**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

-----🙞🙜🕮🙞🙜-----



**BÁO CÁO  
BỘ MÔN: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: “**POWER LAWS AND RICH-GET-RICHER PHENOMENA**”**

*(Luật số lớn và hiện tượng giàu có hơn)*

NHÓM 8:

Phạm Thị Thảo Nguyên - 17020937

Hà Đức Hiệp - 17020732

Lê Hồng Phong – 17020952

Nguyễn Thị Lê – 17020847

Lê Thị Thùy Linh – 17020854

Nguyễn Vũ Tùng Dương - 17020657

[I. Network Dynamics: Population Models: 3](#_Toc28439040)

[1. Giới thiệu chung: 3](#_Toc28439041)

[2. Information Cascades (Tầng thông tin): 4](#_Toc28439042)

[3. Network Effects (Hiệu ứng mạng): 4](#_Toc28439043)

[II. Power Laws and Rich-Get-Richer Phenomena (Luật số lớn và hiện tượng giàu có hơn): 6](#_Toc28439044)

[1. Popularity as a Network Phenomenon (Phổ biến như một hiện tượng mạng): 6](#_Toc28439045)

[2. Power Laws (Luật số lớn): 8](#_Toc28439046)

[3. Rich-Get-Richer Models (Mô hình giàu có hơn): 11](#_Toc28439047)

[4. The Unpredictability of Rich-Get-Richer Effects (Tính không thể dự đoán trước của hiệu ứng giàu có hơn): 14](#_Toc28439048)

[5. The Long Tail (Thị trường ngách): 16](#_Toc28439049)

[6. The Effect of Search Tools and Recommendation Systems (Ảnh hưởng của công cụ tìm kiếm và hệ tư vấn): 20](#_Toc28439050)

[7. Advanced Material: Analysis of Rich-Get-Richer Processes (Phân tích quy trình Rich-Get-Richer): 21](#_Toc28439051)

# Network Dynamics: Population Models:

## Giới thiệu chung:

Network (Mạng máy tính) là [mạng viễn thông](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_vi%E1%BB%85n_th%C3%B4ng) [kỹ thuật số](https://vi.wikipedia.org/w/index.php?title=T%C3%ADn_hi%E1%BB%87u_k%C4%A9_thu%E1%BA%ADt_s%E1%BB%91&action=edit&redlink=1) cho phép [các nút mạng](https://vi.wikipedia.org/wiki/N%C3%BAt_m%E1%BA%A1ng) chia sẻ tài nguyên. Trong các mạng máy tính, [các thiết bị máy tính](https://vi.wikipedia.org/wiki/M%C3%A1y_t%C3%ADnh) [trao đổi dữ liệu](https://vi.wikipedia.org/wiki/Truy%E1%BB%81n_d%E1%BB%AF_li%E1%BB%87u) với nhau bằng các kết nối ([liên kết dữ liệu](https://vi.wikipedia.org/w/index.php?title=Li%C3%AAn_k%E1%BA%BFt_d%E1%BB%AF_li%E1%BB%87u&action=edit&redlink=1)) giữa các nút. Các liên kết dữ liệu này được thiết lập qua [cáp mạng](https://vi.wikipedia.org/w/index.php?title=C%C3%A1p_m%E1%BA%A1ng&action=edit&redlink=1) như dây hoặc [cáp quang](https://vi.wikipedia.org/wiki/C%C3%A1p_quang) hoặc [phương tiện không dây](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_kh%C3%B4ng_d%C3%A2y) như [Wi-Fi](https://vi.wikipedia.org/wiki/Wi-Fi).

Phân loại mạng:

* Đồng bộ
* Không đồng bộ
* Tĩnh
* Động (Dynamic)

Network Dynamics: là một lĩnh vực nghiên cứu các mạng có trạng thái thay đổi theo thời gian. Các biến động có thể đề cập đến cấu trúc của các kết nối của các đơn vị mạng, đến các trạng thái bên trong của mạng, hoặc cả hai. Các hệ thống mạng có thể có trong các lĩnh vực: sinh học, vật lý, xã hội học, kinh tế học, khoa học máy tính…hệ thống mạng thường được mô tả như là hệ thống phức tạp bao gồm nhiều đơn vị, có khả năng thay đổi, cấu trúc liên kết tương tác.

Population Models: mô hình dân số là mô hình toán học được áp dụng để nghiên cứu biến động dân số.

Network Dynamics: Population Models là một hiện tượng dân số tin vào những điều phổ biến mặc dù không biết những điều đó là đúng hay sai.

**Ví dụ**: các trang web truyền thông xã hội có xu hướng cung cấp cho người dùng những tin tức tương tự như các phương tiện truyền thông mà người dùng đã đọc trước đây. Đây cũng chính là sự phổ biến như một hiện tượng mạng và mô hình giàu có hơn. Nếu một người phân tích mức độ phổ biến (tức là có bao nhiêu độc giả) của các bài báo trực tuyến khác nhau, thì người ta có thể sẽ tìm thấy một luật số lớn ở dữ liệu. Lý do luật số lớn xảy ra là hiện tượng giàu có hơn. Điều này có nghĩa là mặc dù một số bài báo tin tức không nhất thiết phải có uy tín, nhưng chúng đủ phổ biến, được biết đến nhiều, thì chúng sẽ tiếp tục thu hút được độc giả và ngày càng trở nên phổ biến hơn. Bảng xếp hạng của Google cung cấp một ví dụ về hiện tượng giàu có hơn. Vì đó là kết quả phổ biến nhất của một truy vấn tìm kiếm sẽ tiếp tục được phổ biến hay được đọc nhiều hơn. Các các bài viết không phổ biến hoặc ít hơn sẽ xếp sau các trang đó.

Theo nghiên cứu, đa số mọi người sẽ tin vào những bài thông tin xuất hiện đầu tiên trên Google, nhưng thực chất thứ hạng đầu trên Google không làm cho bài viết trở nên đáng tin cậy, vì thứ hạng đó được xác định bởi một số yếu tố bao gồm mức độ phổ biến

.

## Information Cascades (Tầng thông tin):

Là một hiện tượng được mô tả trong hành vi kinh tế học ([behavioral economics](https://en.wikipedia.org/wiki/Behavioral_economics)) và lý thuyết mạng (network theory) đó là hiện tượng một số người đưa ra quyết định tương tự theo kiểu liên tiếp với các quyết định của những người khác đã đưa ra trước đó.

Tầng thông tin là quá trình hai bước:

* Một cá nhân phải đưa ra một quyết định, đó là một quyết định nhị phân.
* Có các yếu tố bên ngoài có thể ảnh hưởng đến quyết định này.

**Ví dụ**: Có nên áp dụng một công nghệ mới, mặc một kiểu quần áo mới…

Quyết định có thể là đồng ý hoặc từ chối. Mọi người đưa ra quyết định một cách tuần tự và mỗi người có thể dựa vào các lựa chọn của những người đưa ra quyết định hay hành động trước đó. Hoặc có một số thông tin khác, ảnh hưởng trực tiếp đến việc ra quyết định của họ.

Quyết định được giải thích bằng định lý Bayes (thuật toán phân lớp).

Phân tích mạng xã hội (Social network analysis) phân tích tính tác động của các tầng thông tin trên mạng xã hội có thể dẫn đến tạo ra nhiều ứng dụng giúp ích như xác định các cá nhân có tầm ảnh hưởng trên mạng xã hội. Thông tin này có thể được sử dụng để tối đa hóa hiệu quả thị trường hoặc ảnh hưởng đến dư luận.

## Network Effects (Hiệu ứng mạng):

Để giải thích cho hiện tượng trên, ta có thể đưa ra hai lý do cơ bản như sau:

* Một lý do đưa ra được dựa trên các vấn đề thông tin: vì hành vi của người khác truyền tải thông tin về những gì họ biết, hay quan sát hành vi của người khác và sao chép nó, đôi khi có thể là một quyết định hay và hợp lý.
* Lý do khác dựa trên các hiệu ứng lợi ích, còn được gọi là hiệu ứng mạng: đối với một số quyết định, bạn sẽ nhận được lợi ích khi bạn điều chỉnh hành vi của mình đối với hành vi của người khác.

Hiệu ứng mạng: là hiệu ứng mô tả trong [kinh tế](https://en.wikipedia.org/wiki/Economics) và [kinh doanh](https://en.wikipedia.org/wiki/Business) mà người dùng bổ sung, cung cấp [dịch vụ](https://en.wikipedia.org/wiki/Service_(economics)), sản phẩm tốt, tạo ra [giá trị](https://en.wikipedia.org/wiki/Value_(economics)) của sản phẩm cho người khác. Khi có hiệu ứng mạng, giá trị của sản phẩm hoặc dịch vụ sẽ tăng theo số lượng người sử dụng.

**Ví dụ**: Nhiều [trang web](https://en.wikipedia.org/wiki/Web_site) được hưởng lợi từ hiệu ứng mạng. [eBay](https://en.wikipedia.org/wiki/EBay) sẽ không phải là một trang web đặc biệt hữu ích nếu [đấu giá](https://en.wikipedia.org/wiki/Auction) không cạnh tranh. Tuy nhiên, khi số lượng người dùng tăng lên trên eBay, các cuộc đấu giá ngày càng cạnh tranh hơn, đẩy giá thầu lên các mặt hàng tăng lên. Điều này làm cho việc bán hàng trên eBay trở nên đáng giá hơn và đưa nhiều người bán hơn lên eBay, điều này khiến giá giảm trở lại vì điều này làm tăng nguồn cung, đồng thời đưa nhiều người lên eBay hơn vì có nhiều thứ được bán hơn mà mọi người muốn. Về cơ bản, khi số lượng người dùng eBay tăng lên, giá giảm và nguồn cung tăng lên, và ngày càng nhiều người thấy trang web này hữu ích.

# Power Laws and Rich-Get-Richer Phenomena (Luật số lớn và hiện tượng giàu có hơn):

## Popularity as a Network Phenomenon (Phổ biến như một hiện tượng mạng):

Như phần trước đã nói, hành vi của một người hoặc quyết định của một người đôi khi phụ thuộc vào sự lựa chọn hay quyết định của nhiều người khác - hoặc bởi vì thành quả của mình tạo ra có phụ thuộc vào những gì người khác làm - sự lựa chọn, hay quyết định của người khác mang lại thông tin hữu ích hơn nên làm cho hành vi hoặc quyết định của mình bị thay đổi.

Điều này có thể áp dụng vào hiệu ứng mạng để phân tích khái niệm chung về sự phổ biến (popularity) là hiện tượng đặc trưng bởi sự mất cân bằng.

**Ví dụ:** Một số người có ảnh hưởng rộng tới một xã hội và được biết đến nhiều hơn so với những người hầu như không có ảnh hưởng sẽ được biết đến ít hơn.

Theo wiki: Sự phổ biến của một người, một ý tưởng hay một cái gì đó là sự ưa thích, được biết đến bởi nhiều người, là sự hấp dẫn, thống trị, hay là sự nổi tiếng. Vì vậy, có thể kết luận là: sự phổ biến của đối tượng là kết quả của ảnh hưởng xã hội:

**Information cascades: population** là một thuật ngữ áp dụng cho thời kì hiện đại chủ yếu nhờ vào mạng xã hội. Phổ biến là một hiện tượng xã hội, sự chú ý của tập thể là cách duy nhất để làm cho một cái gì đó trở nên phố biến và các tầng thông tin đóng một vai trò lớn trong sự gia tăng nhanh chóng.

Trong văn hóa đại chúng, như phim ảnh, âm nhạc, sách báo, thường không phản ánh thị hiếu của công chúng, mà là của một vài người sử dụng đầu tiên, vì ảnh hưởng xã hội - mạng xã hội đã làm cho những thứ đó trở nên phổ biến hơn. Tầng thông tin ảnh hưởng mạnh mẽ khiến cho các cá nhân bắt chước hành động của người khác, cho dù họ có đồng ý hay không.

**Ví dụ**: mạng xã hội Facebook, đa số mọi người có khoảng 130 bạn bè, trong đó lại có một số người có đến tận nghìn người bạn. Điều này phản ánh rằng số lượng bạn bè của một người, là cách xác định mức độ phổ biến, hay sự nổi tiếng của một cá nhân.

Trong việc đo mức độ phổ biến, rất khó để ta biết được số lượng người biết đến Obama, Bill Gates. Thay vì tính toán, ta dựa vào số lượng tìm kiếm thông tin trên web về người nổi tiếng như Obama, Bill Gates… và ta chỉ cần đếm số lượng liên kết đến các trang web như: Google, Amazon, Wikipedia…

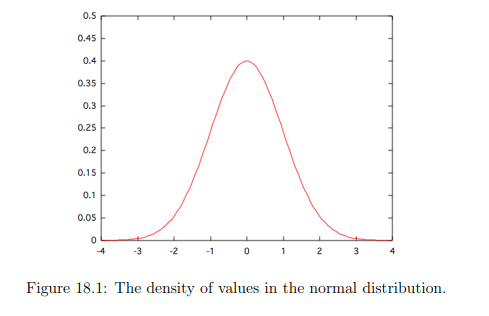
Để xác định độ phổ biến của một trang - ta sử dụng số lượng liên kết k đến một trang làm thước đo của trang phổ biến. Câu hỏi: **Là một hàm của k, phần nào của các trang trên web có k liên kết?**

Nếu ta coi các liên kết đến trang web như một đồ thị, với các trang web là các nốt, và các liên kết là các cạnh nối từ trang web này đến trang web khác. Đây chính là đồ thị phân phối trên tập các trang web.

Với các nốt có số lượng cạnh lớn hay có k lớn, thì độ phổ biến của trang web đó sẽ lớn hơn. Có thể nói đây chính xác là câu hỏi về độ phổ biến được phân phối trên tập hợp các trang.

Có một giả thuyết cho câu hỏi trên: đó chính là **Phân bố chuẩn.**

Đây là một giả thuyết do tác giả cuốn sách đưa ra



Hình 2: Mật độ xác suất của phân phối chuẩn

Năm 1990, Định lý giới hạn trung tâm đưa ra lời giải thích nếu ta lấy bất kỳ chuỗi số lượng ngẫu nhiên độc lập nhỏ nào, sau đó trong giới hạn tổng của chúng (hoặc trung bình) sẽ được phân phối theo phân phối chuẩn.

Làm thế nào để áp dụng điều này trong các trang web?

🡪Nếu chúng ta mô hình cấu trúc liên kết của Web. Bằng cách giả sử rằng mỗi trang web là ngẫu nhiên độc lập, số lượng liên kết đến một trang nhất định là tổng của nhiều trang, và do đó hy vọng nó sẽ là phân phối chuẩn. Đặc biệt, điều này sẽ gợi ý một giả thuyết cho câu trả lời cho câu hỏi ban đầu. Nếu là mô hình này, số lượng trang có k liên kết sẽ giảm theo cấp số nhân tính theo k, khi k ngày càng lớn.

Gọi X là biến ngẫu nhiên biểu thị mức độ phổ biến của một trang.

Công thức tính xác suất Pr[X = k] = A.e-αk trong đó A hằng số.

Kết luận này không được xác thực trong thực tế, vì giả định không hợp lý.

## Power Laws (Luật số lớn):

Theo wiki: Luật số lớn là mối quan hệ giữa hai đại lượng, trong đó sự thay đổi tương đối trong một đại lượng này dẫn đến sự thay đổi tương đối tỷ lệ trong một đại lượng khác.

**Ví dụ**: xem xét diện tích hình vuông theo chiều dài cạnh của nó, nếu chiều dài được nhân đôi, diện tích được nhân với hệ số 4.

Tính chất:

Tính bất biến: cho một hàm f(x) = ax-k, nhân với đối số x một hằng số c không đổi, ta được tỷ lệ của chính hàm đó:

*Tuân theo định luật lũy thừa với số mũ tỷ lệ cụ thể tương đương với các yếu tố không đổi. Tạo ra mối quan hệ tuyến tính khi logarit cả f(x) và x., đường thẳng trên biểu đồ log-log được gọi là kí hiệu của Luật số lớn. Với dữ liệu thực, đây là điều kiện cần, nhưng không đủ, cho dữ liệu theo mối quan hệ Luật số lớn.*

*Trên thực tế, có nhiều cách để tạo ra kí hiệu này nhưng trong giới hạn tiệm cận của chúng ta, không phải là Luật số lớn.*

Thiếu giá trị trung bình được xác định rõ: Luật số lớn được xác định trong khoảng x thuộc [1;∞) với k > 2, và nó có phương sai hữu hạn nếu k > 3, hầu hết các định luật số lớn được xác định trong tự nhiên đều có số mũ sao cho giá trị trung bình được xác định rõ nhưng phương sai thì không.

Tính phổ quát: các hệ thống có cùng số mũ được cho là thuộc cùng một lớp phổ quát.

Phân phối xác suất của Luật số lớn: phân phối xác suất theo Luật số lớn là phân phối có hàm mật độ có dạng:

Với α > 1 và L(x) là một hàm biến đổi chậm, là một hàm thỏa mãn:

Trong thực tế, khi đo lường sự phân phối các liên kết trên web, nghiên cứu trên nhiều ảnh chụp nhanh các trang web khác nhau, tỷ lệ các trang web có k liên kết xấp xỉ tỷ lệ 1/k2 (số mũ của k có thể lớn hơn 2).

Câu hỏi đặt ra là : Tại sao điều này lại khác với Phân phối chuẩn?

Phân phối chuẩn giảm theo cấp số nhân e-k, e-k^2, 2-k

Luật số lớn: k-2

**Ví dụ:** Với k = 1000 thì,

Phân phối chuẩn > 0, còn

Luật số lớn > 10-6

Giải thích: hàm k-2 giảm chậm hơn nhiều khi k tăng, do đó các trang có số lượng lớn liên kết phổ biến hơn. Hàm k-2 giảm dần từ k xuống một số lớn cố định, được sử dụng để đo tỷ lệ các trang web có giá trị k.

**Một số ví dụ về Luật số lớn:**

Tỷ lệ số điện thoại nhận được k liên kết mỗi ngày xấp xỉ k-2.

Tỷ lệ của những quyển sách được mua bởi k người xấp xỉ k-3.

Kết luận: Luật số lớn chiếm ưu thế trong trường hợp số lượng được đo có thể được xem là loại phổ biến.

- Ví dụ: Ta có bảng hiển thị số lượt tải xuống hàng tháng cho mỗi bài hát tại một trang web âm nhạc, một trong những điều đầu tiên đáng làm là kiểm tra xem nó xấp xỉ luật số lớn k-c với một số c hay không.

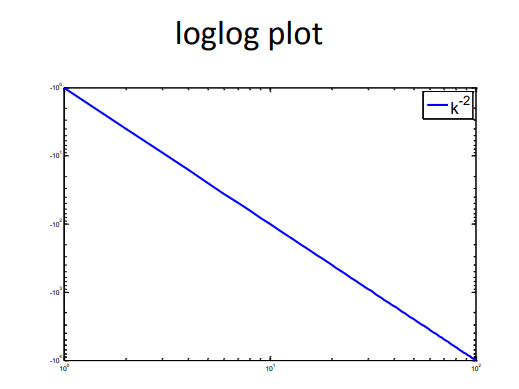
- Làm thế nào để kiểm tra được một tập dữ liệu có thể phân phối theo luật số lớn hay không?

Đặt P(k) là tỷ lệ (xác xuất) của dữ liệu cần được kiểm tra có giá trị k.

Giả sử ta muốn kiểm tra xem P(k) = a k-c với  a,c là một hằng số.

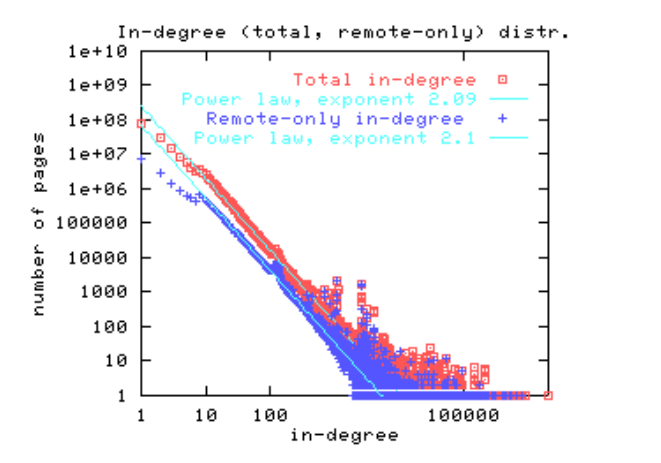
Sau đó lấy logarit của cả hai vế, ta được phương trình

LogP(k) = log(a) – c log(k)



- Nếu chúng ta vẽ hàm LogP(k) là hàm của Log k, chúng ta sẽ nhận được một đường thẳng, một đường thẳng với độ dốc –c, log(a) sẽ bị y chặn.

- Ví dụ:



- Hình ví dụ trên là tỷ lệ của các trang web có k liên kết.

- Điều gì đã gây ra luật số lớn?

- Chúng ta cần một lời giải thích đơn giản cho những gì gây ra luật số lớn?

- Hàm log-log của các liên kết trang web, một đường thẳng cho phần lớn các phân phối. -- Vậy quá trình nào đã giữ cho dòng thẳng như vậy?

## Rich-Get-Richer Models (Mô hình giàu có hơn):

- Lấy ý tưởng từ việc phân tích các tầng thông tin và hiệu ứng mạng cung cấp cơ sở bởi một cơ chế rất tự nhiên để tạo ra Luật số lớn. Giống như các phân phối chuẩn chúng phát sinh từ quyết định ngẫu nhiên trung bình, chúng ta sẽ biết rằng Luật số lớn phát sinh từ các phản hồi được đưa ra bởi quyết định tương quan trên toàn bộ.

- Đây thật sự là một câu hỏi đáng nghiên cứu và rất thú vị để cung cấp một mô hình hoàn toàn thỏa đáng về Luật số lớn bắt đầu từ các mô hình đơn giản của việc ra quyết định cá nhân (như chúng ta đã làm đối với các tầng thông tin). Thay vào đó ta sẽ xây dựng một mô hình của mình không phải từ nội bộ quá trình ra kết luận của từng người, mà là hệ quả của quá trình quan sát được qua việc đưa ra quyết định với sự có mặt của các tầng: ta cho rằng mọi người thường có xu hướng sao chép, hay dựa vào những quyết định của những người thực hiện trước họ.

- Dựa trên ý tưởng này, đây là mô hình đơn giản để tạo liên kết giữa các trang web.

Các trang được tạo theo thứ tự và được đặt tên là 1, 2, 3, …, N.

Khi trang j được tạo, nó tạo ra một liên kết đến một trang web trước đó theo quy tắc xác suất (được điều chỉnh bởi một số p duy nhất trong khoảng từ 0 đến 1).

Với xác suất p, trang j chọn một trang i một cách ngẫu nhiên trong số tất cả các trang trước đó và tạo ra một liên kết đến trang i.

 Với xác suất 1-p, trang j chọn một trang i ngẫu nhiên trong số tất cả các trang trước đó và tạo một liên kết đến trang mà người dùng trỏ đến.

Điều này mô tả việc tạo ra một liên kết đơn từ trang j: người ta có thể lặp lại quá trình này để tạo ra nhiều liên kết độc lập từ trang j. (Tuy nhiên, để đơn giản hóa, ta sẽ cho rằng mỗi trang chỉ tạo một liên kết ngoài.)

- Phần (2b) là chìa khóa của quá trình này: sau khi tìm thấy một trang i ngẫu nhiên trước đó trong toàn bộ, tác giả của trang j không liên kết với i, mà thay vào đó sao chép quyết định của tác giả trang i – liên kết đến cùng trang mà ta đã làm.

- Kết quả chính của mô hình này là nếu ta chạy nó trên nhiều trang, một phần nhỏ các trang có k liên kết sẽ được phân phối xấp xỉ theo công xuất 1/kc, trong đó giá trị của c phụ thuộc vào sự lựa chọn của p. Sự phụ thuộc này đi theo hướng tự nhiên tự giác: khi p càng nhỏ, việc sao chép trở nên thường xuyên hơn, c cũng nhỏ hơn, khiến nhiều người có thể nhìn thấy các trang cực kỳ phổ biến.

- Chứng minh kết quả này đòi hỏi chúng ta phải làm các phân tích phức tạp hơn ở đây, nhưng sẽ hữu ích khi làm việc thông qua một số ý tưởng không chính thức đằng sau phân tích này. Trước hết, cơ chế sao chép trong (2b) thực sự là một triển khai của mô hình “Rich-get Richer” sau đây: Khi bạn sao chép quyết định của một trang L là tỉ lệ thuận với tổng số trang hiện đang liên kết đến L. Do đó, một cách  tương đương để viết quy trình sao chép của chúng tôi sẽ là cụm từ (2b)

( 2) ...

Với xác suất 1-p, trang j chọn một trang L với xác suất tỷ lệ thuận với số lượng liên kết hiện tại của L và tạo liên kết đến L

- Tại sao chúng ta gọi đây là quy tắc “Rich-get-Richer”? Bởi vì xác suất mà trang L trải qua sự gia tăng mức độ phổ biến tỷ lệ thuận với mức độ phổ biến hiện tại của L. Hiện tượng này được gọi là Preferential Attachment (tệp đính kèm ưu tiên), các liên kết được hình thành ưu tiên là các trang có tính phổ biến cao. Và mô hình sao cung cấp một hoạt động về lý do tại sao sự phổ biến thể hiện ở “Rich-get-Richer” như vậy: về cơ bản, một người nào đó càng nổi tiếng chúng ta càng có khả năng nghe thấy tên của họ xuất hiện trong cuộc trò chuyện, và do đó càng có nhiều khả năng cuối cùng bạn cũng biết về họ. Điều tương tự cũng xảy ra với trang web, trọng tâm cụ thể của mô hình của chúng tôi là ở đây.

- Trực giác còn lại đằng sau phân tích chạy như sau. Với “Rich-get-Richer”, mô hình của chúng tôi dự đoán rằng mức độ phổ biến sẽ tăng theo cùng một quy tắc chi phối sự phát triển: mức độ phổ biến của trang web tăng theo tỉ lệ thuận với giá trị hiện tại của nó và theo cấp số nhân theo thời gian. Do đó, một số trang có được một khách hàng tiềm năng nhỏ hơn những người người khác sẽ có xu hướng mở rộng khách hàng tiềm năng này; trong khi mấu chốt của Định lý giới hạn trung tâm là các giá trị ngẫu nhiên độc lập nhỏ có xu hướng triệt tiêu lẫn nhau, “Rich-get-Richer” của việc sao chép thực sự khuếch đại các giá trị lớn, làm cho chúng còn lớn hơn.

- Như với bất kỳ mô hình đơn giản nào, mục tiêu không phải là nắm bắt tất cả các lý do tại sao mọi người tạo liên kết trên web hoặc trong bất kỳ mạng nào khác, mà chỉ ra rằng một nguyên tắc đơn giản và rất tự nhiên đằng sau việc tạo liên kết dẫn trực tiếp đến Luật số lớn - và do đó , người ta không nên ngạc nhiên như lần đầu tiên họ xuất hiện.

- Thật vậy, mô hình “Rich-Get-Richer” có thể đề xuất một cơ sở cho Luật số lớn trong một loạt các cài đặt, bao gồm cả các mô hình không liên quan đến việc ra quyết định của con người.

- Ví dụ: Nếu chúng ta giả định rằng các thành phố được hình thành vào thời điểm khác nhau và một khi được hình thành, một thành phố phát triển tương xứng với quy mô hiện tại của nó chỉ đơn giản là kết quả của sự tăng dân số, thì chúng ta có mô hình “Rich-Get-Richer” gần như chính xác - và do đó chúng ta không nên ngạc nhiên khi thấy Luật số lớn thực tế có mặt trong thực tế.

- Để lấy một ví dụ rất rõ ràng, các nhà nghiên cứu sinh học đã lập luận (mặc dù dữ liệu vẫn còn quá ít để chắc chắn) rằng số lượng bản sao của một gen trong bộ gen xấp xỉ theo phân phối theo Luật số lớn . Nếu chúng ta tin rằng các bản sao gen phát sinh phần lớn thông qua các sự kiện đột biến trong đó một đoạn DNA ngẫu nhiên được sao chép ngẫu nhiên, thì một gen có nhiều bản sao có khả năng nằm trong một đoạn DNA ngẫu nhiên được sao chép - vì vậy các gen “rich”(những người có nhiều bản sao) nhận được nhiều “richer”, một lần nữa, chúng ta không nên ngạc nhiên khi thấy một Luật số lớn.

- Một nguyên tắc tương tự, các luật tương tự điều chỉnh mức độ phổ biến của trang web, dân số thành phố và bản sao gen là khá bí ẩn: nhưng nếu người ta xem tất cả những điều này là kết quả của các quá trình thể hiện các tác phẩm “Rich-Richer”, thì bức tranh bắt đầu trở nên rõ ràng hơn. Một lần nữa, người ta phải nhấn mạnh rằng đây vẫn là những mô hình đơn giản được thiết kế chỉ để ước tính những gì đang diễn ra: và có các lớp mô hình đơn giản khác được thiết kế để nắm bắt hành vi luật số lớn mà chúng ta chưa thảo luận ở đây.

- Ví dụ, một chuỗi nghiên cứu song song đã lập luận làm thế nào các luật số lớn có thể phát sinh từ các hệ thống đang được tối ưu hóa khi có các ràng buộc. Nhưng điều mà tất cả các mô hình đơn giản này gợi ý là khi người ta nhìn thấy một Luật số lớn trong dữ liệu, những lý do có thể khiến nó có thể thường quan trọng hơn thực tế đơn giản là nó ở đó.

## The Unpredictability of Rich-Get-Richer Effects (Tính không thể dự đoán trước của hiệu ứng giàu có hơn):

Do tính chất của hiệu ứng phản hồi tạo ra các định luật số lớn, nó dự đoán một cách ngẫu nhiên rằng đối với một trang web, một cuốn sách, một bài hát hoặc bất kỳ đối tượng nào khác được chú ý, giai đoạn ban đầu của sự nổi tiếng (hay phổ biến) là một điều tương đối khó. Một khi bất kỳ một trong những hạng mục này được thiết lập tốt (tồn tại trong một thời gian dài), động lực giàu trở nên giàu hơn của sự phổ biến có nhiều khả năng để đẩy nó lên thậm chí còn cao hơn, nhưng làm cho quá trình làm giàu trở nên ngày càng giàu có hơn này được đốt cháy ở giai đoạn đầu có vẻ như là một quá trình khó khăn, đầy những nguy cơ tiềm ẩn.

Sự nhạy cảm này đối với các dự đoán ban đầu không thể đoán trước là điều mà chúng ta đã thấy trong hai chương trước: các tầng thông tin có thể phụ thuộc vào kết quả của một số lượng nhỏ các quyết định ban đầu của dân số, và một công nghệ tồi tệ hơn có thể giành chiến thắng vì nó đạt đến một quy mô lớn, quan trọng trước khi đối thủ cạnh tranh khác của nó làm ra. Tính biến động của sự phổ biến cho thấy rằng hiệu ứng ngẫu nhiên trong tiến trình cũng sẽ đóng một vai trò ở đây.

Ví dụ, nếu chúng ta có thể quay lại thời gian 15 năm, và sau đó chạy lại lịch sử, liệu những cuốn sách Harry Potter sẽ bán được hàng trăm triệu bản như bây giờ, hay chúng sẽ bị lu mờ trong khi một số tác phẩm tiểu thuyết của trẻ em khác đã đạt được thành công lớn? Một trực giác gợi ý là cái sau. Tổng quát hơn, nếu lịch sử được phát lại nhiều lần, có vẻ như sự phổ biến của luật số lớn sẽ phân bố theo mỗi thời điểm, nhưng điều đó không rõ ràng rằng các mặt hàng phổ biến nhất sẽ luôn giống nhau.

Các thí nghiệm kiểu này rất hữu ích trong việc xem xét hậu quả của các mô hình của chúng tôi, nhưng không cần phải nói, nó rất khó để thực hiện chúng như những thí nghiệm thực. Tuy nhiên, gần đây, Salgankik, Dodds và Watts đã thực hiện một thí nghiệm bắt đầu cung cấp một số hỗ trợ theo kinh nghiệm cho trực giác này. Họ đã tạo ra một trang web tải nhạc, với 48 bài hát khó hiểu với chất lượng khác nhau được viết bởi các nhóm biểu diễn thực tế. Khách truy cập vào trang web đã được trình bày một danh sách các bài hát và có cơ hội nghe chúng. Mỗi khách truy cập cũng được hiển thị một bảng liệt kê số lượt tải xuống hiện tại cho mỗi bài hát - số lần được tải xuống từ trang web cho đến nay. Vào cuối phiên, khách truy cập có cơ hội tải xuống các bản sao của các bài hát mà mình thích.

Bây giờ, không biết đến khách truy cập, khi truy cập tới, họ thực sự được chỉ định ngẫu nhiên cho một trong tám bản sao có sự tương đồng của trang. Các bản sao tương đồng bắt đầu giống hệt nhau, với cùng một bài hát và với mỗi bài hát có số lượt tải xuống bằng không. Tuy nhiên, mỗi bản sao tương đồng sau đó phát triển khác nhau khi người dùng truy cập đến. Sau đó, trong một môi trường cài đặt quy mô nhỏ có kiểm soát, thí nghiệm này đã cung cấp một cách để quan sát những gì xảy ra với mức độ phổ biến của 48 bài hát khi bạn chạy lịch sử về phía trước tám lần. Và trên thực tế, người ta đã phát hiện ra rằng thị phần của các bài hát khác nhau đã thay đổi đáng kể trên các bản sao tương đồng khác nhau, mặc dù những bài hát hay nhất không bao giờ kết thúc ở cuối và những bài hát tệ nhất không bao giờ kết thúc ở đầu.

Salganik và cộng sự cũng sử dụng phương pháp này để chỉ ra rằng, về tổng thể, phản hồi tạo ra sự bất bình đẳng lớn hơn trong kết quả. Đặc biệt, họ đã chỉ định một số người dùng cho phiên bản thứ chín của trang web trong đó không có phản hồi nào về số lượng tải xuống được cung cấp cả. Trong phiên bản này của trang web, không có cơ hội trực tiếp cho người dùng đóng góp vào động lực làm giàu phong phú hơn, và thực sự, có sự thay đổi đáng kể về thị phần của các bài hát khác nhau. Có những hàm ý rõ ràng về sự phổ biến trong các môi trường ít kiểm soát hơn, song song với một số kết luận được rút ra từ các mô hình của chúng tôi - cụ thể, rằng sự thành công trong tương lai của một cuốn sách, phim, người nổi tiếng hoặc trang web bị ảnh hưởng mạnh mẽ bởi các loại hiệu ứng phản hồi, và do đó có thể ở một mức độ nào đó vốn không thể đoán trước được.

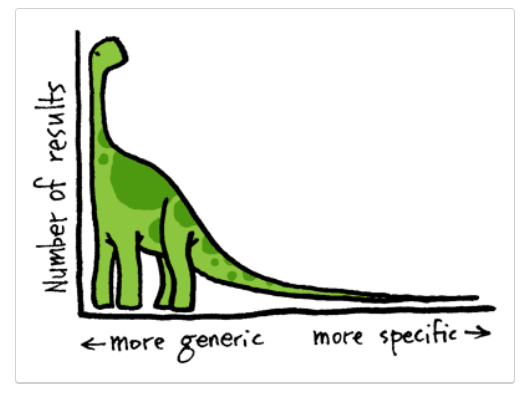
Mối quan hệ chặt chẽ giữ luật số lớn và các tầng thông tin? Xem xét về vấn đề này có một câu hỏi quan trọng để nghiên cứu thêm đó là: hiểu mối quan hệ giữa các quy luật số lớn và các tầng thông tin ở mức độ sâu hơn. Khi chúng tôi xem xét các tầng thông tin, chúng tôi thấy một bộ phận dân số trong đó mọi người nhận thức được các quyết định trước đó được đưa ra giữa hai lựa chọn (ví dụ như chấp nhận một ý tưởng hoặc từ chối nó) có thể kết thúc thành một tầng, ngay cả khi mỗi người đưa ra quyết định tối ưu cho những gì họ đã quan sát. Mô hình sao chép của chúng tôi cho luật số lớn dựa trên trực giác của mô hình này, nhưng nó khác nhau ở một số khía cạnh. Đầu tiên, một mô hình cho sự phổ biến nên bao gồm các lựa chọn trong số nhiều tùy chọn có thể (ví dụ: tất cả các trang web có thể) thay vì chỉ có hai tùy chọn. Thứ hai, mô hình sao chép liên quan đến một nhóm người tham gia vào việc quan sát dân số rất hạn chế: khi bạn tạo một trang web mới, mô hình giả định rằng bạn tham khảo quyết định của chỉ một người được chọn một cách ngẫu nhiên khác. Và thứ ba, mô hình sao chép dựa trên ý tưởng rằng những người sau đó bắt chước quyết định của những người trước đó, nhưng nó không bắt nguồn mô phỏng này từ một mô hình cơ bản hơn về việc ra quyết định hợp lý.

Hai trong số những sự kiện này chỉ đơn giản là gợi lại đặc điểm của vấn đề đang được mô hình hóa ở đây - cách mà sự phổ biến phát triển theo thời gian. Nhưng nó sẽ rất thú vị khi vượt qua phần ba của những điểm khác nhau này, xây dựng một mô hình kiểu sao chép để tạo ra các luật số lớn trên một mô hình cơ sở của việc ra quyết định cá nhân. Cách tiếp cận như vậy có thể làm sáng tỏ thêm các cơ chế đằng sau các động lực làm giàu - giàu hơn và cung cấp một bức tranh về sự phổ biến khi phát sinh từ các tầng thông tin cạnh tranh có cường độ thay đổi theo quy luật số lớn mà chúng ta quan sát được trong các hệ thống thực.

## The Long Tail (Thị trường ngách):

Định nghĩa The Long Tail là Gì?

- Thuật ngữ Long tail (tên gọi: cái đuôi dài) Long tail ở đây được biết đến là hình thức kinh doanh dựa trên 80% bị đánh giá kém còn lại của thị trường, cái mà trước đây các doanh nghiệp không lựa chọn vì cho rằng những giá trị này đem về hiệu quả rất ít hoặc thậm chí không có hiệu quả.



Định nghĩa của thị trường ngách (niche market) là một phân đoạn nhỏ của thị trường mà bạn đặt mục tiêu vào một nhóm khách hàng riêng biệt. Chúng ta không tập trung vào phân khúc của toàn bộ thị trường mà tập trung vào một phân đoạn chuyên biệt." Cùng với hình tượng: "Con cá lớn trong cái hồ nhỏ, còn hơn là con cá nhỏ trong đại dương bao la".

- Theo quy luật Pareto (hay còn gọi là quy luật 80/20): "chỉ có 20% sản phẩm đem lại 80% lợi nhuận cho doanh nghiệp". Đây được xem là quy luật phổ biến trong rất nhiều doanh nghiệp trong nhiều năm qua. Nhưng điều này liệu còn chính xác trong thời đại thế giới phẳng hiện nay chăng? - Câu trả lời là KHÔNG! Từ khi có sự ra đời của Internet, thế giới phẳng hình thành đã mở ra nhiều cơ hội cho Thị trường ngách phát triển… thuật ngữ Long tail (tên gọi: cái đuôi dài) xuất hiện. Long tail ở đây được biết đến là hình thức kinh doanh dựa trên 80% bị đánh giá kém còn lại của thị trường, cái mà trước đây các doanh nghiệp không lựa chọn vì cho rằng những giá trị này đem về hiệu quả rất ít hoặc thậm chí không có hiệu quả. Ngày càng có nhiều Doanh nghiệp tìm đến những sản phẩm nằm ở phần đuôi (sản phẩm ngách) và điều này cho thấy sự ảnh hưởng của sản phẩm ở phần đuôi tăng lên. Mặc dù không một loại sản phẩm nào được bán với số lượng lớn nhưng khi gộp nhiều sản phẩm lại với nhau sẽ đem về lợi nhuận khổng lồ. Vì vậy, sự thống trị của các sản phẩm dẫn đầu sẽ không còn nữa. Tuy nhiên, các sản phẩm dẫn đầu vẫn chiếm một phần của đường nhu cầu giống như các sản phẩm ngách.

- Việc phân phối những mặt hàng phổ biến có thể có những kết quả kinh doanh quan trọng, đặc biệt là trong ngành truyền thông. Cụ thể, hãy tưởng tượng một công ty truyền thông có hàng tồn kho lớn.

- Ví dụ: như một nhà bán lẻ sách và xem xét câu hỏi sau: hầu hết doanh số được tạo ra bởi một tập nhỏ các mặt hàng rất phổ biến hay từ những mặt hàng không phổ biến nhưng đa dạng với số lượng lớn phù hợp với riêng cho từng nhóm người? Trong trường hợp trước đây, công ty đang dựa trên thành công của mình khi bán các “hits” - một số lượng nhỏ các bộ phim bom tấn tạo ra doanh thu khổng lồ. Trong trường hợp thứ hai, công ty đang dựa trên sự thành công của mình trên vô số các sản phẩm, trong đó mỗi sản phẩm đều thu hút một bộ phận nhỏ khán giả.

 - Trong một bài báo được đọc rộng rãi năm 2004 có tên là “Thị trường ngách”, Chris Anderson đã lập luận rằng sự trao đổi hàng hóa với sự hỗ trợ của Internet và các yếu tố khác đang thúc đẩy ngành công nghiệp truyền thông và giải trí hướng tới một thế giới mà sự thay thế sau này sẽ chiếm ưu thế, với sự phát triển của thị trường ngách, các sản phẩm tưởng chừng ít được để ý sẽ thúc đẩy phần lớn sự quan tâm của khách hàng. Như anh ấy đã viết, bạn có thể tìm thấy mọi thứ ngoài kia ở trong những thị trường ngách. Có nhiều danh mục cần lựa chọn phía sau, các album cũ vẫn được người hâm mộ lâu năm nhớ đến hoặc được phát hiện lại bởi những người nghe mới. Có các bản nhạc trực tiếp, những bài hát trong album, bản phối lại. Có rất nhiều những ngách nhỏ trong sở thích của hàng ngàn người.

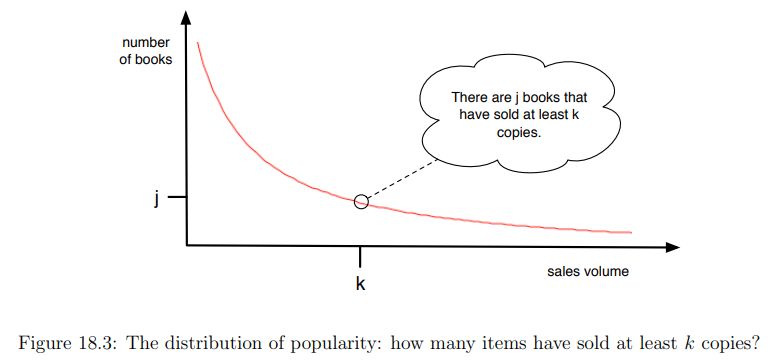
- Mặc dù dữ liệu bán hàng cho thấy xu hướng trên thực tế hơi phức tạp, sự cạnh tranh này giữa các hits, bom tấn và sản phẩm thích hợp làm cho một khung tổ chức hấp dẫn. Nó cũng phù hợp với các mô hình cơ bản của các công ty như Amazon hay Netflix, nơi khả năng mang theo lượng phim, dữ liệu cũ khổng lồ - mà không bị hạn chế bởi các cửa hàng vật lý khiến việc bán sản phẩm đa dạng thiên văn trở nên khả thi ngay cả khi rất ít trong số đó tạo ra doanh thu lớn cho họ. Và cuối cùng, định lượng tầm quan trọng của thị trường ngách đi đến phân tích về mối quan hệ giữa hai đại lượng.

- Hình dung về thị trường ngách. Điều đầu tiên cần chú ý, khi chúng ta so sánh cuộc thảo luận về thị trường ngách với phân tích trước đây về mối quan hệ giữa hai đại lượng, là ở một khía cạnh nào đó, chúng ta hiện đang nhìn mọi thứ dường như nhỏ đi. Ban đầu, chúng tôi bắt đầu từ một đường cơ sở mà chúng tôi dự kiến ​​sẽ thấy các phân phối Gaussian và độ tập trung chặt chẽ quanh mức trung bình, và sau đó chúng tôi quan sát thấy rằng số lượng các mặt hàng rất phổ biến cao hơn nhiều so với đường cơ sở này sẽ gợi ý. Mặt khác, chúng ta đang bắt đầu từ một cái nhìn rất khác về thế giới - một kiểu mẫu của ngành kinh doanh truyền thông trong đó chỉ có các bộ phim bom tấn - và chúng tôi đang quan sát rằng tổng doanh số của các mặt hàng không phổ biến, đã được thực hiện cùng nhau, thực sự rất có ý nghĩa. Về nội dung hình 18.2, khung nhìn mới này đang tập trung vào phần trên bên trái của cốt truyện, trong khi trước khi chúng tôi chủ yếu tập trung vào phía dưới bên phải. Một khi bạn nhận ra rằng sự tương phản này đang diễn ra, nó không khó để điều hòa hai quan điểm [4]. Đầu tiên, hãy để sửa đổi định nghĩa ban đầu của chúng ta về đường cong phổ biến, theo cách mà không thay đổi về cơ bản những gì chúng ta đo lường. - Thay vì hỏi

- Là một hàm của k, phần nhỏ của các sản phẩm có mức độ phổ biến chính xác là k?

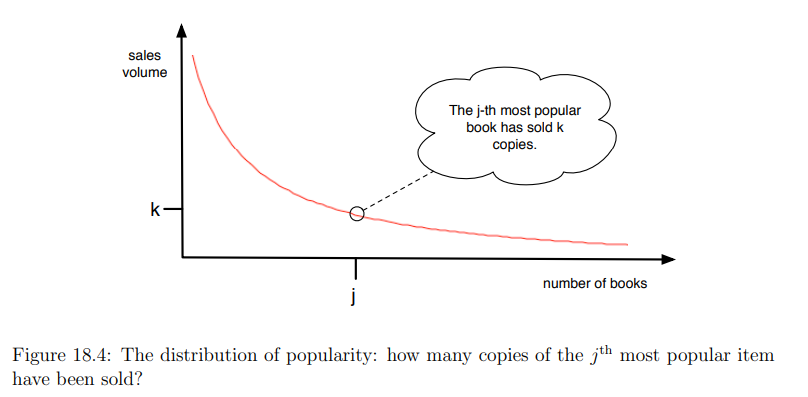
Thay vào đó hãy hỏi

- Là một hàm của k, số lượng sản phẩm có mức độ phổ biến ít nhất là k?



- Lưu ý rằng có hai sự khác nhau: “phần nhỏ” thành “số lượng” (một thay đổi hoàn toàn không quan trọng), và “chính xác” thành “ít nhất”. Thay đổi thứ hai này sửa đổi chú thích mà chúng tôi đang xem xét; nhưng kể cả khi chúng ta không đi qua đạo hàm ở đây, có thể chỉ ra rằng nếu chức năng ban đầu là một phần phối lũy thừa, sau đó cái mới này cũng vậy. Chúng tôi hiển thị một sơ đồ nguyên lý của phân phối mới này trong Hình 18.3; nếu chúng ta nói về sự phổ biến của một số mặt hàng như sách, sau đó một điểm tọa độ (k, j) trên đường cong này có nghĩa là có những cuốn sách j đã bán được ít nhất k bản.

- Cho đến nay, đây vẫn là cách hiểu theo khái niệm từ phần trước: khi chúng ta đi theo trục x của đường cong sang phải, về cơ bản chúng ta đang hỏi, khi bạn nhìn vào doanh số bán hàng lớn và rất lớn, bạn tìm ra được bao nhiêu quyển sách? Để có thể nắm bắt trực tiếp các đuôi của đồ thị của thị trường ngách, chúng tôi muốn đặt câu hỏi sau khi chúng ta đi theo trục x sang phải: Bạn nhìn vào các mặt hàng ít phổ biến hơn, bạn thấy khối lượng bán hàng là bao nhiêu ?



- Nếu chúng ta nghĩ về nó, điều này chỉ đơn giản liên quan đến việc chuyển đổi hai trục. Đó là, giả sử rằng chúng ta vẽ chính xác cùng một đường cong, nhưng chúng ta trao đổi vai trò của các trục x và y, như được hiển thị trong hình 18.4. Giải thích đường cong mới này theo nghĩa đen từ định nghĩa của nó, một điểm (j, k) trên đường cong nói”Cuốn sách phổ biến thứ j đã bán được k bản.” . Đây là chính xác những gì chúng tôi muốn: chúng tôi đặt mua những cuốn sách theo thứ hạng doanh số bán hàng, và sau đó chúng tôi xem xét mức độ phổ biến của sách khi chúng tôi tăng dần thứ hạng doanh số bán hàng- vào các sản phẩm thích hợp.1 Và hình dạng đặc trưng của đường cong này, kéo dài từ từ xuống bên phải, là cơ sở trực quan cho thuật ngữ “The Long Tail."

- Giờ đây, người ta có thể dễ dàng thảo luận về các xu hướng trong khối lượng bán hàng và hệ quả của chúng, về đường cong trong Hình 18.4. Về cơ bản, khu vực dưới đường cong từ một số điểm j trở ra là tổng khối lượng bán hàng được tạo ra bởi tất cả các mặt hàng bán hàng thứ hạng j trở lên; và do đó những cách hỏi khác nhau giữa sản phẩm mang tính nổi bật thịnh hành với những sản phẩm ít được chú ý nhưng đa dạng đối với một bộ sản phẩm cụ thể, là liệu có nhiều khu vực hơn đáng kể dưới phần bên trái của đường cong này (sản phẩm thịnh hành) hoặc bên phải (sản phẩm ít chú ý). Và cuộc tranh luận về xu hướng đối với các sản phẩm ít được chú ý  trở thành một câu hỏi liệu đường cong này có thay đổi hình dạng theo thời gian hay không, thêm nhiều khu vực bên phải và giảm bớt khu vực bên trái.

- Điều đáng chú ý là các đường cong trong Hình 18.4 - với các trục được sắp xếp sao cho biến trên trục x là thứ hạng thay vì mức độ phổ biến - có một lịch sử lâu dài. Chúng thường được gọi là âm mưu Zipf theo nhà ngôn ngữ học George Kingsley Zipf, người đã tạo ra những đường cong như vậy cho một số hoạt động của con người [423]. Nổi tiếng nhất, ông đã xác định nguyên tắc thực nghiệm được gọi là Định luật Zipf, rằng tần suất của từ phổ biến thứ j trong tiếng Anh (hoặc hầu hết các ngôn ngữ phổ biến khác của con người) tỷ lệ thuận với 1/j. Do đó, có lẽ không có gì đáng ngạc nhiên, cuộc tranh luận trong ngành truyền thông về những đường cong như thế này lặp lại sự mê hoặc trước đó trong các lĩnh vực khác.

## The Effect of Search Tools and Recommendation Systems (Ảnh hưởng của công cụ tìm kiếm và hệ tư vấn):

Chúng tôi kết luận bằng cách thảo luận về một câu hỏi khác dần trở nên quan trọng khi mọi người xem xét đến sự phổ biến và phân phối của nó: công cụ tìm kiếm internet tạo nên [rich-get-richer dynamics] của sự phổ biến là tích cực hay tiêu cực? [Điều thú vị là có hai [compelling but juxtaposed sides] với câu hỏi này], và giải pháp cuối cùng của nó có thể sẽ liên quan đến các quyết định về cách các cá nhân hay tập thể thiết kế và triển khai các công cụ tìm kiếm trong tương lai.

Một trong những mặt của câu hỏi này, chúng ta thấy rằng mô hình trong đó mọi người sao chép các liên kết từ các trang web ngẫu nhiên đồng đều đã mang lại lợi thế cho các trang phổ biến. Tuy nhiên, một khi mọi người sử dụng phương pháp tìm kiếm như là Google để tìm các trang, thì kết quả sao chép đó dường như phần nào cũng bị sai lệch về độ chính xác. Chúng ta có thể thấy được google xếp hạng các trang dựa trên độ phổ biến, những trang web xếp hạng cao là những trang web mà người dùng thường quyết định để truy cập (càng truy cập nhiều thì càng có khả năng xuất hiện đầu tiên). Một góc độ tương tự cũng có thể đưa ra cho các phương tiện truyền thông khác, trong đó một số ít các mục phổ biến có khả năng để trở nên vượt trội so với các mục khác. Trong mô hình đơn giản, loại thông tin phản hồi này làm nổi bật [rich-get-richer dynamics], tạo ra nhiều bất bình đẳng hơn trong sự phổ biến.

Tuy nhiên, nó có nhiều tác dụng khác trong công việc. Người dùng bắt đầu với việc gõ một truy vấn với phạm vi rất rộng vào google, và kết quả trả về không phải là một danh sách duy nhất của “các trang hàng đầu trên Google” – thay vào đó, ta nhận được kết quả là tương đối mơ hồ (tức là kết quả trả về rộng), khi đó, người dùng được dẫn tới các trang mà họ có thể không bao giờ phát hiện thông qua trình duyệt đơn. Các công cụ tìm kiếm thường được sử dụng theo cấu trúc này, mục tiêu tìm kiếm sẽ chặt chữ hơn với sở thích của người dùng, có thể trong thực tế sẽ cung cấp các đường dẫn xung quanh các trang phổ biến trên toàn cầu, cho phép mọi người tìm thấy các trang ít phổ biến hơn một cách dễ dàng hơn và hạn chế [rich-get-richer dynamics]. Tương tự, các mô hình toán học đơn giản trong đây có thể chứng minh được những kĩ thuật đó có thể hoạt động.

Quan điểm sau cũng là một phần quan trọng của lập luận [Long-Tail] của Anderson: Để kiếm tiền từ một lượng lớn sản phẩm, công ty cần phải cho khách hàng biết về các sản phẩm và một số cách để trải nghiệm chúng. Dưới góc độ này, các hệ thống tư vấn mà các công ty như Amazon hay Net đã triển khai có thể coi là tối quan trọng trong chiến lược kinh doanh của họ: về cơ bản, chúng là các công cụ tìm kiếm được thiết kể để đưa mọi người tới với các mặt hàng ít được phổ biến nhưng lại phù hợp với miêu tả của người dùng như cách được suy ra từ lịch sử mua hàng trong quá khứ của họ.

Cuối cùng, việc thiết kế các công cụ tìm kiếm là một ví dụ về một hệ phản hồi cao cấp: bằng cách khiến mọi người xử lý các tuỳ chọn có sẵn của mình bằng cách này hay cách khác, chúng ta có thể hạn chế hiện tượng [rich-get-richer] hoặc là tăng cường nó, hoặc là điều khiển nó theo một hướng nhất định. Đây là một trong những kết quả tinh tế khi chúng ta đưa các hệ thống thông tin tinh vi vào hệ thống xã hội phức tạp.

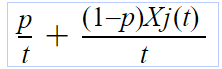
## Advanced Material: Analysis of Rich-Get-Richer Processes (Phân tích quy trình Rich-Get-Richer):

- Trước hết, ta mô tả tính gần đúng của mô hình, thảo luận về một vài tính chất đơn giản của mô hình xác suất nguyên bản của nó.

- Đầu tiên, số liên kết đến biến j tại t > j là biến ngẫu nhiên Xj(t). Quan sát 2 tính chất của Xj(t) :

Điều kiện ban đầu: tại thời điểm j nút j được khởi tạo với 0 liên kết ban đầu (Xj(j) = 0).

Xj theo thời gian t: Tại t+1, nút j đạt được 1 liên kết khi và chỉ khi liên kết từ nút vừa tạo có t+1 điểm tới nó. Với xác suất p, Nút t+1 liên kết với nút trước đó đều được chọn ngẫu nhiên, còn với xác suất 1-p thì nó liên kết với các nút trước đó với xác suất tỷ lệ thuận với số liên kết tại nút đó. Trong trường hợp trước, nút t+1 liên kết với nút j với xác suất 1/t. Đối với trường hợp sau, quan sát tại thời điểm tại thời điểm nút t+1 được tạo thấy rằng, tổng số liên kết trong mạng lưới này là t (1 trong mỗi liên kết với nút trước). và trong đó. Xj(t) trỏ đến nút j. Do vậy, trong trường hợp sau thì nút t+1 liên kết đến nút j với xác suất Xj(t)/t. Vì vậy, Công thức xác suất tổng quát mà nút t+1 liên kết tới nút j là:



- Kế hoạch cơ bản trong việc xây dựng mô hình gần đúng là phân tích các mục khác nhau, chặt chẽ tương tự, nhưng quy trình rich-get-richer thì đơn giản hơn. Trong đó nó tương ứng dễ dàng hơn so với việc khám phá luật số lớn (Power Laws). Một lần nữa, điều này không trực tiếp ám chỉ rằng mô hình ban đầu hoạt động cùng một các thức. Nhưng sự tương đồng giữa 2 mô hình đưa ra bằng chứng mà sau đó có thể xác minh bằng cách phân tích sâu hơn về mô hình ban đầu.

- Ý tưởng chính trong việc xây dựng mô hình đơn giản là làm cho nó có tính xác định - nghĩa là, một mô hình mà trong đó không có xác suất, nhưng trong đó thay vì phát triển cố định theo thời gian, giống như một hệ thống vật lý được lý tưởng hóa theo một phương trình chuyển động bắt đầu từ một tập hợp các điều kiện ban đầu. Để làm được điều này chúng ta cho thời gian chạy từ 0 đến N (thay vì các bước riêng biệt 1,2,3….) và ước tính Xj(t) - số lượng liên kết của nút j - bởi hàm liên tục theo thời gian xj(t). Hàm xjđặc trưng bởi 2 thuộc tính cách tính gần đúng với điều kiện ban đầu và thay đổi dự kiến theo thời gian Xj(t).

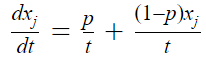
- Hai thuộc tính của hàm xj(t):

The initial condition (Điều kiện ban đầu): Với Xj(j) = 0, thì xj(j) = 0;

The growth equation (Phương trình tăng tiến): Khi đạt tới nút t+1, số lượng liên kết tới nút j tăng với xác suất

https://lh6.googleusercontent.com/YguyOyelQiKlIWrh3GI35-zj0UZDf8A2wobsrtCaiQZxwICEsiZto7OvaYQ-NIXwmSmOA54xQ5s-pTlM2rBmV8_9IiEm0oqEAQJ5pXd8FLzETHj-mrUJA3S4yXAql3zsTktcNMk

- Trong phép xác định gần đúng được cung cấp bởi hàm Xj, chúng ta mô hình tốc độ tăng trưởng này bằng phương trình vi phân:

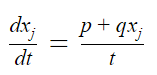


- Sử dụng phương trình vi phân chỉ định hành vi của xj, và xác định gần đúng số liên kết tới nút j theo thời gian. Về bản chất, thay vì xử lý với biến ngẫu nhiên Xj(t) di chuyển trong vùng xác suất nhỏ nhảy vào các điểm riêng biệt (rời rạc) trong miền thời gian t, thì chúng ta làm việc với số lượng xj tăng hoàn toàn xuyên suốt trong miền thời gian, với tốc độ được điều chỉnh phù hợp với thay đổi dự kiến (kỳ vọng) trong các biến ngẫu nhiên tương ứng.

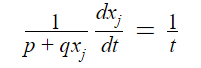
Bây giờ, tìm hiểu về những hệ quả của việc sử dụng phương trình vi phân xác định xj; điều này nhanh chóng dẫn đến sự chi phối luật số lớn mà ta muốn.

- Giải các phép tính gần đúng xác định:

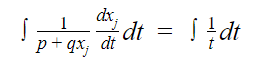
Chúng ta bắt đầu giải quyết phương trình tương đương theo xj, thay q = 1-p ta được phương trình tương đương:

 (1)

Chia cả 2 vế cho p + qxj ta được:



Lấy tích phân 2 vế:



Ta được

ln(p + qxj) = qln(t) + c

Cho hằng số c. Số mũ và cho A = ec ta được

https://lh3.googleusercontent.com/PDoKgkiB4EebNOAiaQ7HlAc7MgSBjW1p7TcIeZuYAJtLJ_DX9hBtGhYhSV03l8R1hI9kFs8QZGHfiTRCiO4aAsZYhHbjyHI2AzUBIiuMQqpcDjK135J4x4kHClB_VI6KxkeBVv4

và vì thế

https://lh3.googleusercontent.com/ql0stWx4OJ28QeqyPWHlMT6oSGfidGaTcU_cYkdq2DvxIvZfQRsF8i_C_hTk8YaAp4kBxhemBPex5X9XSGgyBbxP8Cdk1Uiz8Iob7Razz_KougCzKosHsyWCayyyT-tKZr-6aYo (2)

Bây giờ, chúng ta có thế xác định giá trị của hằng số A bằng cách sử dụng điều kiện ban đầu xj(j) = 0

Điều kiện này cho chúng ta phương trình

https://lh3.googleusercontent.com/2tbwXnMvEwve70cX7AD9RDWrRK-dj9h3JTy8FTzADFf3uOAv2gBDrN_jcX1cSkpJGsnY4HhPD06YVN-EtrC-DHvzNJwm4FL0q8c1CWpAZ5n3mw_IXX-_rqqscfMHbldBNcoRWoQ

vì vậy A = p/jq. Thay A = p/jq vào pt (2) ta được

https://lh6.googleusercontent.com/Azw4m8Wm7dT3ilXX9lWQAdLrTo7rQmAc3ZIjlM1_qQ_PUtOPCGOef1QsFRqy05FWcyqnvTZ6lxa79s6qj5GpMmLJYtzpM5_xpg6kvkHs5vWP0XfP8ySRNwSSj-qKQ3m0H9xVjzw (3)

- Nhận biết luật số lớn trong phép xác định gần đúng.

- Phương trình (3) là một bước trung gian quan trọng để phân tích, vì nó đưa ra biểu thức dạng đóng cho mỗi xj tăng tiến theo thời gian. Bây giờ, ta muốn sử dụng điều này để đặt các câu hỏi sau: Giá trị đã cho của k và thời gian t, tỷ lệ nào của tất cả các nút có ít nhất k liên kết tại thời điểm t? Vì xj xấp xỉ số liên kết của nút j, tương tự với câu hỏi này mà ta đánh giá trong mô hình đơn giản hóa là: Giá trị đã cho k và thời gian t, phần nào của tất cả các hàm xj thỏa mãn Xj(t) >= k?

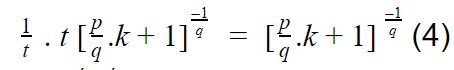
Từ (3) ta có

https://lh3.googleusercontent.com/F3ENijy5Q3ZJNSTWJHL-gDMSonKLeqG9ka5BNhZzfj4rwjNJqVuo0reo1jHOlEy0JwegEZtGSKEd77tovIV5xTlPBaciQI2MEd1UqvH90NSPAvfaD5MPSLrgTy6Y41eDc47gwuc

Suy ra, điều kiện của j

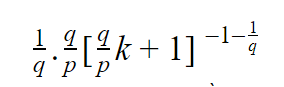
https://lh6.googleusercontent.com/AXLW3b43A_iyEoH8i9DvkpQ60khHgGw4uOUpaQSJwqUq4ZEvur0hXzbIkEJl5B5_qKI3OwBZNO88A4t0tJSY7qGKaJLEeB9nHu_wBY7JKfiLUiwIYOps5HA8CrnREdy39qjK864

- Trong tất cả các hàm x1,x2,....,xt tại thời gian t, thì phần của giá trị j thỏa mãn điều kiện này là đơn giản.



- Chúng ta có thế thấy luât số lớn được hình thành ở đây: vì p và q là hằng số, các biểu thức trong ngoặc bên vế phải tỷ lệ với k, và vì vậy tỷ lệ xj có ít nhất k tỷ lệ với k-1/q.

- Đối với bước cuối cùng, lưu ý ở bước này là về tỷ lệ của các F(k) với ít nhất k liên kết. Nhưng từ điều này, chúng ta có thể xấp xỉ trực tiếp tỷ lệ với nút f(k) với chính xác k liên kết đơn giản bằng cách lấy đạo hàm. Nói cách khác, xác định gần đúng f(k) bằng –dF/dk. Lấy vi phân (4) ta được



- Nói cách khác, dự đoán mô hình xác định rằng tỷ lệ của các nút với k liên kết là tỷ lệ với k-(1+(1/q)) . Luật số lớn với lũy thừa:

https://lh4.googleusercontent.com/8p5CimVpOsypjbiYrG5Tku-GVzs1zl-gY8xYQbXCd-oPp545V7JUcXM42Zib_8-emjRtR2ML1j40sFsY2FxOxTYjvt1hmDCebBjawv1JhKb4G8i-j6OEryrNjiQRFRnP_KeK434

- Phân tích tiếp theo của mô hình xác suất ban đầu cho thấy rằng, với xác suất cao trên sự hình thành ngẫu nhiên của liên kết, tỷ lệ của các nút với k liên kết là tỷ lệ thuận với https://lh6.googleusercontent.com/1U3A33iJ9rO2o8AJEiZwJf2HGzC6IpDnUFr5dtFfo2wA-PpI5UugerDiR6kFoZt6l8tZW2juFyUEBOCdT04hDFU2DvCxFw4NIjCmB3kf4xk_XXUMpBvkVi7KzCD0SUJGG2sxiZA.Đối số được cung cấp bởi phép tính xác định gần đúng cho mô hình, do đó cung cấp 1 cách đơn giản để xem vị trí lũy thừa https://lh6.googleusercontent.com/FRb0T2QHNgjzBW9KY42csLAwL1Nm4YQip1lX1-_L_4lC5qjeIuiMZMaA4BurZ1ZuFwk5FfvlA88idR80oll6uz4u0eBPjXnF0Y2gWg-FjAhqIoTurFI3fkTg2FHubJ5PJsXsqO0 này theo luật số lớn đến từ đâu.

- Hành vi của số mũ này cũng có ý nghĩa bằng trực giác khi thay đổi p. Khi p gần đến 1 cảm nhận bằng trực giác, hình thành liên kết chủ yếu dựa trên sự lựa chọn ngẫu nhiên thống nhất. Tương tự, số mũ luật số lớn có xu hướng tiến tới vô cùng, cho thấy các nút có số lượng liên kết lớn ngày càng hiếm. Mặt khác, khi p dần về 0, sự phát triển của mạng lưới bị chi phối mạnh mẽ bởi hành vi của Rich-Get-Richer, và số mũ của luật số lớn giảm về 2, cho phép nhiều nút có số lượng liên kết rất lớn. Thực tế thì 2 là giới hạn tự nhiên cho số mũ như phân tích Rich-Get-Richer trở nên mạnh mẽ hơn cũng cung cấp 1 cách tốt để suy nghĩ về thực tế rằng nhiều số mũ luật cống suất trong mạng thực có xu hướng cao hơn 2. Một tính năng hấp dẫn cuối cùng của phép xác định gần đúng là nó rất dễ uốn - nó có thể dễ dàng sửa đổi để bao phủ các phần mở rộng của mô hình.