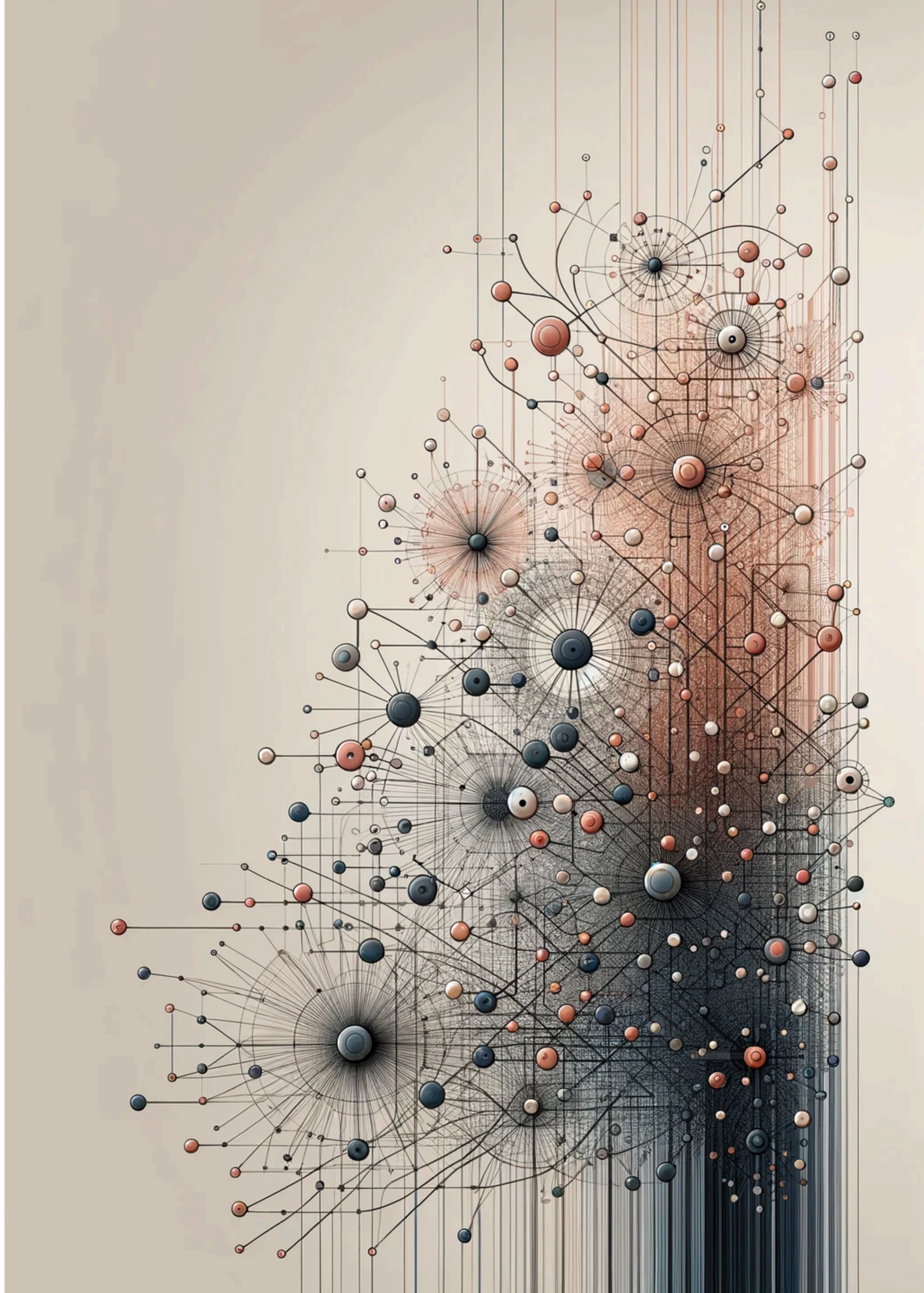
UNIVERSITÀ
DI TORINO

Analisi e Visualizzazione delle Reti Complesse

NS10 - Power Laws and Rich-Get-Richer Phenomena

Prof. Rossano Schifanella



Agenda

- Popularity as a Network Phenomenon
- Rich-Get-Richer Models
- The Unpredictability of Rich-Get-Richer Effects
- The Long Tail
- The Effect of Search Tools and Recommendation Systems

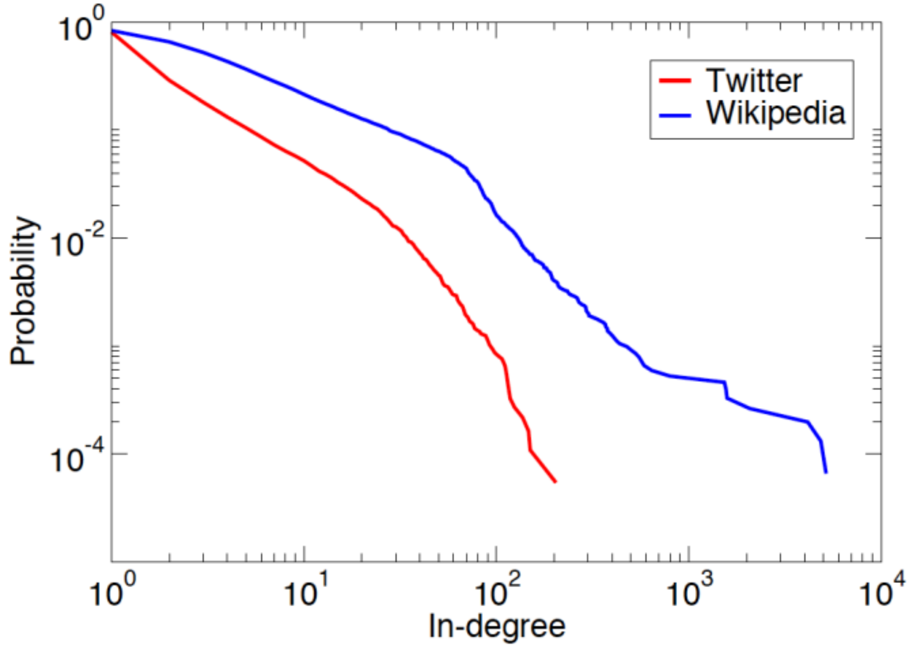
Popularity as a Network Phenomenon

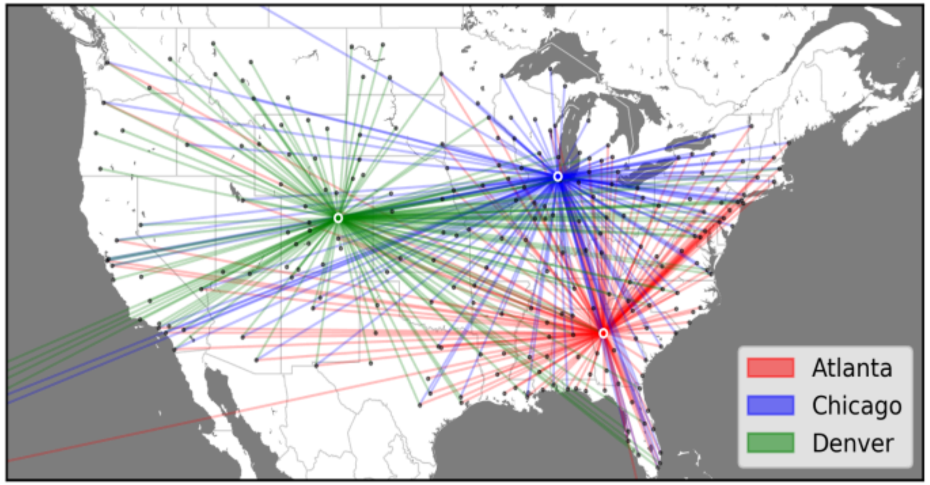
Popularity, heterogeneity and networks

Recap:

- Real networks are **heterogeneous**
- **In-links as a measure of popularity**
- The **heterogeneity** parameter as a **measure of distribution's broadness**

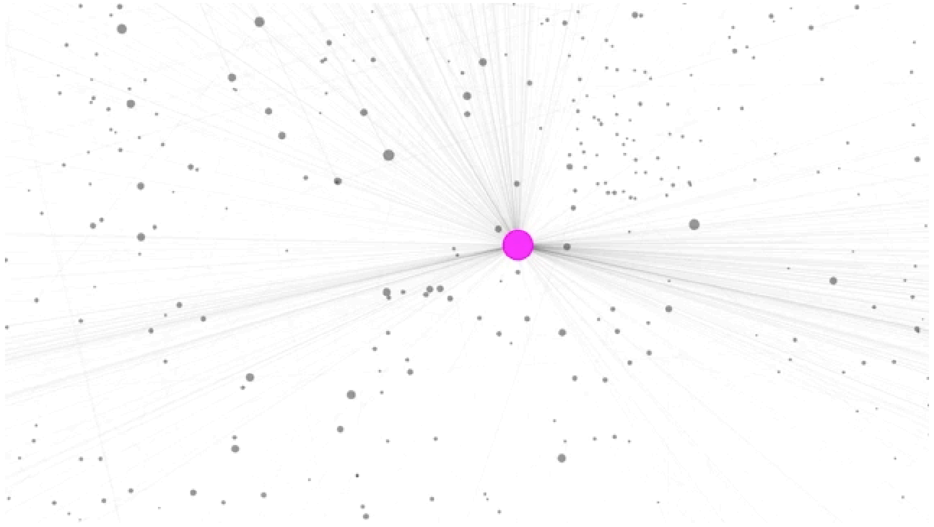
$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$$





Case study: the Web

- Characterizing popularity reveals imbalances (inequalities)
  - almost everyone is popular for very few people
  - very few people achieve high popularity
  - very, very few people achieve global popularity
- Why? Is this phenomenon intrinsic to the whole idea of popularity itself?
  - Have a look at the video for a dynamic view



Why do hubs emerge?

- Let's accept that power laws represent many phenomena
- Why?
  - We are observing a kind of "order" emerging from chaos
  - Is there an underlying process that keeps the line of the log-log fit so straight?
  - Like normal distributions arise from many independent random decisions averaging out, can we find something similar in this context?
    - we will find that power laws arise from the feedback introduced by correlated decisions across a population

Understanding Rich-Get-Richer Models

- **Core principle**: New nodes tend to connect to already popular nodes ("the rich get richer")
- **Fundamental assumption**: People tend to copy decisions made by others who acted before them
- This creates a powerful **feedback mechanism** that amplifies initial advantages

Rich-Get-Richer Models: A Framework

- Nodes (e.g., pages) are created sequentially: $1, 2, \ldots, N$
- For each new node j that joins the network:
  - With probability p: **random attachment** (uniform selection)
  - With probability $(1 - p)$: **copying mechanism**

This elegantly captures how individuals balance:

- **Exploration** (finding new options)
- **Exploitation** (copying successful precedents)

Why Copying Creates Rich-Get-Richer

When new node j uses the copying mechanism:

1. It randomly selects an existing node i (uniform probability $\frac{1}{j-1}$)
2. Then connects to wherever i points to (node l)

This naturally produces preferential attachment because:

- If node l has in-degree k_l, then k_l different nodes point to it
- The probability node l receives this new link is $\frac{k_l}{j-1}$

Without requiring global knowledge of popularity, this simple copying behavior produces the rich-get-richer effect where:

$$P(\text{node } j \text{ connects to } l) \propto k_l$$

The Unpredictability of Rich-Get-Richer Effects

The fragility of popularity

- Power laws are produced by **feedback effects**
- The initial stages of the process that gives rise to popularity is a **relatively fragile state**
- Focusing on **cultural market**:
    - **Can we predict the popularity of a song, a movie, a book, etc.?**
- We can expect initial fluctuations
    - this brings **unpredictability**

Predicting hubs emergence

- We can predict that a **power law** and **hubs will emerge**
- **But which hubs?**
    - **Predicting the success of an individual item is not like predicting that some individual will have global success!**

The MusicLab experiment

Science   Contents   News   Careers   Journals

SHARE

REPORT

Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market

Matthew J. Salganik[1,2,*], Peter Sheridan Dodds[2,*], Duncan J. Watts[1,2,3,*]

+ See all authors and affiliations

Science 10 Feb 2006:
Vol. 311, Issue 5762, pp. 854-856
DOI: 10.1126/science.1121066

[The MusicLab experiment paper] [pdf]

UNIVERSITÀ DI TORINO

The MusicLab Experiment

- Study conducted by Salganik, Dodds, and Watts (2006) published in Science
- **Key question**: How much is popularity determined by quality versus social influence?
- **Experimental design**:
  - Web-based music market with 14,341 participants
  - 48 songs by unknown bands available for listening and downloading
  - Participants could listen, rate, and download songs
  - Manipulated visibility of others' choices across different "worlds"

Experimental Design: The Different Worlds

- Participants randomly assigned to one of **nine parallel "worlds"**:

1. **Independent world** (control group):
   - No social influence - participants saw only song names and bands
   - No information about others' downloads
2. **Social influence worlds** (eight separate worlds):
   - Identical initial conditions
   - Participants could see how many times each song had been downloaded
   - Songs were displayed in order of popularity (most downloaded at top)

Key Innovations of the Experiment

- First large-scale **controlled experiment** on cultural market dynamics
- Created **multiple independent "histories"** of the same market
- Allowed researchers to:
  - Observe how the same pool of songs performed in different "worlds"
  - Measure the impact of social influence on popularity
  - Quantify the **unpredictability** of success

Key Findings

The experiment revealed three distinct categories of outcomes:

1. **High-Quality Winners**: A small subset of songs consistently performed well across all worlds
    - These songs possessed intrinsic qualities that transcended social influence effects
    - Demonstrates a threshold effect where truly exceptional content resists randomness
2. **Low-Quality Underperformers**: Some songs consistently failed regardless of social context
    - Even positive social signals couldn't rescue fundamentally weak content
    - Suggests a quality floor below which social influence is ineffective
3. **The Unpredictable Middle**: The majority of songs (~70%) showed highly variable success
    - Neither consistently good nor bad, but subject to path-dependent dynamics
    - Early random fluctuations became amplified through feedback loops
    - **These songs demonstrate how small initial perturbations in nonlinear systems can lead to divergent trajectories**

The MusicLab: Experimental Results

- **Inter-world correlation analysis revealed true unpredictability**
  - Success correlation between worlds was only 0.27 (vs. expected ~0.90 for quality-driven markets)
  - For middle-quality songs, correlations dropped to near-random levels (below 0.15)
  - This indicates that **success was not primarily determined by inherent quality**
  - Same songs had dramatically different outcomes in different worlds
- **Inequality metrics showed social influence amplified popularity differences**
  - Social influence worlds: Gini coefficient = 0.31
  - Independent world (no social signals): Gini coefficient = 0.20
  - Top songs in social influence worlds received disproportionately more downloads
  - Market share prediction error was 2.1× higher in social influence worlds
  - Early random fluctuations became self-reinforcing through feedback
- **"Success breeds success" dynamics quantified:**
  - First 100 listens significantly predicted final market shares ($p < 0.01$)
  - Early advantages persisted and amplified throughout the experiment

MusicLab: Influence Strength and Implications

- **Variable influence experiment demonstrated causality**
    - Researchers created additional worlds with different levels of social signals:
        - Strong influence: Complete download counts visible
        - Weak influence: Download counts shown as coarse categories
        - No influence: Download information hidden
    - **Results scaled systematically with influence strength:**
        - Inequality (Gini): 0.31 → 0.24 → 0.20 as influence decreased
        - Unpredictability (σ^2): 0.0088 → 0.0064 → 0.0033 as influence decreased

MusicLab: Influence Strength and Implications

- **Key theoretical implications**
  - Empirically validates rich-get-richer mechanisms in cultural markets
  - Shows how simple copying behavior creates power-law distributions
  - Demonstrates why predicting individual outcomes is fundamentally limited
  - **Quality matters, but its influence is mediated by social dynamics**
  - Even experts would struggle to predict hits due to inherent path dependency

The Long Tail

The Long Tail Concept

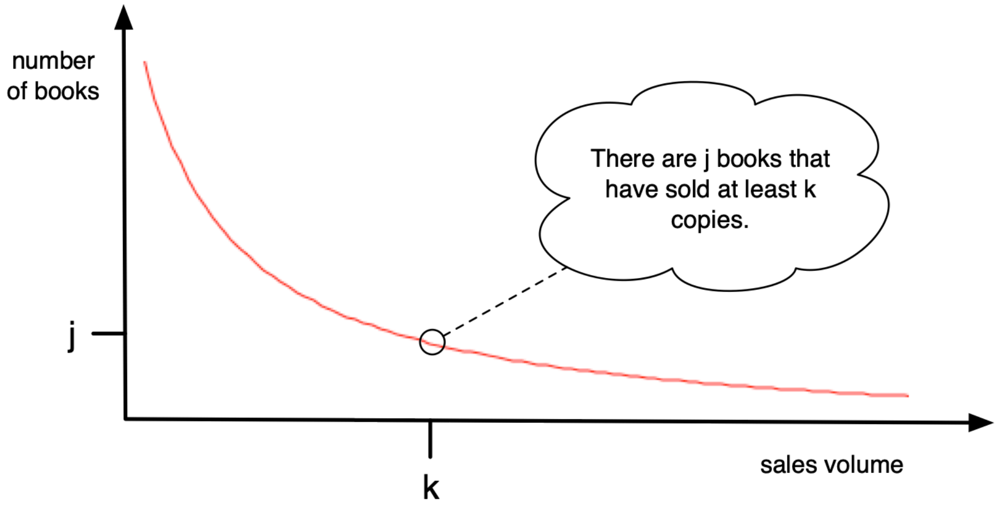
- **Definition**: The strategy of selling a large number of unique items with relatively small quantities sold of each
- **Traditional business model**: Focus on "hits" (high-volume products)
  - Example: Bookstores stocking bestsellers
- **Long Tail business model**: Capture the total value of countless niche markets
  - Example: Amazon offering millions of books that physical stores can't stock
- **Key insight**: In many markets, the accumulated sales/popularity of niche products can exceed that of hits

Looking at Two Different Perspectives

- **Hits vs. Niches**: Two ways to visualize the same distribution
- **The key question**: Is there more value in the head or the tail?
  - Are most sales generated by a small set of enormously popular items?
  - Or by a much larger population of items that are individually less popular?
- **Chris Anderson's argument**: The internet economy fundamentally changes the equation
  - Digital distribution reduces inventory costs
  - Search and recommendation systems help consumers find niche products
  - This unlocks tremendous value previously hidden in the "long tail"

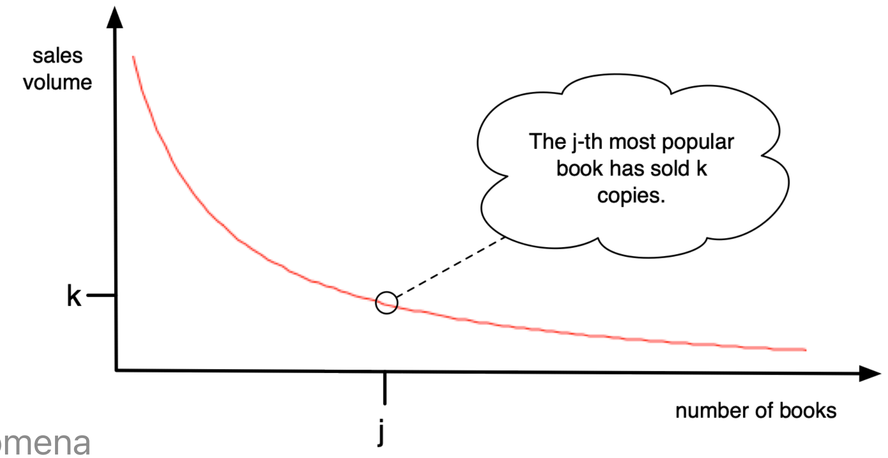
Focus on Hits: The Traditional View

- Stereotype of traditional media business focused on **hits**
- Power law visualization: items ranked by popularity
  - X-axis: Product rank
  - Y-axis: Number of sales/downloads/views
- In this view, popular products appear dominant
  - The "head" of the distribution looks significant
  - Tail appears to quickly drop to negligible levels



Focus on Niches: The Long Tail View

- **Changing perspective**: Switch the axes to reveal the long tail
- Now plotting:
  - X-axis: Product popularity (sales volume)
  - Y-axis: Number of products at each popularity level
- This reveals the massive number of niche products
  - While each sells little individually, collectively they represent substantial value
  - The "long tail" extends far to the right



The Long Tail and Digital Markets

- **Digital technologies enable the long tail by**:
  - **Eliminating physical constraints**: Virtually unlimited "shelf space" online
  - **Reducing discovery costs**: Search and recommendation systems connect users to niche content
  - **Lowering production costs**: More creators can produce and distribute content
- **Economic implications**:
  - Markets becoming less hit-driven
  - Diversity of products increasing
  - Consumer tastes fragmenting into micro-cultures
- **Examples**: Netflix, Spotify, Amazon, YouTube – all leverage the long tail as major part of their business model

The Long Tail and Rich-Get-Richer Dynamics

- **Seemingly contradictory phenomena**:
  - Rich-get-richer processes create extreme popularity inequality
  - Yet the long tail creates significant value from non-hits
- **Resolving the paradox**:
  - Rich-get-richer explains the shape of the distribution (power law)
  - The long tail focuses on the aggregate value hidden in that distribution
  - Both can coexist in the same markets
- **Key insight**: In power law distributions with exponents near 2, both the head and tail contain substantial value

The Effect of Search Tools and Recommendation Systems

Search and Recommendation in the Age of Information Abundance

- **Central paradox**: Search engines and recommendation systems simultaneously **amplify and counteract** rich-get-richer dynamics
- These technologies serve as **gatekeepers of attention** in our digital environment
- Key question: Do they democratize access to content or reinforce existing popularity hierarchies?

How Search Tools Amplify Rich-Get-Richer Dynamics

The Feedback Loop of Search Rankings

- **PageRank and popularity-based algorithms**:
  - Search engines often rank results based partly on existing popularity metrics
  - Higher-ranked pages receive more visibility → more clicks → more links → higher rankings
  - Creates a self-reinforcing cycle that benefits already-popular content
- **Example**: Google's original innovation was using link structure to determine importance
  - Pages with many inbound links rank higher
  - This algorithmically implements a rich-get-richer mechanism
- **Empirical evidence**: Top search results receive disproportionate attention
  - ~60% of clicks go to top three results
  - ~90% of users never go beyond the first page

The Attention Economy's Matthew Effect

- **Winner-takes-all dynamics**:
  - Limited user attention + abundance of content = extreme competition
  - Search engines create "superstar economics" where top results capture most value
- **Position bias studies**:
  - Users trust higher-ranked results even when rankings are deliberately inverted
  - Experiments show quality perceptions follow position, not actual content quality
- **Traffic concentration**:
  - Studies show increasing concentration of web traffic to dominant platforms
  - Top 10 websites now capture >50% of all internet traffic (vs. ~30% in 2010)

How Search Tools Counteract Rich-Get-Richer Effects

- **Specialized queries enable niche content discovery**:
  - Long-tail search queries (3+ words) now represent >70% of all searches
  - Specific queries bypass popularity contests, connecting users directly to relevant content
  - This creates "micro-markets" where smaller players can compete
- **Personalization reduces popularity biases**:
  - Modern search engines consider user history, location, and preferences
  - Results vary by individual, reducing the universal advantage of popular content
  - Creates multiple parallel "popularity contests" instead of one global competition
- **Example**: Small local businesses can outrank global corporations for location-specific queries

Collaborative Filtering's Inherent Biases

- **Traditional recommendation approaches**:
    - "People who liked X also liked Y" inherently favors popular items
    - Items with more interactions receive more recommendations
    - New or niche content faces "cold start" problem
- **Popularity bias in practice**:
    - Netflix study: Recommender increased plays of popular movies by 30%
    - Amazon's early recommendation systems showed 30-40% bias toward bestsellers

Countermeasures and Evolution

- **Modern recommendation algorithms increasingly incorporate**:
    - **Diversity metrics** to avoid recommendation homogeneity
    - **Novelty factors** that reward exposing users to new content
    - **Serendipity parameters** that intentionally surface unexpected items

Case Studies: Rich-Get-Richer in Digital Platforms

YouTube's Recommendation Engine

- **Initial design**: Heavily optimized for engagement (watch time)
  - Led to increasing concentration of views among top creators
  - Created "superstar economy" where top 3% of channels received 90% of views
- **Recent evolution**: Introduction of "Up-and-coming Creator" recommendations
  - Deliberately surfaces newer channels with potential
  - Creates "on-ramps" to counteract incumbent advantages

Spotify's Discovery Weekly

- **Balanced approach**:
  - Algorithm specifically designed to surface music from both established and emerging artists
  - Combines popularity signals with novelty and personalization
  - Has become a significant pathway for independent artists to gain attention
- **Reported outcomes**:
  - 40% of Spotify streams now come from algorithmic recommendations
  - Over 16 billion discoveries of new artists facilitated annually

Exercise: Analyzing Rich-Get-Richer Phenomena in Web Graphs

Part 1: Data Collection and Initial Analysis

- Download a Web graph sample from SNAP Stanford Web Graphs or WebGraph datasets
- Create a directed graph using NetworkX or a similar library
- Calculate the in-degree distribution (number of pages that point to each page)

Part 2: Comparing Web Graph to Random Models

- Generate an Erdős–Rényi random graph with the same number of nodes and edges
- Plot the in-degree distributions of both graphs on a log-log scale
- Calculate the heterogeneity parameter κ for both networks
- Does the real web graph show significantly more heterogeneity than the random graph? Why?

Part 3: Testing for Power Law Properties

- Use statistical tools (e.g., powerlaw Python package) to fit the web graph's in-degree distribution
- Calculate the power law exponent α and test the goodness of fit
- Generate a synthetic network using preferential attachment (Barabási–Albert model)
- Compare how well each model (random vs. preferential attachment) approximates the real data

Part 4: Temporal Analysis (Optional Extension)

- If available, use a temporal web dataset to analyze how links evolve over time
- Track the growth rate of high-degree nodes vs. low-degree nodes
- Calculate the probability that new links attach to nodes based on their current degree
- Does this empirically confirm the "rich get richer" principle?
- **Academic temporal datasets**:
  - Stanford SNAP Temporal Networks
  - KONECT Dynamic Networks
  - Network Repository Temporal Graphs

Reading material

[ns2] Chapter 18 (18.1 - 18.7) Power Laws and Rich-Get-Richer Phenomena

Q&A