Khai Thác Dữ Liệu Đồ Thị

MẪU ĐỒ THỊ

Giảng viên: Lê Ngọc Thành

Email: lnthanh@fit.hcmus.edu.vn

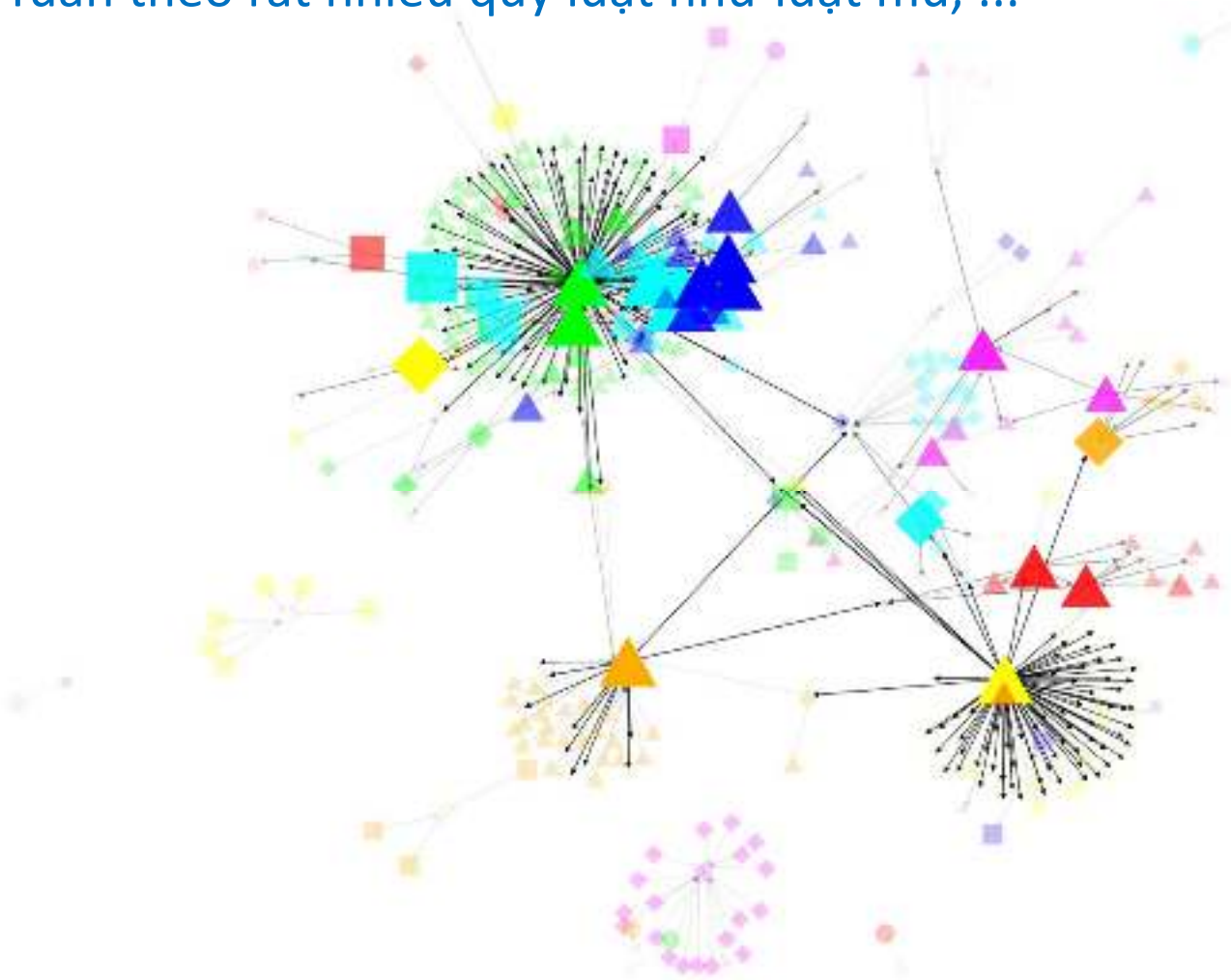
fit@hcmus

Nội dung

- **Mẫu đồ thị**
- Mẫu trong đồ thị tĩnh
- Mẫu trong đồ thị động
- Mẫu trong đồ thị có trọng số
- Chi phí tính toán

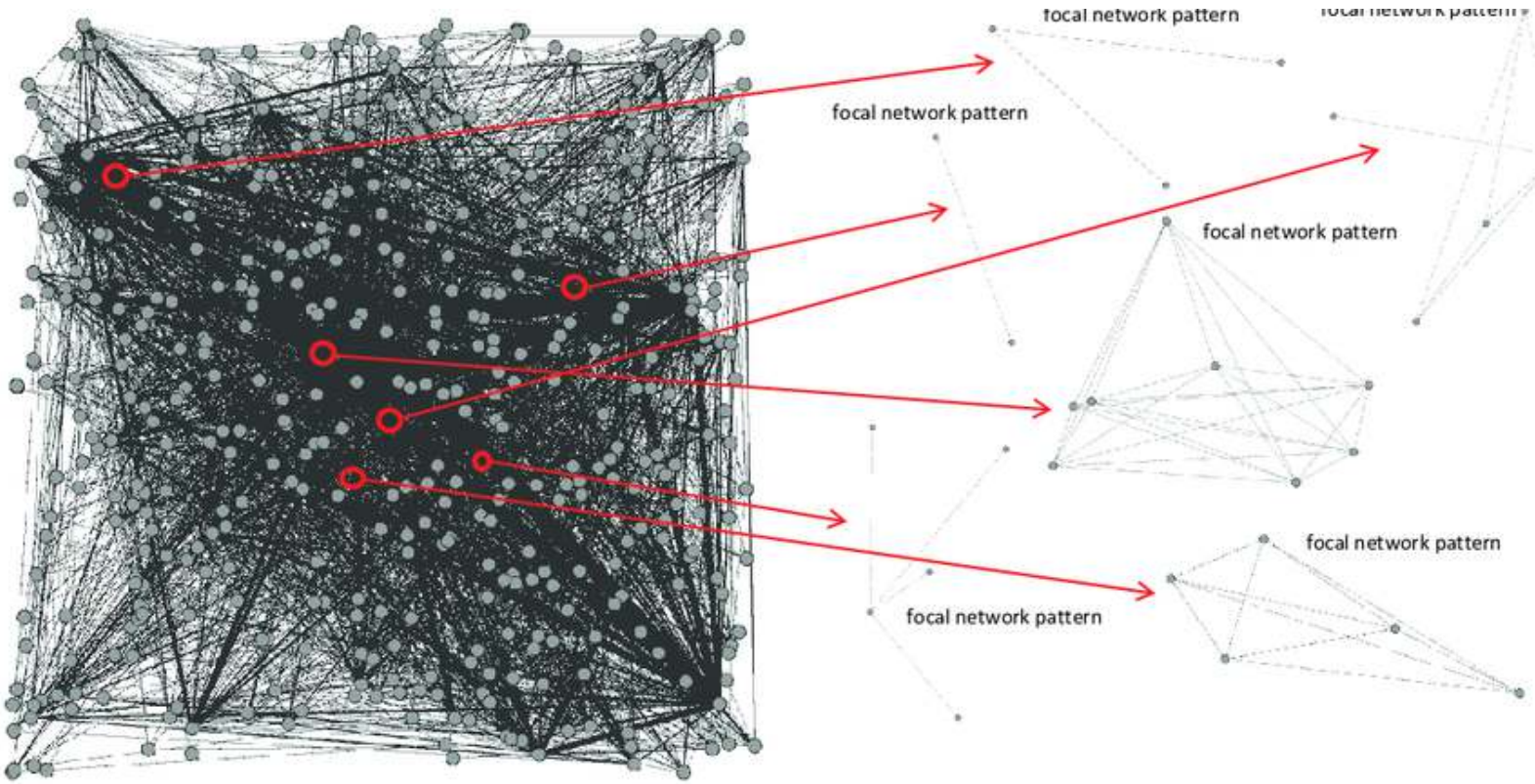
Đồ thị thế giới thật

- Có phải đồ thị thế giới thật ngẫu nhiên?
  - Không!!!
    - Tuân theo rất nhiều quy luật như luật mũ, ...



Mẫu đồ thị

- Mẫu đồ thị là dạng thuộc tính hoặc đồ thị con thường xuất hiện trong đồ thị thế giới thật.



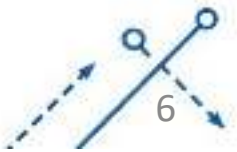
Các mẫu trong mạng xã hội gồm 442 đỉnh và 3171 cạnh được thu thập từ http://www.livejournal.com.

Nghiên cứu mẫu đồ thị

- Tại sao phải xem xét các mẫu đồ thị?
  - Hiểu các thuộc tính thú vị của đồ thị thế giới thật
  - Các mẫu thể hiện thông tin cô đọng về đồ thị
  - Sử dụng để phát sinh ra đồ thị tương tự với đồ thị thế giới thật phục vụ để nghiên cứu
  - Hỗ trợ phát hiện các bất thường và điểm ngoại lai.

Phát hiện bất thường

- Trong đồ thị thế giới thật có thể chứa:
  - Cạnh bất thường
  - Đỉnh bất thường
  - Đồ thị con bất thường
- "Bất thường" thường khác với các mẫu "bình thường". => Cả outlier lẫn noise
  - Do đó, hiểu các mẫu xảy ra một cách tự nhiên là điều kiện tiên quyết để xác định mẫu bất thường.

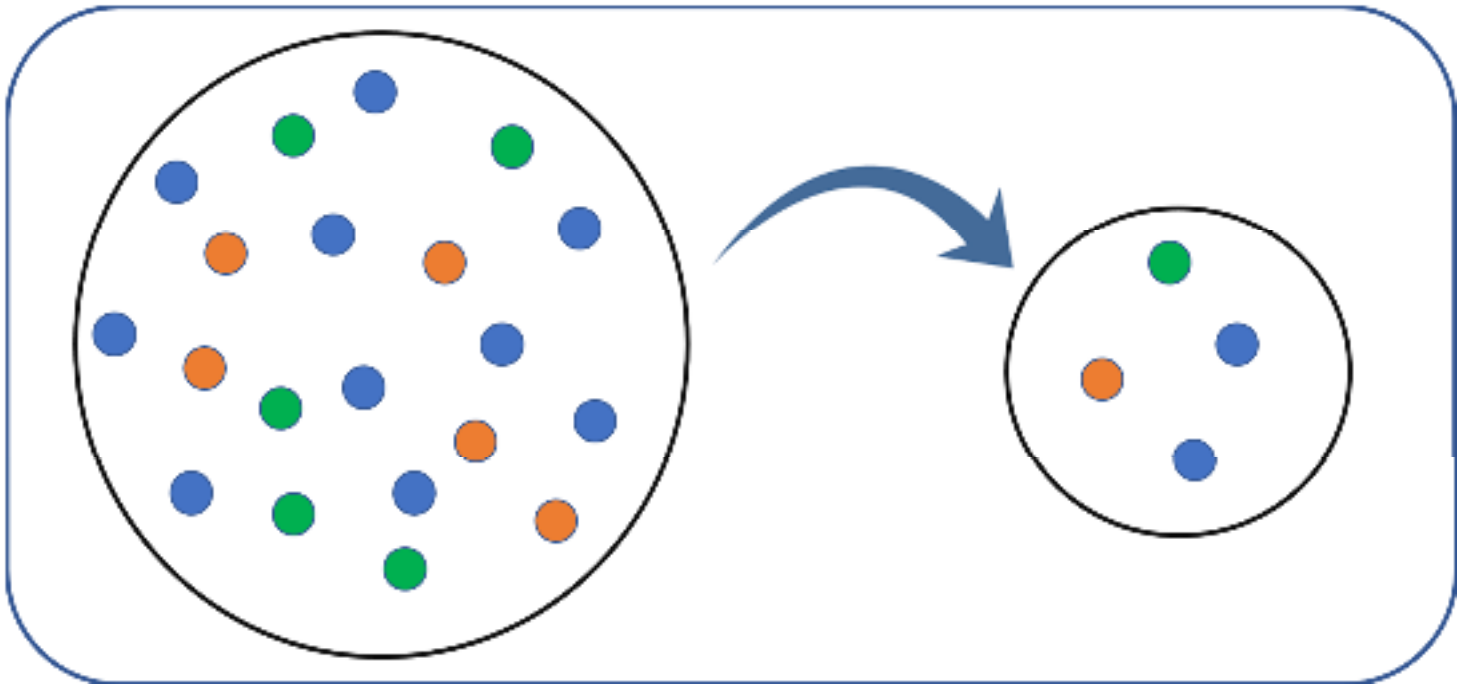


Giả lập

- Kiểm tra các thuật toán chạy trên đồ thị thế giới thật khổng lồ cần các đồ thị được phát sinh giống với nó:
  - Một số tổ chức không chia sẻ dữ liệu
  - Cần kiểm tra trước khi nó xảy ra
    - Ví dụ: kiểm tra giao thức Internet thế hệ mới, cần giả lập một đồ thị tương tự với thế giới Internet trong vài năm tới.

Lấy mẫu

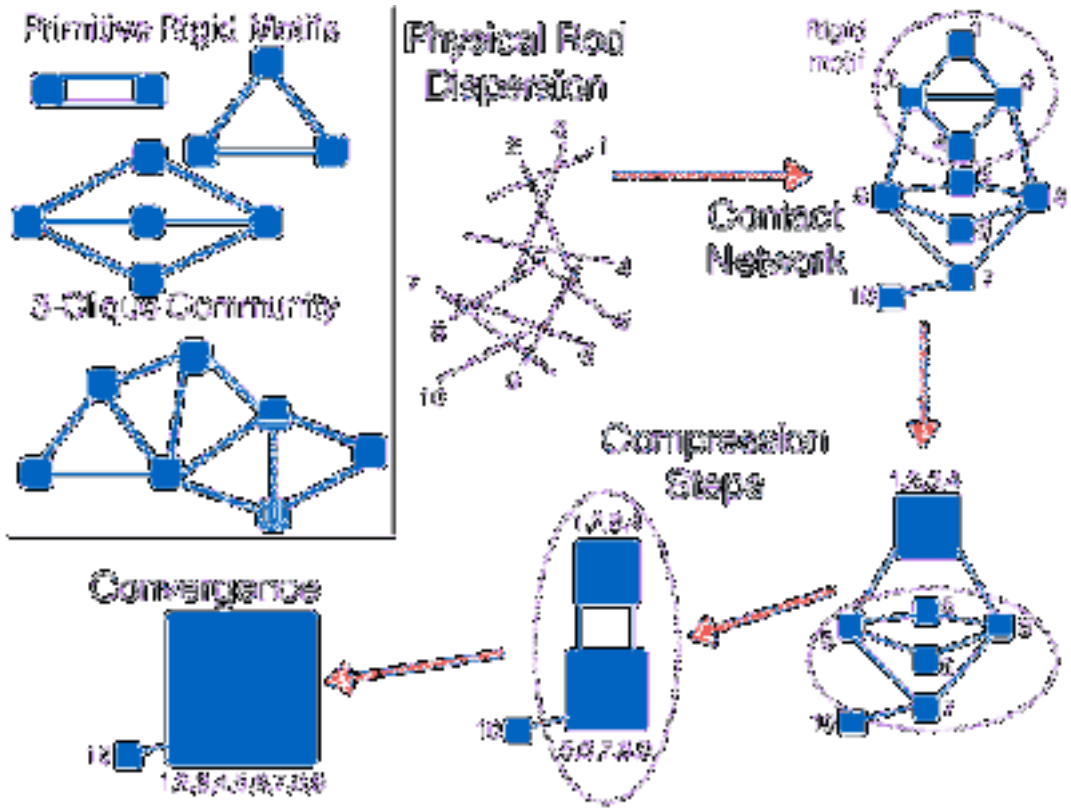
- Khi xây dựng/trích xuất một đồ thị mẫu nhỏ, ta thường trông đợi đồ thị tương tự với dạng đồ thị lớn hơn.



Nén đồ thị

- Các mẫu đồ thị thể hiện các quy tắc trong dữ liệu.

  Các quy tắc này có thể được sử dụng để nén dữ liệu đồ thị.

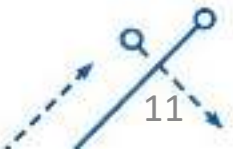


Bài toán phát hiện mẫu đồ thị

- Bài toán dễ hay khó?
  - Khó!!!
  - Tại sao?
    - Cần xác định mẫu "tốt"?
      - Mẫu tốt là mẫu giúp phân biệt giữa đồ thị thế giới thật và mẫu được tạo giả.
      - Vậy nó là mẫu gì?
    - Có phải chỉ cần 1 mẫu là đủ để có thể phân biệt được?
    - Có tính toán hiệu quả trên đồ thị lớn?
      - Một mẫu mất $O(N^3)$ hay $O(N^2)$ với N là số đỉnh của đồ thị trở nên không thực tế.

Nội dung

- Mẫu đồ thị
- **Mẫu trong đồ thị tĩnh**
  - Phân phối bậc
  - Luật mũ
  - Luật mũ giá trị riêng
  - Luật mũ tam giác
- Mẫu trong đồ thị động
- Mẫu trong đồ thị có trọng số
- Chi phí tính toán

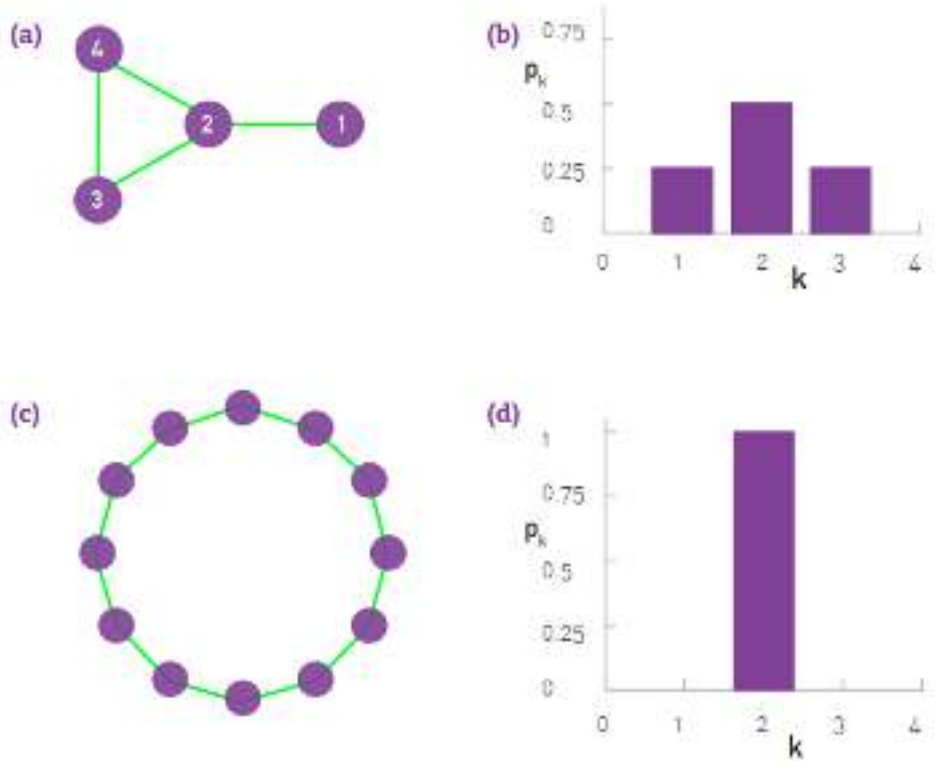


Phân phối bậc trong đồ thị

- Cho đồ thị có N đỉnh, phân phối bậc của đồ thị là một histogram được chuẩn hóa:

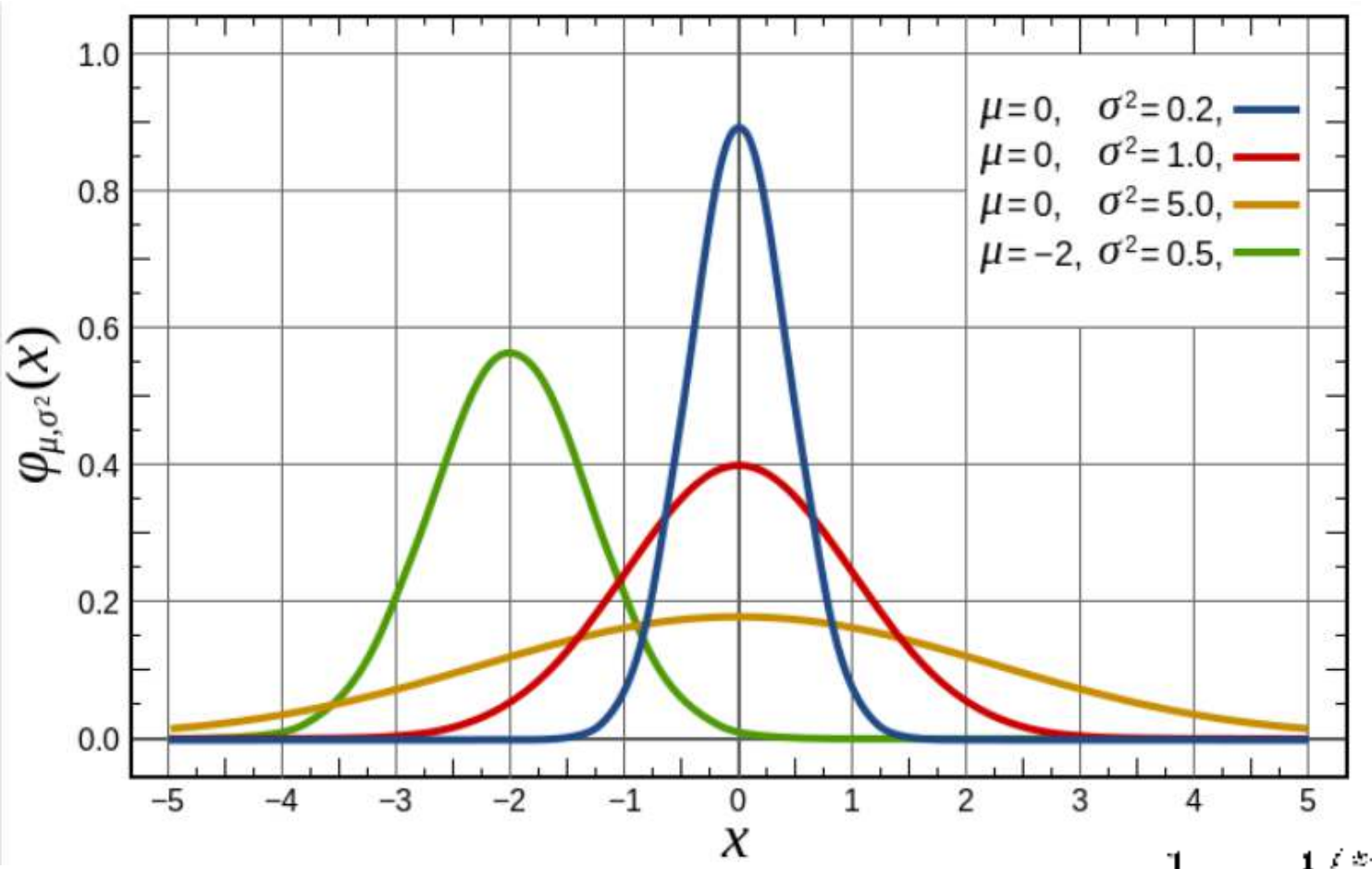
$$p_k = \frac{N_k}{N}$$

Với N_k là số đỉnh có bậc k



Phân phối bậc trong đồ thị

- Có phải đồ thị trong thế giới thật có phân phối theo phân phối tự nhiên (normal/gaussian distribution)?



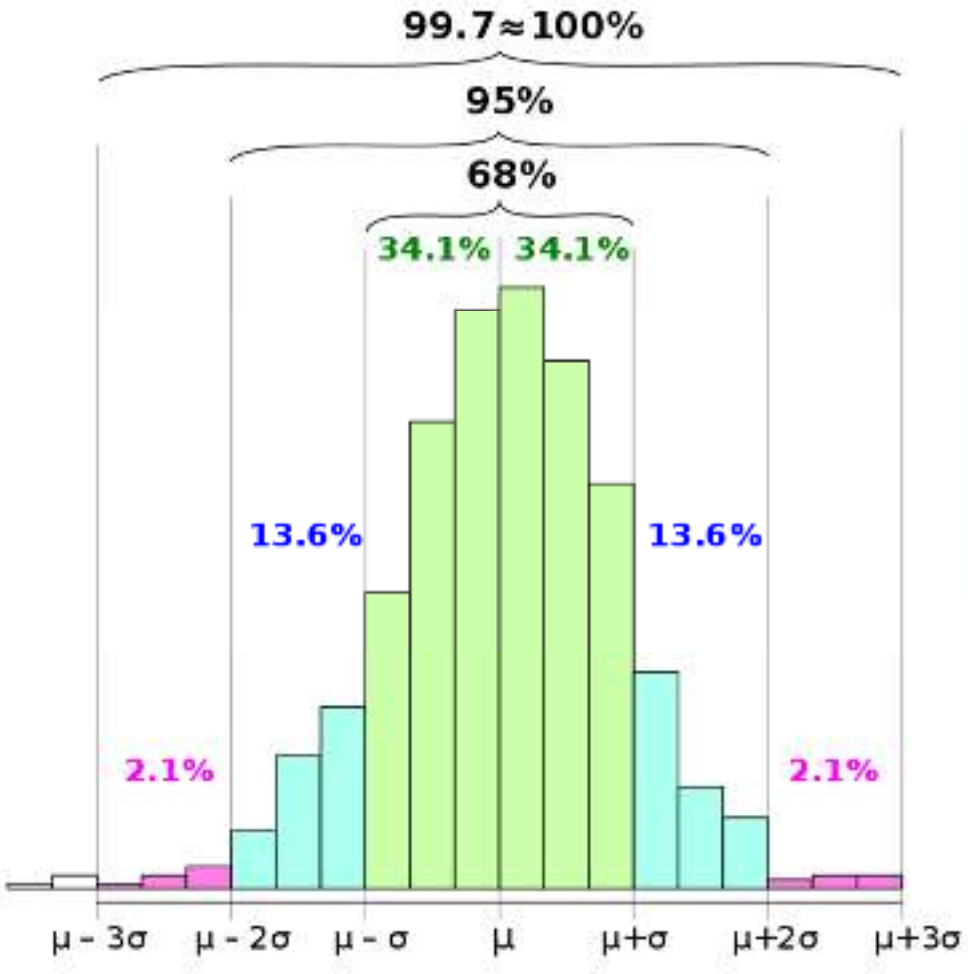
Ví dụ:
- Chiều cao con người
- Điểm số của sinh viên
- …

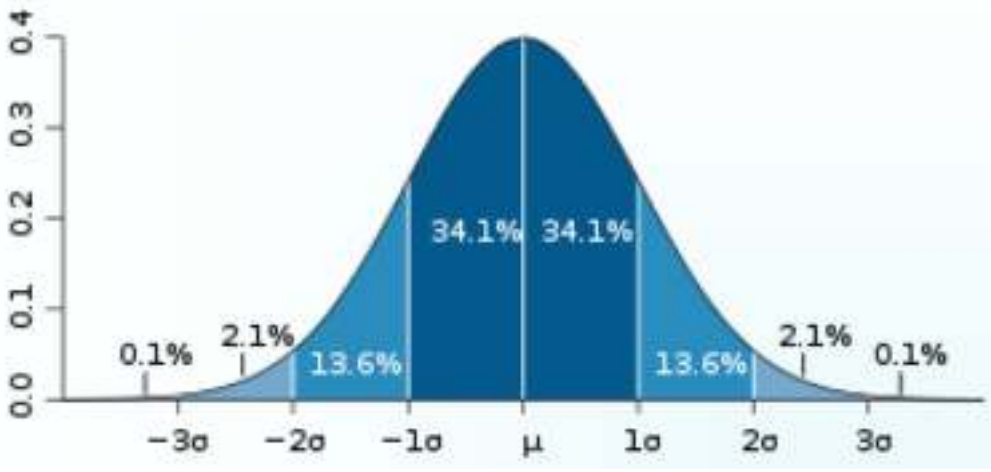
Đặc điểm:
- Số lượng mẫu có xu hướng hội tụ (tăng lên) tại điểm trung tâm (kỳ vọng) …

Phân phối tự nhiên/chuẩn/hình chuông $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Luật 3-sigma

- Luật 3-sigma hay luật 68-95-99.7 chỉ ra rằng khoảng 99.7% dữ liệu rơi vào trong khoảng 3 lần độ lệch chuẩn xung quanh giá trị kỳ vọng.





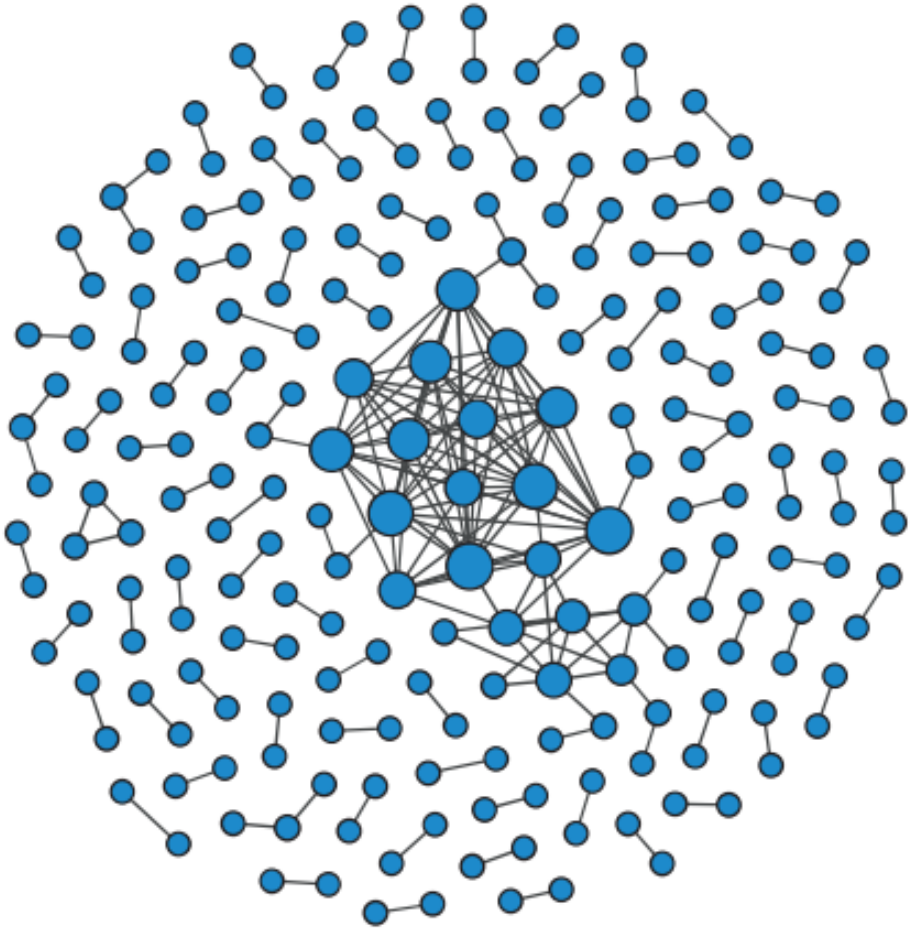
Phân phối bậc trong đồ thị

- Có phải đồ thị trong thế giới thật có phân phối theo phân phối tự nhiên (normal/gaussian distribution)?
  - Ví dụ 1: trong mạng xã hội có 1 triệu đỉnh, mỗi đỉnh có trung bình 50 kết nối (bạn bè).
    - Nếu chọn 1 đỉnh ngẫu nhiên, bạn đoán xem số lượng bạn bè họ có là bao nhiêu?
    - Bạn có ngạc nhiên nếu tôi nói trong mạng xã hội này có người có 10,000 bạn bè?

Phân phối bậc trong đồ thị

- Có phải đồ thị trong thế giới thật có phân phối theo phân phối tự nhiên (normal/gaussian distribution)?
  - Ví dụ 1: trong mạng xã hội có 1 triệu đỉnh, mỗi đỉnh có trung bình 50 kết nối (bạn bè).
    - Nếu chọn 1 đỉnh ngẫu nhiên, bạn đoán xem số lượng bạn bè họ có là bao nhiêu?
      - Gần bằng 1
      - Hầu hết mọi người hiếm khi kết nối, nguyên nhân có thể: không dễ dàng truy cập máy tính, internet, hoặc chi phí cao, quá phức tạp, mang lại ít thuận lợi cho họ, …
    - Bạn có ngạc nhiên nếu tôi nói trong mạng xã hội này có người có 10,000 bạn bè?
      - :o
      - Một số người muốn có rất nhiều bạn như diễn viên, ca sĩ hoặc họ đang tham gia cuộc thi mức phổ biến (popularity contest) trên Facebook, …

Phân phối bậc trong đồ thị



Phân phối bậc trong đồ thị

- Có phải đồ thị trong thế giới thật có phân phối theo phân phối tự nhiên (normal/gaussian distribution)?
  - Ví dụ 2: một cuốn từ điển Oxford tiếng Anh có khoảng 100,000 từ. Qua khảo sát tài liệu hàng ngày, người ta tính được trung bình mỗi từ được dùng khoảng 5,000 lần
    - Liệu có từ nào xuất hiện >> 5,000
    - Có từ nào xuất hiện = 0?

Phân phối bậc trong đồ thị

- Có phải đồ thị trong thế giới thật có phân phối theo phân phối tự nhiên (normal/gaussian distribution)?
  - Ví dụ 2: một cuốn từ điển Oxford tiếng Anh có khoảng 100,000 từ. Qua khảo sát dữ liệu âm thanh hàng ngày, người ta tính được trung bình mỗi từ được nói khoảng 5,000 lần
    - Liệu có từ nào xuất hiện >> 5,000
      - Có: a, an, the, …
    - Có từ nào xuất hiện = 0?
      - Có: người ta nhận thấy một người bản địa trưởng thành không dùng nhiều hơn 2,000 từ vựng trong cuộc sống hàng ngày của họ.

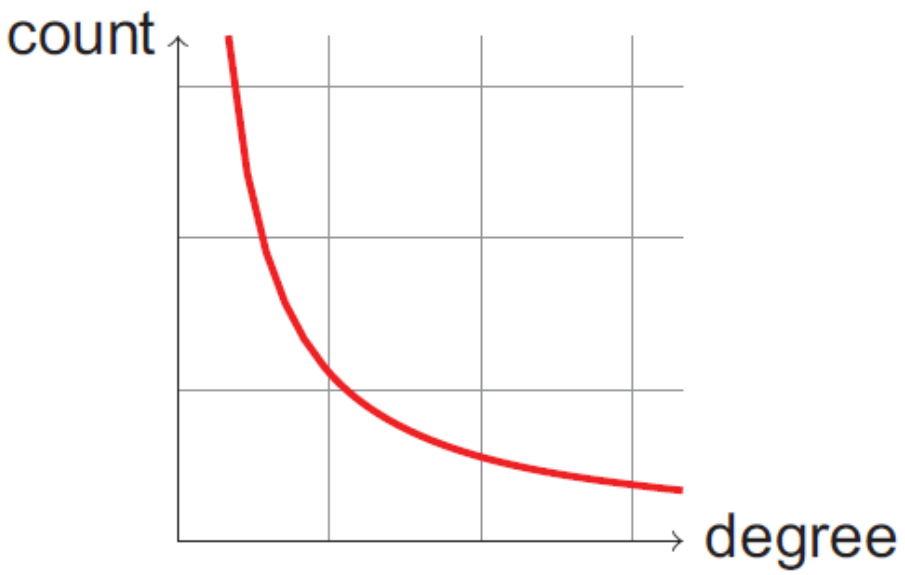
Phân phối bậc trong đồ thị

- Có phải đồ thị trong thế giới thật có phân phối theo phân phối tự nhiên (normal/gaussian distribution)?
  - Ví dụ 3: các router khi tham gia kết nối Internet?
    - Liệu có router nào kết nối cực lớn?
    - Liệu có router nào kết nối rất ít?

Phân phối bậc trong đồ thị

- Phân phối bậc trong đồ thị/mạng thế giới thật thường bị lệch.

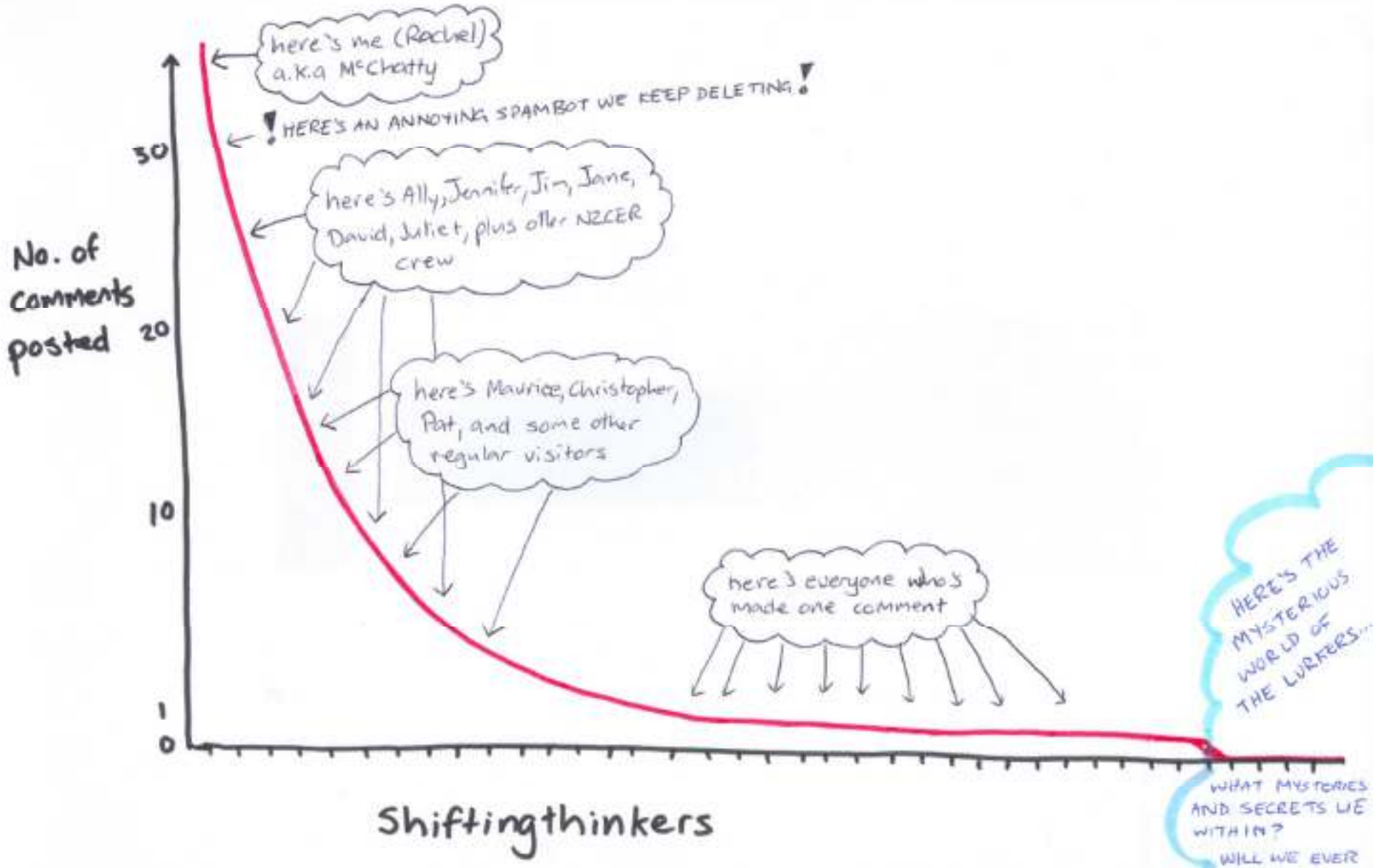
(a) WRONG intuition



(b) Reality

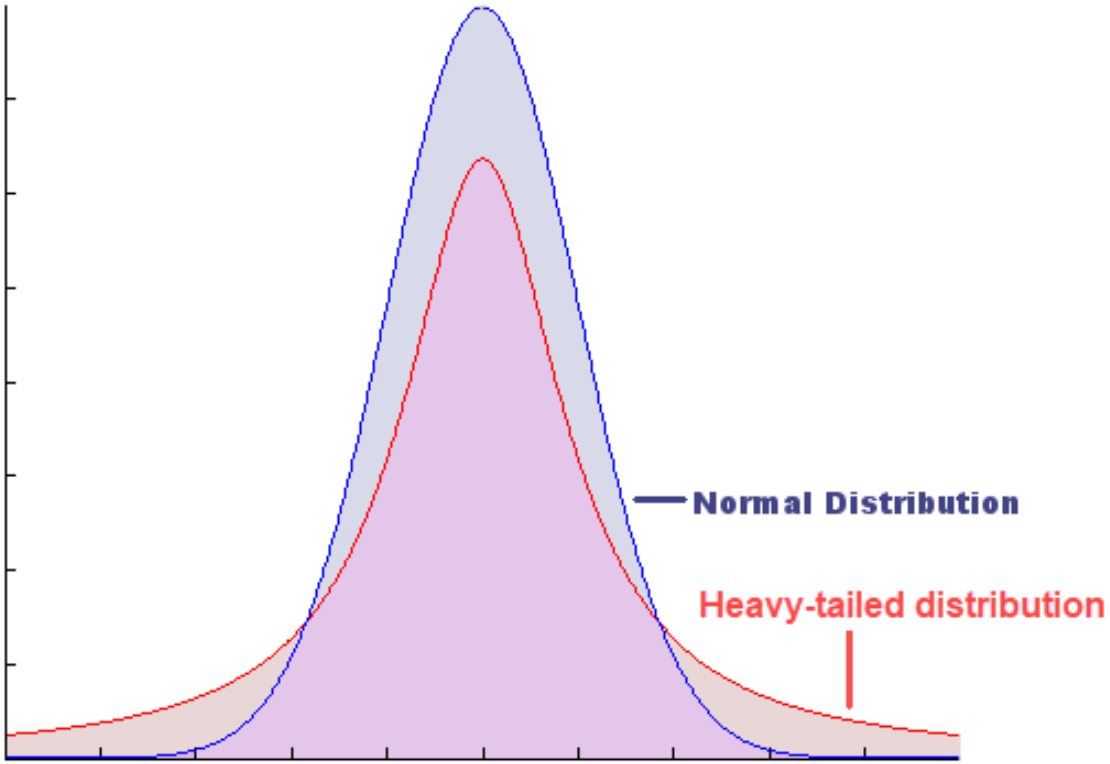
Các đỉnh trong phân phối chuẩn có bậc gần với giá trị trung bình, giá trị lệch lớn so với nó gần như không xảy ra. Trong khi đó, đối với đồ thị thế giới thật, rất nhiều đỉnh hiếm khi được kết nối.

Phân phối số lượng comment trên một blog



Phân phối bậc trong đồ thị

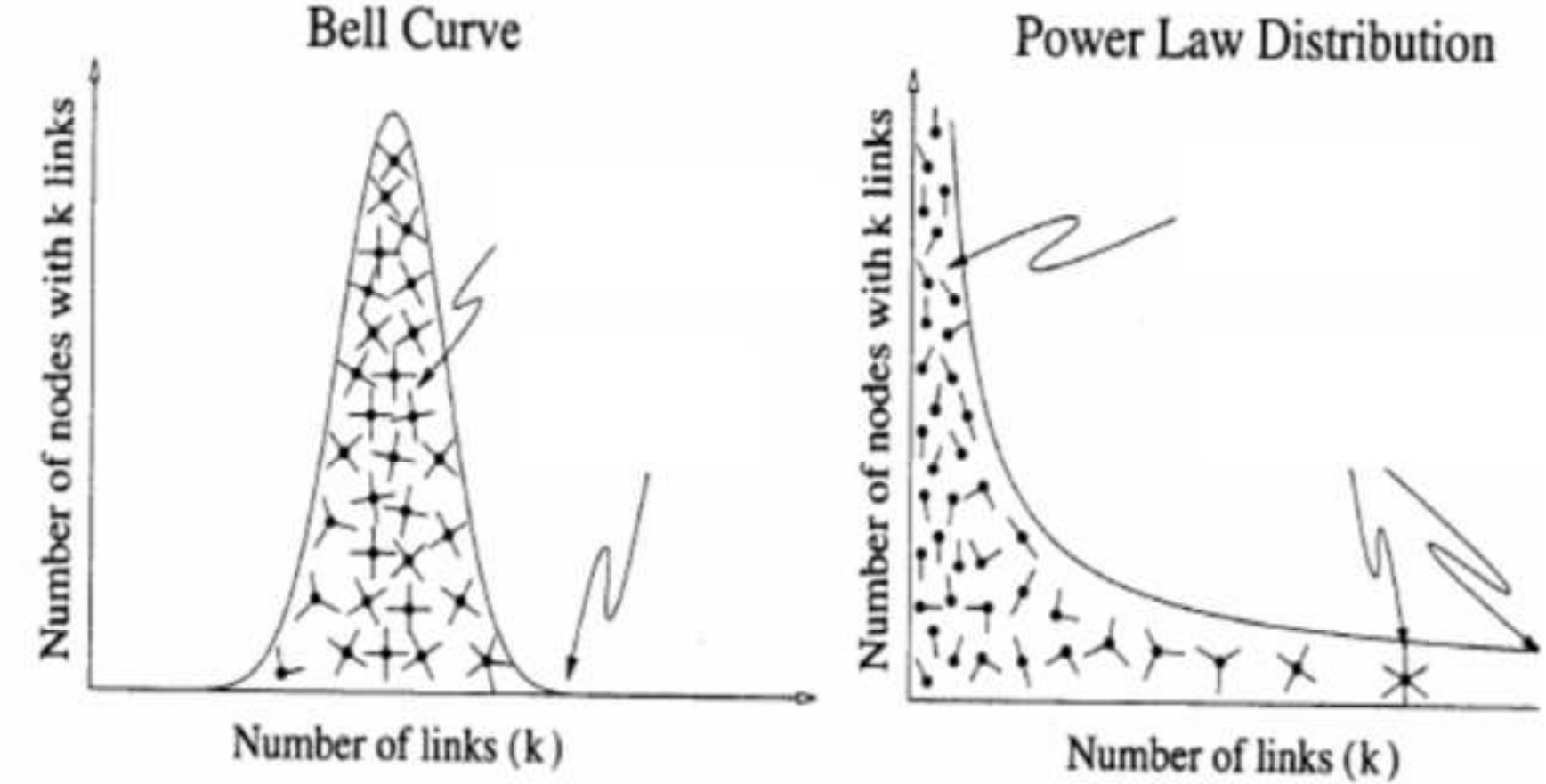
- Phân phối bậc trong đồ thị thường có dạng đuôi "nặng" (heavy-tailed distribution)
  - Đặc điểm: tồn tại các phần tử có giá trị rất lớn so với giá trị trung bình. Nói cách khác, nó giảm chậm hơn khi $x \to \infty$



Phân phối đuôi nặng

- Một số hàm phân phối có dạng đuôi nặng:
  - Phân phối luật mũ (Power-law distribution)
  - Phân phối chuẩn log (Log-normal distribution)
  - Phân phối Weibull (Weibull distribution)
- Thông thường dữ liệu đồ thị tuân theo phân phối luật mũ.

So sánh phân phối chuẩn và đuôi nặng



Phân phối luật mũ

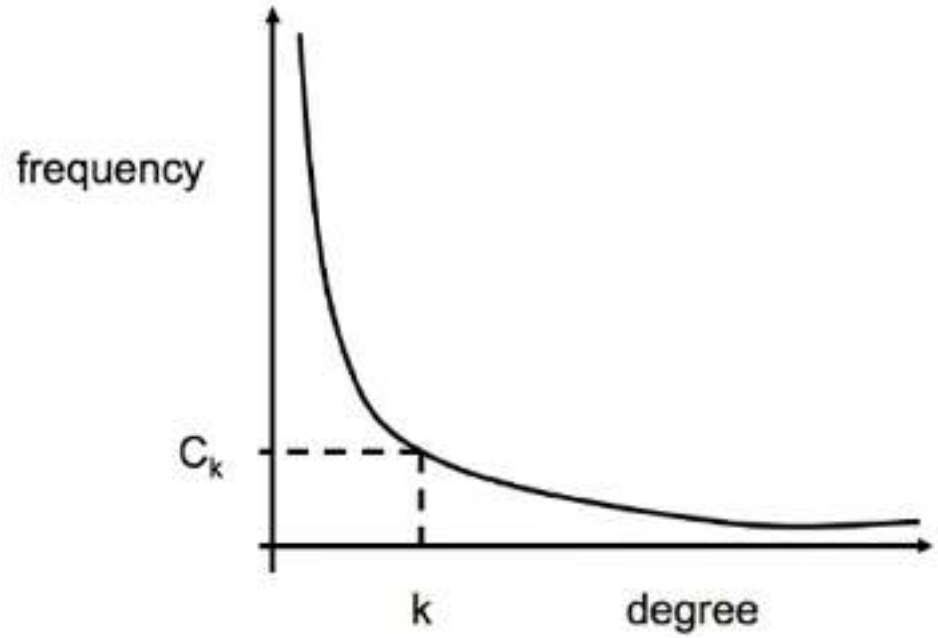
- Phân phối bậc trong đồ thị thường tuân theo luật mũ (power law):

$$f(d) \propto d^{-\alpha}$$

Với d là bậc

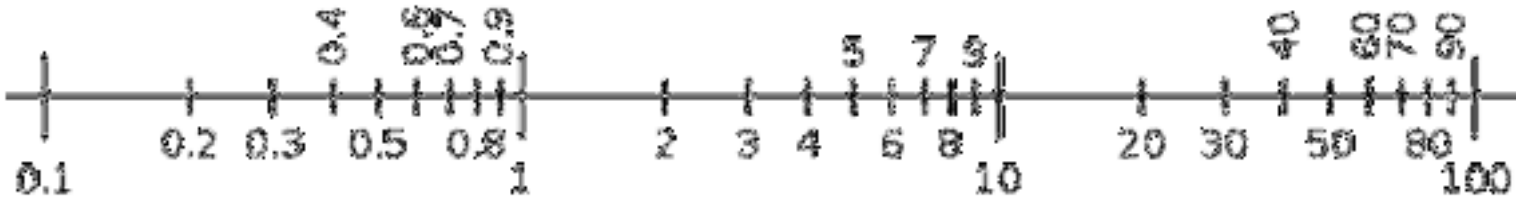
α là hằng số dương, được gọi là lũy thừa của luật mũ (power-law exponent)

+ d càng lớn (bậc càng cao) thì xuất hiện càng ít và bậc nhỏ (ít liên kết) xuất hiện nhiều



Phân phối luật mũ

- Giá trị phân phối mũ thường rất nhỏ khi bậc càng lớn
- Để biểu diễn phân phối mũ trên hệ trục tọa độ, người ta thường chuyển giá trị về không gian logarit:
  - Thể hiện một miền giá trị lớn trong một không gian nhỏ gọn.
  - Ví dụ: 10, 100 sẽ biểu diễn độ rộng như nhau

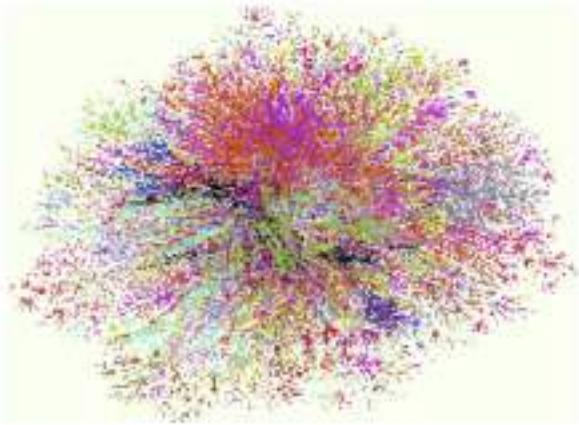


Phân phối luật mũ

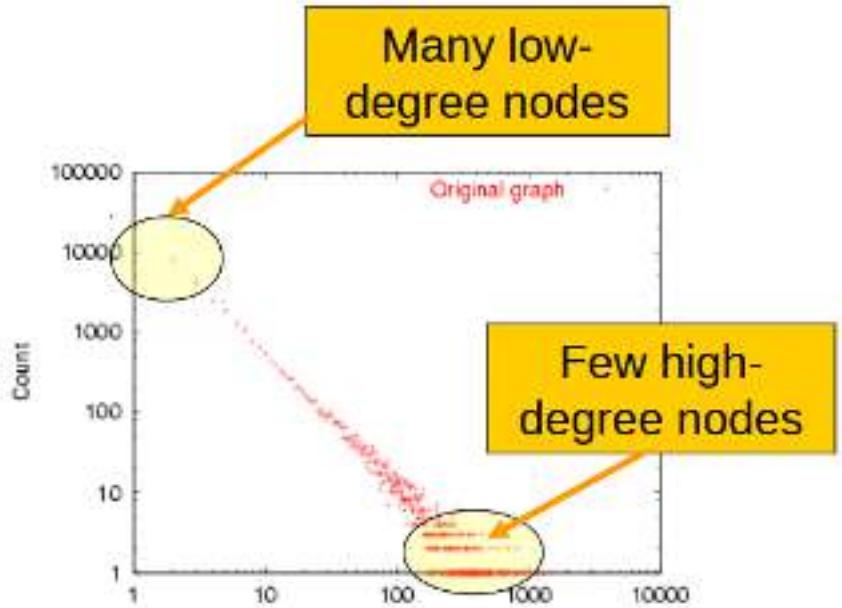
$$f(d) = Ad^{-\gamma}$$

$$\log(f(d)) = \log(A) - \gamma \log(d)$$

- Đồ thị biểu diễn phân phối mũ dạng này được gọi là đồ thị log-log.



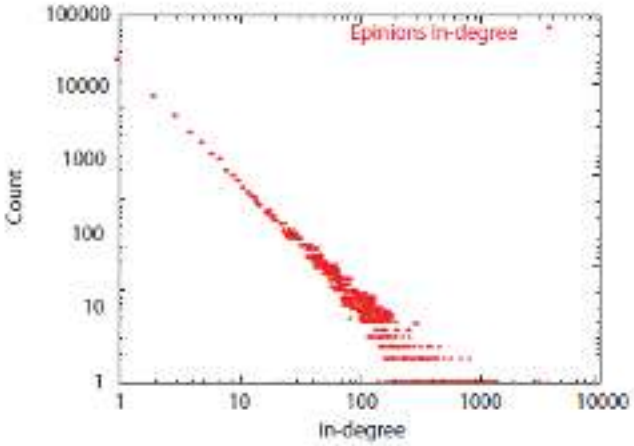
Internet in December 1998



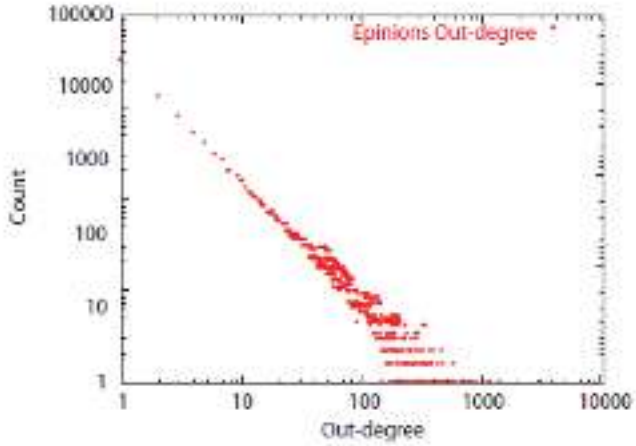
log(Count) vs. log(Degree)

Ví dụ phân phối luật mũ

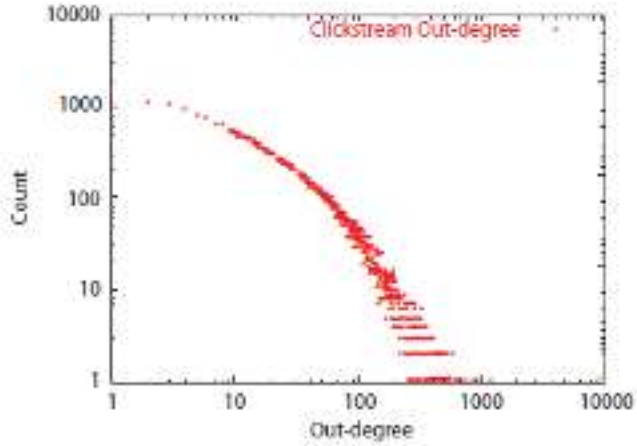
- Trong đồ thị mạng WWW và mạng xã hội, phân phối bậc trong, bậc ngoài cũng tuân theo luật mũ.



(a) Epinions In-degree



(b) Epinions Out-degree

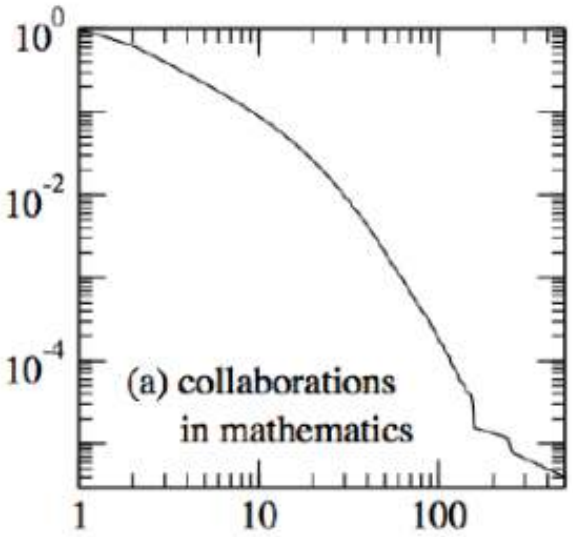


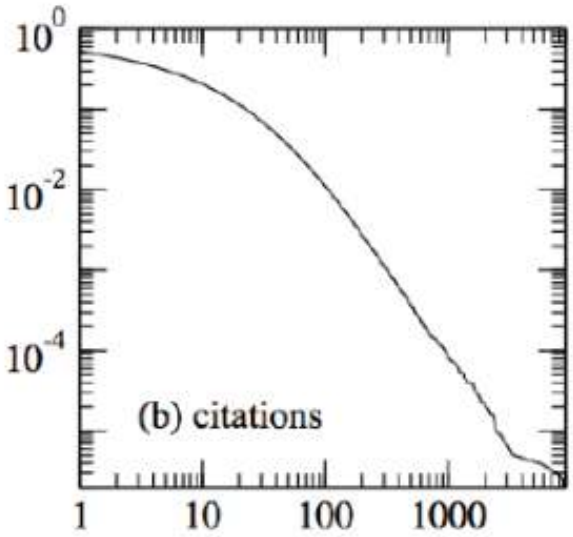
(c) Clickstream Out-degree

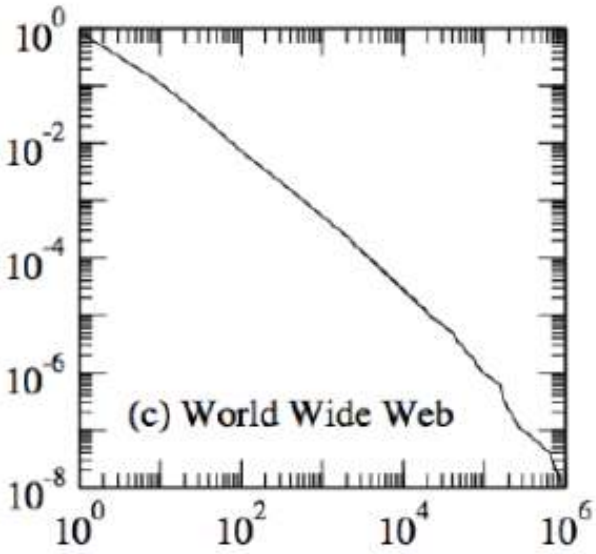
Figure: *Power laws:* Plots (a) and (b) show the in-degree and out-degree distributions on a log-log scale for the *Epinions* graph (an online social network of $75K$ people and $508K$ edges). Plot (c) shows the out-degree distribution of a *ClickStream* graph (a bipartite graph of users and the websites they surf).

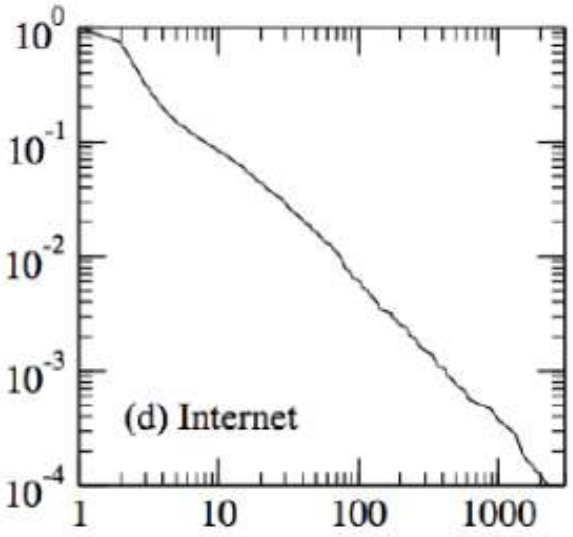
=> Hầu hết các mạng có xuất hiện luật mũ nhưng cũng có 1 vài mạng ko tuân theo luật mũ

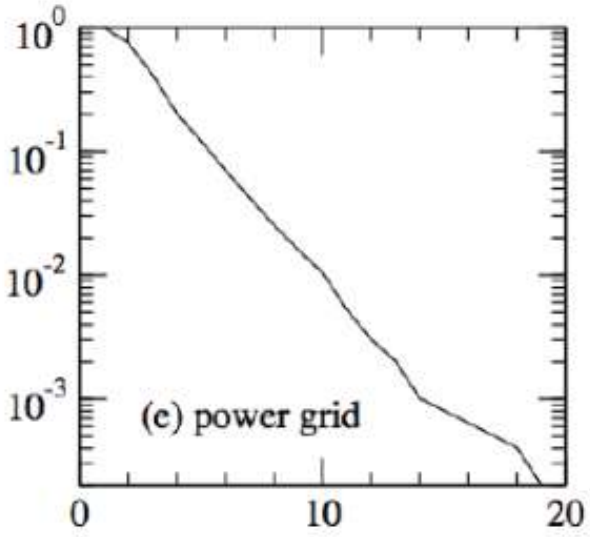
Ví dụ phân phối luật mũ

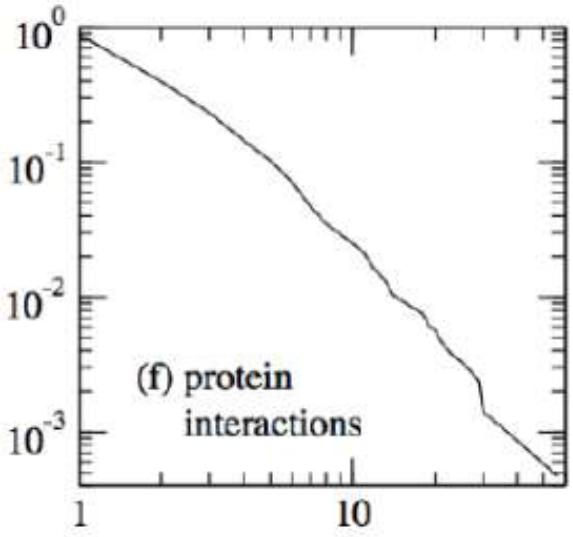












Bài tập

- Bài tập 1: Sử dụng thư viện NetworkX để tạo ra một bộ dữ liệu tuân theo phân phối luật mũ với hệ số gamma lần lượt là 1 2,2 3
- Bài tập 2: Sử dụng thư viện Matplotlib để vẽ biểu đồ phân phối trên

Đồ thị scale-free

- Đồ thị với phân phối bậc theo luật mũ cũng được gọi là đồ thị scale-free.
  - Đặc trưng của nó độc lập với kích thước của đồ thị. Nghĩa là khi đồ thị phình ra, cấu trúc bên dưới vẫn tương tự.
  - Với luật mũ: $y(x) = Ax^{-\gamma}$ thì $y(ax) = by(x)$

+ x là bậc

Luật mũ giá trị riêng

- Siganos đã nhận thấy quang phổ (một tập các giá trị riêng được sắp xếp giảm dần về độ lớn) của một ma trận kề biểu diễn đồ thị mạng Internet cũng tuân theo phân phối luật mũ.

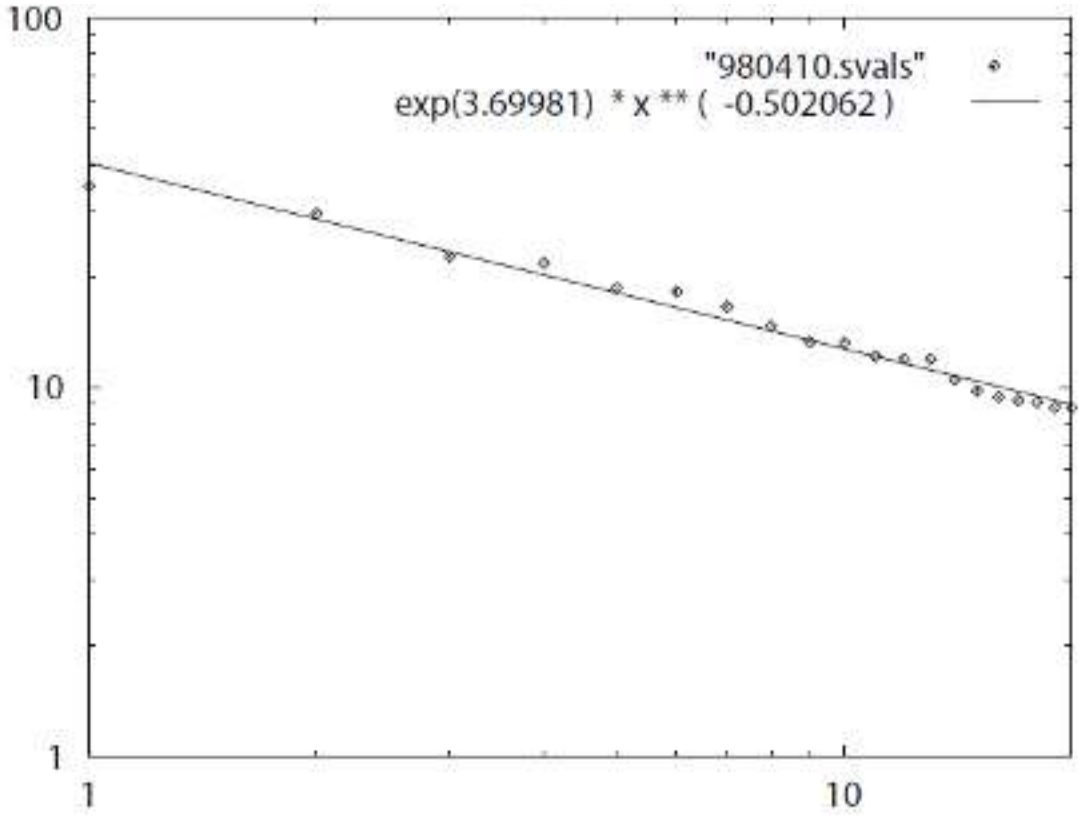
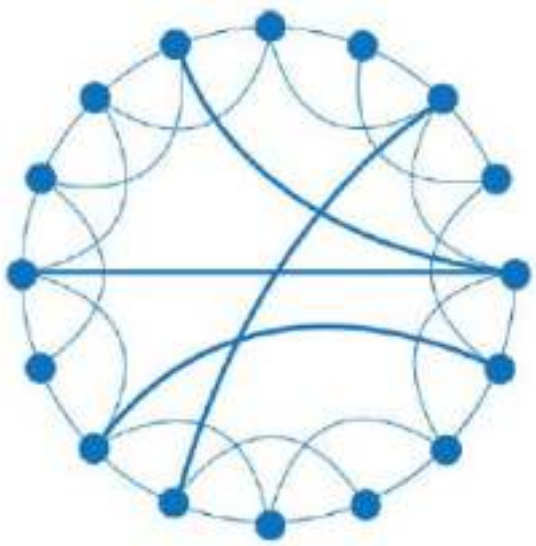


Figure: Scree plot obeys a power law: Eigenvalue λ_i versus rank i for the adjacency matrix of autonomous systems (April 10, '98).

=> Đường kính là đường đi ngắn nhất của 2 điểm xa nhau nhất

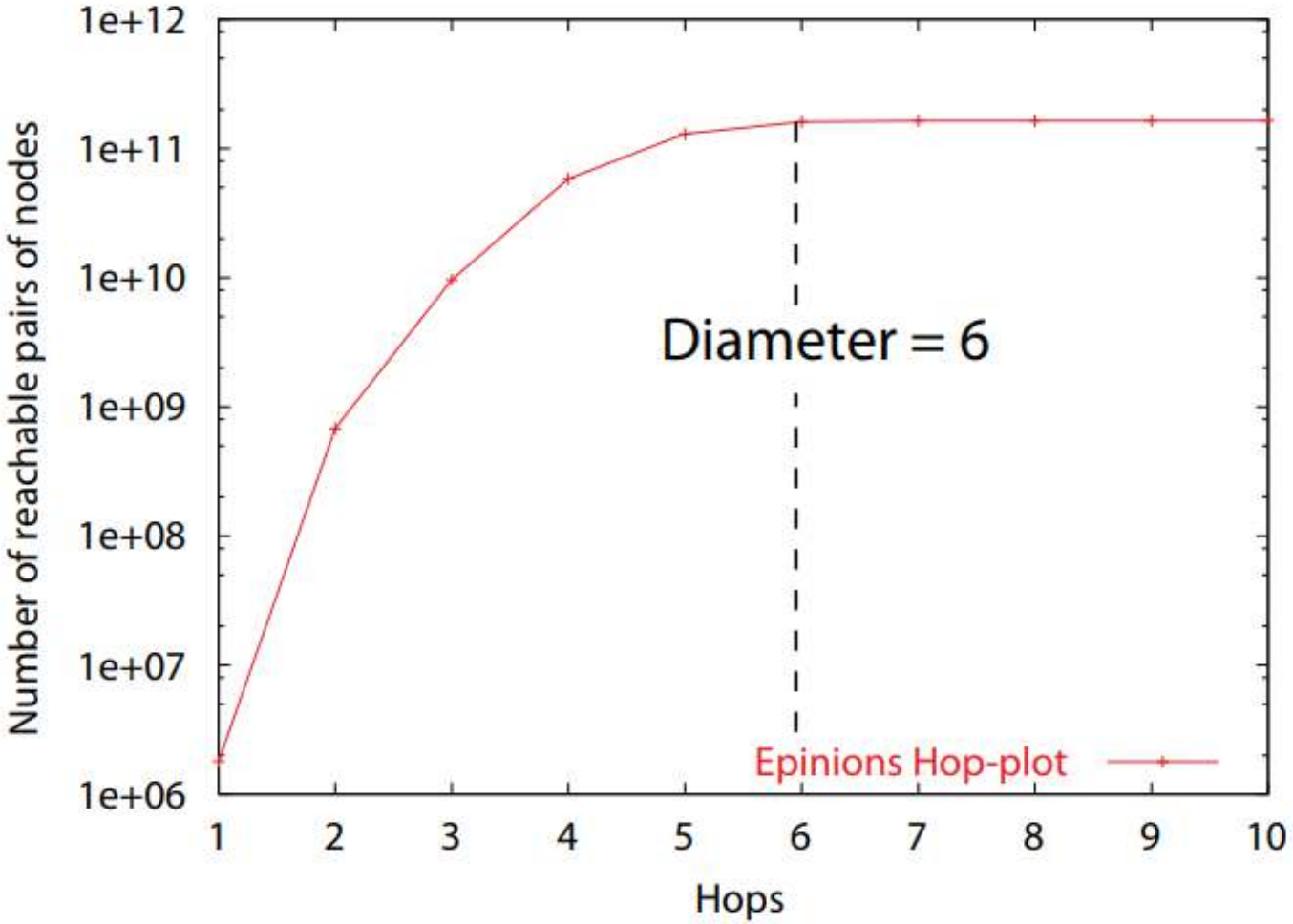
Đường kính nhỏ

- Đường kính của một đồ thị là khoảng cách lớn nhất giữa bất kỳ cặp đỉnh nào.
  - Khoảng cách được tính trên đường đi ngắn nhất giữa hai đỉnh, không quan tâm đến hướng.
- Đồ thị thế giới thật cũng thường có đường kính nhỏ
  - Cũng được gọi là hiện tượng thế giới nhỏ (small-world phenomenon)
  - Trong mạng xã hội, nó cũng được gọi là sáu bậc chia (six degrees of separation)



Đường kính nhỏ

Ví dụ đường kính nhỏ



Đồ thị thể hiện số hop (cạnh) và số cặp đỉnh có thể tới được trong dữ liệu mạng Epinions

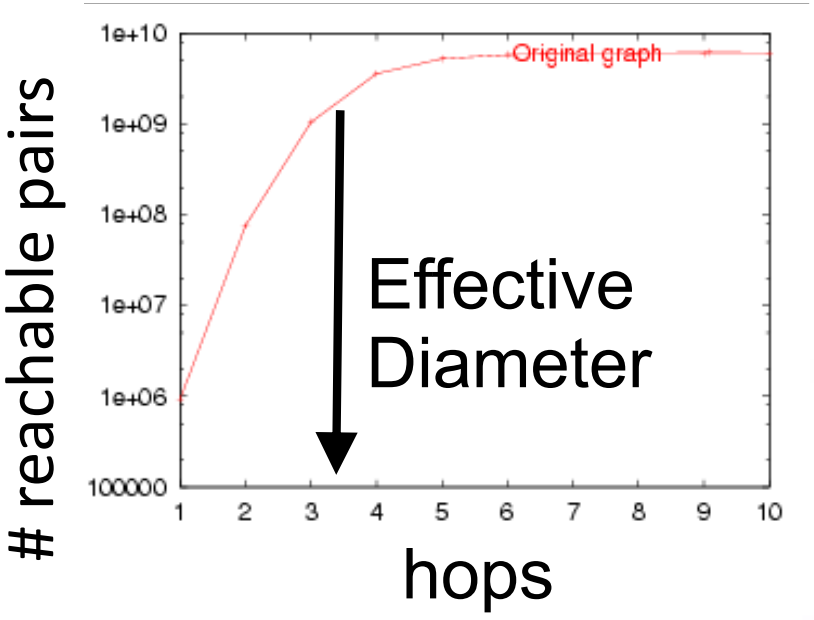
=> Chỉ cần 6 cạnh thì mạng hội tụ (ko còn thay đổi nữa)

Vấn đề trong đường kính đồ thị

- Đường kính của đồ thị có thể bị ảnh hưởng bởi một số ít cặp xảy ra nhưng lại có chuỗi đường đi dài.
- Thay vì đó, sử dụng độ đo đường kính hiệu dụng (effective diameter)
  - Số lượng tối thiểu cạnh mà một tỉ lệ (ví dụ 90%) tất cả các cặp đỉnh có thể đạt đến.

=> Cạnh nhiễu là cạnh đáng lẽ ko thuộc mạng đó nên sự kết nối giữa cạnh đó vs các cạnh rất ít

=> ra đời đg kính hiệu dụng=> vd: nếu 4 hops đã chiếm 90% dữ liệu thì lấy 4 hops là đường kính hiệu dụng



Vấn đề chi phí tính toán

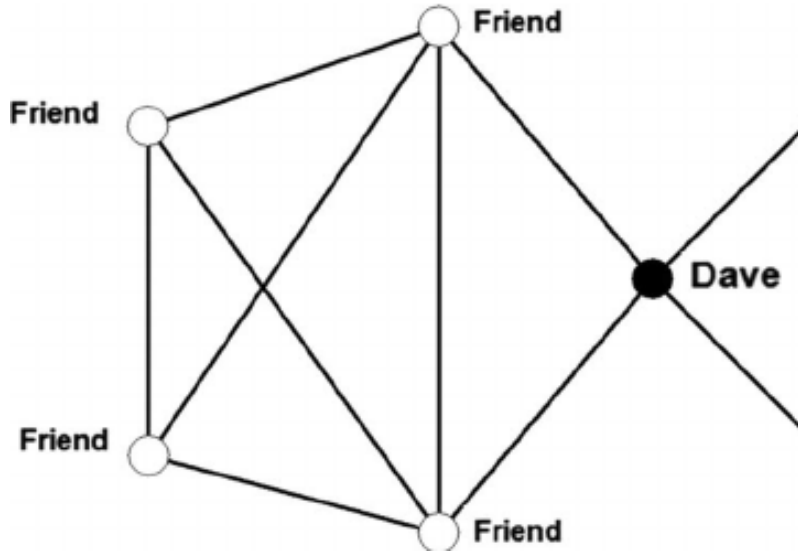
- Chi phí để tính toán chiều dài đường đi ngắn nhất giữa tất cả các cặp đỉnh trong đồ thị lớn thường tốn chi phí cao (ít nhất $O(N^2)$)
  - Giải pháp: áp dụng hàm láng giềng xấp xĩ (approximate neighborhood function)

Read more
https://www.researchgate.net/publication/2565684_ANF_A_Fast_and_Scalable_Tool_for_Data_Mining_in_Massive_Graphs

Luật mũ tam giác

- Đồ thị thế giới thật thường có nhiều kết nối dạng tam giác.
  - Ví dụ: bạn của bạn thường cũng là bạn của mình.

=> Nếu kết nối mở tam giác mà đóng tam giác thì sẽ có 1 liên kết bị triệt tiêu



=> Muốn đóng tam giác: vì kết nối ko mờ và thắt chặt mạng để tránh thay thế hay lk bị đứt
=> Có 2 kiểu đóng tam giác:
+ Cố tình: đc giới thiệu bởi trung gian
+ Vô tình
=> Người nào đang ở trong đóng tam giác thì có xu thế muốn đóng tam giác (Chọn C,D,E (vì có nhiều liên kết sau B) để đóng tg với B thì chọn C hoặc D vì C,D đag ở trong tg mà E thì đag có xu hướng mở tam giác (E ko nằm trong tam giác)

Luật mũ tam giác

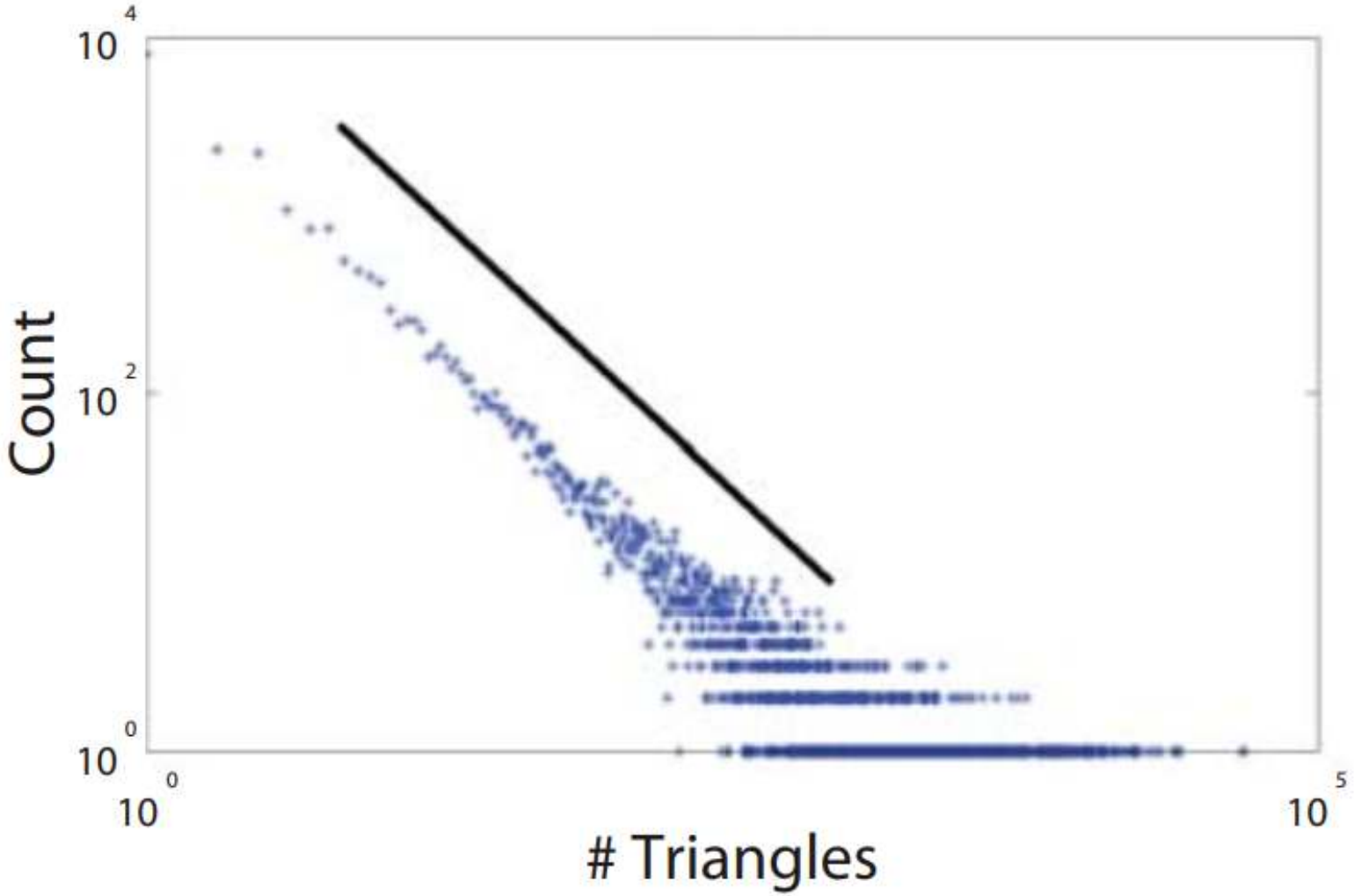
- Đặt Δ là số kết nối tam giác trong đồ thị

Δ_i là số kết nối tam giác có đỉnh i tham gia

Luật tham gia tam giác (Triangle participation law – TPL) thể hiện phân phối của Δ_i cũng tuân theo luật mũ với lũy thừa σ.

- Rất nhiều đỉnh có ít kết nối tam giác
- Tồn tại đỉnh tham gia vào số lượng lớn các kết nối tam giác

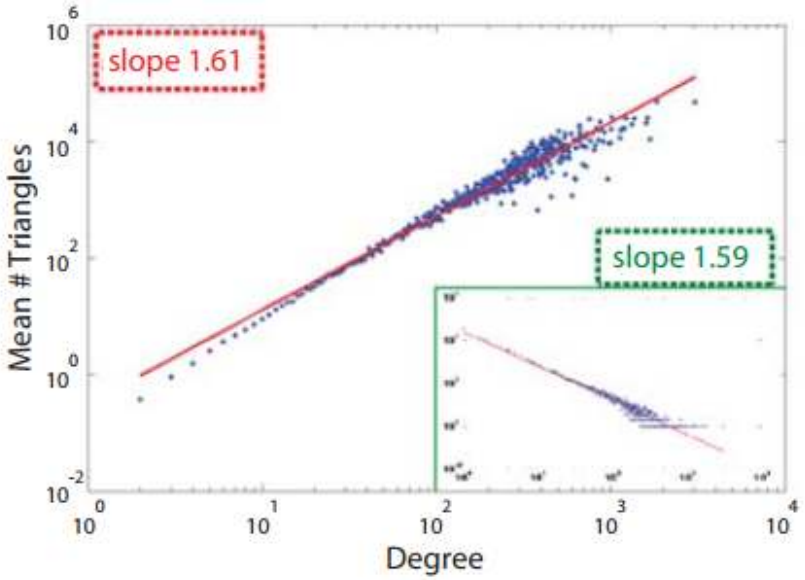
Luật mũ tam giác



Thể hiện phân phối cho TPL trong tập dữ liệu Epinion gồm 70 ngàn đỉnh, 500 ngàn cạnh

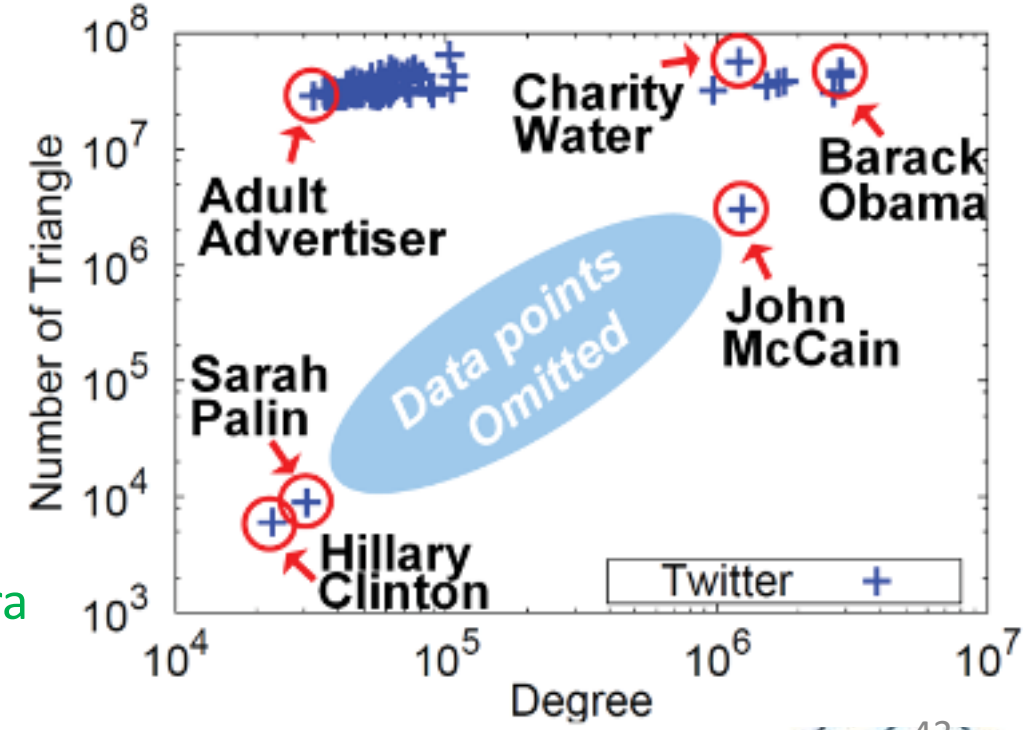
Luật mũ tam giác

- Khi xem xét mối tương quan giữa số đỉnh d_i của đồ thị và số kết nối Δ_i:
  - Tuân theo luật mũ với hệ số góc dương: $\Delta_i \propto d_i^s$ với $s \approx 1.5$
  - Ví dụ: một người càng có nhiều bạn, càng có nhiều kết nối tam giác được hình thành
  - Được gọi là luật tham gia "bậc-tam giác" (degree-triangle participation law – DTPL)



Luật mũ tam giác

- Tác giả Kang khám ra ra mạng Twitter cũng có tính chất DTPL.
  - Những người nổi tiếng có bậc cao và có các kết nối của những follower họ tuân theo luật bậc-tam giác.
  - Tuy nhiên có một số dữ liệu bất thường:
    - Một số tài khoản có bậc tương đối nhỏ nhưng lại tham gia vào số lượng lớn kết nối tam giác
    - Có thể một số người có nhiều tài khoản, mỗi tài khoản follow cái khác để đẩy bậc lên cao
    - Đây thường là những nhà quảng cáo phim người lớn, hoặc spammer
    - Sử dụng DTPL ta có thể khám phá ra các dữ liệu bất thường này.

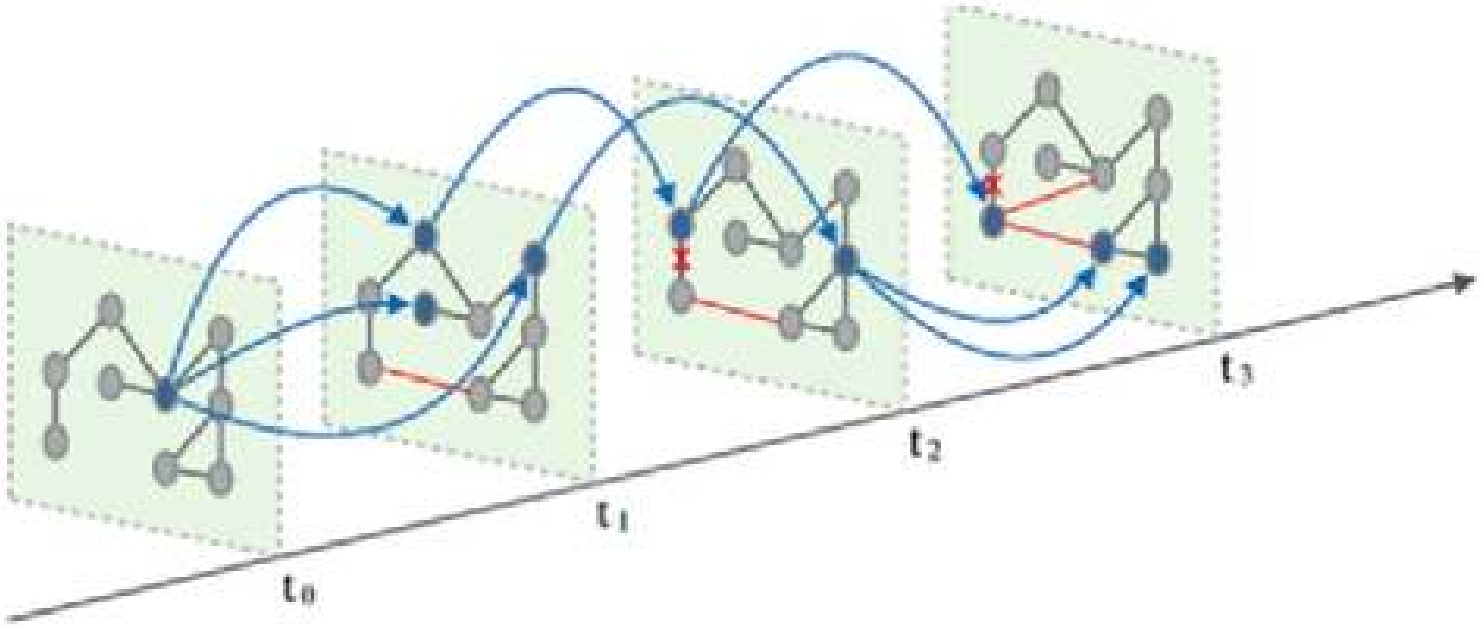


Nội dung

- Mẫu đồ thị
- Mẫu trong đồ thị tĩnh
- **Mẫu trong đồ thị động**
  - Đường kính thay đổi
  - Luật mũ tăng trưởng
  - Điểm kết dính
  - Trị riêng chính qua thời gian
- Mẫu trong đồ thị có trọng số
- Chi phí tính toán

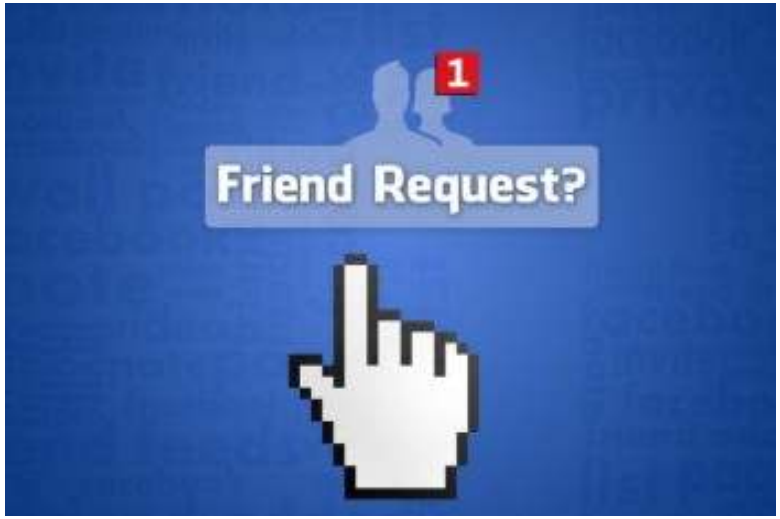
Đồ thị động

- Đồ thị với các kết nối được thêm hay bỏ đi theo thời gian được gọi là đồ thị động (dynamic, time-varying, evolving graph).



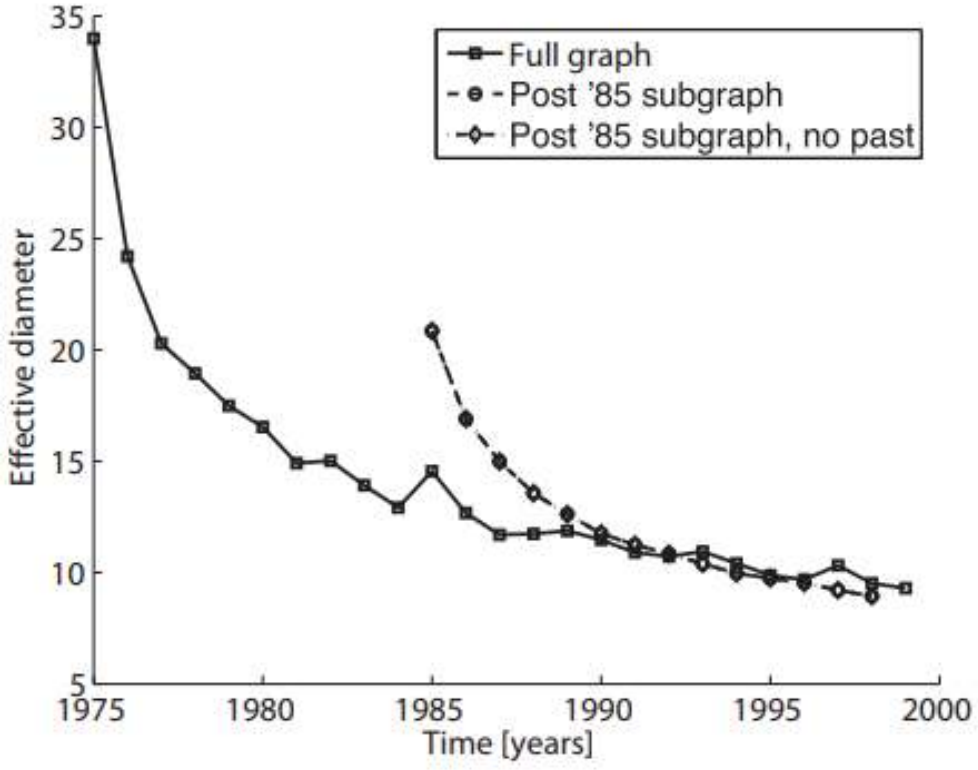
Hiện tượng đường kính thay đổi

- Khi đồ thị phát triển qua thời gian, đường kính của đồ thị sẽ thay đổi như thế nào?
  - Tăng hay giảm?
  - Ví dụ: thêm bạn bè trong mạng xã hội



Hiện tượng đường kính co lại

- Các nghiên cứu chỉ ra rằng khi đồ thị phát triển, đường kính của đồ thị có xu hướng co lại (shrink) thậm chí khi thêm đỉnh mới.



Đường kính hiệu dụng của đồ thị trích dẫn bằng sáng chế qua thời gian

Luật mũ tăng trưởng

- Giả định tại thời điểm t, ta có số đỉnh của đồ thị là $N(t)$ và số cạnh là $E(t)$.
- Tại thời điểm tiếp theo $t+1$, số đỉnh của đồ thị tăng gấp đôi $N(t+1) = 2 * N(t)$
  - Vậy số cạnh của đồ thị tại thời điểm này là bao nhiêu $E(t+1)$?
    - Gấp đôi đỉnh, gấp đôi cạnh?
      - Sai

Luật mũ tăng trưởng

- Trong đồ thị thế giới thật, người ta nhận thấy số lượng đỉnh và số lượng cạnh cũng tuân theo luật mũ với lũy thừa dương.

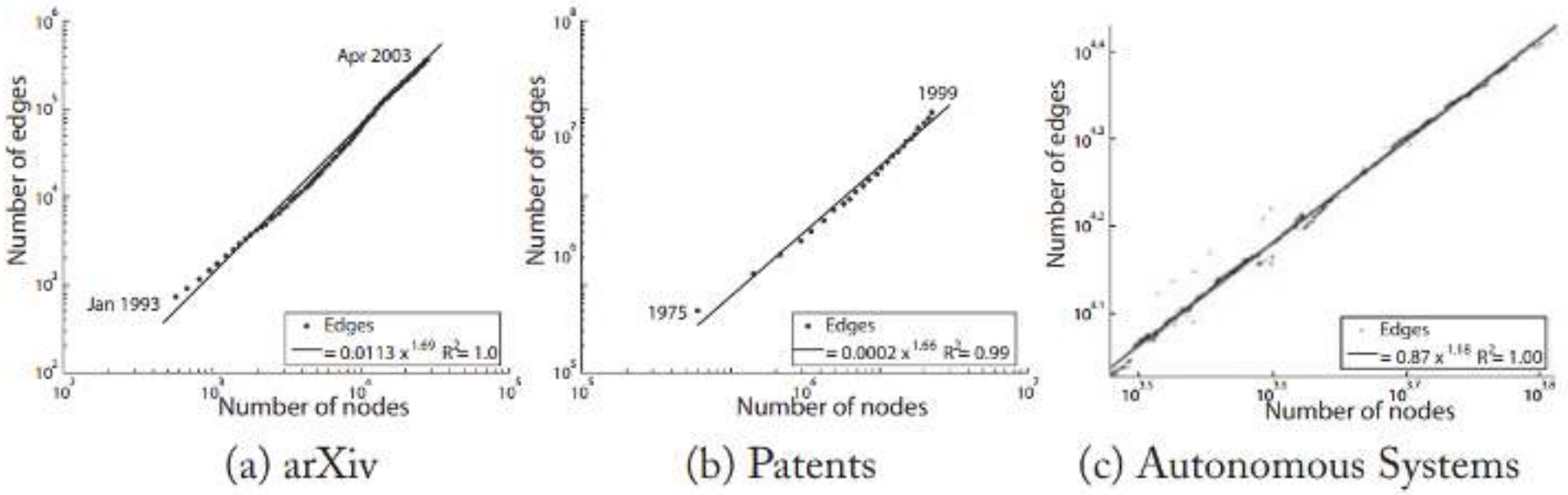
$$E(t) \propto N(t)^{\beta}$$

với β là mũ tăng trưởng

- Đây được gọi là luật mũ tăng trưởng (desification/growth power law)

=> đánh giá tăng trưởng để quyết định có đầu tư ko dựa vào beta

Luật mũ tăng trưởng



(a) arXiv (b) Patents (c) Autonomous Systems

Luật mũ tăng trưởng trong 3 tập dữ liệu đồ thị thực tế
Lũy thừa nằm trong khoảng 1.03 đến 1.7

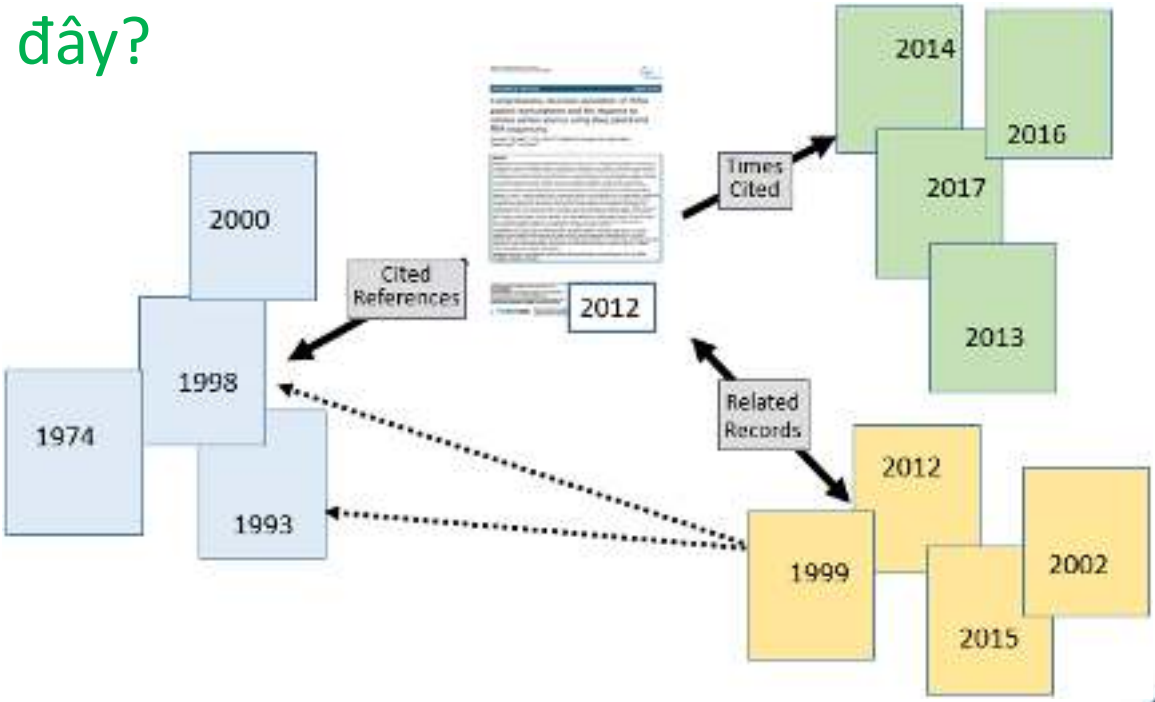
Luật mũ tăng trưởng

- Nếu lũy thừa $\beta > 1$ (thông thường):
  - Khi số đỉnh tăng gấp đôi, số cạnh của đồ thị sẽ tăng nhiều hơn gấp đôi.

    → Qua thời gian, bậc trung bình của các đỉnh tăng lên do có càng nhiều cạnh mới hơn đỉnh mới.
  - Đây là nguyên nhân làm cho đường kính co lại.

Luật mũ tăng trưởng

- Xét ví dụ mạng trích dẫn (citation network)
  - Theo luật mũ tăng trưởng, qua thời gian, bậc trung bình sẽ tăng lên hay số lượng trích dẫn trung bình cho mỗi bài báo tăng lên.
  - Liệu có phải mọi người viết bài báo ngày nay trích dẫn nhiều hơn trước đây?



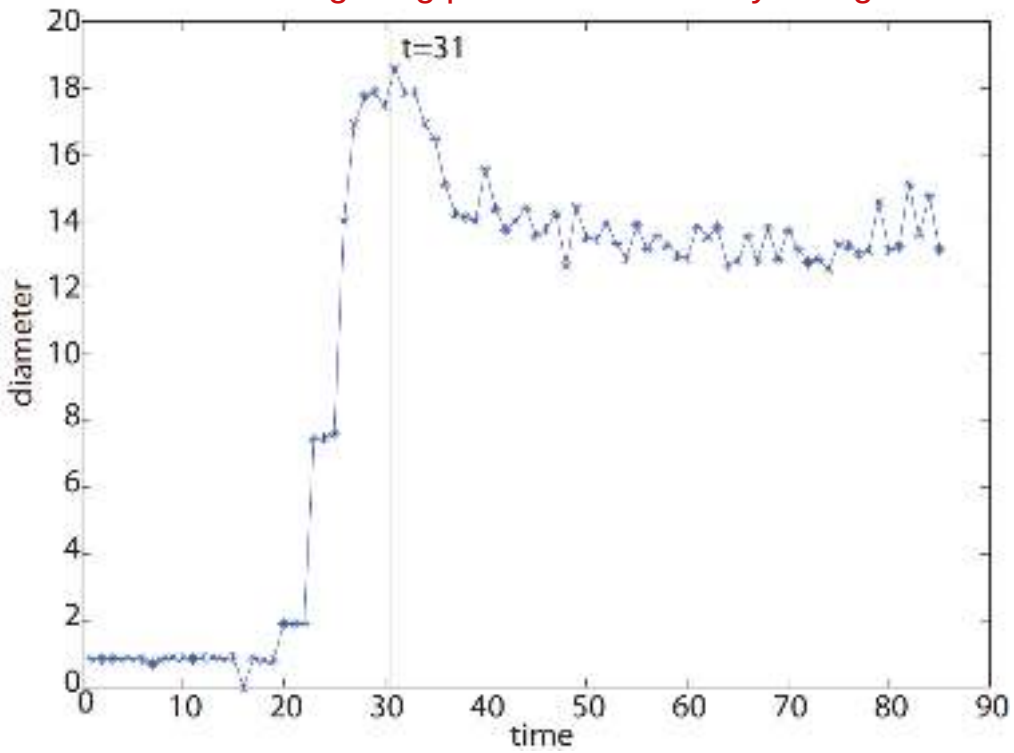
Luật mũ tăng trưởng

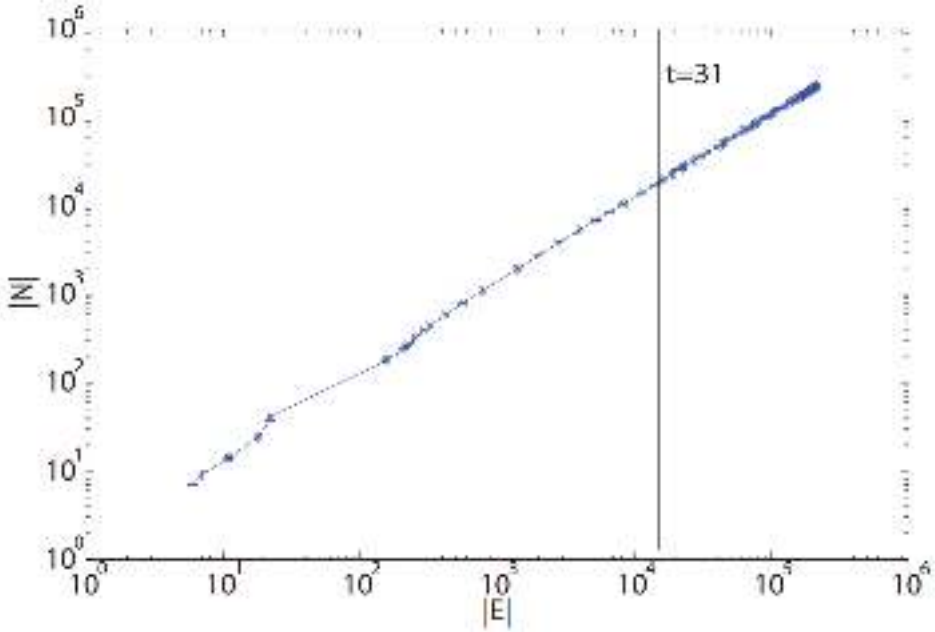
- Xét ví dụ mạng trích dẫn (citation network)
  - Theo luật mũ tăng trưởng, qua thời gian, bậc trung bình sẽ tăng lên hay số lượng trích dẫn trung bình cho mỗi bài báo tăng lên.
  - Liệu có phải mọi người viết bài báo ngày nay trích dẫn nhiều hơn trước đây?
    - Không. Hầu hết chúng ta viết báo trích dẫn thông thường từ 10 đến 30 bài báo từ trước tới giờ.
    - Vậy điều gì đã khiến cho bậc trung bình tăng lên theo thời gian?
      - Super paper, textbook càng về sau tham chiếu hàng ngàn bài báo trước đây

Điểm kết dính

- McGlohon nhận thấy rằng có một thời điểm đường kính của đồ thị tăng vọt lên.

gelling point: sau điểm này thì đg kính có xu hướng co lại





Dữ liệu theo thời gian của mạng PostNet

Biểu đồ (a) thể hiện thay đổi đường kính theo thời gian

Biểu đồ (b) thể hiện tương quan giữa số cạnh và số đỉnh theo thời gian (vẫn tuân theo luật mũ tăng trưởng)

Điểm kết dính

+ SCC (khối liên thông mạnh): bất kì đỉnh nào trog khối đều có đg đi đến nhau, xét cả hướng (1 đỉnh cx là SCC)
+ GCC: khối chứa nhiều đỉnh nhất. Những khối còn lại là NLCC
=>Để kiếm đg đi giữa 2 điểm thì xem các khối SCC là 1 đỉnh và tiến hành vét

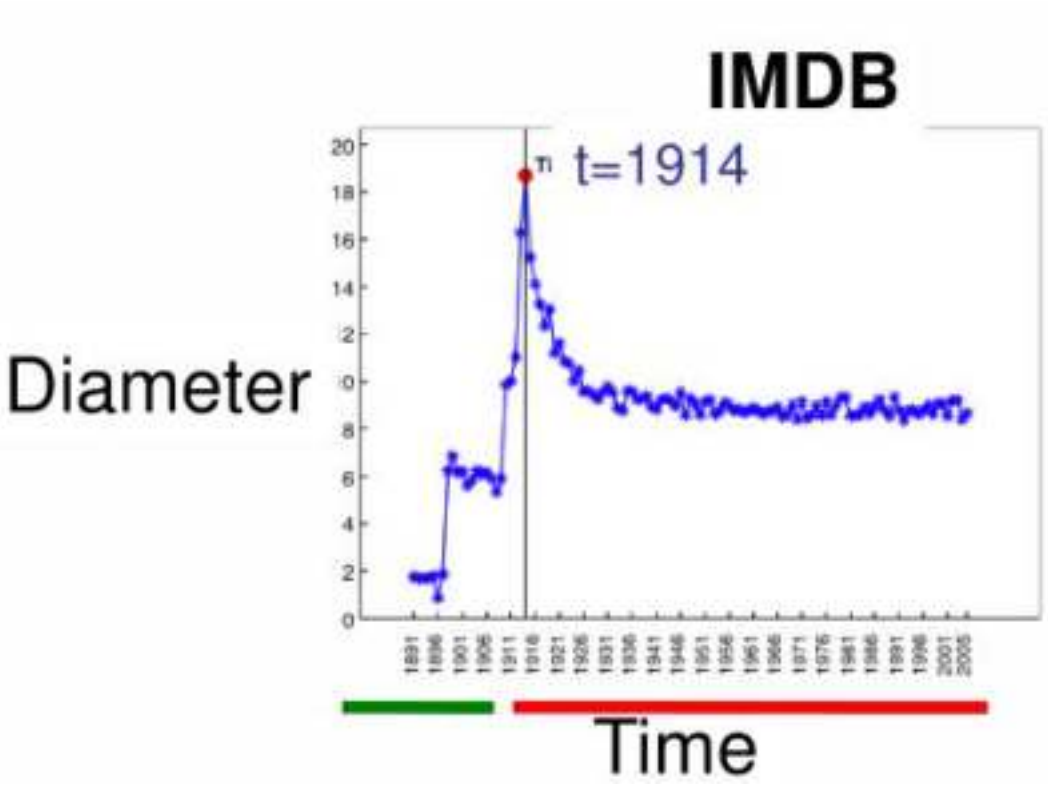
- McGlohon nhận thấy rằng có một thời điểm đường kính của đồ thị tăng vọt lên.

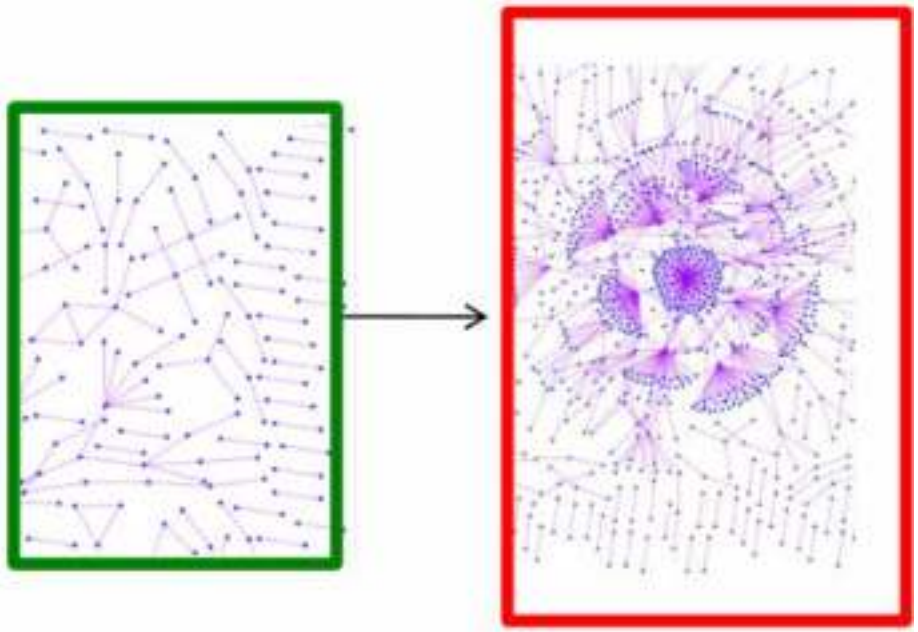
+ reachability: kiểm tra sự đạt đc của các khối

  - Thời điểm này được gọi là điểm kết dính (gelling point)
    - Trước thời điểm đó, đồ thị bao gồm một tập các thành phần không liên thông nhỏ.
    - Tại điểm kết dính, một thành phần liên thông khổng lồ (Giant Connected Component – GCC) bắt đầu hình thành và chiếm phần lớn các đỉnh. Khi đỉnh mới hình thành, nó cũng có xu hướng gia nhập khối GCC này.
    - Các thành phần liên thông còn lại được gọi là NLCC (non-largest connected component) => những nhóm nhỏ

Điểm kết dính

- Ví dụ điểm kết dính trong mạng IMDB





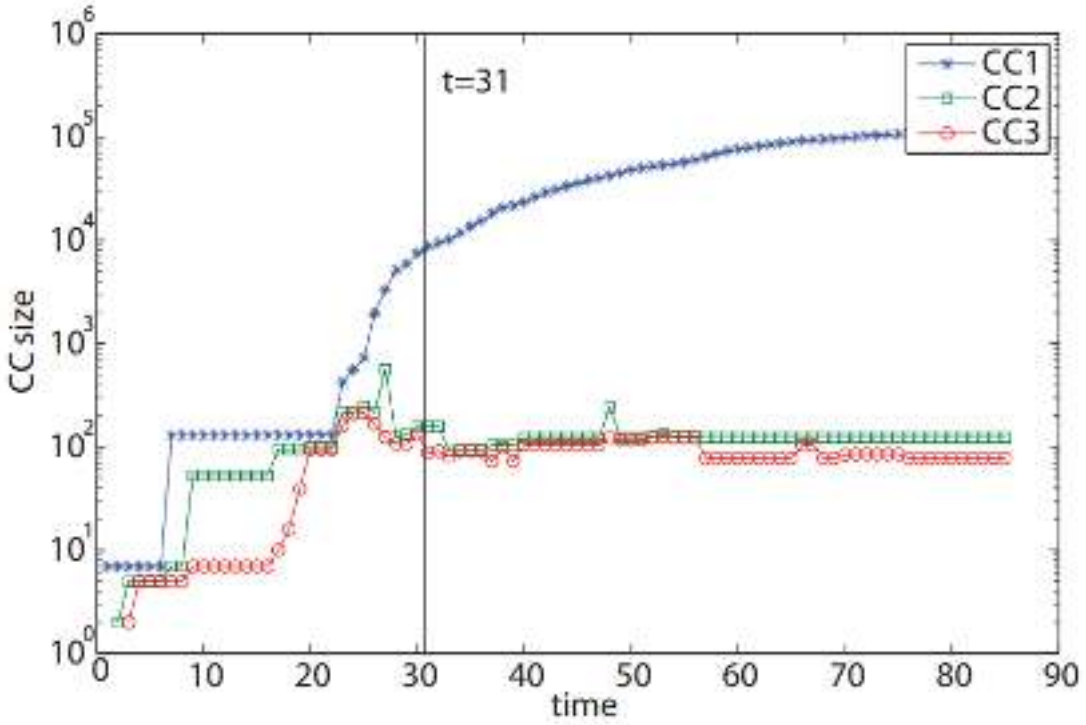
Điểm kết dính

- Chuyện gì sẽ xảy ra sau điểm kết dính?
  - Có phải khối GCC sẽ tiếp tục tăng?
  - Khối NLCC sẽ tăng nhẹ hay không thay đổi?

Hiện tượng xã hội qua điểm kết dính
=> Bài toán chọn nơi mở cửa hàng: 2 công ty nên mở ở vị trí nào trong 6 điểm
+ An toàn thì chọn C hoặc D

Điểm kết dính

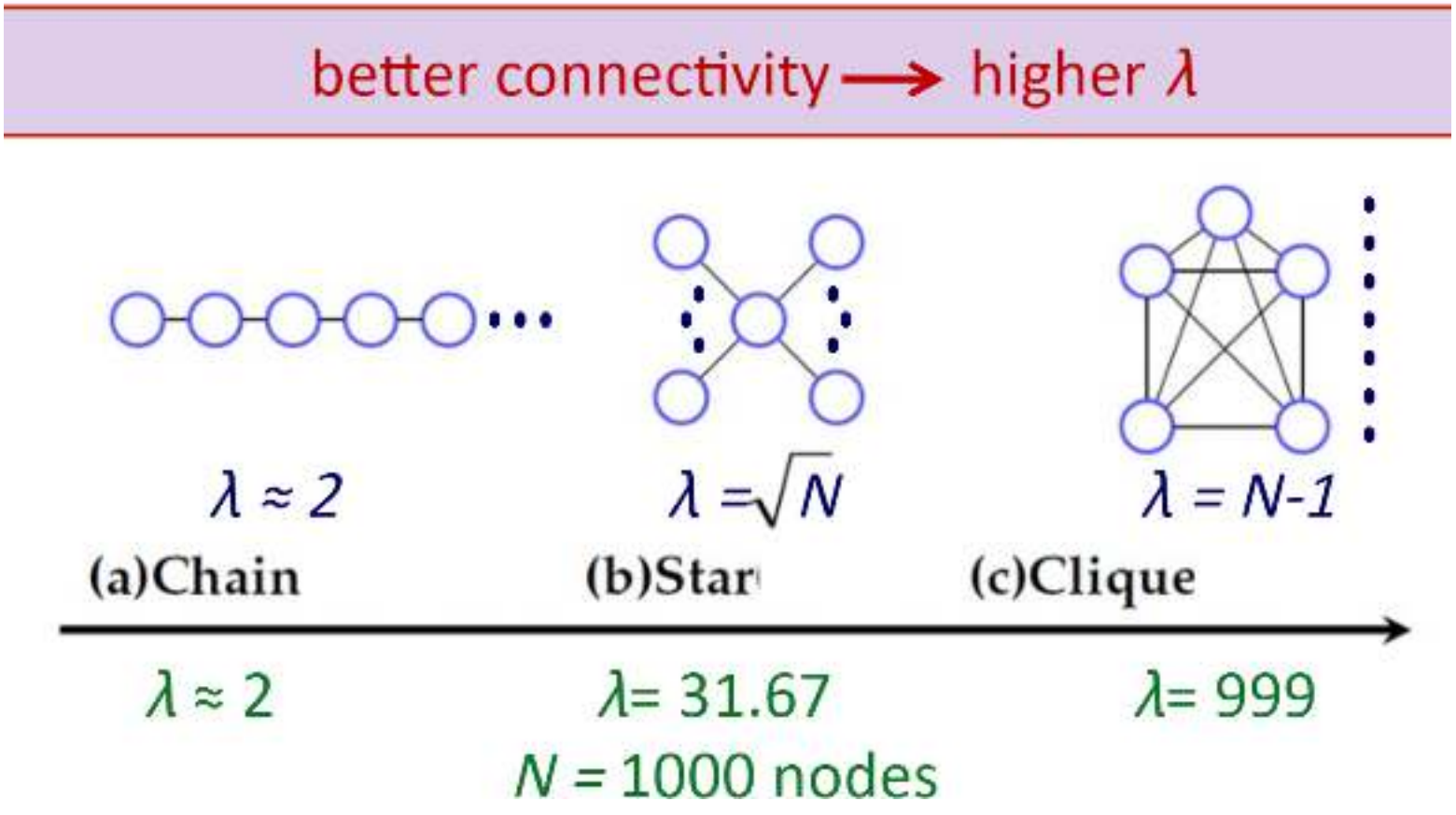
- Sau điểm kết dính:
  - Khối GCC vẫn tiếp tục tăng trưởng về kích thước
  - Khối NLCC hầu như giữ nguyên hoặc dao động quanh một khoảng.



GCC, CC2, and CC3 (log-lin)

Trị riêng chính qua thời gian

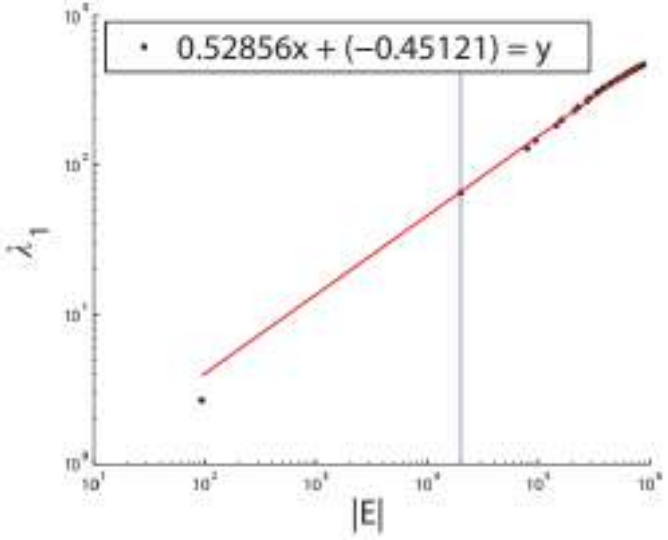
- Trị riêng chính (cực đại) đầu tiên λ_1 là một trong những độ đo quan trọng về tính liên thông của đồ thị.



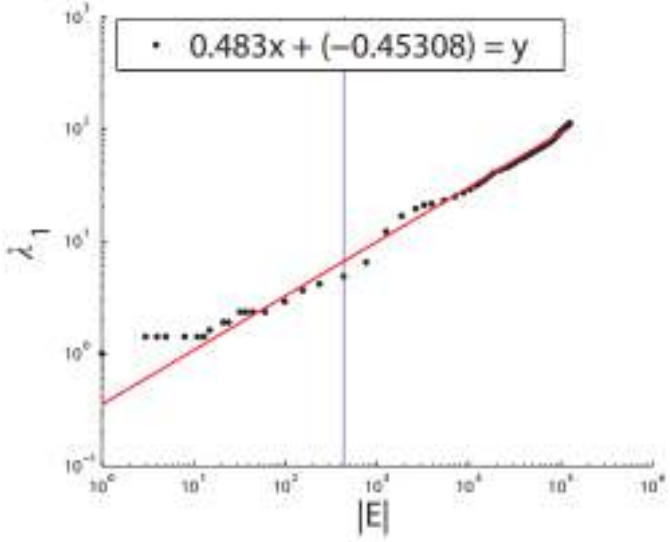
Trị riêng chính qua thời gian

- Trong đồ thị thế giới thật, trị riêng cực đại $\lambda_1(\mathrm{t})$ và số cạnh $E(t)$ thay đổi theo luật mũ qua thời gian với lũy thừa nhỏ hơn 0.5

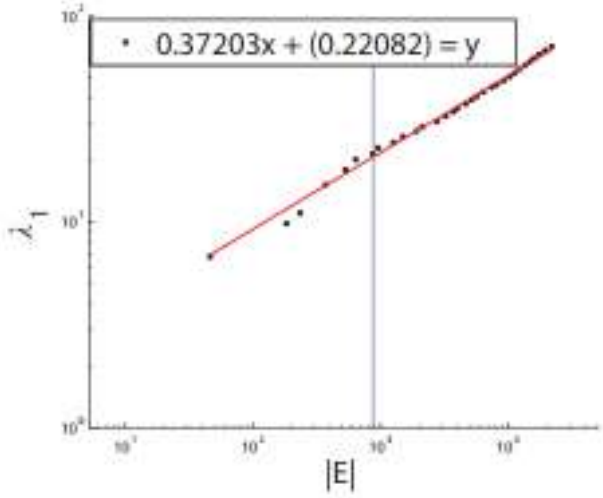
$$\lambda_1(t) \propto E(t)^{\alpha}, \alpha \leq 0.5$$



(a) *CampOrg*



(b) *BlogNet*



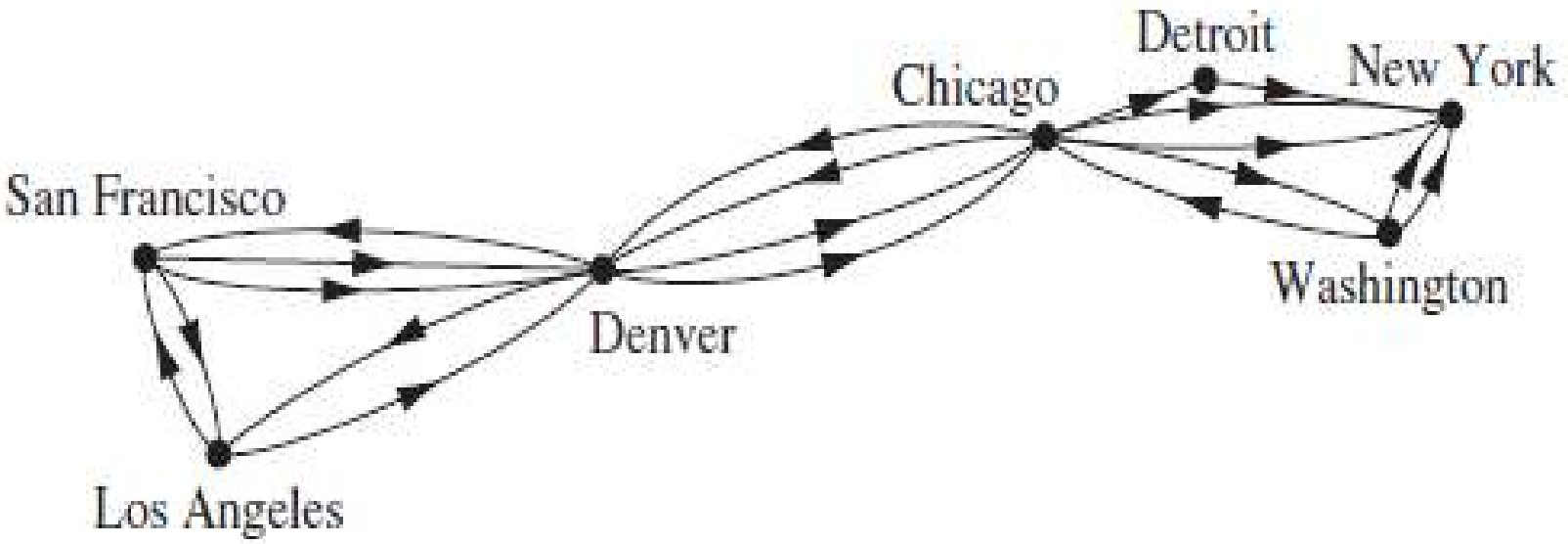
(c) *Auth-Conf*

Nội dung

- Mẫu đồ thị
- Mẫu trong đồ thị tĩnh
- Mẫu trong đồ thị động
- **Mẫu trong đồ thị có trọng số**
  - Luật mũ snapshot
  - Luật mũ trọng số
  - Trị riêng chính có trọng số
- Chi phí tính toán

Đồ thị trọng số thay đổi

- Xét đồ thị có trọng số thay đổi theo thời gian
  - Ví dụ: đồ thị mô tả lượng gói tin truyền tải trong mạng với khoảng thời gian xem xét là mỗi 30 phút
  - Gọi $W(t)$ là tổng trọng số đến thời điểm t (tổng gói tin được trao đổi trong mạng)
  - $E(t)$ là số cạnh phân biệt đến thời điểm t
  - n(t) là số đa cạnh (multi-edge) đến thời điểm t



Luật mũ snapshot

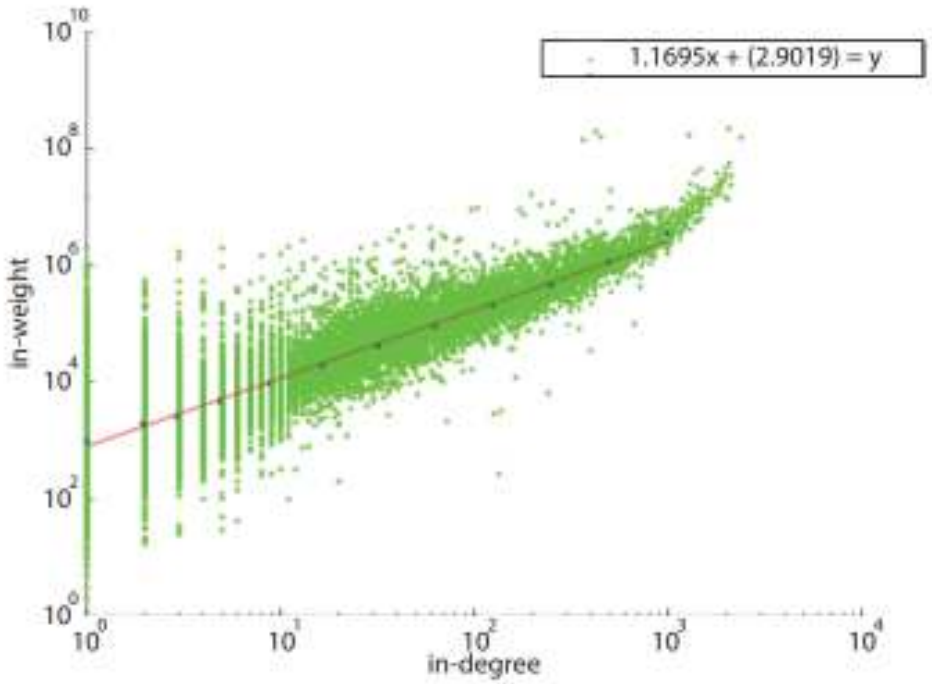
- Người ta nhận thấy tại một thời điểm xác định, bậc ngoài (trong) của một đỉnh i và trọng số ngoài (trong) tại đỉnh đó tuân theo luật mũ.

$$outw_i = out_i^{ow}$$

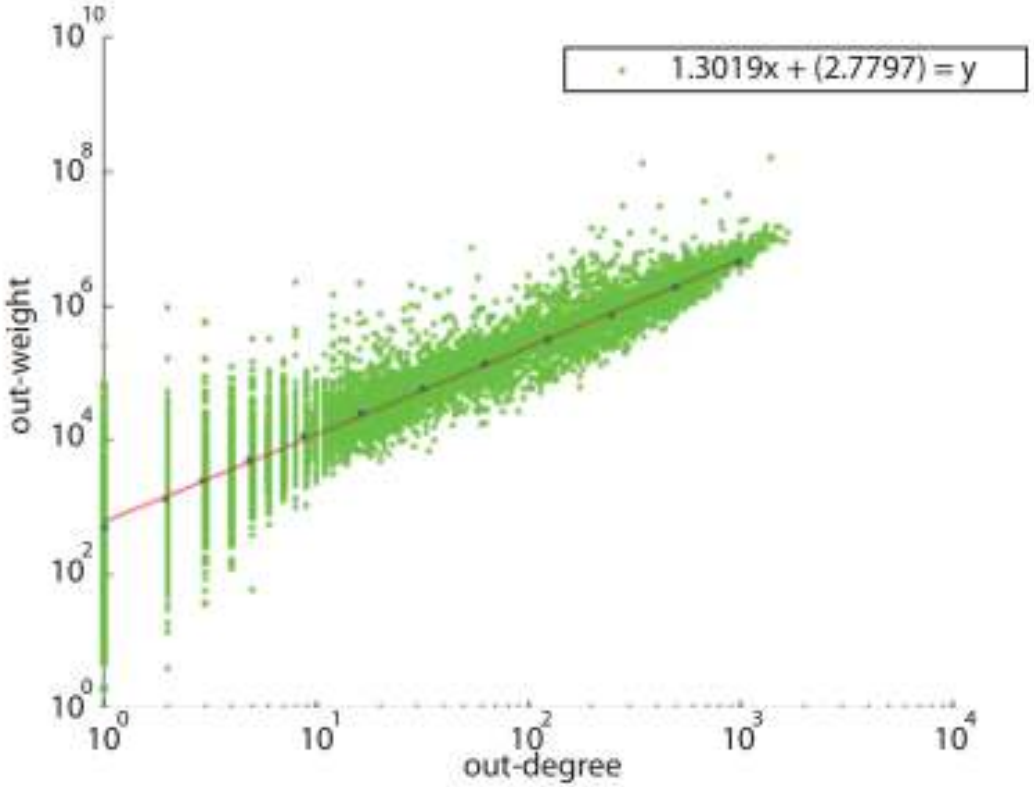
với ow là lũy thừa trọng số ngoài (out-weight-exponent) và thường không đổi theo thời gian.

- Tương tự cho bậc trong và trọng số trong.
- Đây được gọi là luật mũ snapshot (snapshot power law – SPL)

Luật mũ snapshot



(a) inD-inW snapshot



(b) outD-outW snapshot

Mối tương quan giữa trọng số và bậc trong dữ liệu của CampOrg

- Tổ chức hỗ trợ càng nhiều chiến dịch thì số tiền bỏ ra càng nhiều
- Một ứng viên càng nhận nhiều hỗ trợ thì số tiền nhận được càng nhiều

Luật mũ trọng số

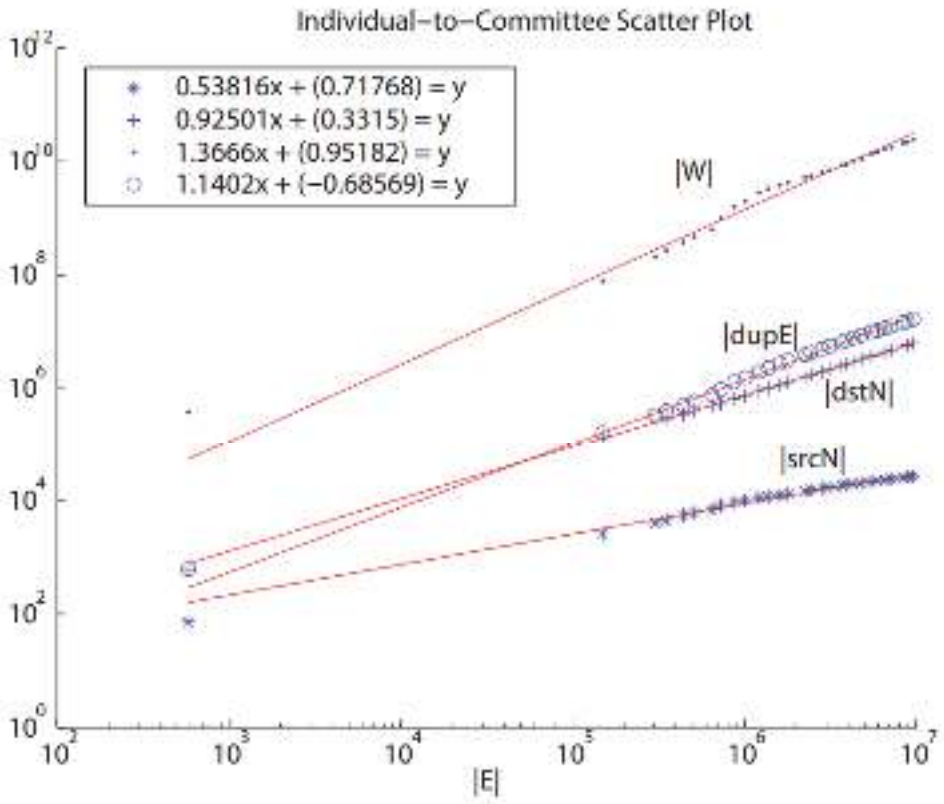
- Với $E(t)$ là tổng số cạnh phân biệt, $W(t)$ là tổng trọng số, $N(t)$ là tổng số đỉnh, $n(t)$ là tổng số đa cạnh:
- Người ta nhận thấy giữa tổng số cạnh và tổng trọng số của đồ thị tại thời điểm t tuân theo luật mũ:

$$W(t) = E(t)^w$$

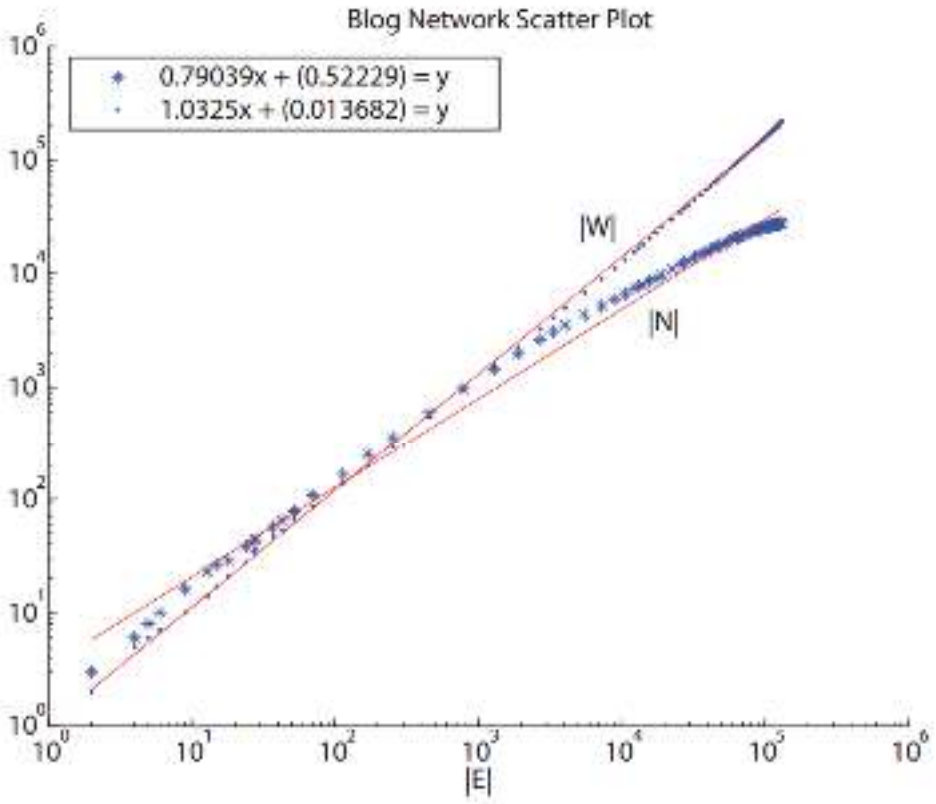
với w là lũy thừa trọng số, thường từ 1.01 đến 1.5.

- Tương tự cho $N(t) - E(t), n(t) - E(t)$.

Luật mũ trọng số



(a) *CampIndiv* WPLs



(b) *BlogNet* WPLs

Mối tương quan giữa tổng trọng số (tổng đa cạnh, tổng đỉnh) và tổng số cạnh trong các dữ liệu. Mỗi điểm dữ liệu tương ứng với 1 thời điểm t.

- Tổ chức hỗ trợ cho nhiều chiến dịch có xu hướng trả nhiều tiền hơn cho mỗi chiến dịch

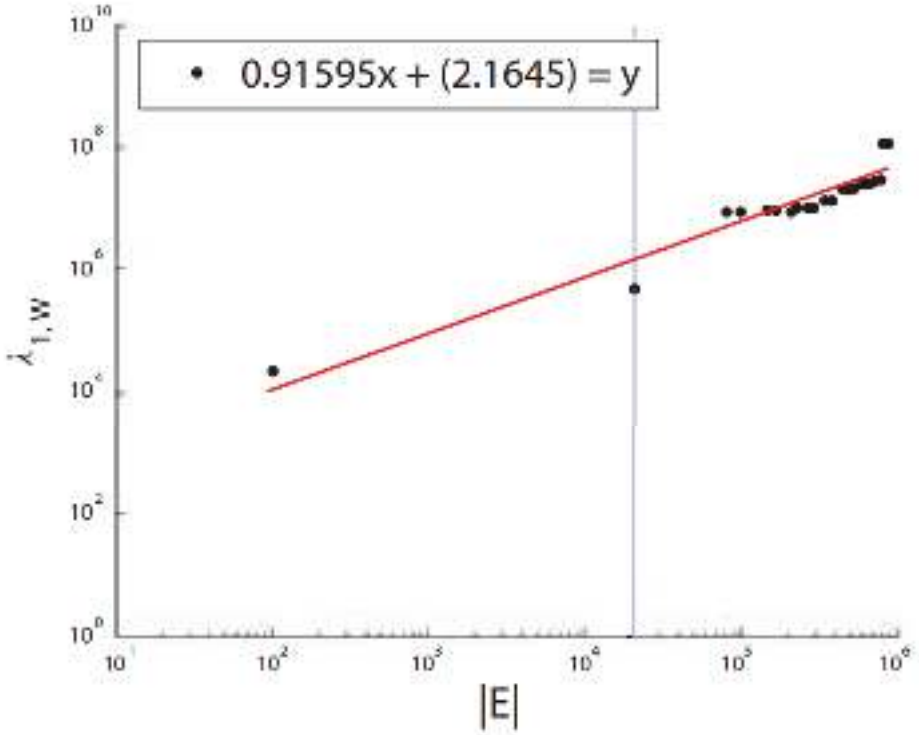
Trị riêng chính có trọng số qua thời gian

- Đặt $\lambda_{1,w}$ là trị riêng lớn nhất (chính) của ma trận trọng số $\boldsymbol{A}_w$
- Trị riêng chính của ma trận trọng số và số cạnh của đồ thị tuân theo luật mũ:

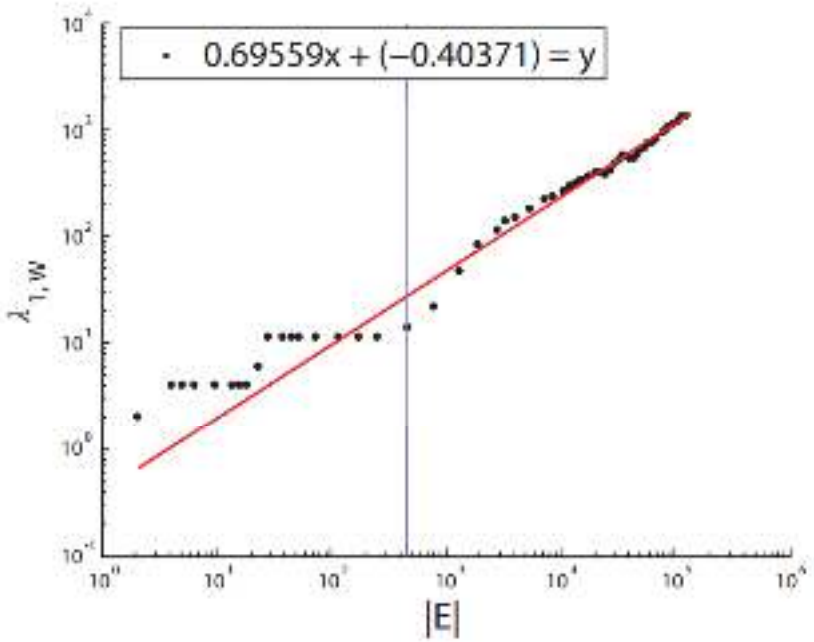
$$\lambda_{1,w}(t) \propto E(t)^{\beta}$$

với lũy thừa β thường trong khoảng 0.5 đến 1.6 (cao hơn so với luật mũ trị riêng trong đồ thị không trọng số).

Trị riêng chính có trọng số qua thời gian



(a) *CampIndiv*



(b) *BlogNet*

Mối tương quan giữa trị riêng chính của ma trận trọng số và tổng số cạnh trong các dữ liệu qua thời gian (đường thẳng đứng là điểm kết dính)

- Tổ chức hỗ trợ cho nhiều chiến dịch có xu hướng trả nhiều tiền hơn cho mỗi chiến dịch

Nội dung

- Mẫu đồ thị
- Mẫu trong đồ thị tĩnh
- Mẫu trong đồ thị động
- Mẫu trong đồ thị có trọng số
- **Chi phí tính toán**

Chi phí tính toán

- Tiến trình tìm mẫu có thể được chia làm 3 phần:
  - Tạo ra đồ thị phân bố
  - Xác định lũy thừa cho luật mũ
  - Kiểm tra mẫu có khớp luật mũ không

Tạo ra đồ thị phân bố

- Thường được đánh giá là thao tác đơn giản
- Giả sử đồ thị được thể hiện dưới dạng một bảng có lược đồ $Graph(fromnode, tonode)$, ta có thể sử dụng SQL để thực hiện:

```
SELECT outdegree, count(*)
FROM
   (SELECT count(*) AS outdegree
    FROM Graph
    GROUP BY fromnode)
GROUP BY outdegree
```

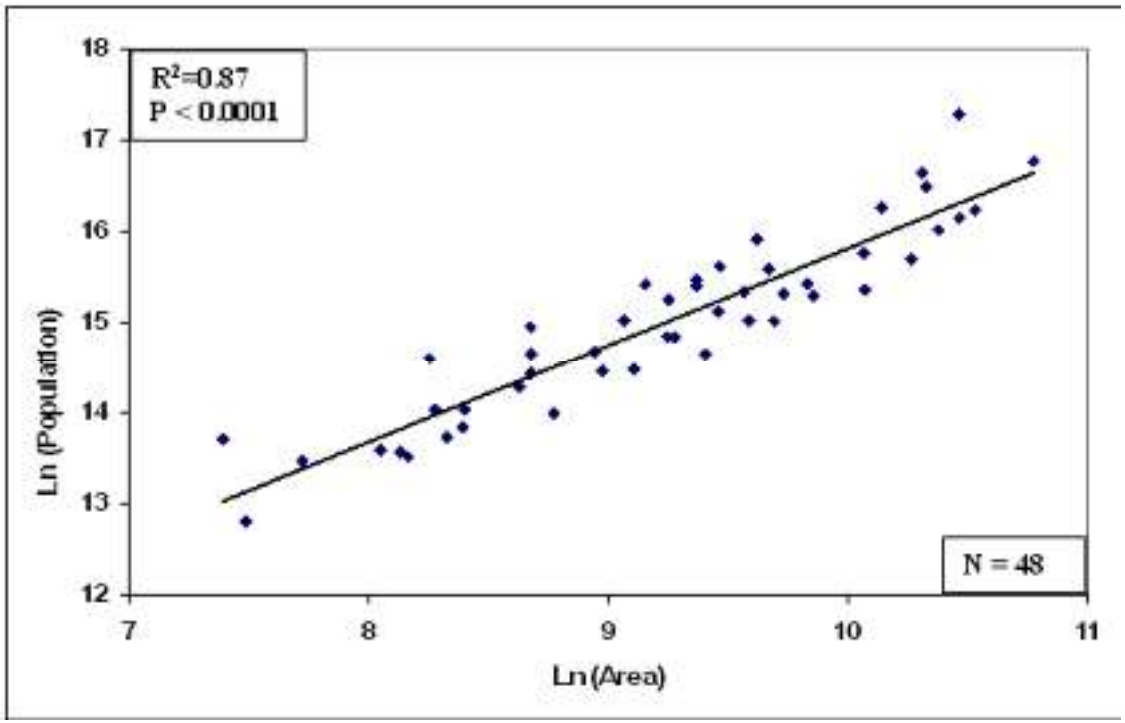
```
SELECT indegree, count(*)
FROM
   (SELECT count(*) AS indegree
    FROM Graph
    GROUP BY tonode)
GROUP BY indegree
```

Xác định lũy thừa cho luật mũ

- Bài toán khó do:
  - Luật mũ có thể chỉ xuất hiện ở phần đuôi ở phân bố, không phải trên toàn phân bố
  - Một số giả định tiên quyết có thể không thỏa
  - …
- Một số phương pháp được đề xuất nhưng chỉ mang tính tương đối (không rõ ai là “winner” hiện tại).

Xác định lũy thừa cho luật mũ

- Hồi quy tuyến tính trên biểu đồ log-log:
  - Vẽ dữ liệu ở thang log-log
  - Tối ưu từng khoảng vào những miền kích thước bằng nhau
  - Tìm hệ số góc để khớp



Xác định lũy thừa cho luật mũ

- Hồi quy tuyến tính trên biểu đồ log-log
- Một số vấn đề phát sinh:
  - Có thể dẫn đến ước lượng bị lệch
  - Luật mũ có thể chỉ xuất hiện ở phần đuôi và điểm nào bắt đầu để xem xét cần phải xác định thủ công
  - Đầu cuối bên phải của phân bố sẽ rất nhiễu
- Tuy nhiên, đây được xem là kỹ thuật đơn giản và được sử dụng phổ biến nhất.

Xác định lũy thừa cho luật mũ

- Một số phương pháp khác cũng được sử dụng:
  - Hồi quy tuyến tính sau chia khoảng theo logarit
  - Hồi quy trên phân phối tích lũy
  - Ước lượng cực đại khả năng (bởi Goldstein)
  - Thống kê Hill (bởi Hill)
  - Khớp chỉ dữ liệu giá trị cực (bởi Feuerverger và Hall)
  - Ước lượng không tham số (bởi Crovella và Taqqu)

Kiểm tra mẫu có khớp luật mũ

- Hệ số tương quan là một cách đo tương đối một phân phối bậc có khớp với luật mũ không.
- Ngoài ra, có một số phương pháp dựa trên thống kê của Beirland hay Goldstein được phát triển.

Tài liệu tham khảo

- Our Networked World. 2020. *Heavy-Tailed Degree Distributions*.
- Chakrabarti, D. and Faloutsos, C., 2012. Graph mining: laws, tools, and case studies. Synthesis Lectures on Data Mining and Knowledge Discovery, 7(1), pp.1-207.
- Sidney Redner. How popular is your paper? an empirical study of the citation distribution. The European Physics Journal B, 4:131–134, 1998.
- G.Siganos,M.Faloutsos,P.Faloutsos, and C.Faloutsos. Power laws and the AS-level internet topology, 2003.
- U. Kang, Brendan Meeder, and Christos Faloutsos. Spectral analysis for billion-scale graphs: Discoveries and implementation. In PAKDD (2), pages 13–25, 2011.
- Mary McGlohon. Structural analysis of networks: Observations and applications. Ph.D.thesis CMU-ML-10-111, Machine Learning Department, Carnegie Mellon University, December 2010.