Mining Graph Data

GRAPH PATTERNS

Lecturer: Le Ngoc Thanh

Email: lnthanh@fit.hcmus.edu.vn

TRUONG DAI HOC KHOA HOC TU NHIEN
TP. HO CHI MINH

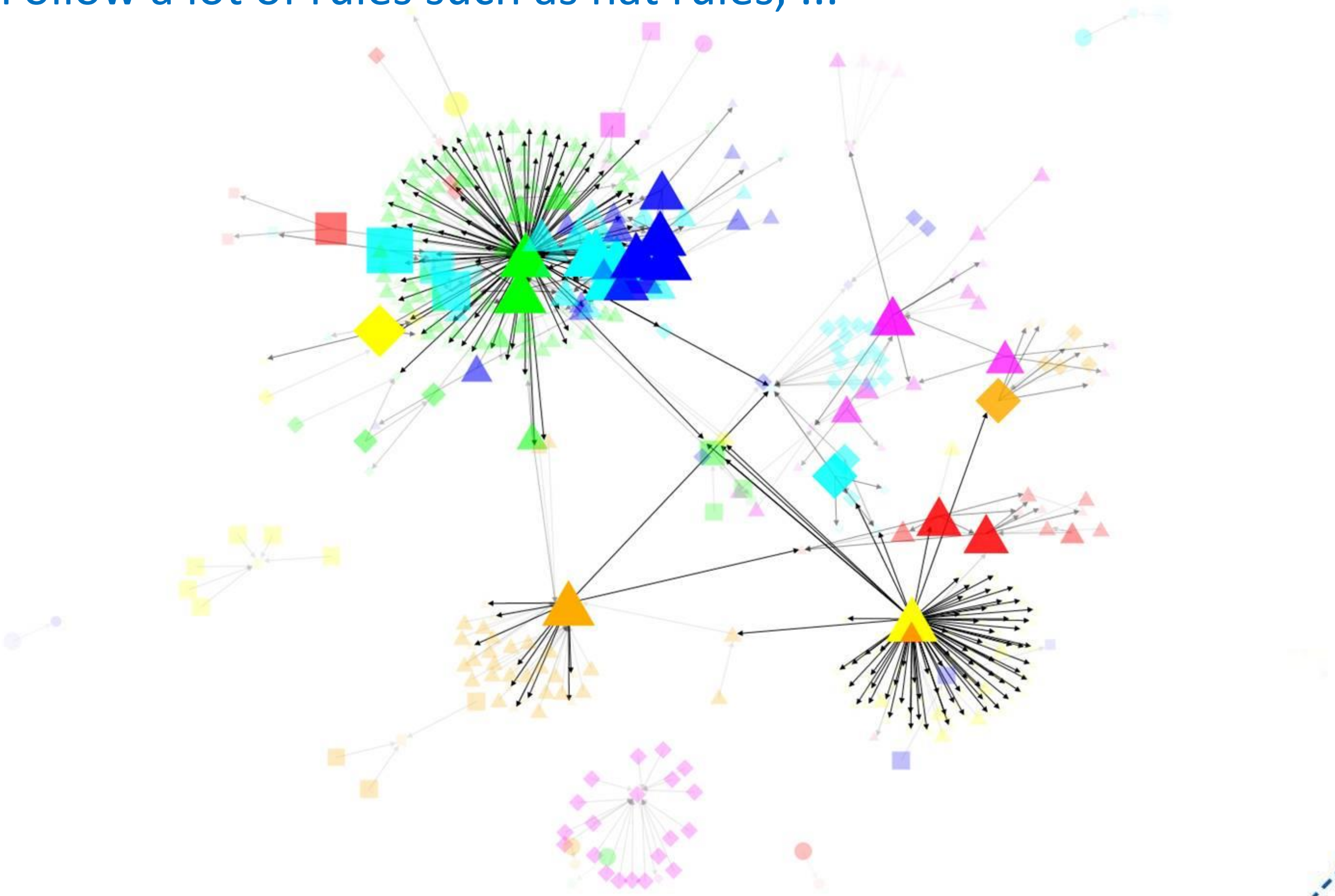
fit@hcmus

Contents

- **Graph patterns**
- Sample in static graphs
- Templates in dynamic graphs
- Sample in weighted graphs
- Cost calculation

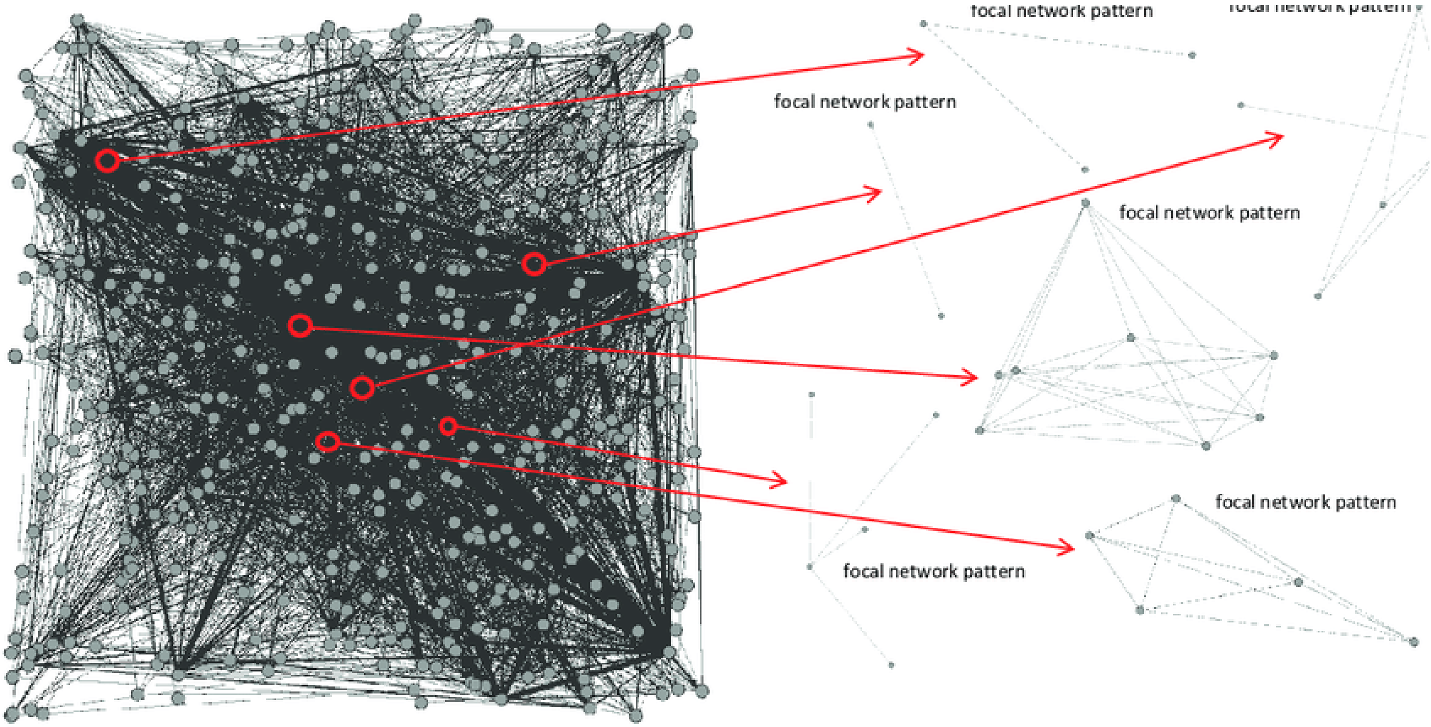
Real world graph

- Is the world graph truly random?
  - – No!!!
    - Follow a lot of rules such as hat rules, ...



Graph pattern

- A graph patterns is an attribute or subgraph that often appears in real-world graphs.
-



Samples in the social network of 442 vertices and 3171 edges were collected from http://www.livejournal.com.

Study graph patterns

- Why consider graph patterns?
- Understand the interesting properties of real world graphs
- Templates that show condensed information about graphs
- Use to generate graphs similar to real world graphs for research
  - Supports the detection of anomalies and outliers.

Abnormal detection

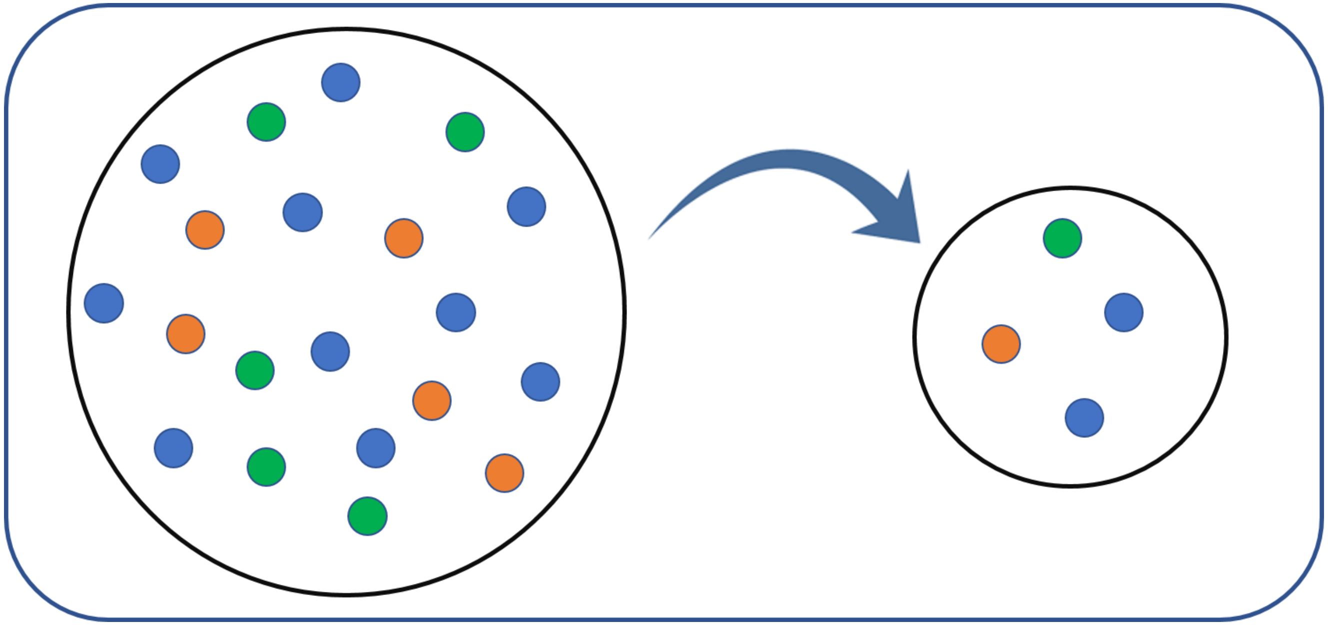
- In real world graphs can contain:
  - Abnormal edge
  - Abnormal peak
  - Abnormal subgraphs
- "Abnormal" often differs from "normal" patterns.
- Therefore, understanding patterns that occur naturally is a prerequisite for identifying abnormal patterns.

Simulation

- Testing algorithms that run on huge real world graphs needs graphs that are derived that resemble it:
  - Some organizations don't share data
  - Need to check before it happens
    - For example, testing the new generation Internet protocol, it is necessary to simulate a graph similar to the Internet world in the next few years.

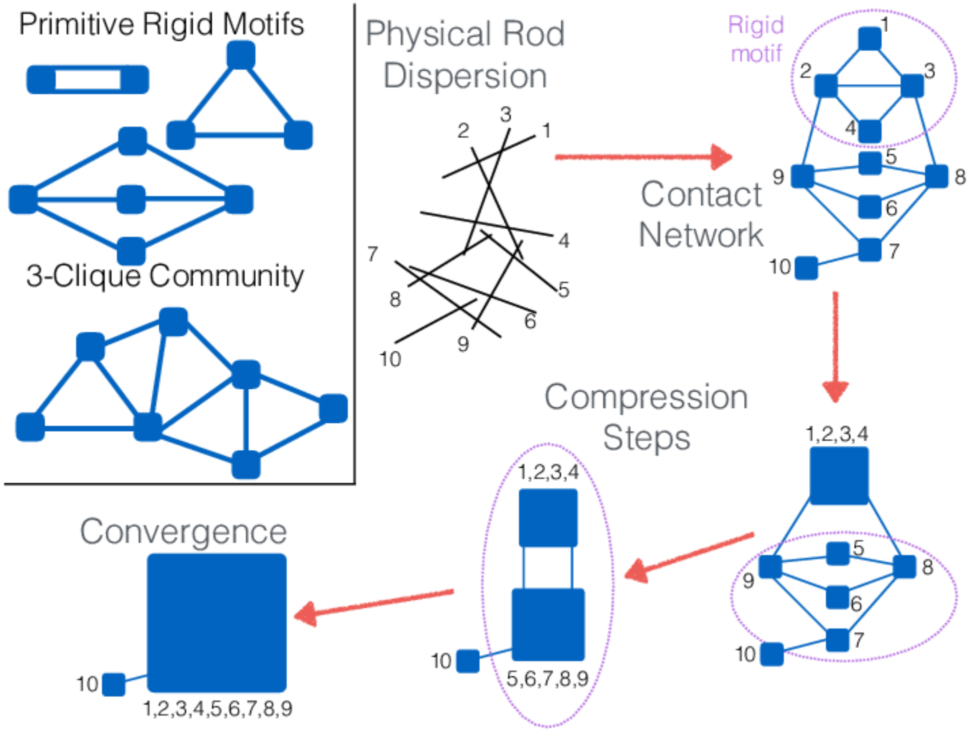
Sampling

- When constructing/extracting a small sample graph, we usually expect the same graph with a larger graph.



Graph compression

- Graph patterns represent rules in data. These rules can be used to compress graph data.



Graph pattern detection problem

- How easy or difficult is the math?
  - Hard!!!
  - Why?
    - Need to identify the "good" template?
      - A good template is one that distinguishes between a real-world graph and a dummy-generated template.
      - So what sample is it?
    - Is just 1 sample enough to be distinguishable?
    - Is there effective calculation on large graphs?
      - A sample that loses O(N^3) or O(N^2) where N is the vertex number of the graph becomes impractical.

Content

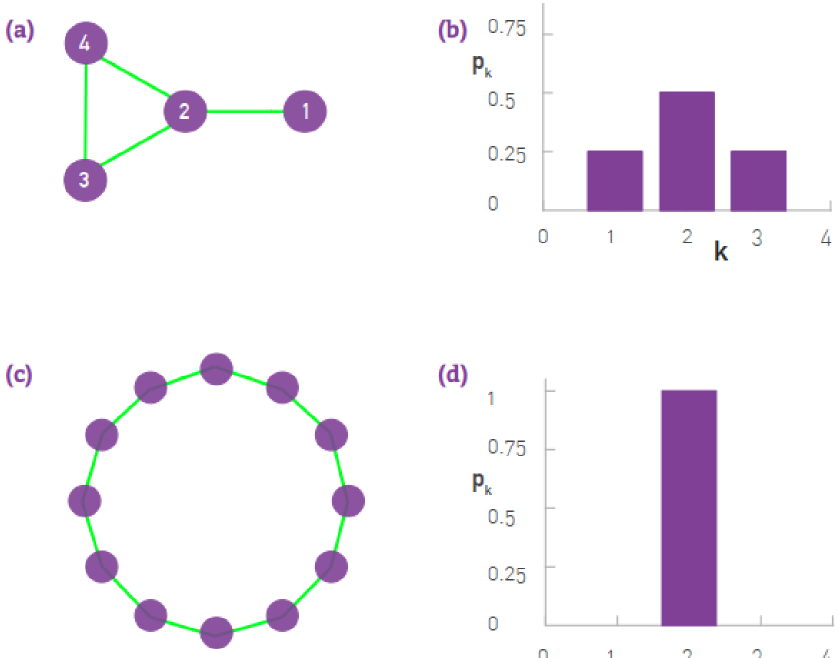
- Graph patterns
- Patterns in static graphs
  - Degree distribution
  - Law of hats
  - Law of exponential eigenvalues
  - Triangle Hat Law
- Patterns in dynamic graphs
- Patterns in weighted graphs
- Cost calculation

Degree distribution in the graph

- Given a graph with N vertices, the order distribution of the graph is a normalized histogram:

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

with N_k is the number of vertices of degree k

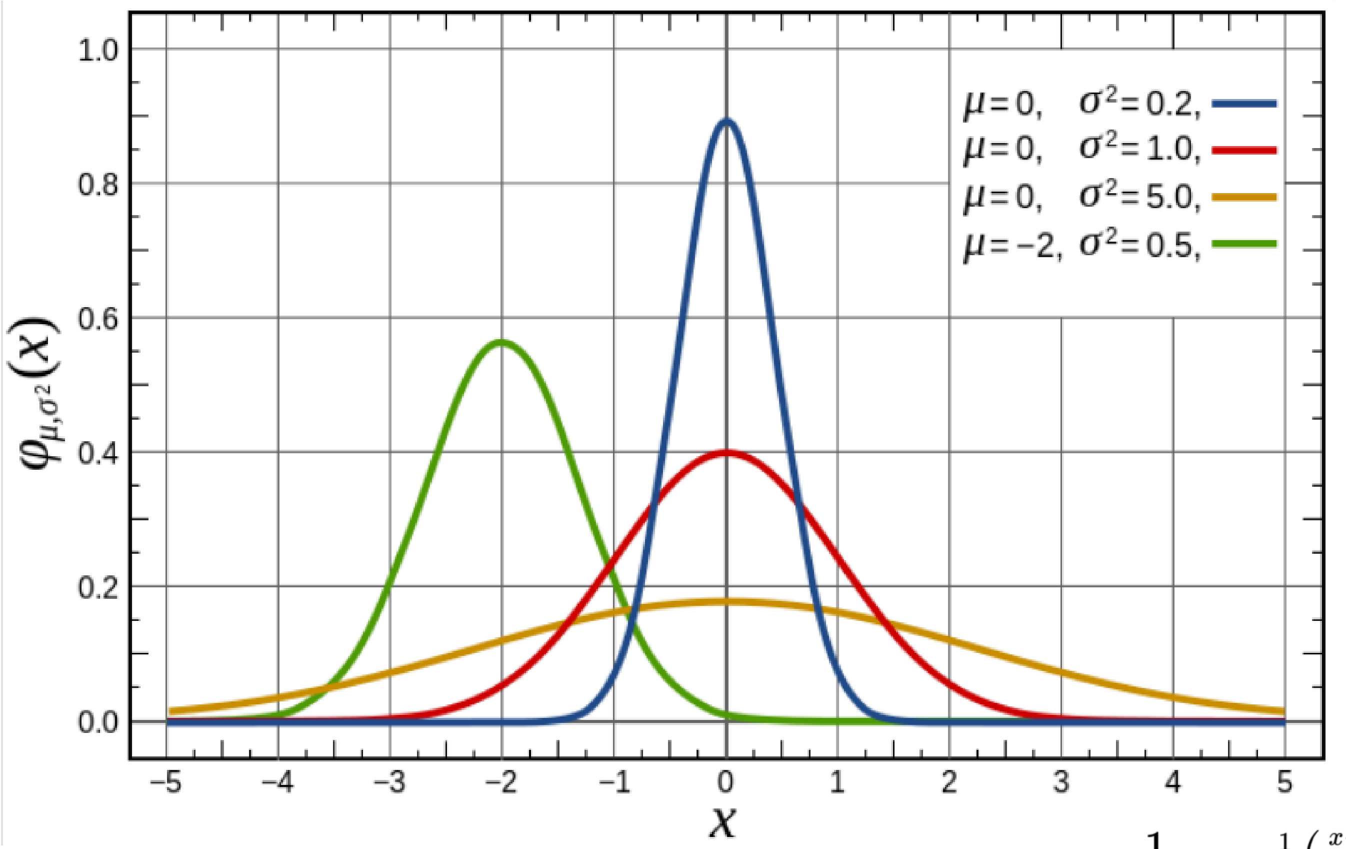


Quiz

- What is degree distribution in a graph, and what does it describe?
- Why is degree distribution important in graph analysis?
- How do you calculate the degree of a vertex in a graph?
- Differentiate between degree distribution in directed and undirected graphs.

Degree distribution in the graph

- Do real-world graphs have distributions by natural distribution (normal/gaussian distribution)?



Ví dụ:

- Chiều cao con người
- Điểm số của sinh viên
- …

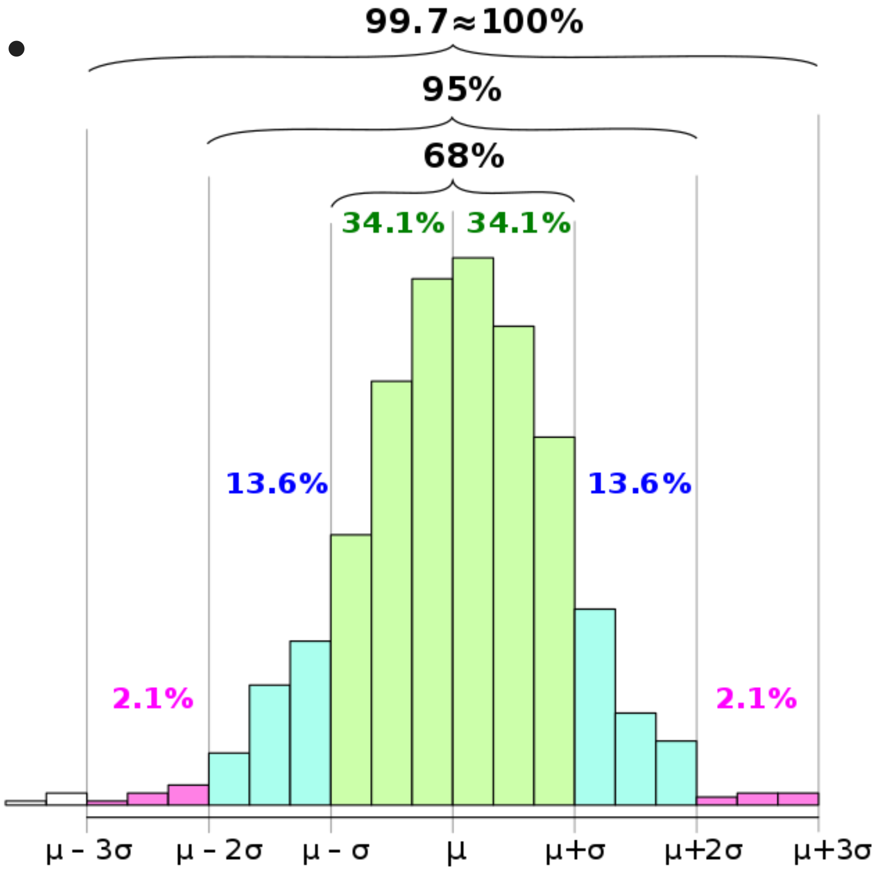
Đặc điểm:

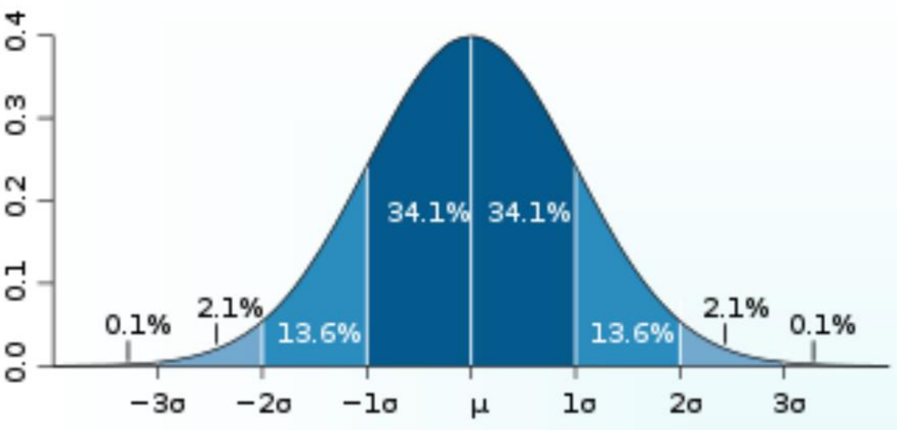
- Số lượng mẫu có xu hướng hội tụ (tăng lên) tại điểm trung tâm (kỳ vọng) …

Phân phối tự nhiên/chuẩn/hình chuông

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

3-sigma law

- The 3-sigma law or 68-95-99.7 states that about 99.7% of the data falls within three times the standard deviation around the expected value.
- 



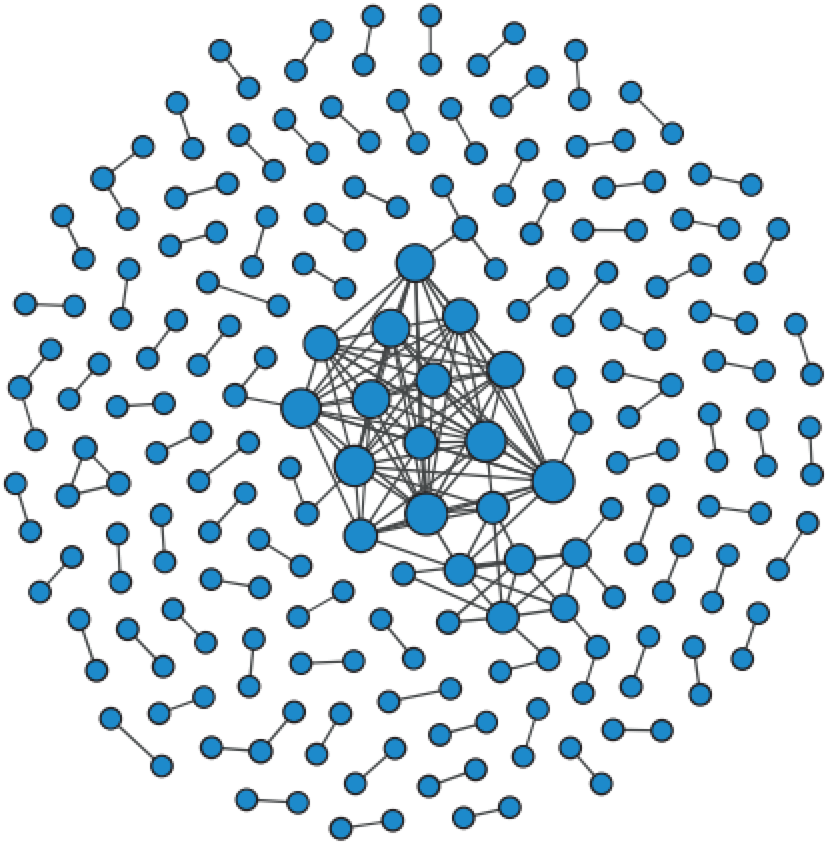
Degree distribution in the graph

- Do real-world graphs have distributions by natural distribution (normal/gaussian distribution)?
  - Example 1: in a social network with 1 million peaks, each peak has an average of 50 connections (friends).
    - If you choose 1 random peak, guess how many friends they have?
    - Would you be surprised if I said in this social network someone has 10,000 friends?

Degree distribution in the graph

- Do real-world graphs have distributions by natural distribution (normal/gaussian distribution)?
  - Example 1: in a social network with 1 million peaks, each peak has an average of 50 connections (friends).
    - If you choose 1 random peak, guess how many friends they have?
      - Close to 1
      - Most people rarely connect, possible reasons: not easy access to computers, internet, or high cost, too complicated, bringing little convenience to them, ...
    - Would you be surprised if I said in this social network someone has 10,000 friends?
      - :o
      - Some people want to have a lot of friends like actors, singers or they are participating in the popularity contest on Facebook, ...

Degree distribution in the graph



Degree distribution in the graph

- Do real-world graphs have distributions by natural distribution (normal/gaussian distribution)?
  - Example 2: an Oxford English dictionary has about 100,000 words. Through a daily document survey, it is estimated that each word is used about 5,000 times on average
    - Will any words appear >> 5,000
    - Does any word appear = 0?

Degree distribution in the graph

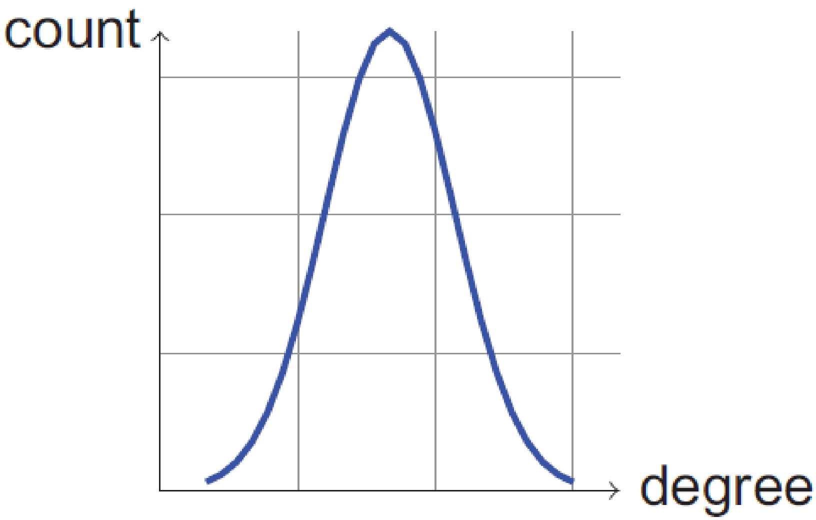
- Do real-world graphs have distributions by natural distribution (normal/gaussian distribution)?
  - Example 2: an Oxford English dictionary has about 100,000 words. Through a daily survey of audio data, it was calculated that each word is spoken about 5,000 times on average
    - Will any words appear >> 5,000
      - Yes: a, an, the, …
    - Any words appear = 0?
      - Yes: it has been noticed that an adult native speaker does not use more than 2,000 vocabulary words in their daily life.

Degree distribution in the graph

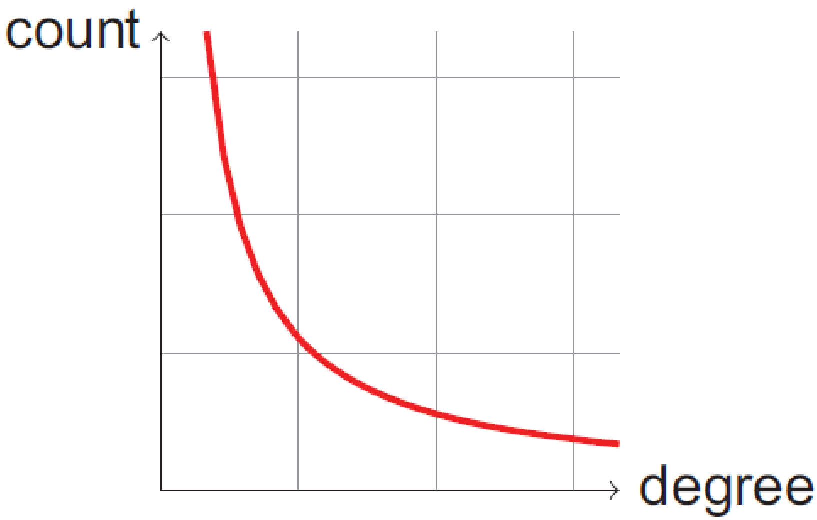
- Do real-world graphs have distributions by natural distribution (normal/gaussian distribution)?
  - Example 3: Routers when connecting to the Internet?
    - Are there any routers with huge connections?
    - Are there any routers that connect very little?

Degree distribution in the graph

- The order distribution in the real world graph/network is often skewed.



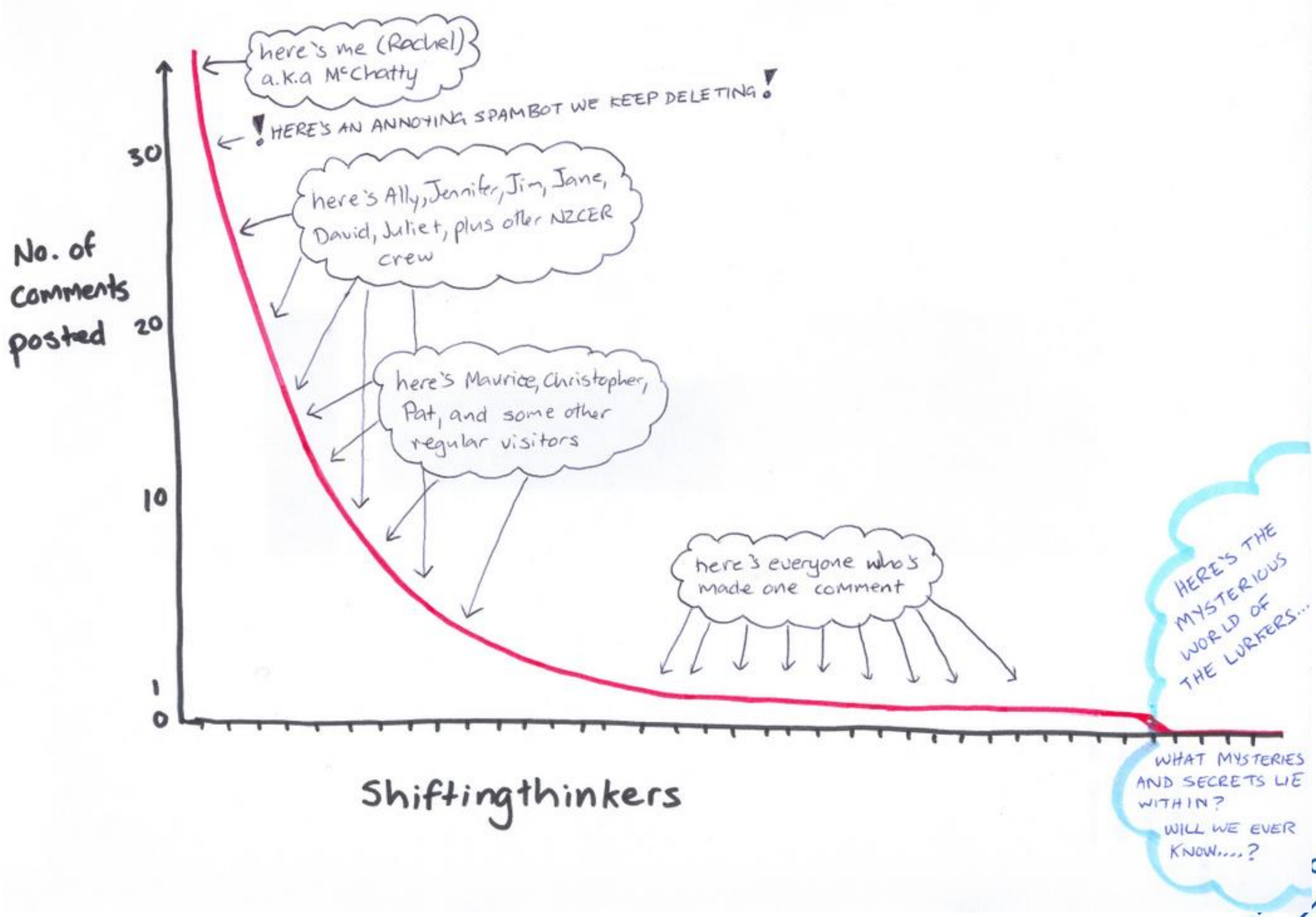
(a) WRONG intuition



(b) Reality

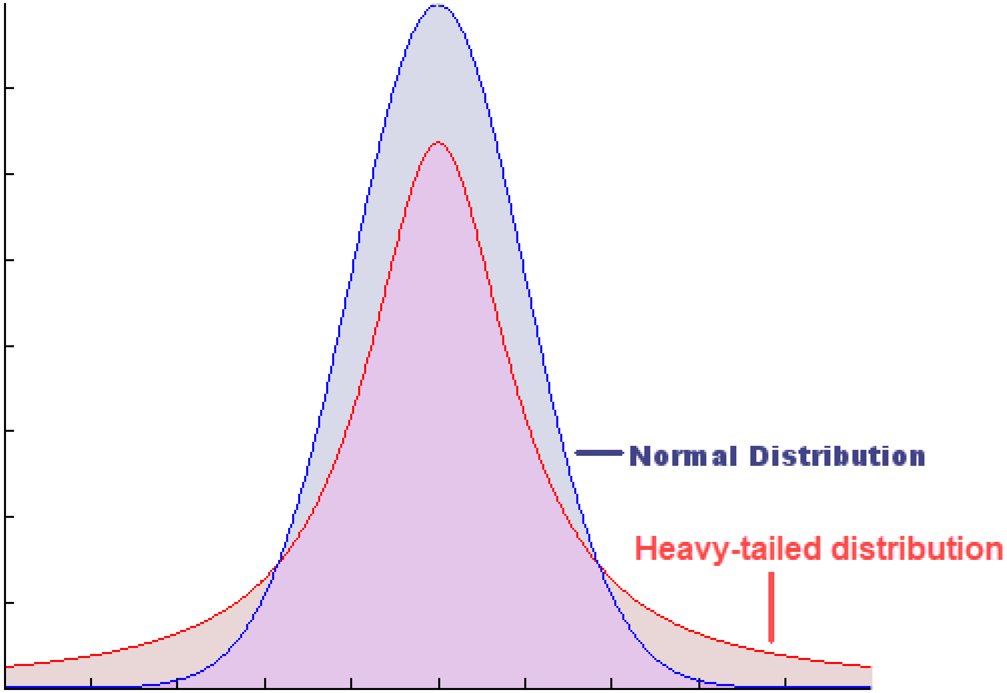
Vertices in the normal distribution have a degree close to the mean, a large deviation from it almost does not occur. Meanwhile, for real-world graphs, a lot of vertices are rarely connected.

Distribution of the number of comments on a blog



Degree distribution in the graph

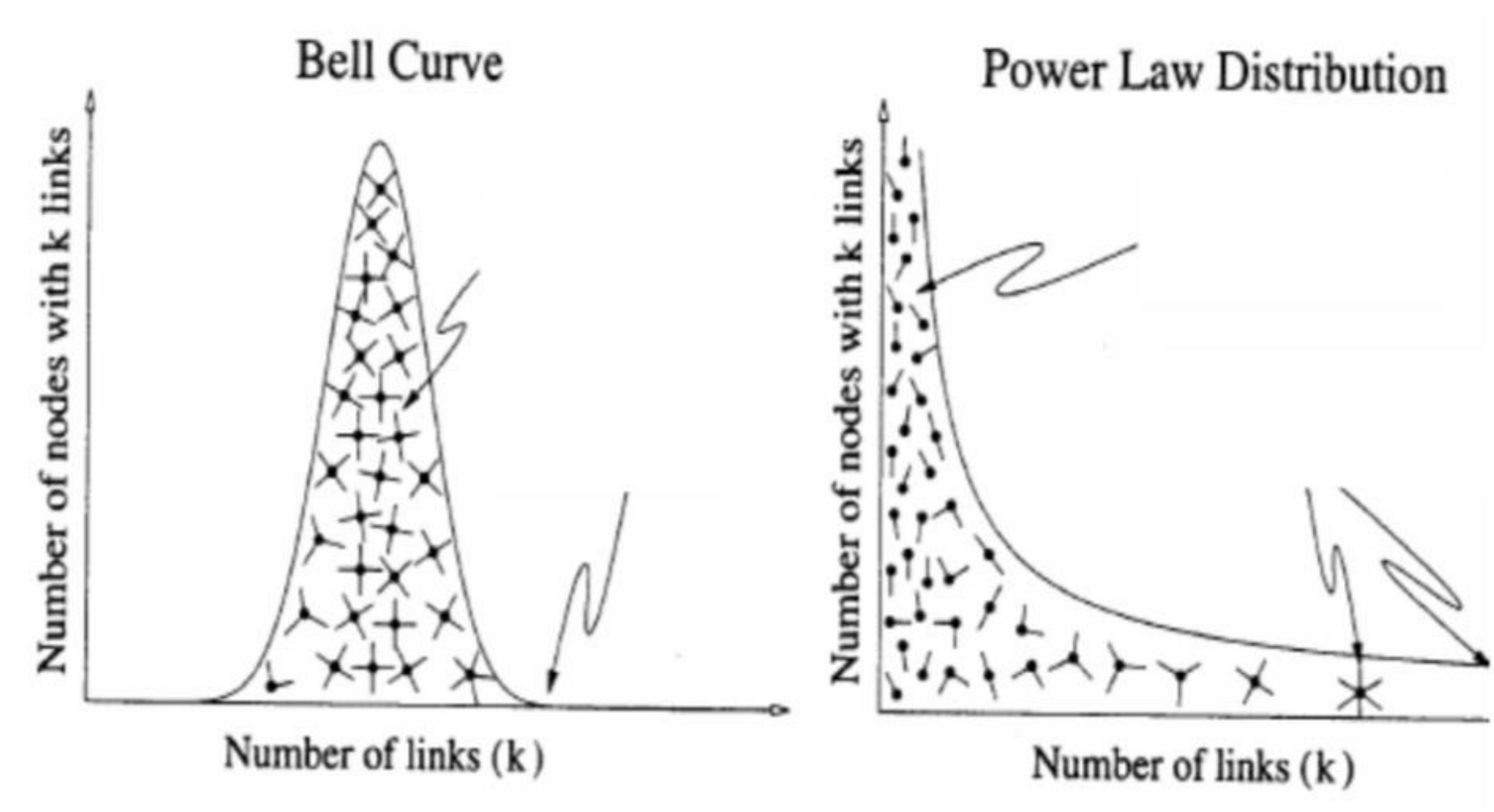
- The order distribution in graphs is usually heavy-tailed
- Characteristics: there exist elements with a very large value compared to the average. In other words, it decreases n



Heavy tail distribution

- Some distribution functions have the form of heavy tails:
  - Power-law distribution
  - Log-normal distribution
  - Weibull distribution
- Typically, graph data follows an exponential law distribution.

Comparison of normal distribution and heavy tail



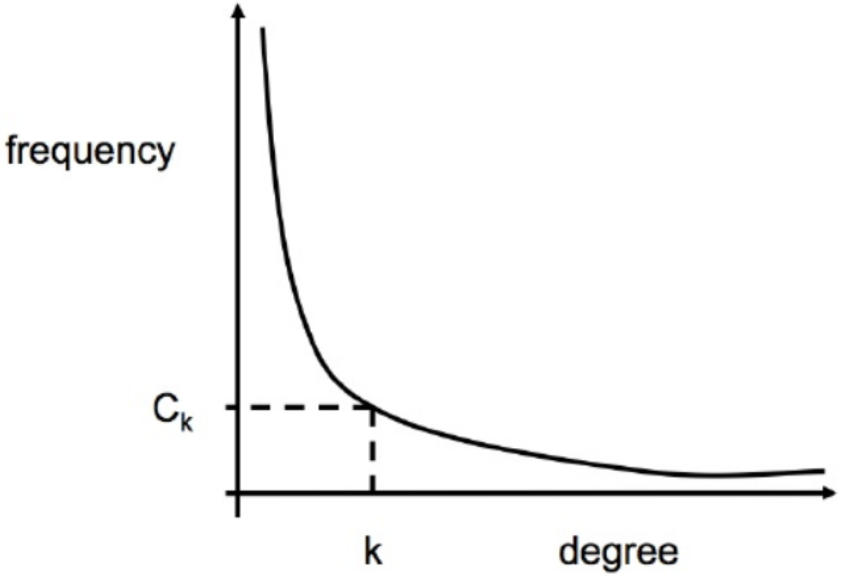
Power law

- The order distribution in a graph usually follows the power law:

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

with d is the order

α *is the positive constant, which is called the power of the exponential law* (power-law exponent)

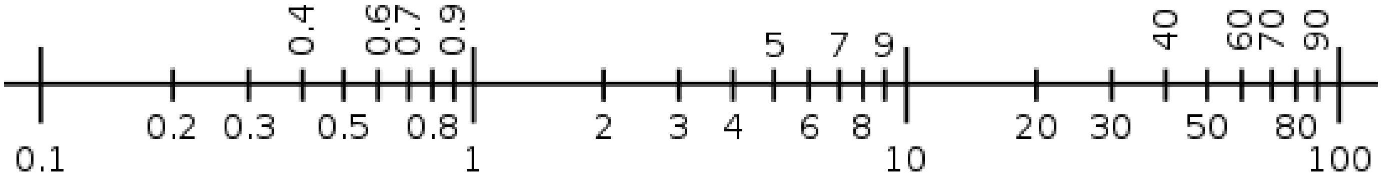


Quiz

- Explain the power law in degree distribution and its applications in real-world scenarios.
- Provide examples of graph types (such as social networks, internet networks) where power law degree distributions are common.
- How does degree distribution affect propagation properties in networks?
- Discuss the impact of degree distribution on optimization algorithms on graphs, such as shortest path algorithms or clustering.
- Discuss methods that can be used to estimate or simulate degree distribution in large graphs without examining the entire structure.

Distribution of exponential laws

- The exponential distribution value is usually very small the larger the order
- To represent the exponential distribution over the coordinate axis, it is customary to convert the value to a logarithmic space:
  - Show a great domain of value in a compact space.
  - For example, 10, 100 would represent the same width

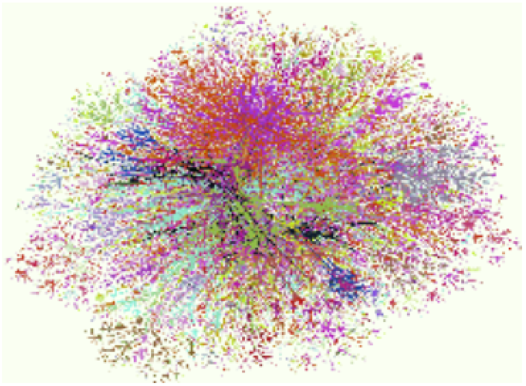


Distribution of exponential laws

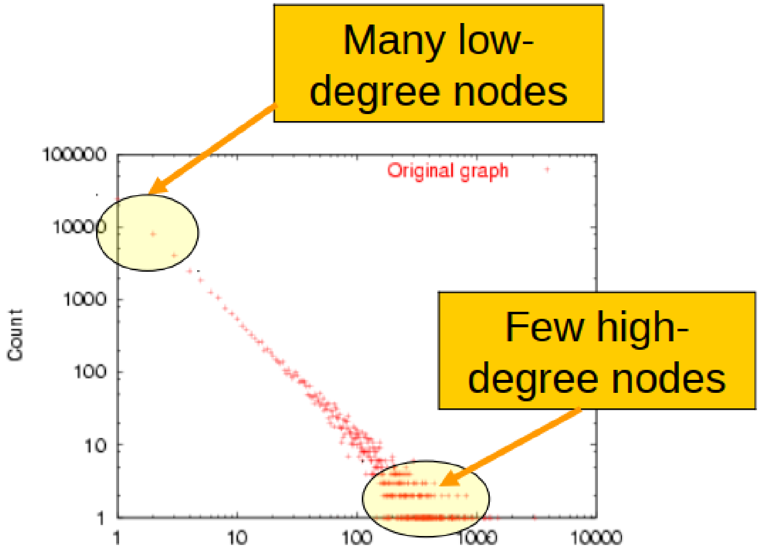
$$f(d) = Ad^{-\gamma}$$

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

- This exponential distribution graph is called a log-log graph.



Internet in December 1998



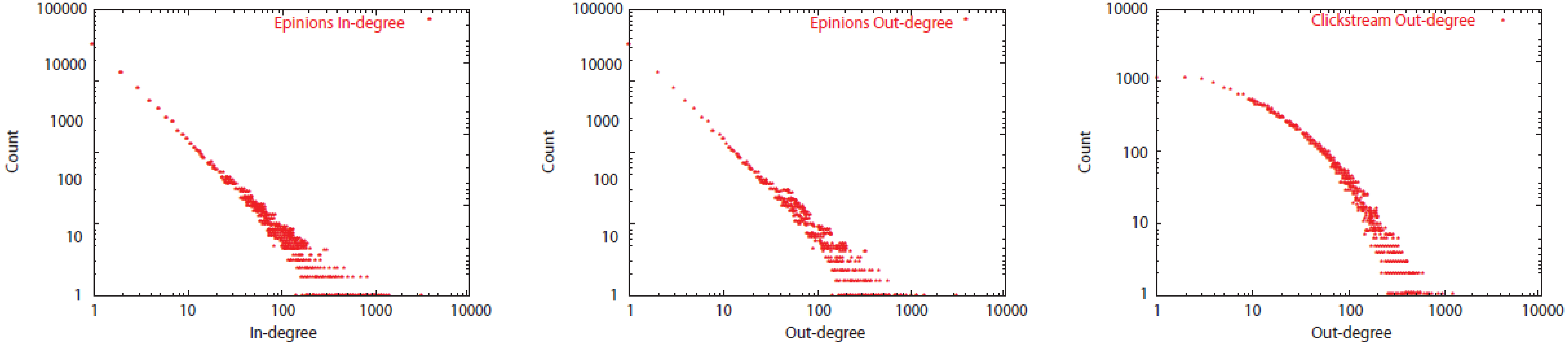
Many low-degree nodes

Few high-degree nodes

log(Count) vs. log(Degree)

Exponential law distribution example

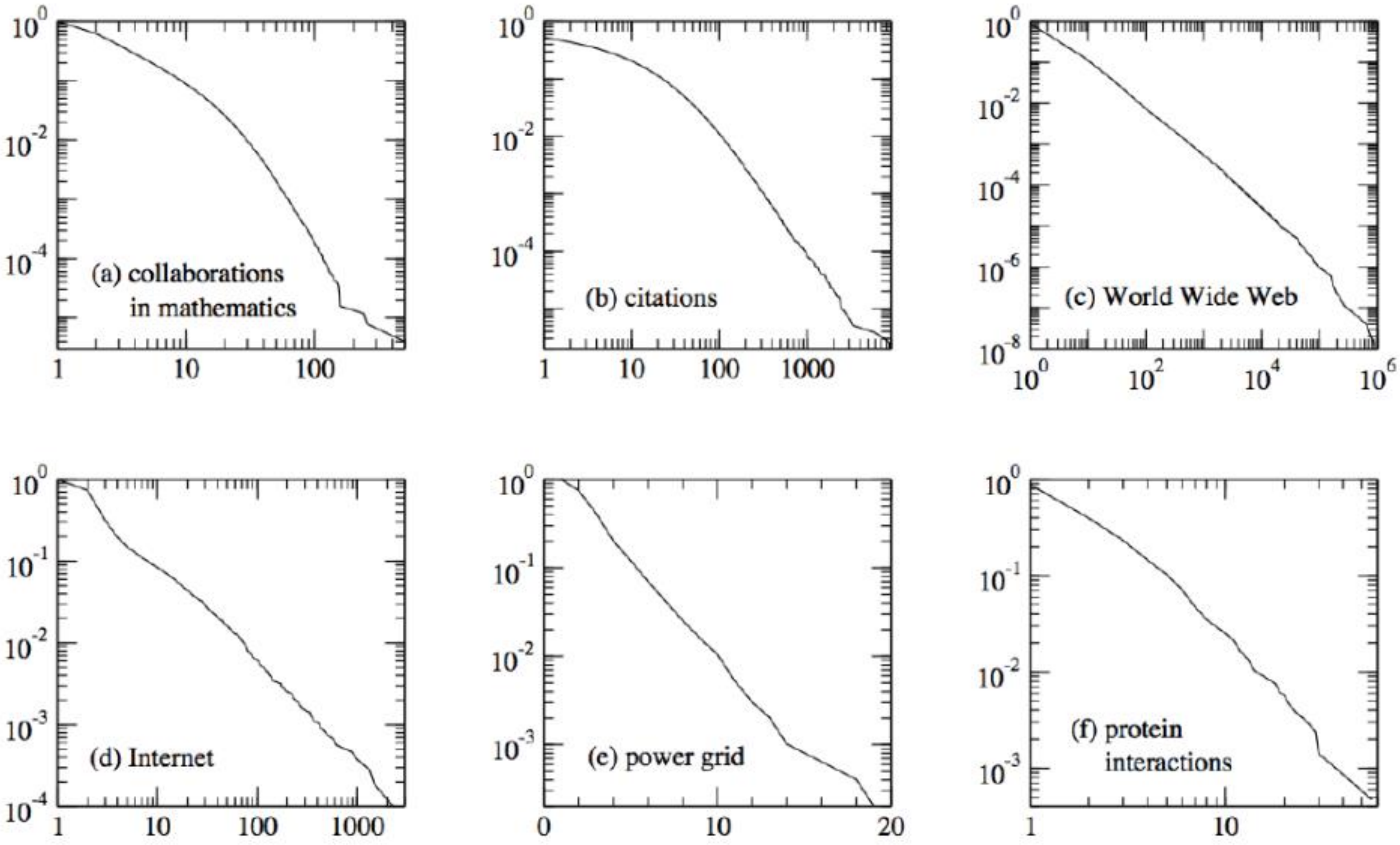
- In the WWW and social network graphs, the inner and outer tier distributions also follow the exponential rule.



(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).

Exponential law distribution example



Scale-free graph

- Graphs with exponential order distributions are also called scale-free graphs.
  - Its characteristic is independent of the size of the graph. That is, when the graph bulges, the underlying structure remains similar.
  - With the hat law: $y(x) = Ax^{-\gamma}$ -> $y(ax) = by(x)$

Law of exponential eigenvalues

- Siganos found that the spectrum (a set of eigenvalues ordered in decreasing order of magnitude) of an adjacency matrix representing a graph of the Internet also follows an exponential distribution.

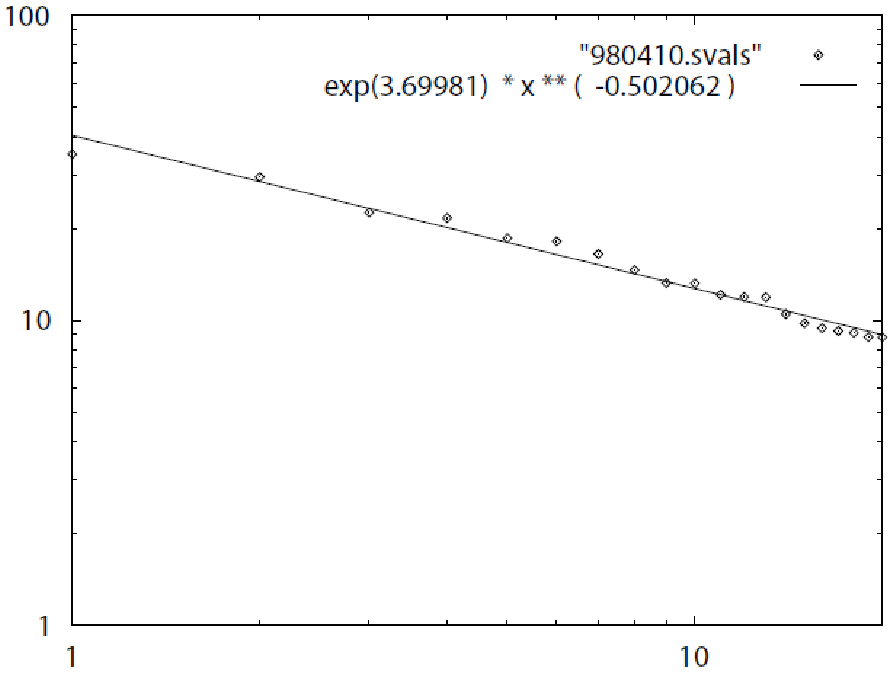
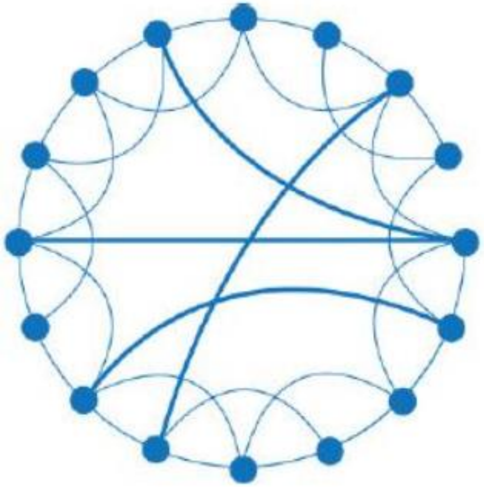


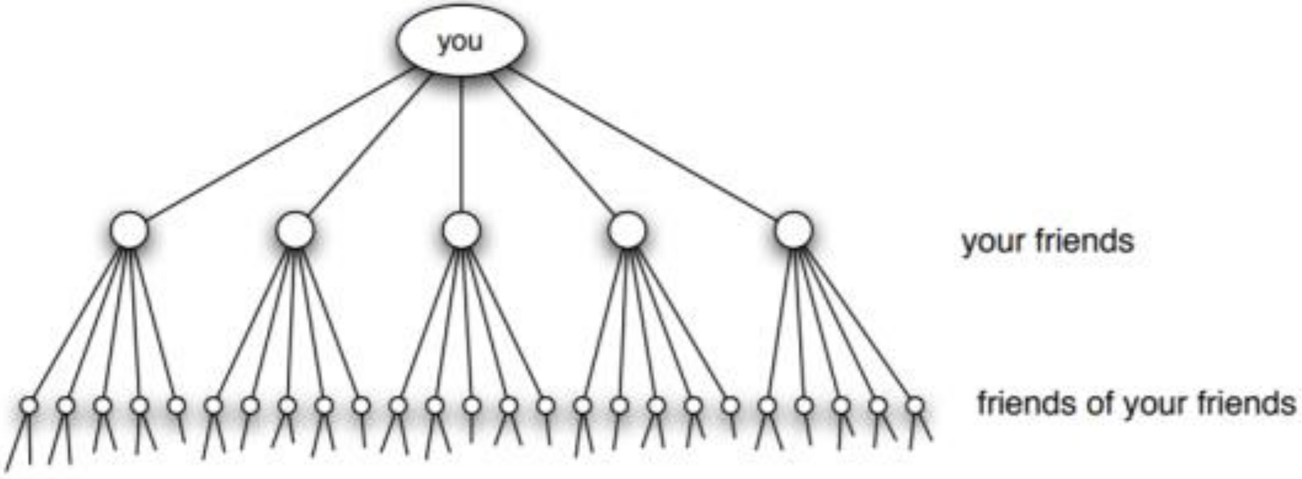
Figure: Scree plot obeys a power law: Eigenvalue λ_i versus rank i for the adjacency matrix of autonomous systems (April 10, '98).

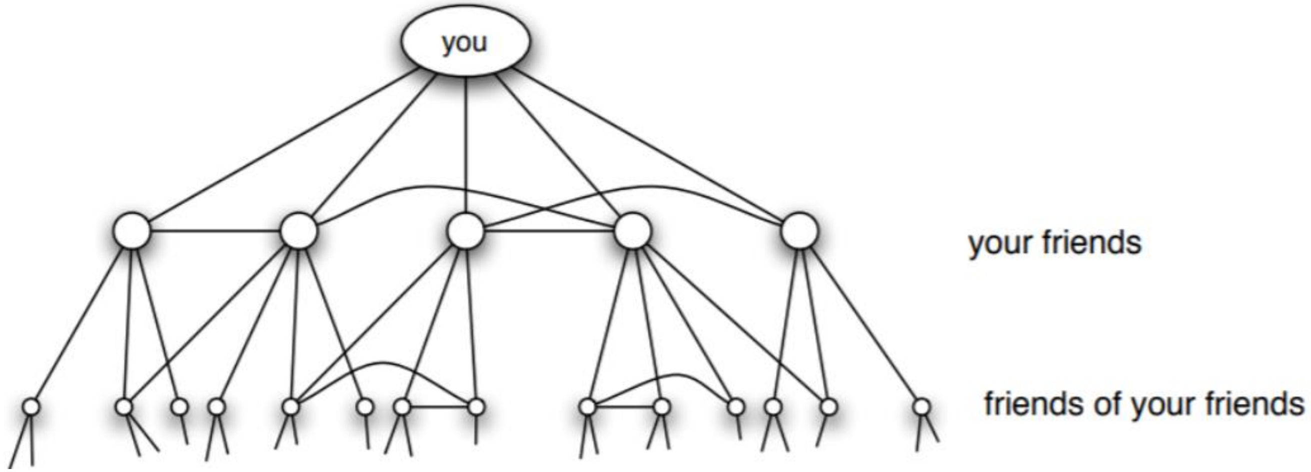
Small diameter

- The diameter of a graph is the greatest distance between any pair of vertices.
  - The distance is calculated on the shortest path between two vertices, regardless of direction.
- Real world graphs also have small diameter
  - Also called a small-world phenomenon
  - In social networking, it is also called six degrees of separation

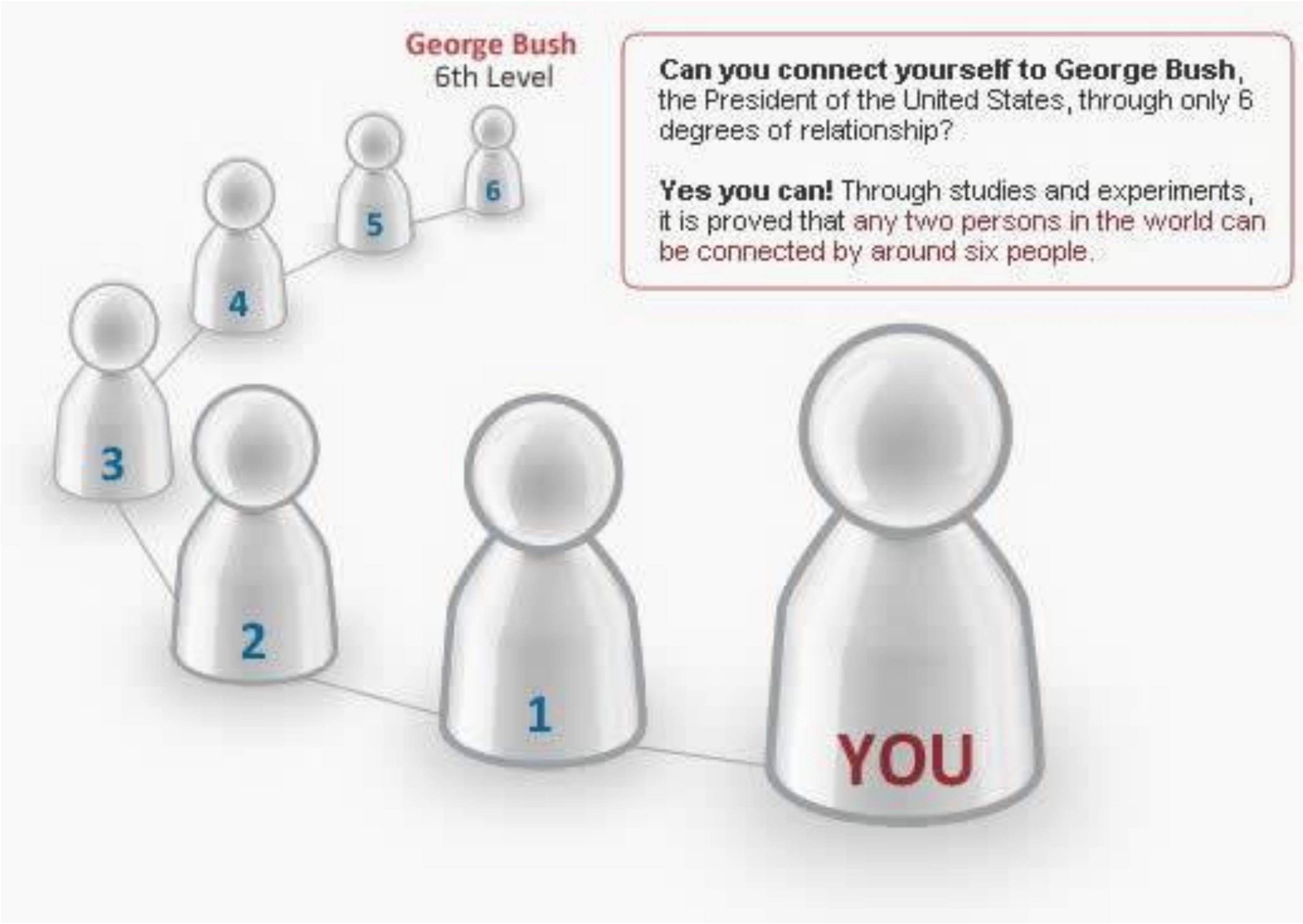


Small diameter





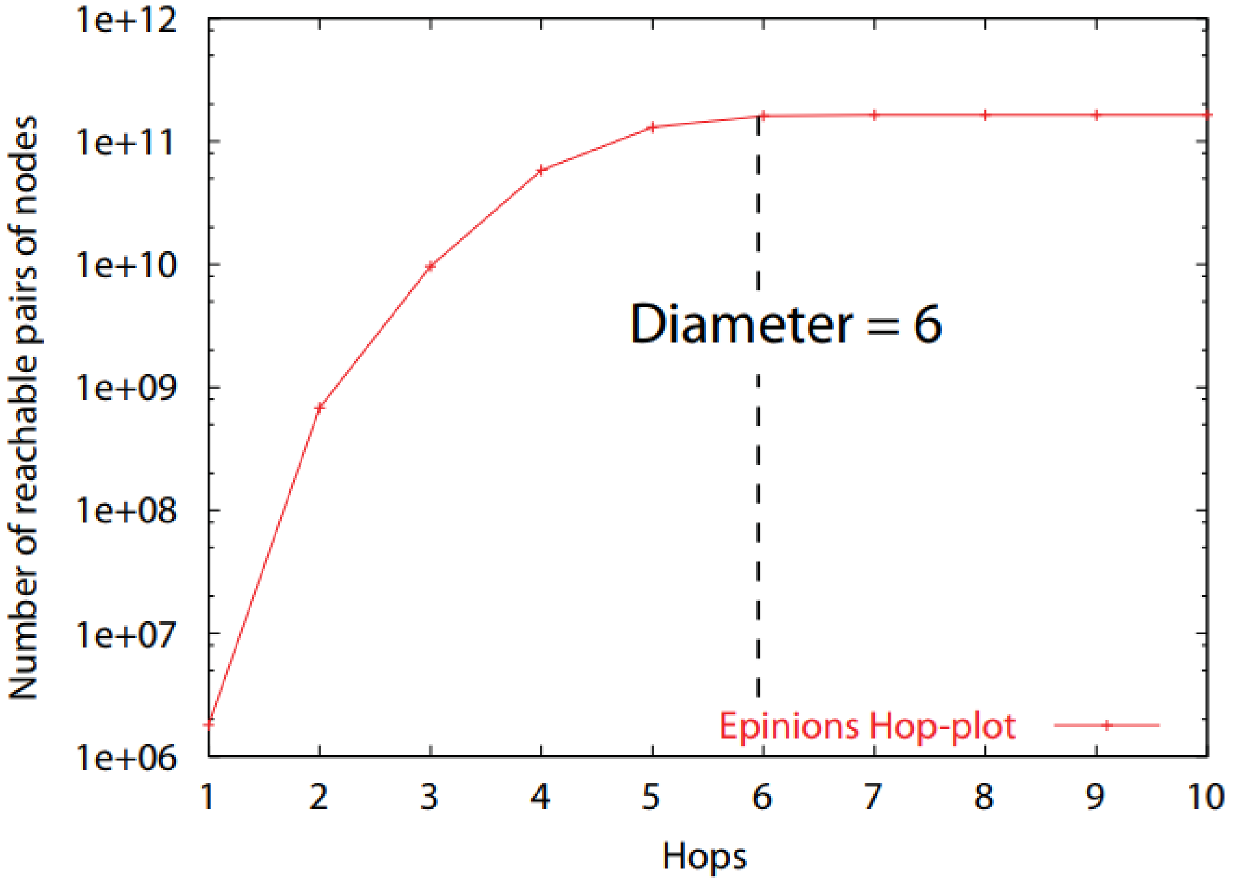
Small diameter



Can you connect yourself to George Bush, the President of the United States, through only 6 degrees of relationship?

Yes you can! Through studies and experiments, it is proved that any two persons in the world can be connected by around six people.

Small diameter examples

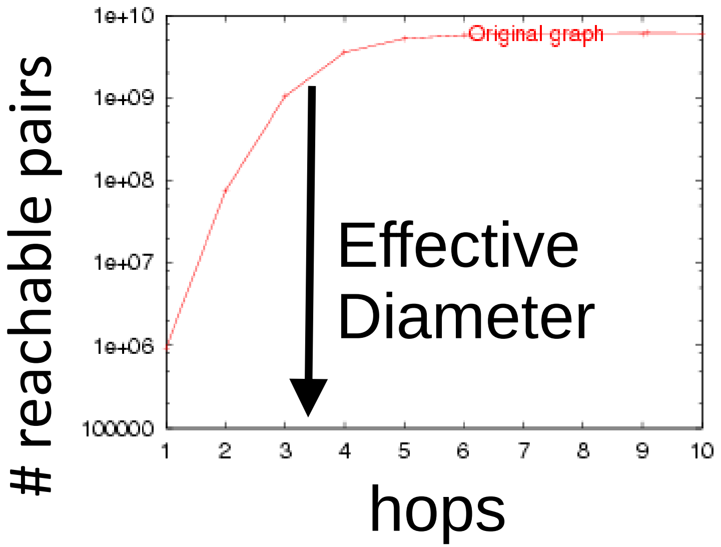


The graph shows the number of hops (edges) and the number of vertex pairs reachable in the Epinions network data

The picture can't be displayed.

Problems in graph diameter

- The diameter of the graph can be affected by a small number of pairs that occur but have a long sequence of paths.
- Instead, use an effective diameter measure
  - – The minimum number of edges that a ratio (e.g. 90%) of all vertex pairs can reach.



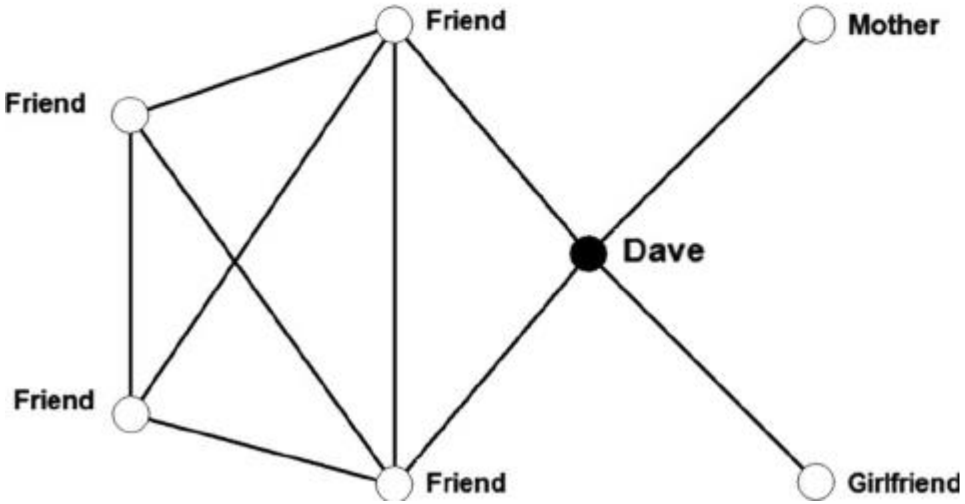
Calculation cost problem

- The cost to compute the shortest path length between all pairs of vertices in a large graph is usually high (at least (N^2))
  - Solution: apply approximate neighborhood function

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

Law of triangle exponent

- Real world graphs often have many triangular connections.
  - For example: your friend is usually also his friend.



Law of triangle exponent

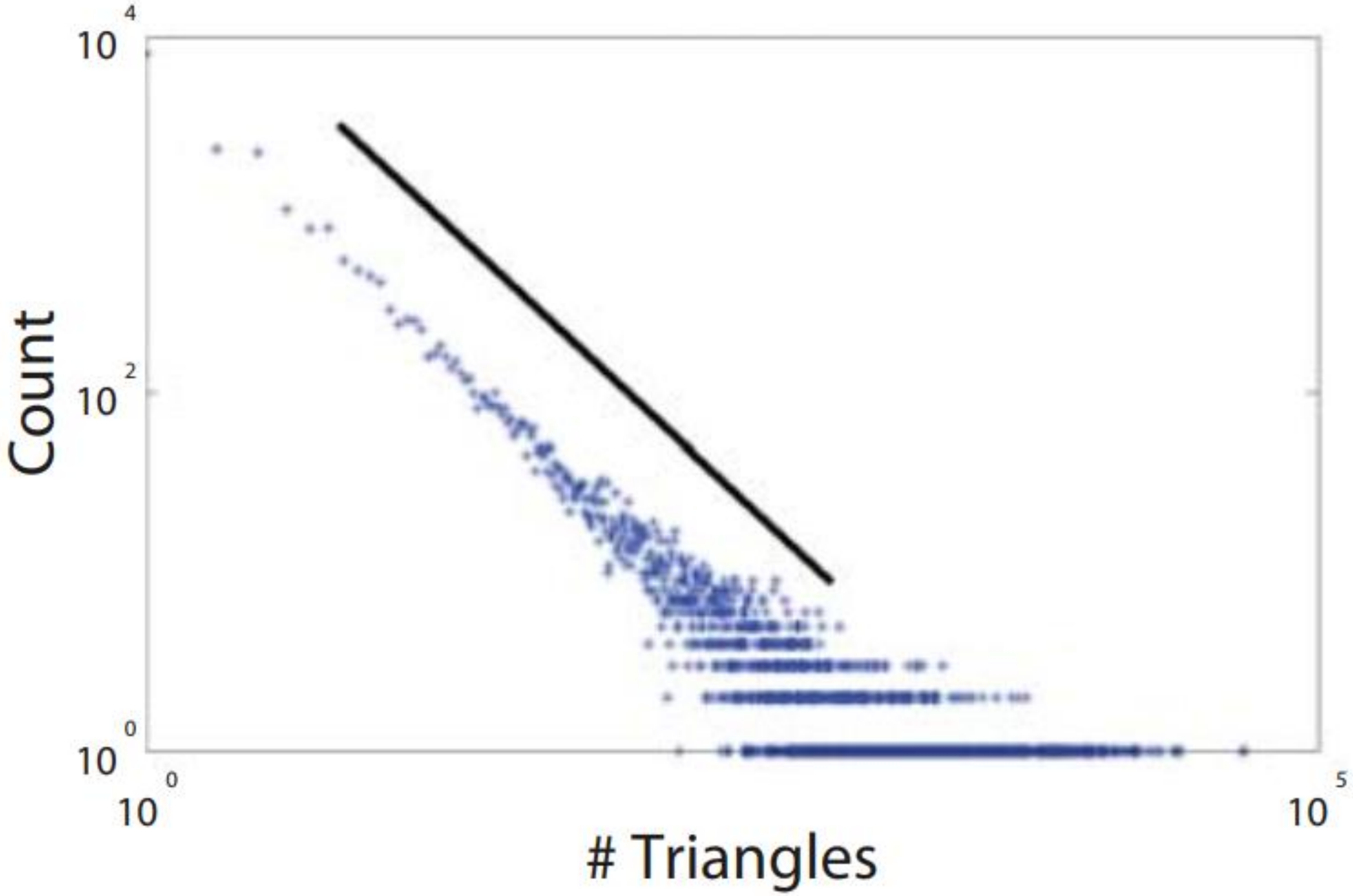
- Set Δ as the number of triangular connections in the graph

$$\Delta_i \mathit{is\ the\ number\ of\ triangle\ connections\ with\ vertices\ participating\ in}$$

The triangle participation law (TPL) shows that the distribution of Δ_i also obeys the exponential law with powers σ.

  - A lot of vertices have few triangular connections
  - Existence vertices participate in a large number of triangular connections

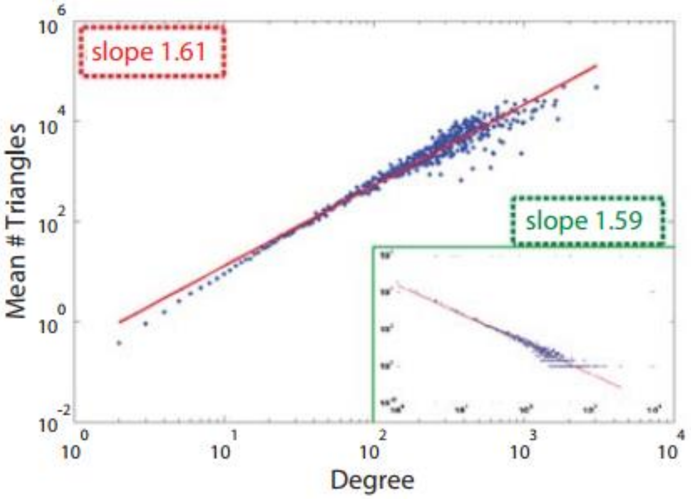
Law of triangle exponent



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

Law of triangle exponent

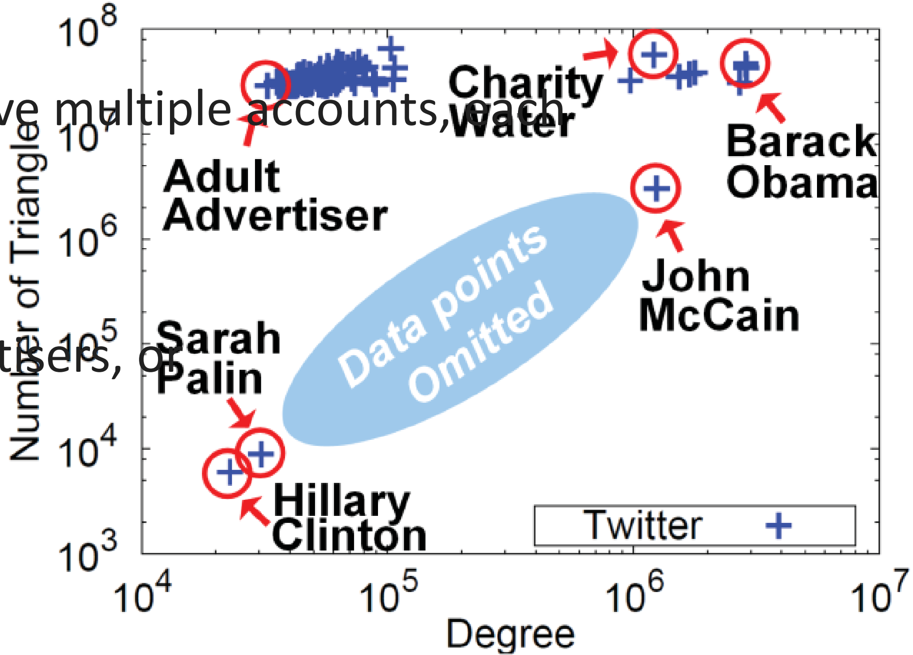
- When considering the correlation between the number of d_i vertices of the graph and the number of connections Δ_i:
  - Obey the exponential rule with a positive angle coefficient: Δ_i∝d_i^s with s≈1.5
  - For example, the more friends a person has, the more triangular connections are formed
  - This is called degree-triangle participation law (DTPL)



DTPL

Law of triangle exponent

- Kang discovered that Twitter is also DTPL.
  - Celebrities are high-ranking and have the connections of their followers who follow the triangular rule.
  - However, there are some unusual data:
    - Some accounts have relatively small tiers but participate in a large number of triangular connections
    - It is possible that some people have multiple accounts, each following account another to push the ladder up
    - These are usually adult film advertisers, or Spammer
    - Using DTPL we can discover these unusual data.

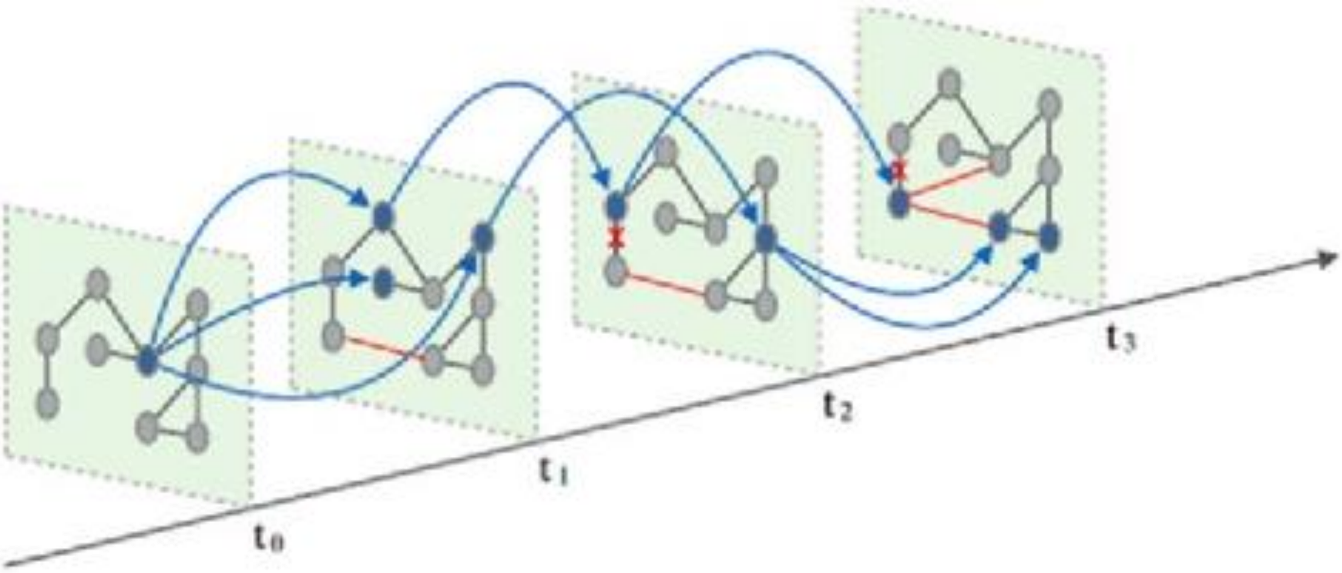


Content

- Graph patterns
- Patterns in static graphs
- Patterns in dynamic graphs
  - Diameter change
  - The law of growth exponents
  - Gelling point
  - Principal eigenvalue over time
- Patterns in weighted graphs
- Cost calculation

Dynamic graph

- Graphs with connections added or removed over time are called dynamic graphs (dynamic, time-varying, evolving graph).

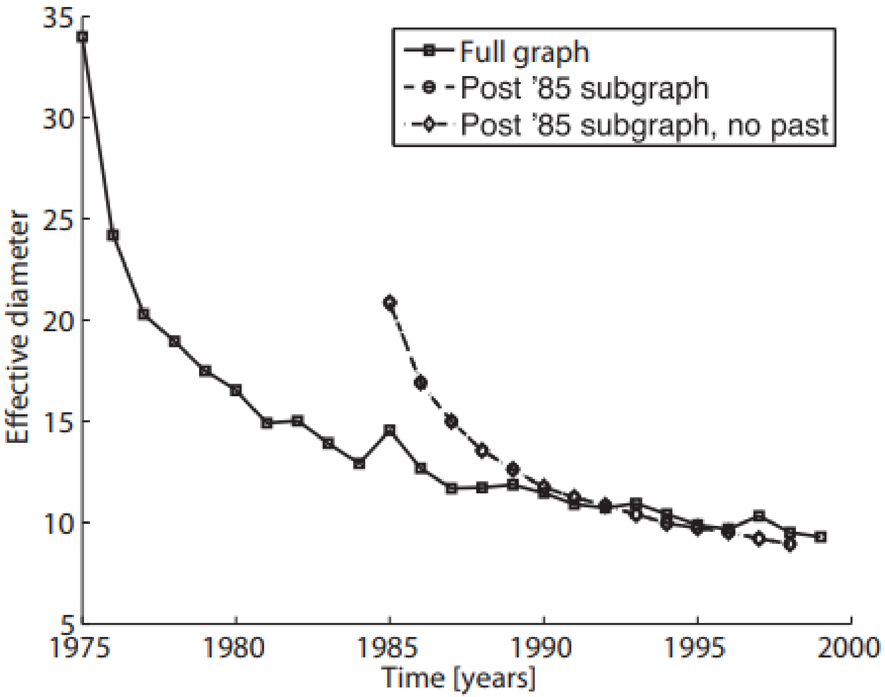


Diameter change phenomenon

- As the graph grows over time, how will the diameter of the graph change?
  - Increase or decrease?
  - For example, adding friends in social networks

The phenomenon of diameter shrinking

- Studies show that as the graph grows, the diameter of the graph tends to shrink even when new vertices are added.



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

The law of growth exponents

- It is assumed that at time t, we have the vertex number of the graph as $N(t)$ and the number of edges as $E(t)$.
- At the next time t+1, the number of vertices of the graph doubles $N(t+1)=2\ast N(t)$
    - So what is the number of edges of the graph at this point $E(t+1)$?
        - Double the top, double the edge?
            - Wrong

The law of growth exponents

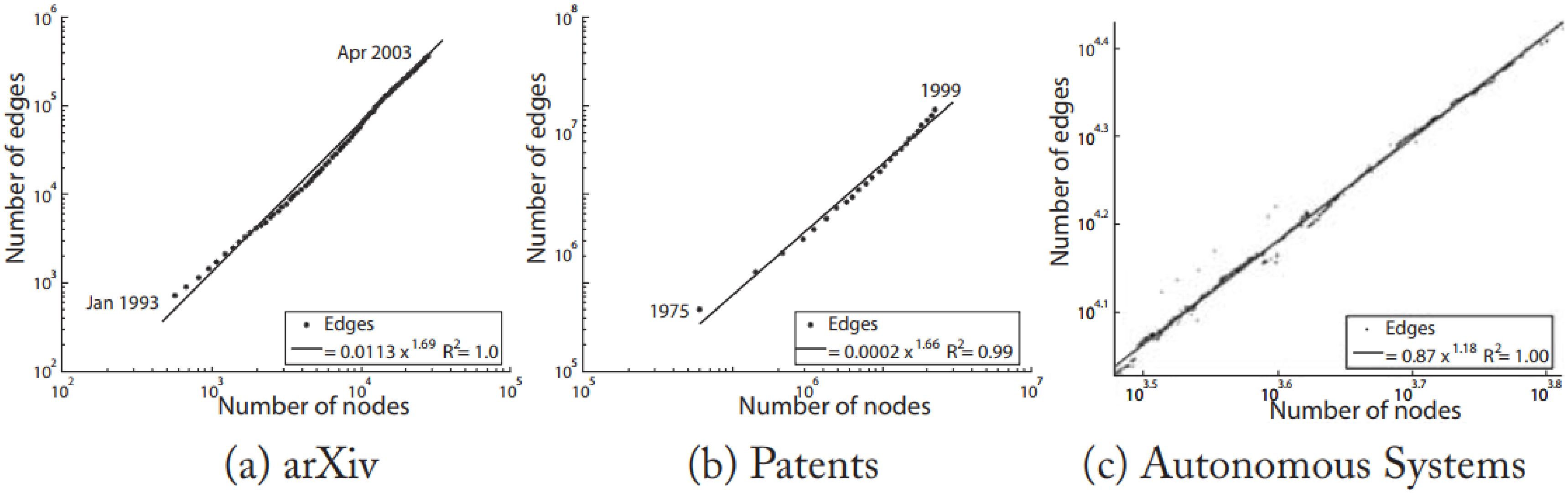
- In the real world graph, it is noticed that the number of vertices and the number of edges also obey the exponential law with positive powers.

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

with β is the growth cap

- This is called the desification/growth power law

The law of growth exponents



(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

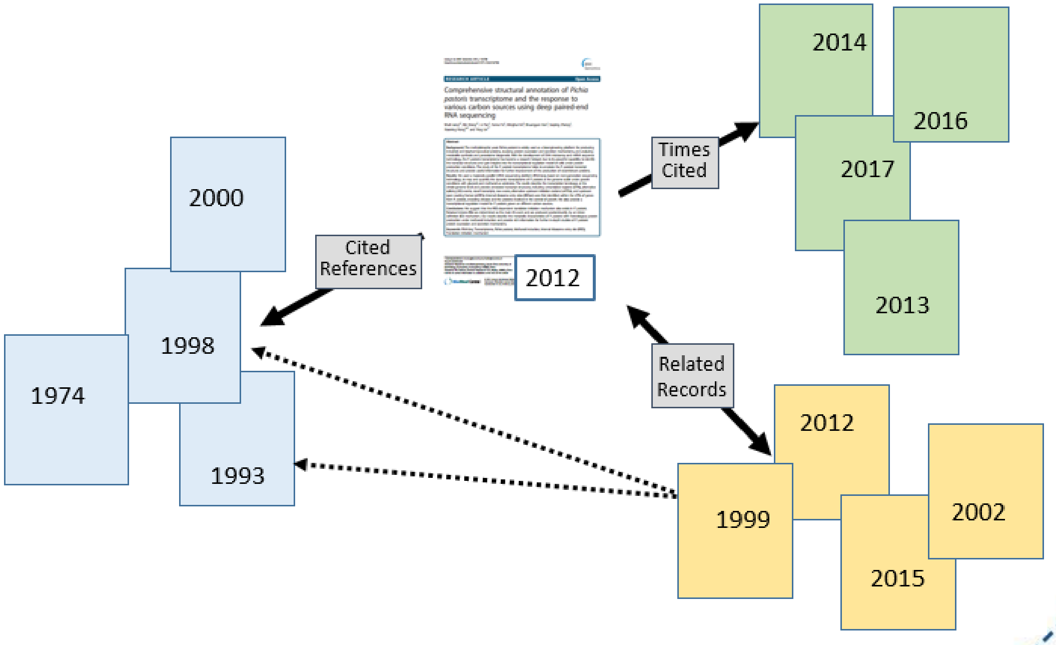
The law of growth exponents

- If the power is β>1 (regular):
  - When the number of vertices doubles, the number of edges of the graph more than doubles.

    → Over time, the average of vertices increases as there are more new edges than new ones.
  - This is what causes the diameter to shrink.

The law of growth exponents

- Consider the citation network example
  - According to the exponential rule of growth, over time, the average tier increases or the average number of citations per article increases.
  - Do people writing articles today cite more than they used to?

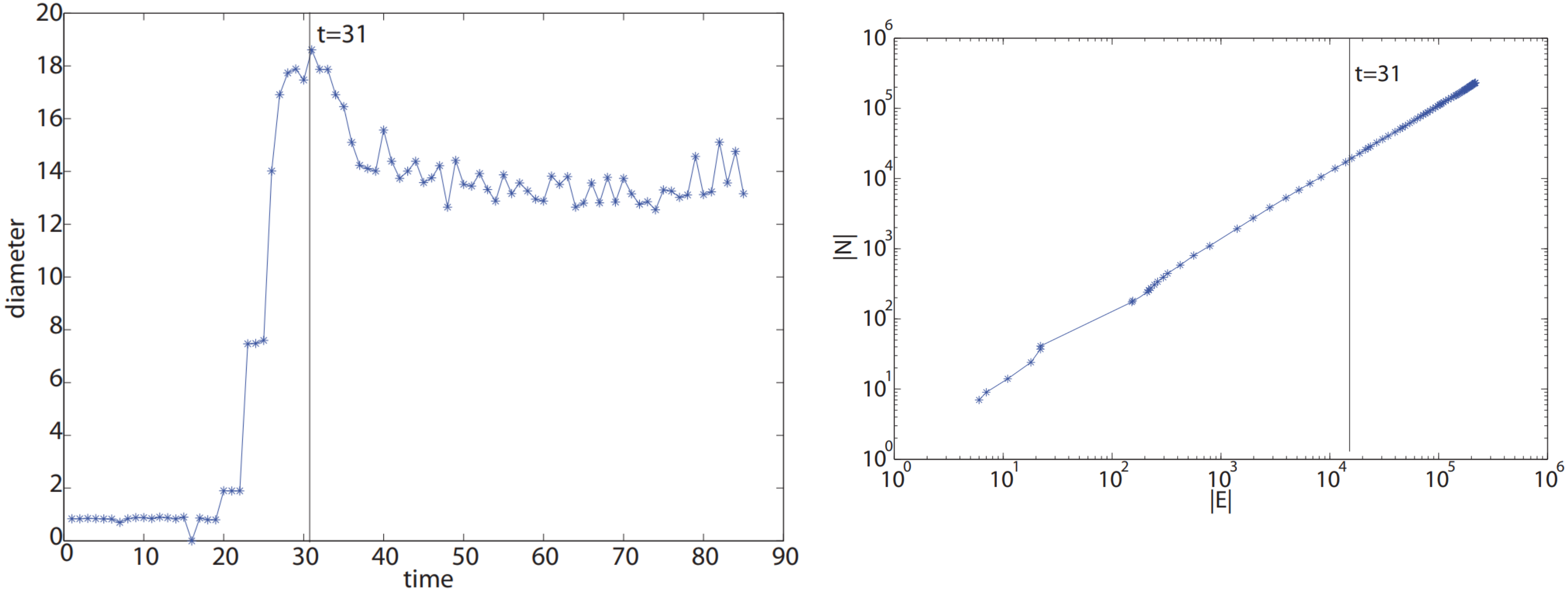


The law of growth exponents

- Consider the citation network example
  - According to the exponential rule of growth, over time, the average tier increases or the average number of citations per article increases.
  - Do people writing articles today cite more than they used to?
    - Not. Most of us write quoted articles with 10 to 30 articles so far.
    - So what causes the average to rise over time?
      - Super paper, textbook later referenced thousands of previous articles

Gelling point

- McGlohon noticed that at one point the diameter of the graph skyrocketed.



Dữ liệu theo thời gian của mạng PostNet

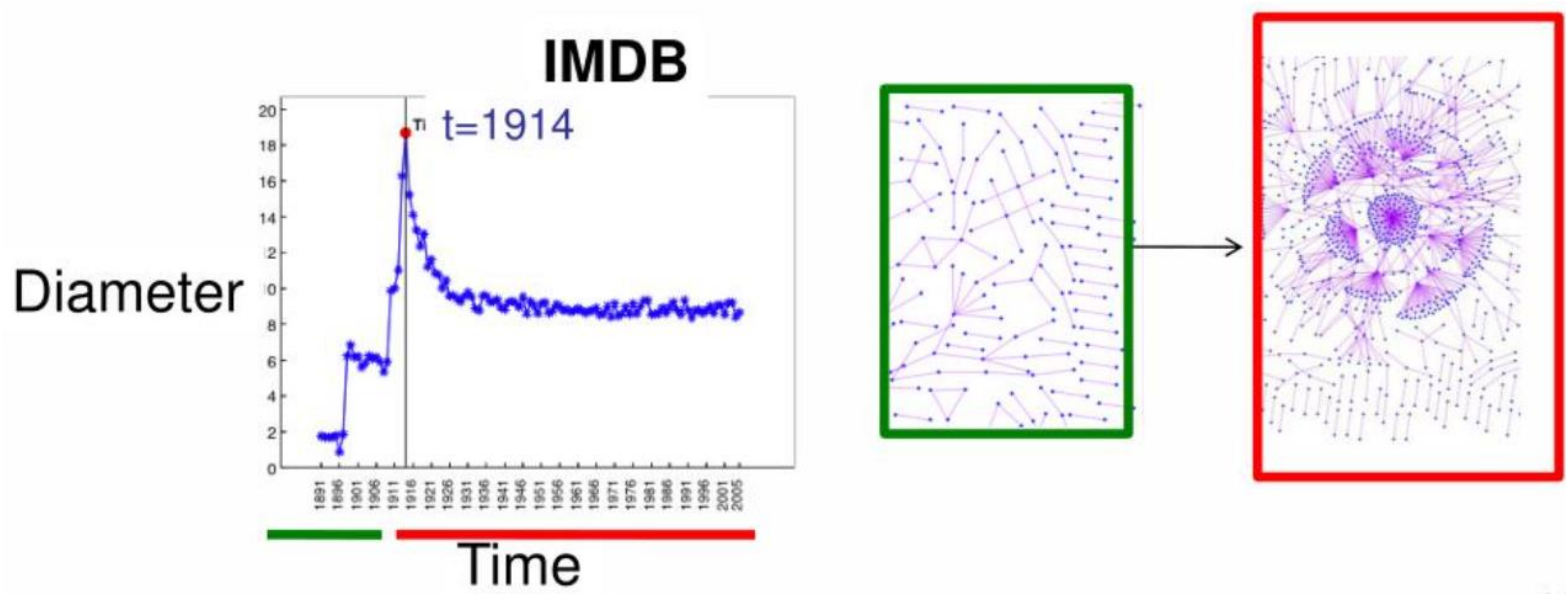
Chart (a) showing changes in diameter over time
Graph (b) shows the correlation between the number of edges and the number of vertices over time (still subject to the exponential rule of growth)

Gelling point

- McGlohon noticed that at one point the diameter of the graph skyrocketed.
  - This point is called the gelling point
    - Prior to that time, graphs consisted of a small set of non-connected components.
    - At the adhesion point, a giant connected component (GCC) begins to form and occupy the majority of the peaks. When new peaks form, it also tends to join this GCC block.
    - The remaining interconnected components are called non-largest connected components (NLCCs)

Gelling point

- Example of a gelling point in an IMDB network

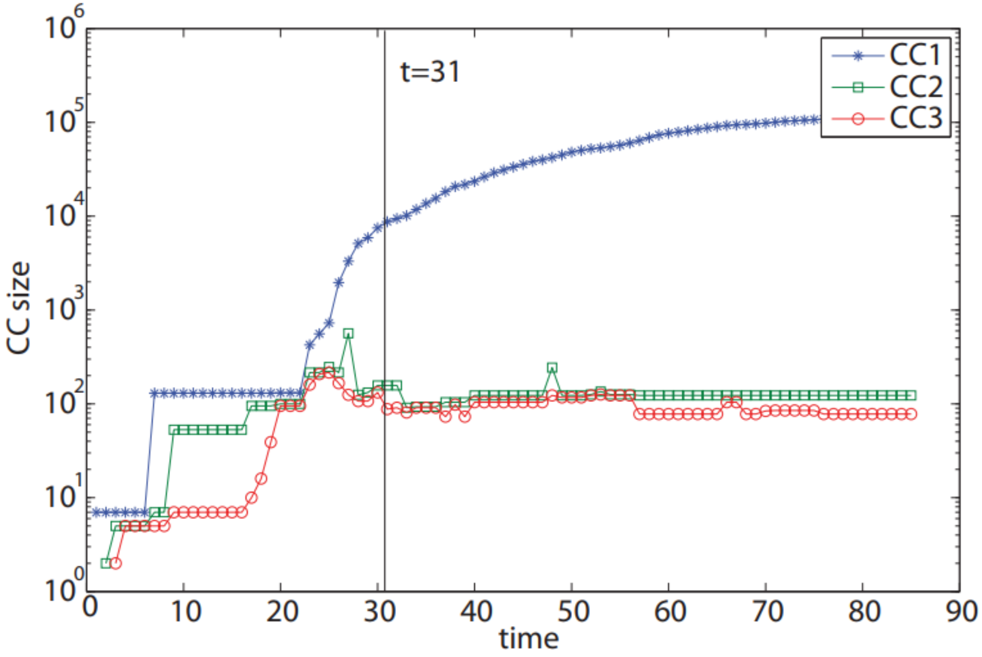


Gelling point

- What happens after the gelling point?
  - Will the GCC continue to rise?
  - Will the NLCC block increase slightly or will it remain unchanged?

Gelling point

- After the gelling point:
  - GCC continues to grow in size
  - The NLCC block mostly stays the same or oscillates around an interval.

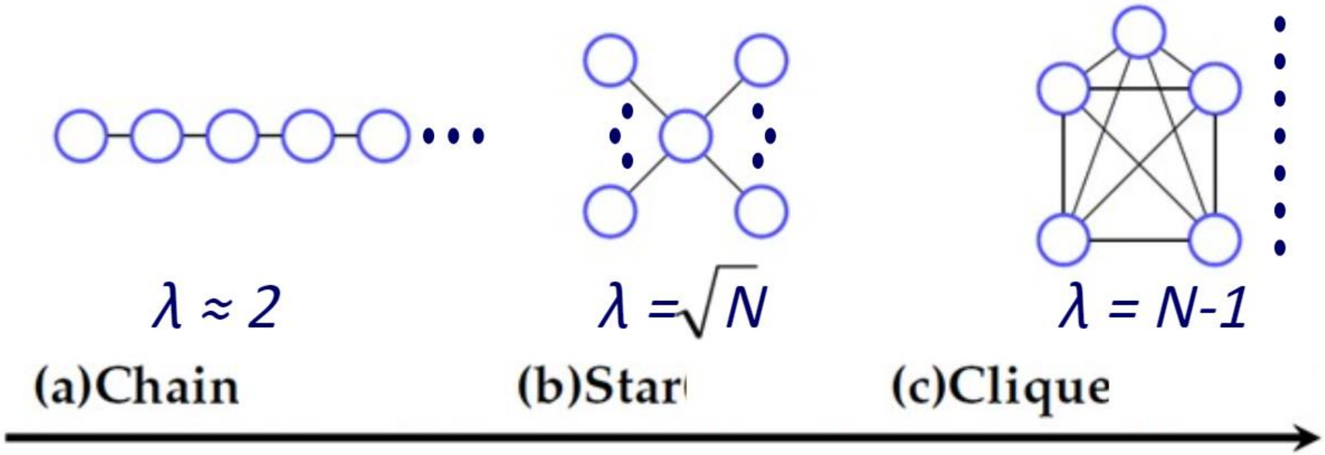


GCC, CC2, and CC3 (log-lin)

Principal eigenvalue over time

- The first principal (maximum) eigenvalue λ_1 is one of the important measures of graph connectivity.

better connectivity ⟶ higher λ



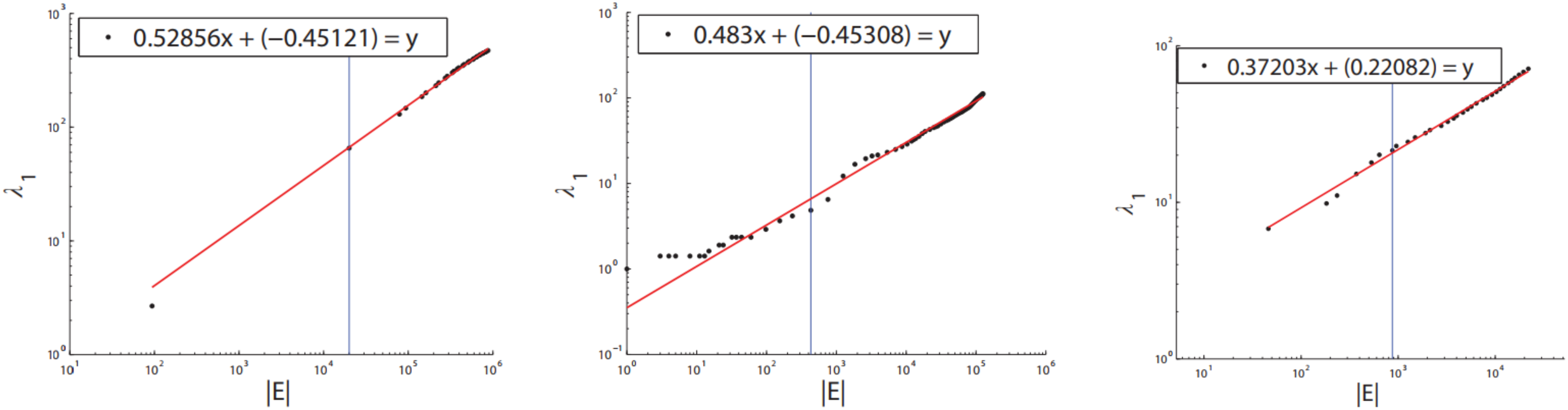
$\lambda \approx 2$ $\lambda = 31.67$ $\lambda = 999$

$N = 1000$ nodes

Principal eigenvalue over time

- In real world graphs, maximum eigenvalue $\lambda_1(\mathrm{t})$ and the number of edges E(t) changes exponentially over time to powers less than 0.5

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



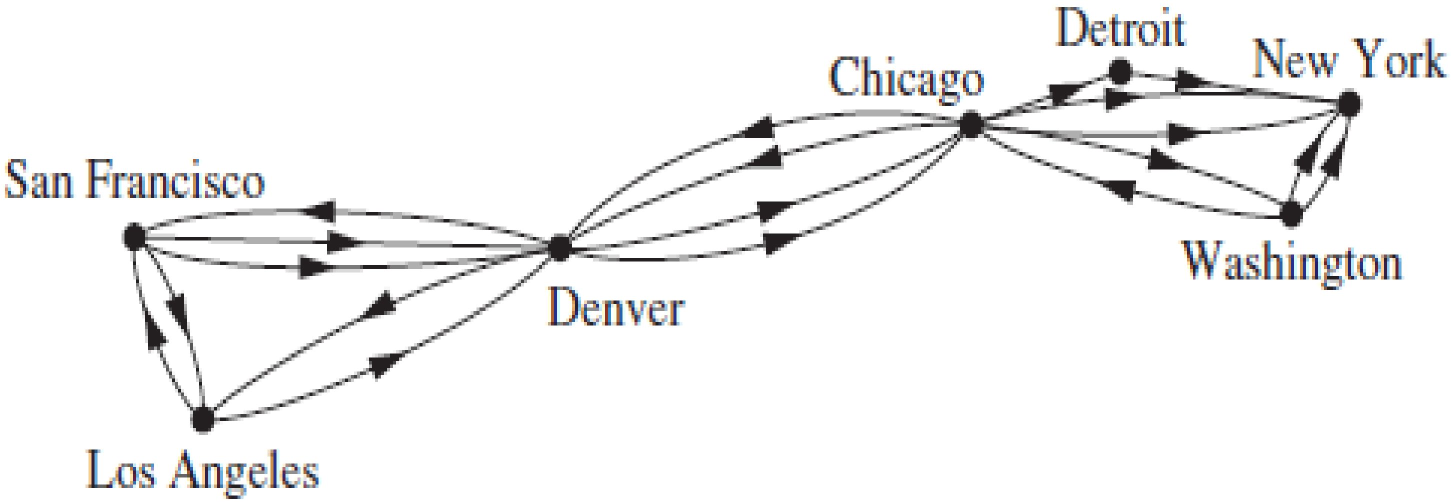
(a) *CampOrg* (b) *BlogNet* (c) *Auth-Conf*

Content

- Graph patterns
- Patterns in static graphs
- Patterns in dynamic graphs
- Patterns in weighted graphs
  - Snapshot Hat Law
  - Weighted exponential law
  - Weighted main eigenvalue
- Cost calculation

Weight change graph

- Consider weighted graphs that change over time
  - For example, the graph depicts the amount of packets transmitted in the network with a review interval of every 30 minutes
  - Let (t) be the total weight up to time t (total packets exchanged in the network).
  - $E(t)$ is the number of distinct edges up to time t
  - n(t) is the number of multi-edges up to time t



Law of Snapshot Hats

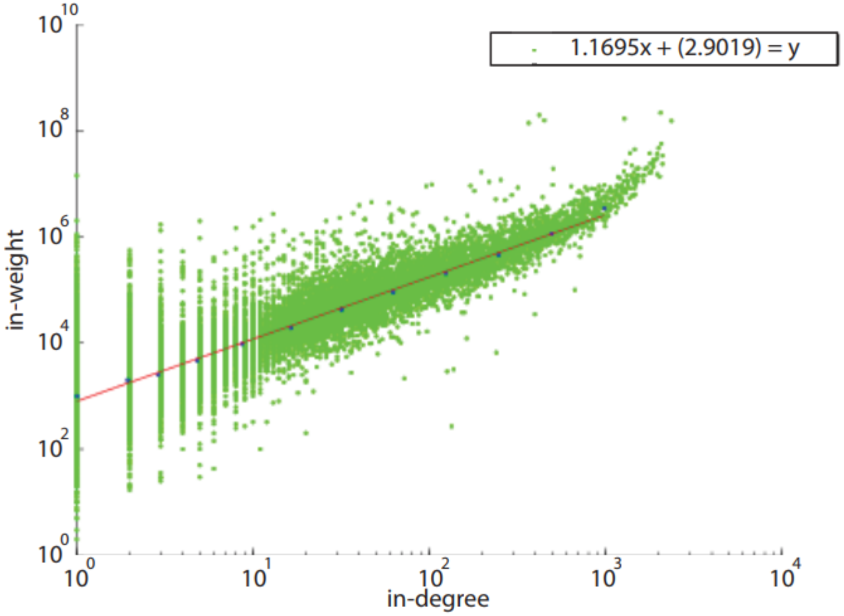
- It is found that at a specified time, the outer (in) degree of a vertex i and the outer (in) weight at that vertex obey the exponential law.

$$outw_i = out_i^{ow}$$

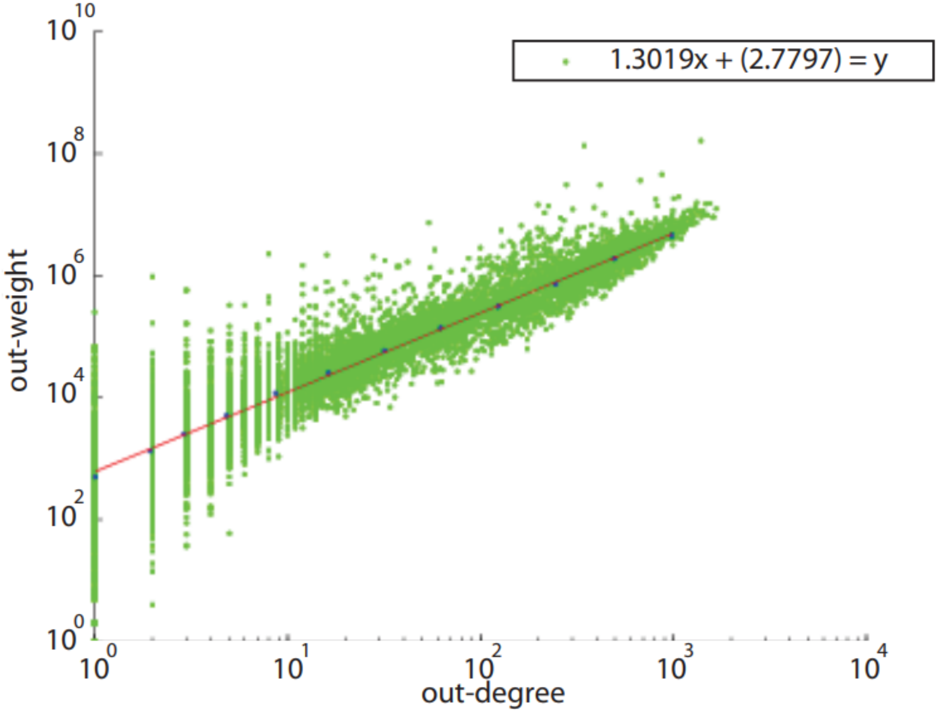
where ow is an out-weight-exponent and is generally constant over time.

- Same for inner degree and inner weight.
- This is called the snapshot power law (SPL)

Law of Snapshot Hats



(a) inD-inW snapshot



(b) outD-outW snapshot

Correlation between weights and tiers in CampOrg data

- The more campaigns the organization supports, the more money it spends
- The more support a candidate receives, the more money it will receive

Law of exponential weights

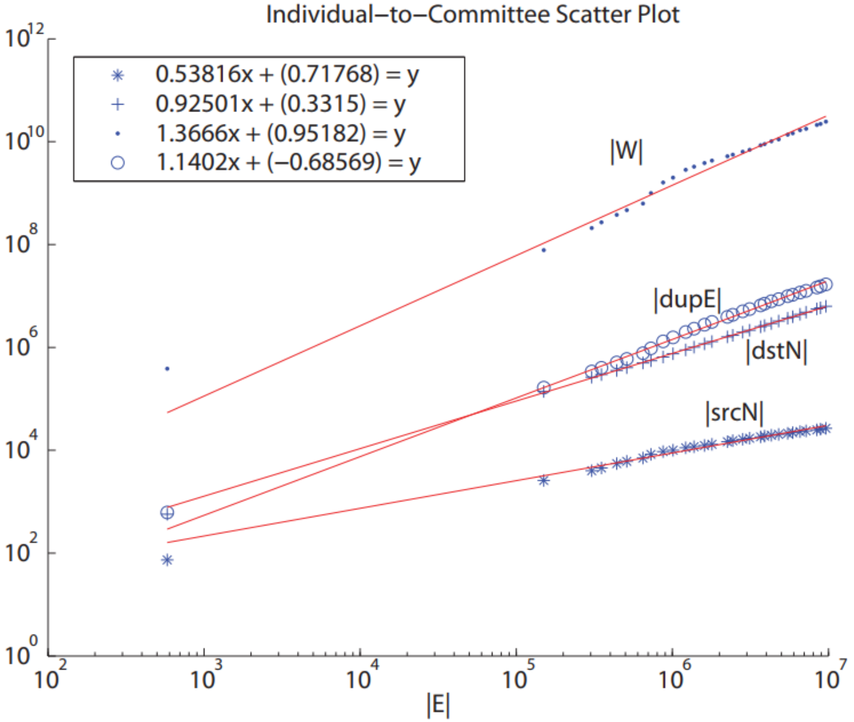
- where E(t) is the total number of distinct edges, W(t) is the total weight, N(t) is the total number of vertices, n(t) is the sum of multiple edges
- It is noticed that between the total number of edges and the total weight of the graph at time t follows the exponential rule:

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

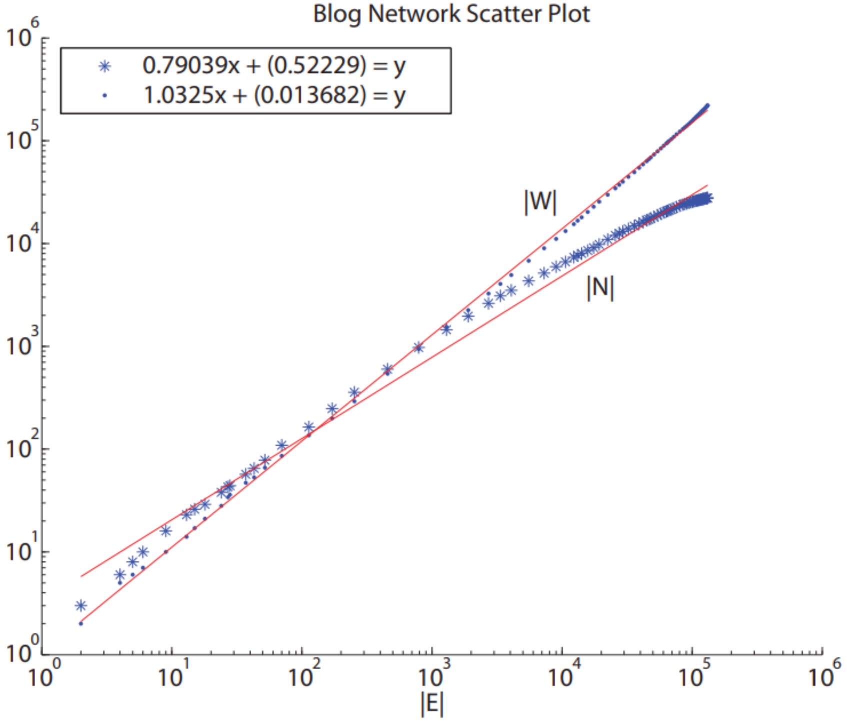
where w is a weighted power, usually between 1.01 and 1.5.

- The same goes for N(t)−E(t), n(t)−E(t).

Law of exponential weights



(a) *CampIndiv* WPLs



(b) *BlogNet* WPLs

Correlation between total weights (total edges, total vertices) and total number of edges in the data. Each data point corresponds to 1 time t.
- Organizations that support multiple campaigns tend to pay more per campaign

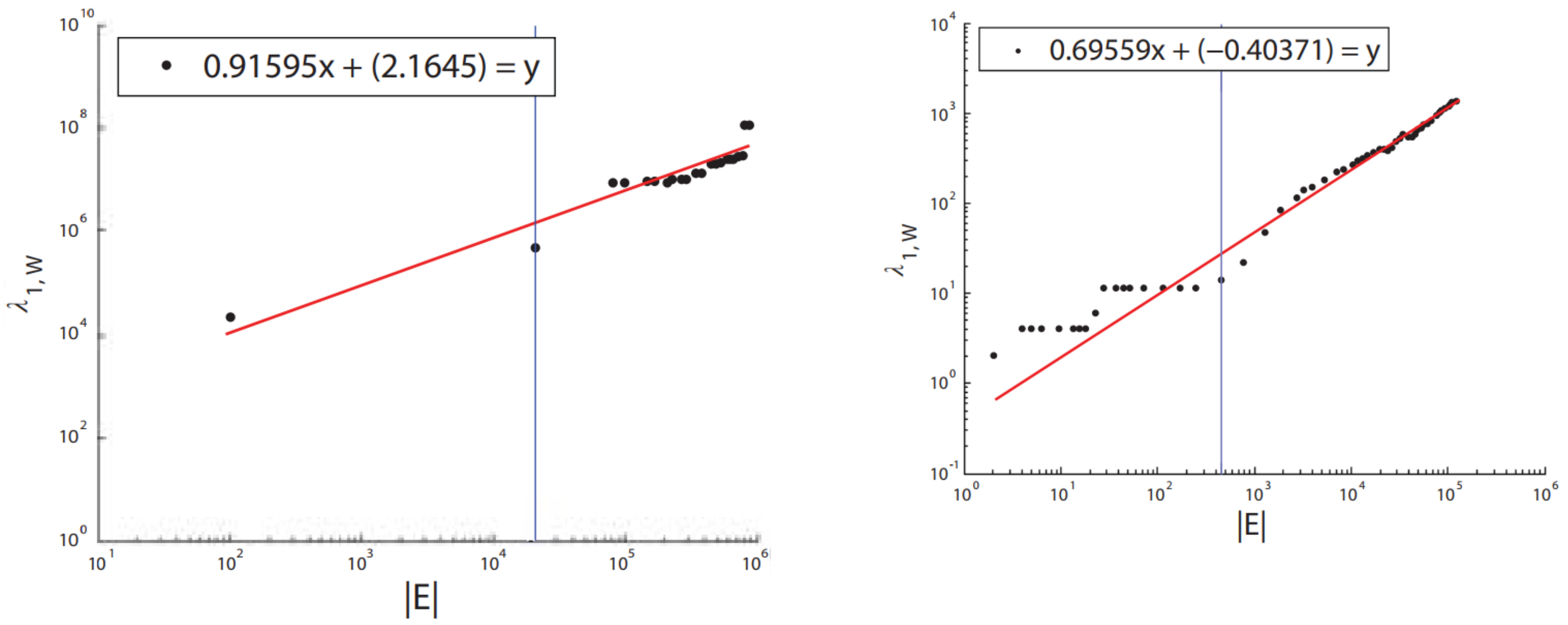
Weighted principal eigenvalues over time

- Set $\lambda_{1,w}$ is the largest (main) eigenvalue of the weighted matrix $\boldsymbol{A}_w$
- The main eigenvalues of the weighted matrix and the number of edges of the graph obey the exponential law:

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

with β powers usually in the range of 0.5 to 1.6 (higher than the exponential law in the unweighted graph).

Weighted principal eigenvalues over time



(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

Content

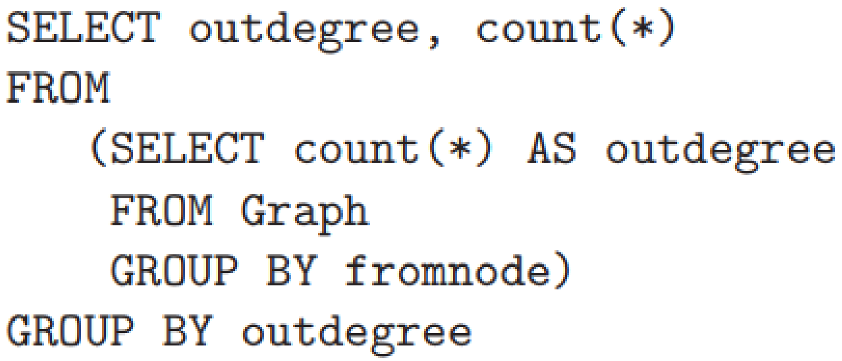
- Graph patterns
- Sample in static graphs
- Templates in dynamic graphs
- Sample in weighted graphs
- **Cost calculation**

Cost calculation

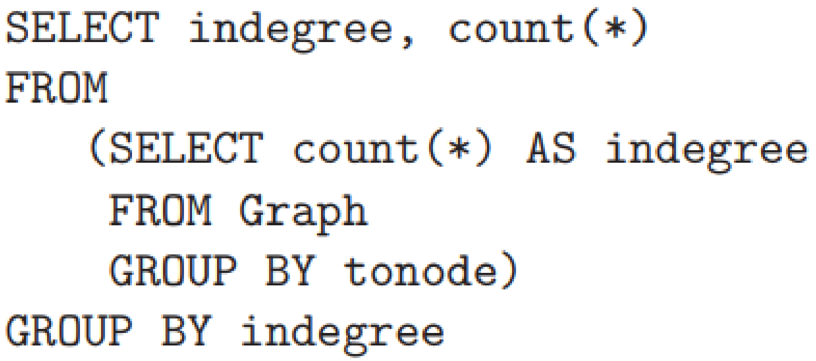
- The process of finding samples can be divided into 3 parts:
  - Create distribution graphs
  - Defining powers for exponential law
  - Check if the sample matches the exponential rule

Create distribution graphs

- Often evaluated as simple operation
- Assuming the graph is represented as a table with a Graph scheme (*fromnode, tonode*), we can use SQL to perform:



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



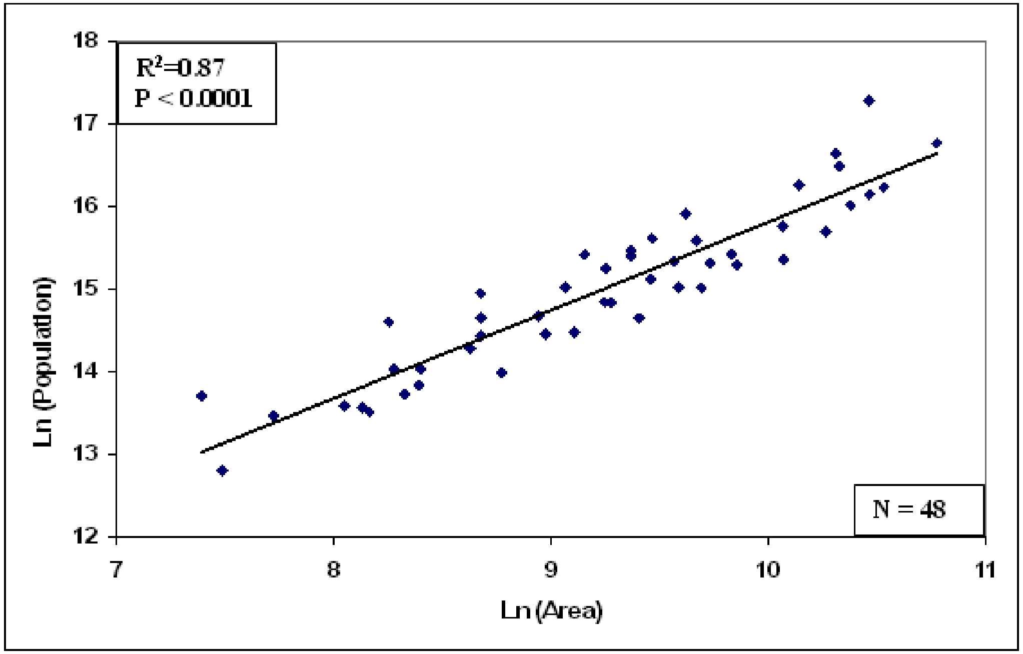
```
SELECT indegree, count(*)
FROM
    (SELECT count(*) AS indegree
     FROM Graph
     GROUP BY tonode)
GROUP BY indegree
```

Determining the exponentiation of the exponential law

- The problem is difficult due to:
  - The exponential rule may occur only in the tail part of the distribution, not on the whole distribution
  - Some prerequisite assumptions may not be satisfactory
  - ...
- Some methods have been proposed but are approximate (it is not clear who is the current "winner").

Determining the exponentiation of the exponential law

- Linear regression on the log-log chart:
  - – Plot data on a log-log scale
  - – Optimize each interval into equal size domains
  - – Find the angle coefficient to match



Determining the exponentiation of the exponential law

- Linear regression on a log-log chart
- Some problems arise:
  - May lead to skewed estimates
  - The hat rule may appear only at the tail part, and at what point to begin consideration it is necessary to determine it manually
  - The right end of the distribution will be very noisy
- However, this is considered the simplest and most commonly used technique.

Determining the exponentiation of the exponential law

- Several other methods are also used:
  - Linear regression after logarithmic interval
  - Regression on cumulative distribution
  - Maximum likelihood estimation (by Goldstein)
  - Statistics Hill (by Hill)
  - Matches only extreme value data (by Feuerverger and Hall)
  - Non-parametric Estimation (by Crovella and Taqqu)

Check pattern for exponential matching

- The correlation coefficient is a relative measure of whether an order distribution fits the exponential law.
- In addition, several methods based on Beirland or Goldstein statistics have been developed.

References

- 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.