EXPLORANDO AS REDES COMPLEXAS - DA TEORIA A APLICAÇÕES

ANA PAULA APPEL
ESTEVAM RAFAEL HRUSCHKA JUNIOR

IBM RESEARCH BRAZIL
SERVICE SYSTEM

DEPARTAMENTO DE COMPUTAÇÃO
UNIVERSIDADE FEDERAL DE SÃO CARLOS

Outline

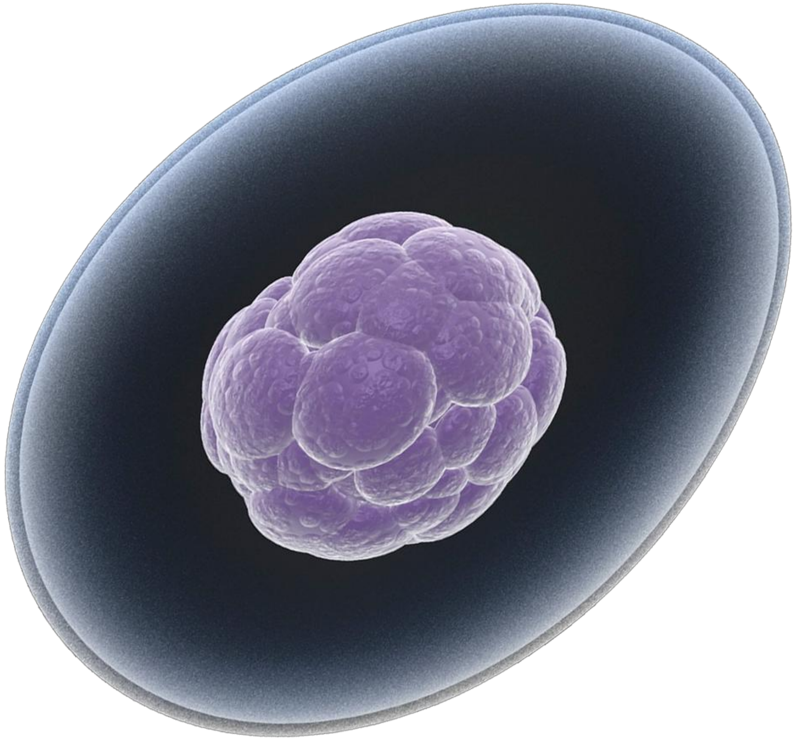
- Part 1: Statistical properties of static and evolving networks.
  - Power law degree distributions found in static networks
  - Small world phenomena and six degrees of separation
  - Densification of time evolving networks
  - Shrinking diameters of growing networks
  - Communities and clusters in networks
- Part 2: Link predictions in complex networks.
  - Link Prediction
    - Link existence
    - Link weight
    - Link type
    - Link cardinality
  - Applications

Outline

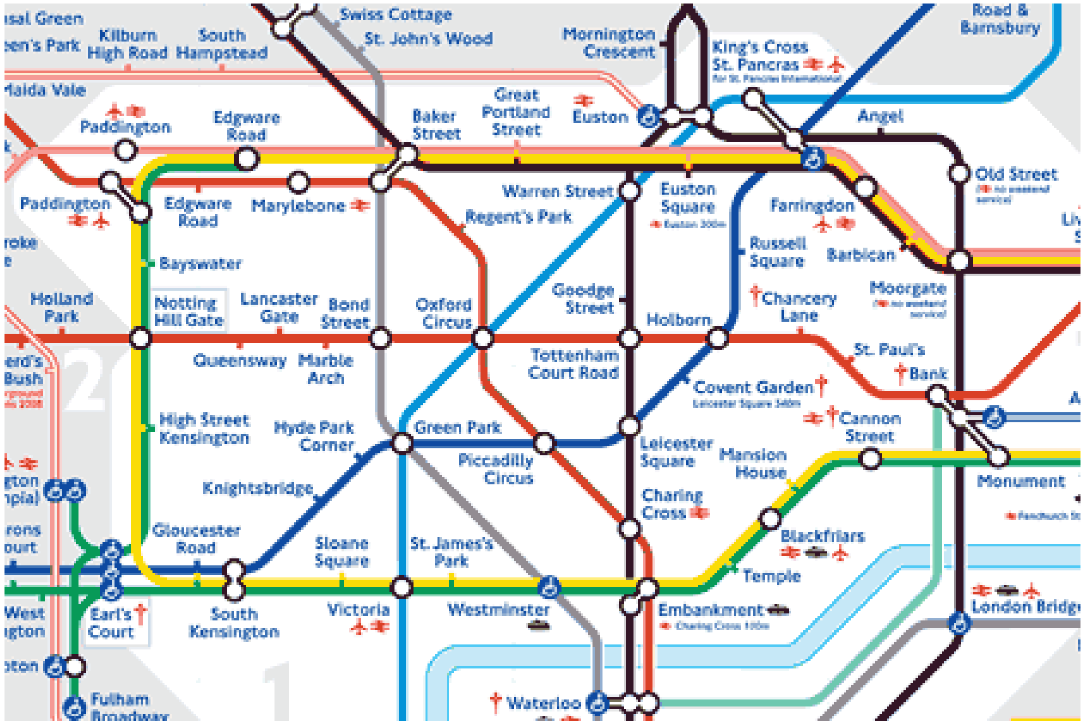
- Part 1: Statistical properties of static and evolving networks.
  - Power law degree distributions found in static networks
  - Small world phenomena and six degrees of separation
  - Densification of time evolving networks
  - Shrinking diameters of growing networks
  - Communities and clusters in networks
- Part 2: Link predictions in complex networks.
  - Link Prediction
    - Link existence
    - Link weight
    - Link type
    - Link cardinality
  - Applications

What do the following things have in common?

World economy

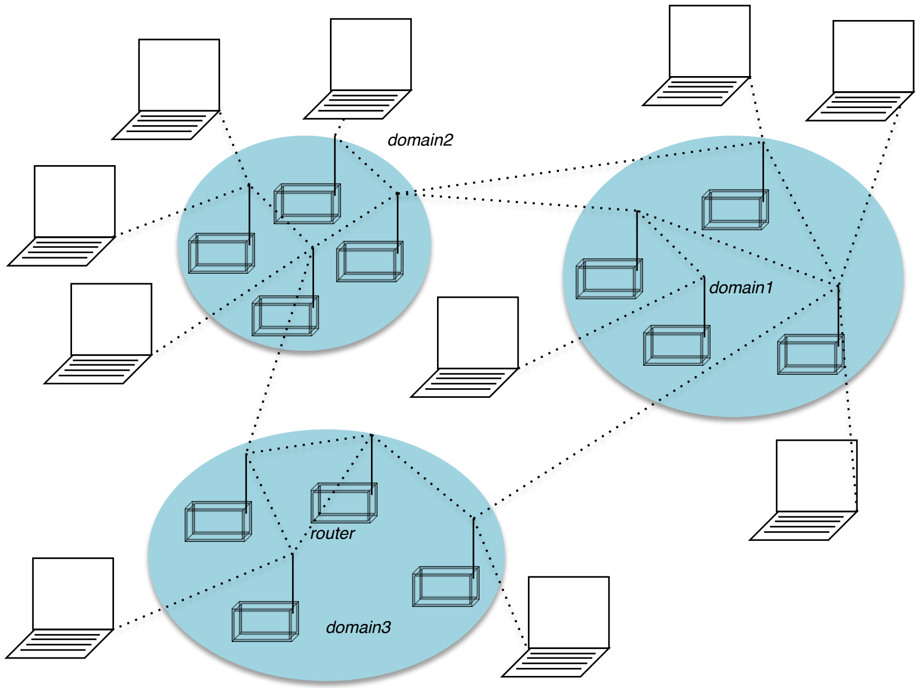


Human cell


Swiss Cottage
St. John's Wood
Kilburn
High Road
South
Hampstead
Mornington
Crescent
King's Cross
St. Pancras
Road &
Barnsbury
Paddington
Edgware
Road
Baker
Street
Great
Portland
Street
Euston
Angel
Old Street
Paddington
Edgware
Road
Marylebone
Warren Street
Regent's Park
Euston
Square
Farringdon
Russell
Square
Barbican
Bayswater
Holland
Park
Notting
Hill Gate
Lancaster
Gate
Bond
Street
Oxford
Circus
Goodge
Street
Holborn
Chancery
Lane
Moorgate
Queensway
Marble
Arch
Tottenham
Court Road
St. Paul's
Bank
Covent Garden
Cannon
Street
High Street
Kensington
Hyde Park
Corner
Green Park
Leicester
Square
Mansion
House
Piccadilly
Circus
Monument
Knightsbridge
Charing
Cross
Gloucester
Road
Sloane
Square
St. James's
Park
Blackfriars
Temple
Earl's
Court
South
Kensington
Victoria
Westminster
Embankment
London Bridge
Fulham
Broadway
Waterloo

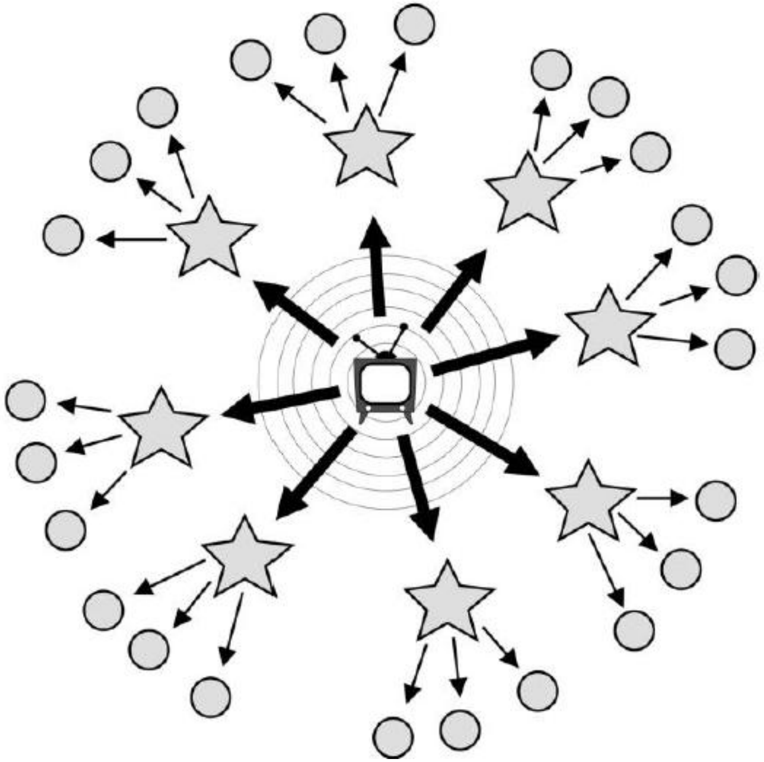
Roads

Brain

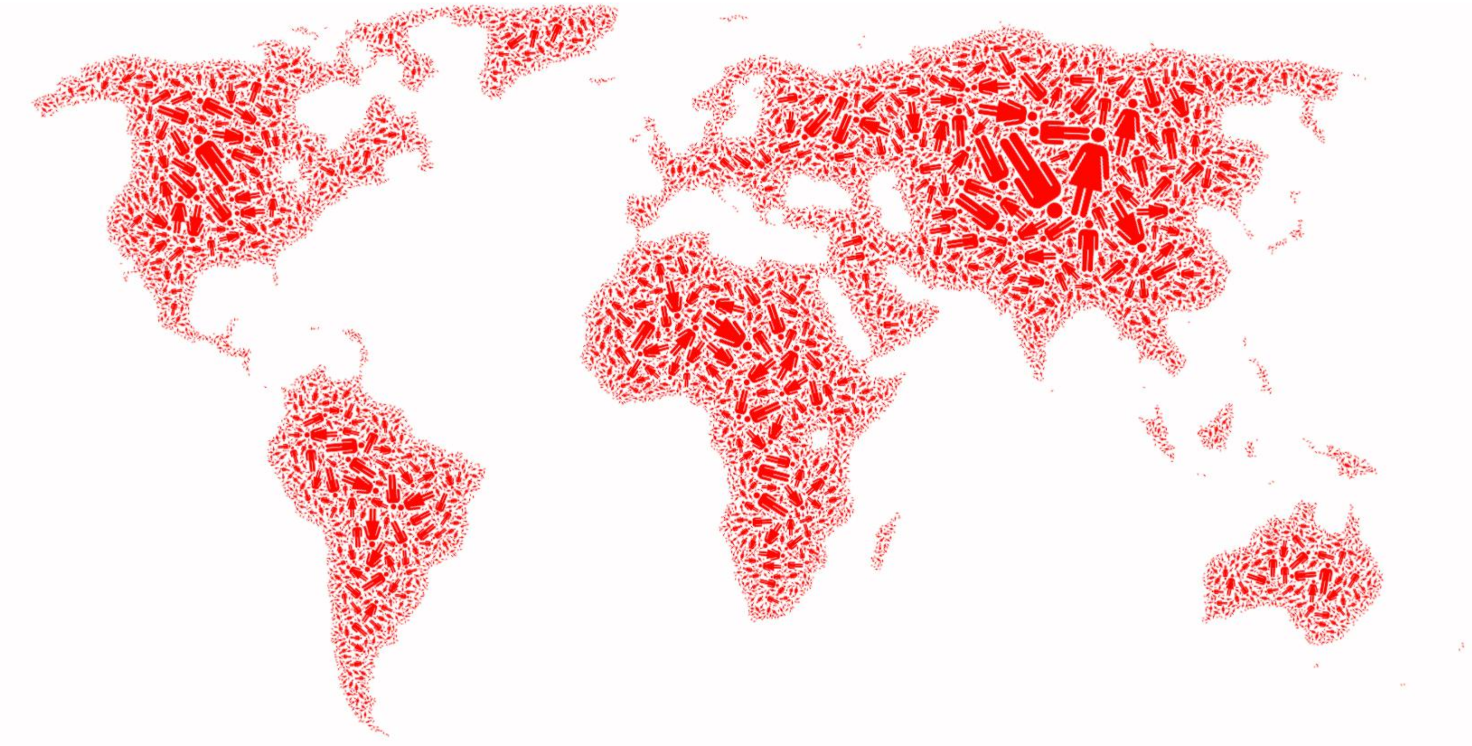


Internet

Friends & Family



Media & Information



Society

The Network

Network

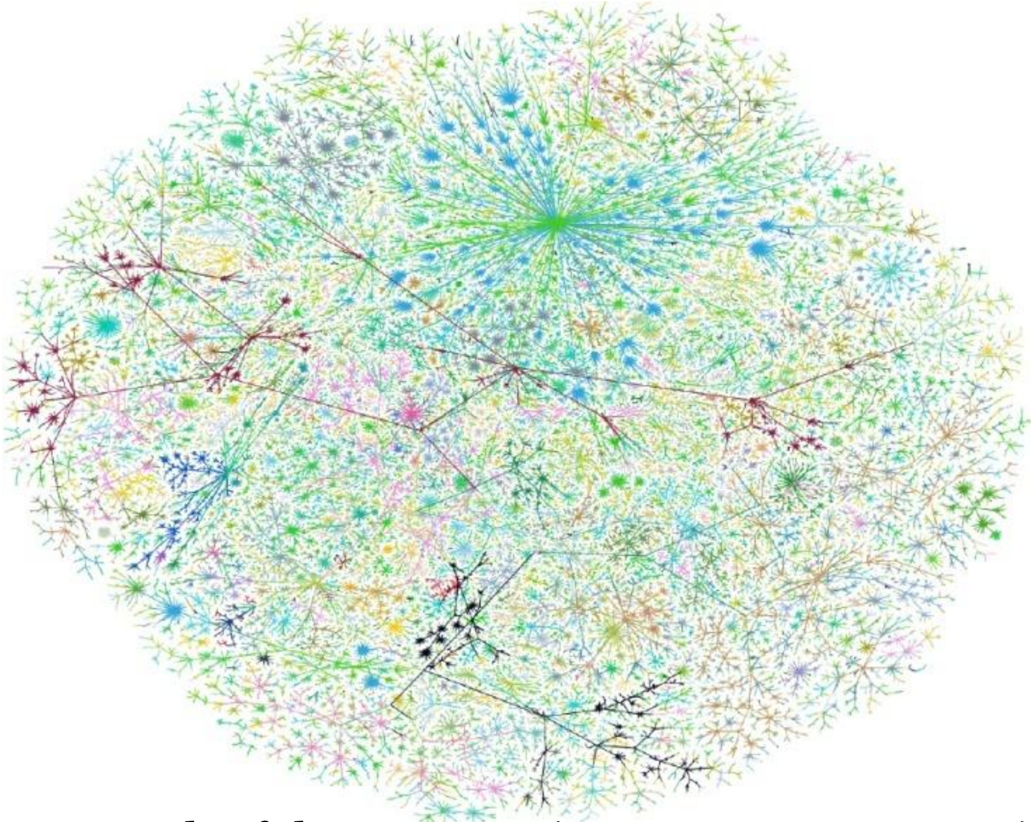
- Behind each such system there is an intricate wiring diagram, a network, that defines the interactions between the components

- We will never understand these systems unless we understand the networks behind it

Networks: Social

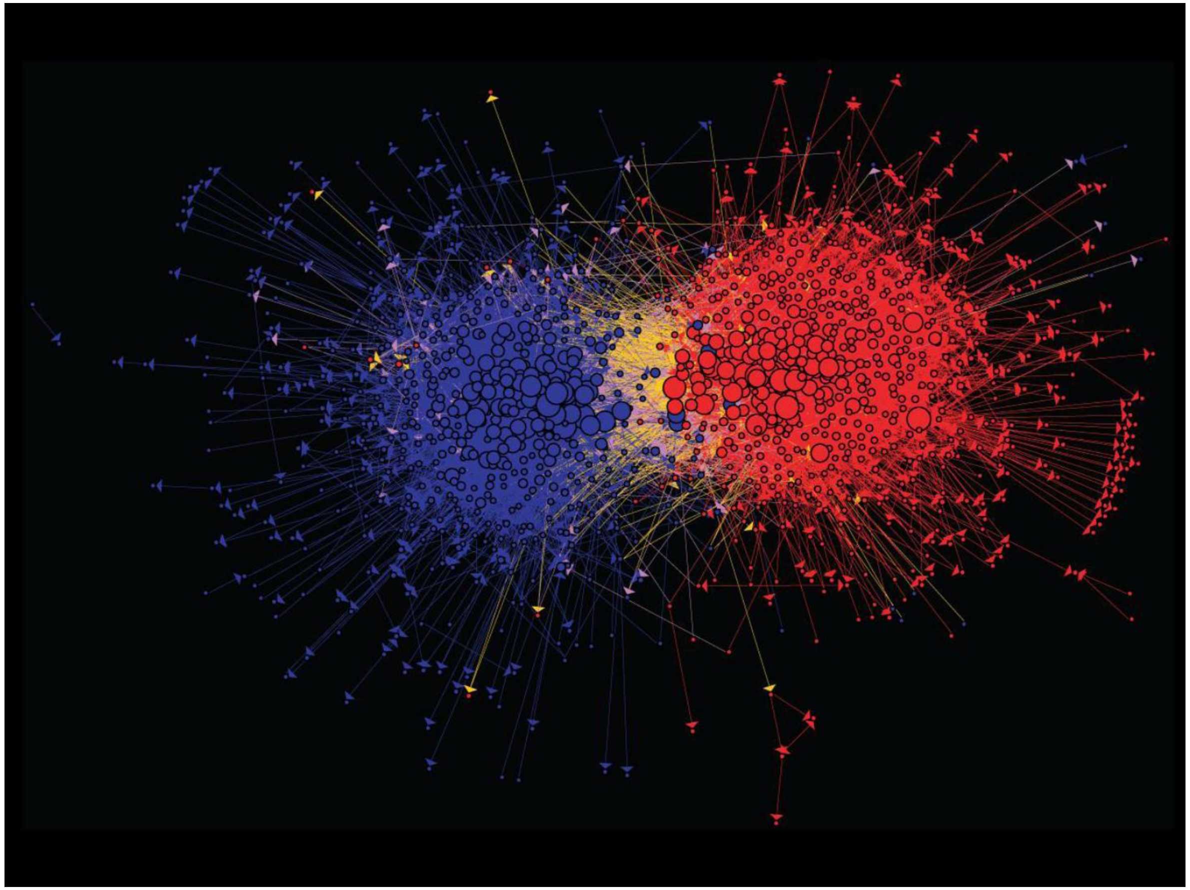
Facebook social graph
4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Networks: Communication



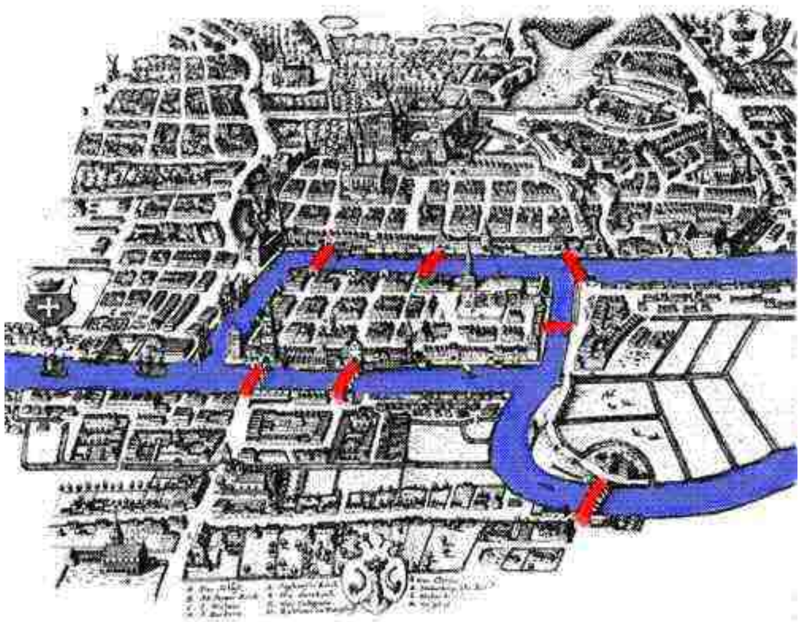
Graph of the Internet (Autonomous Systems)
Power-law degrees [Faloutsos-Faloutsos-Faloutsos, 1999]
Robustness [Doyle-Willinger, 2005]

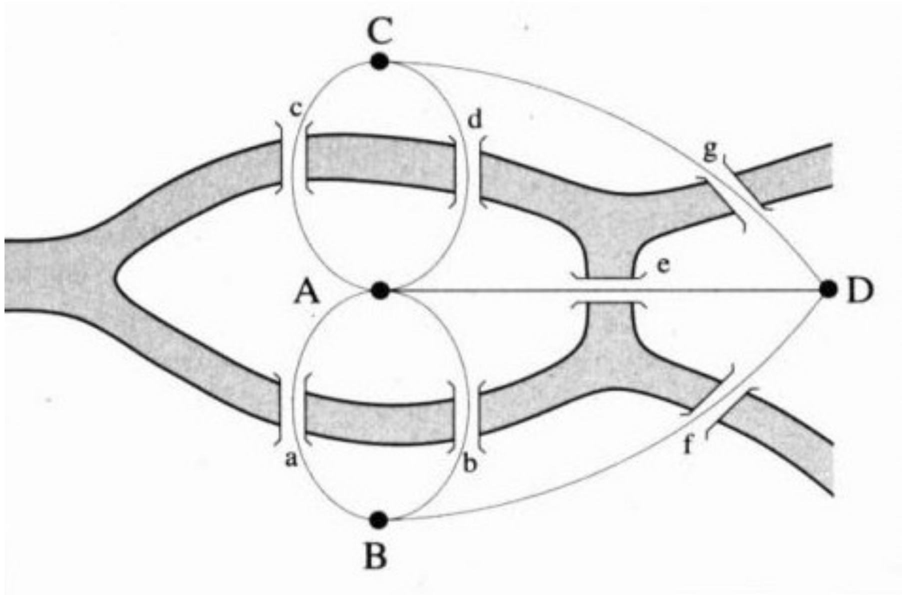
Networks: Media



Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]

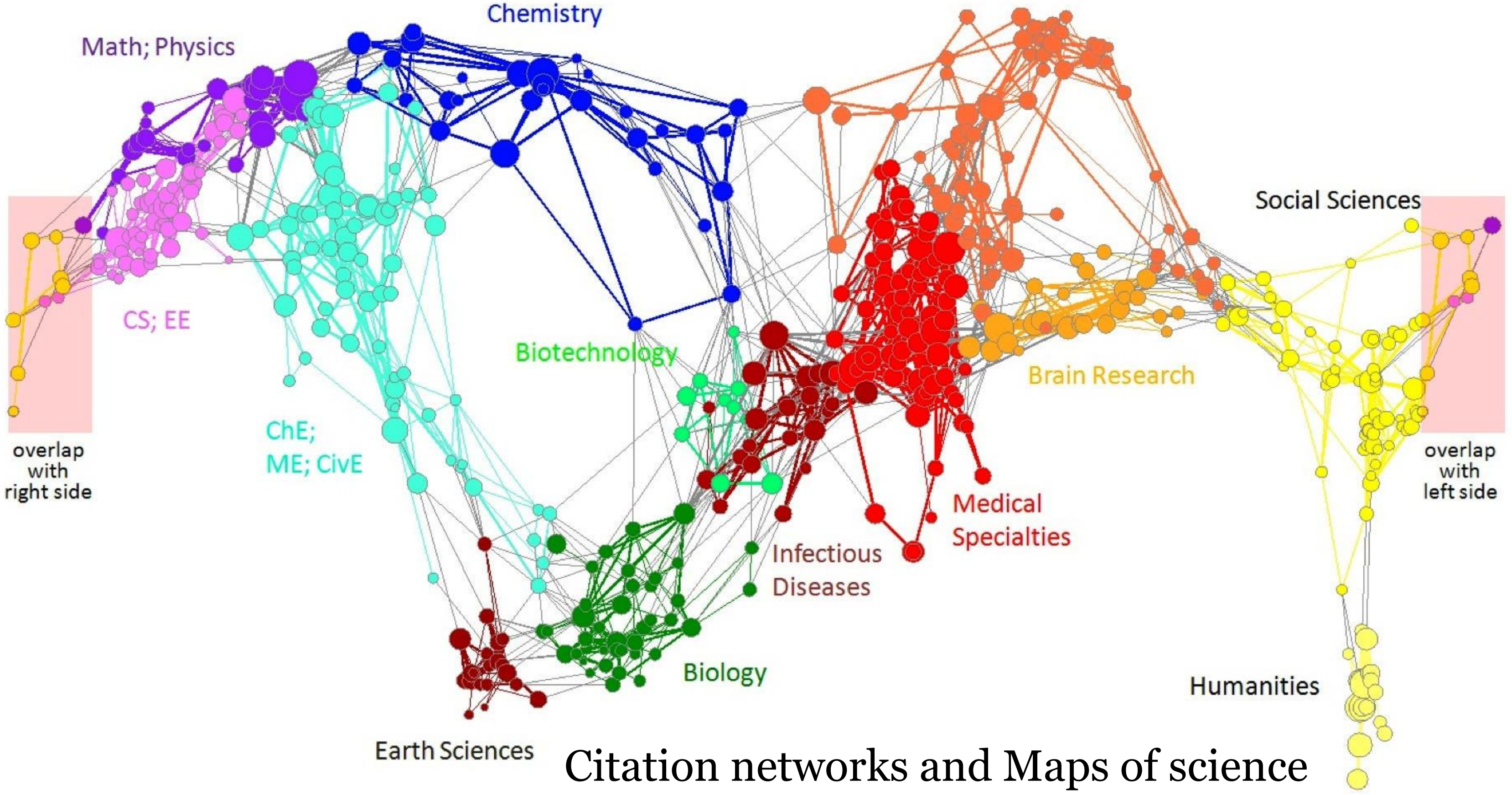
Networks:Technology





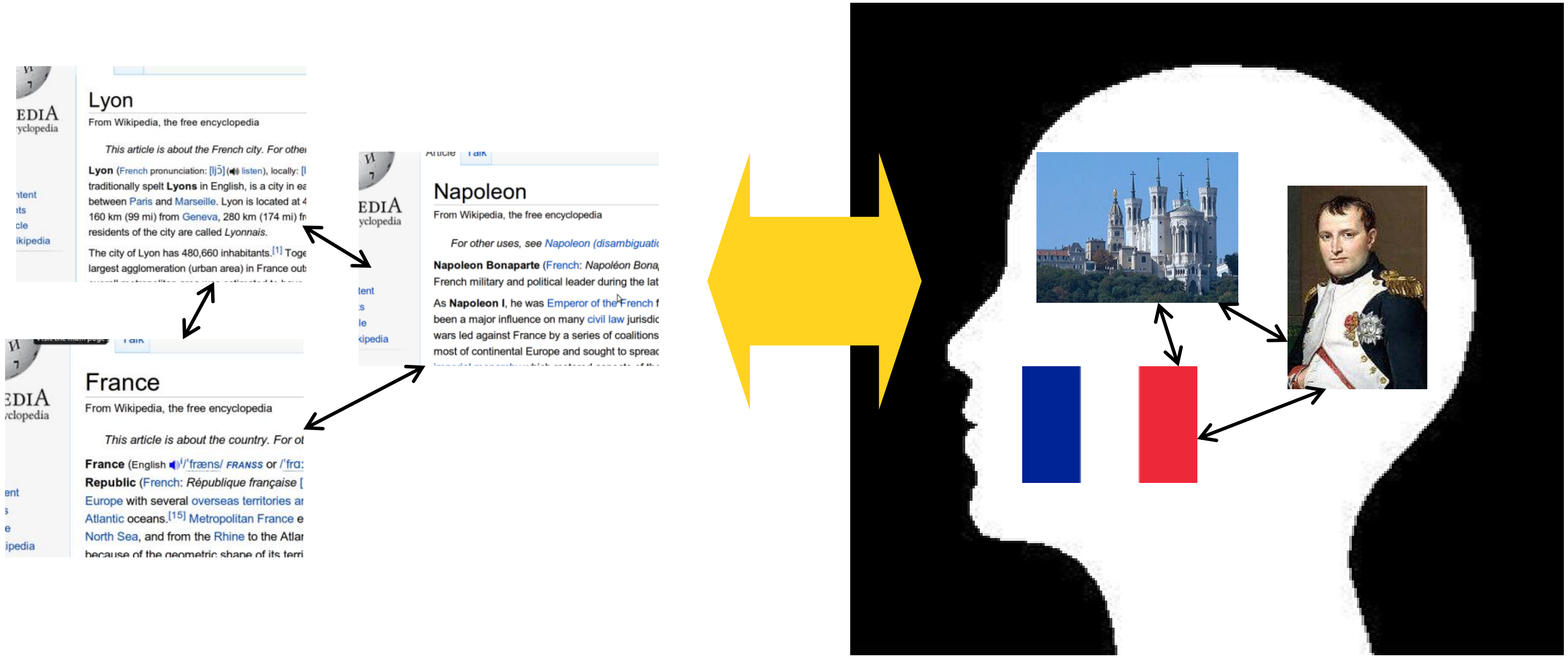
Seven Bridges of Königsberg [Euler, 1735]
Return to the starting point by traveling each link of the graph once and only once.

Networks: Information



Citation networks and Maps of science
[Börner et al., 2012]

Networks: Knowledge

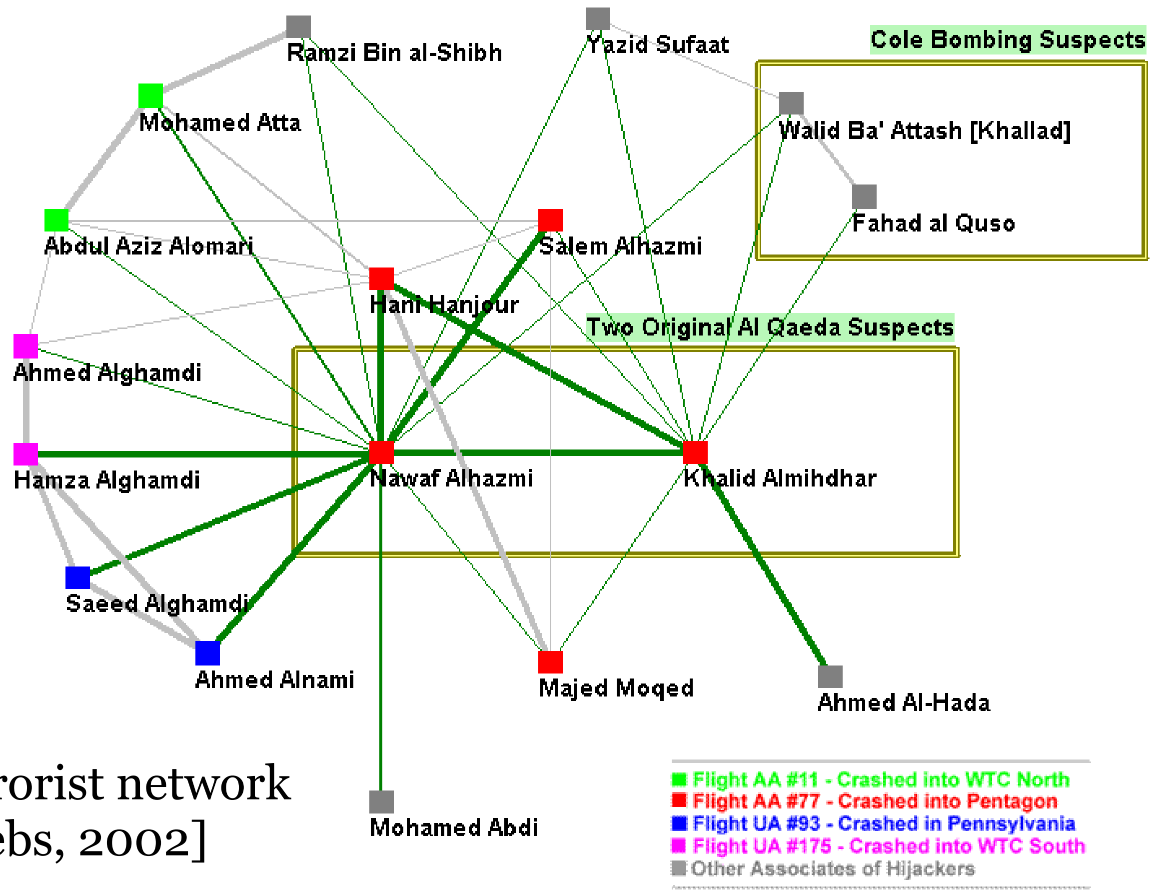


Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

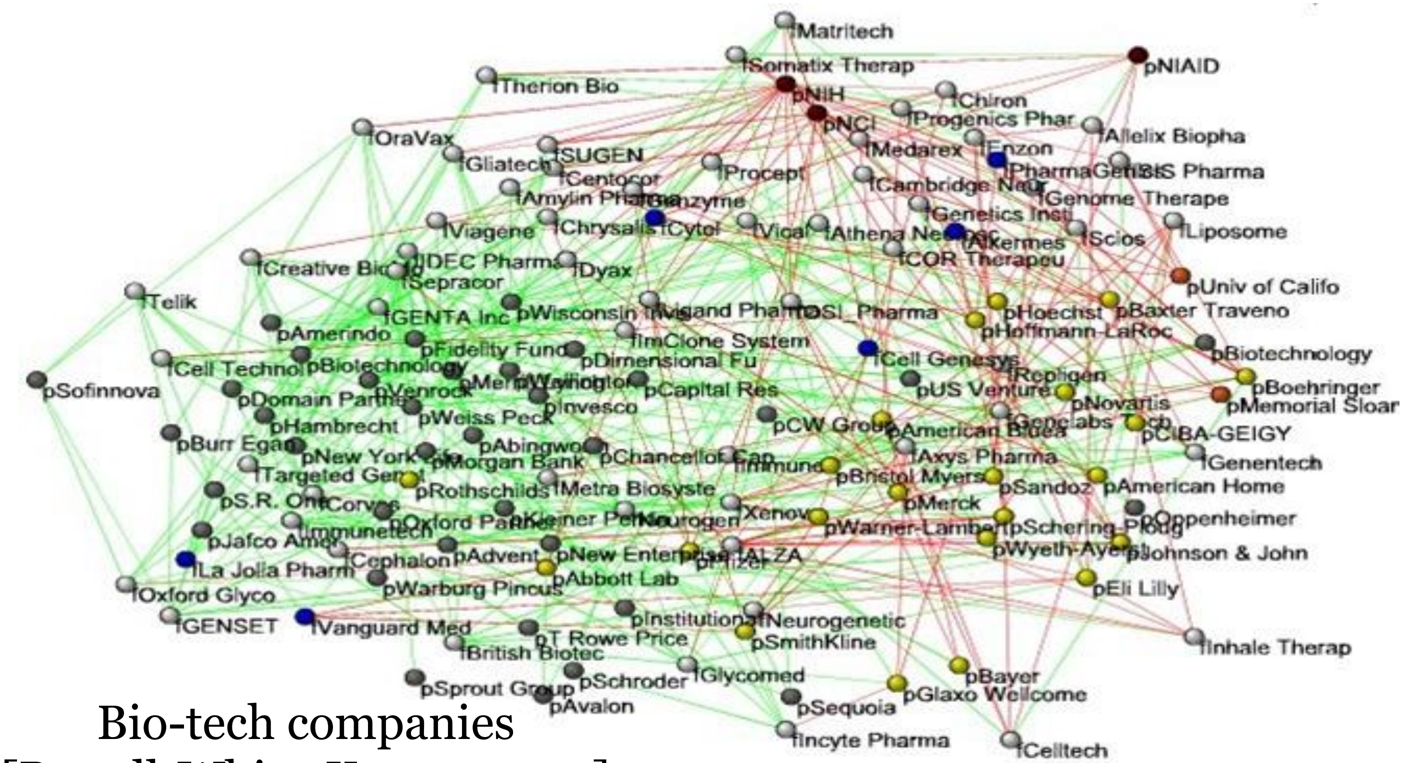
[West-Leskovec, 2012]

Network: Organizations



9/11 terrorist network
[Krebs, 2002]

Networks: Economy



Bio-tech companies
[Powell-White-Koput, 2002]

Nodes:

- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

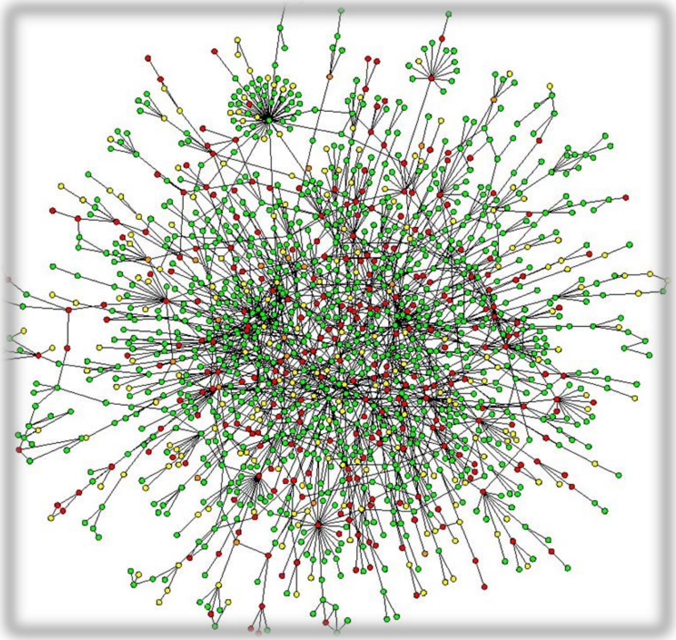
Links:

- Collaborations
- Financial
- R&D

Networks: Brain

Human brain has between 10-100 billion neurons [Sporns, 2011]

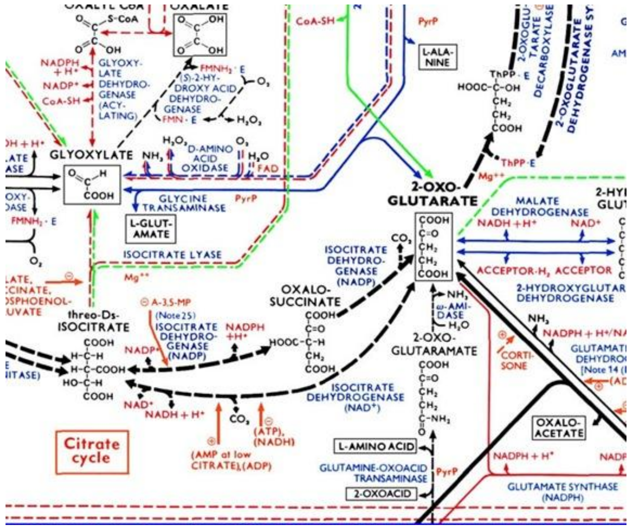
Network: Biology



Protein-Protein Interaction Networks:

Nodes: Proteins

Edges: ‘physical’ interactions



Metabolic networks:

Nodes: Metabolites and enzymes

Edges: Chemical reactions

Reasoning about Networks

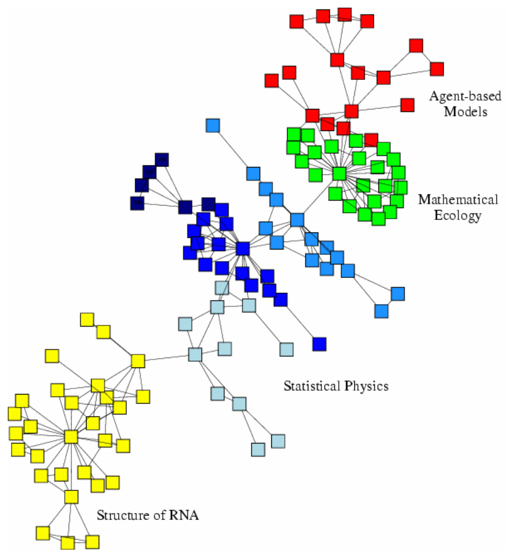
- How do we reason about networks?
  - Empirical: Study network data to find organizational principles
  - Mathematical models: Probabilistic, graph theory
  - Algorithms for analyzing graphs
- What do we hope to achieve from studying networks?
  - Patterns and statistical properties of network data
  - Design principles and models
  - Understand why networks are organized the way they are (Predict behavior of networked systems)

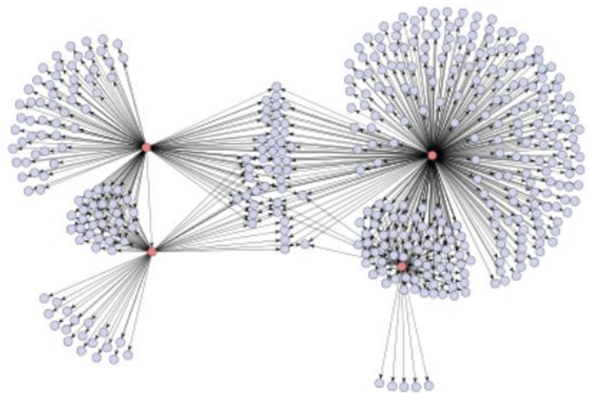
Motivation

- How do large network “look like”?
  - Empirical: statistical tools to quantify structure networks
  - Models: mechanisms that reproduce such properties (models also make “predictions” about other properties)
- 3 parts/goals:
  - Large scale statistical properties of large networks
  - Models that help understand these properties
  - Predict behavior of networked systems based on measured structural properties and local rules governing individual nodes

Motivation

- What do we study in networks?
  - Structure and evolution:
    - What is the structure of a network?
    - Why and how did it became to have such structure?
- Processes and dynamics:
  - Networks provide “skeleton” for spreading of information, behavior, diseases
    - How do information and diseases spread?





Why Networks? Why Now?

- Why is the role of networks expanding?
  - Data availability
    - Rise of Mobile, Web 2.0 and Social media
  - Universality
    - Networks from science, nature, and technology are more similar than one would expect
  - Shared vocabulary between fields
    - Computer Science, Social science, Physics, Economics, Statistics, Biology
  - Impact!
    - Social networking, Social media, Drug design

Networks: Size Matters

- Network data: Orders of magnitude
  - 436-node network of email exchange at a corporate research lab [Adamic-Adar, SocNets ‘03]
  - 43,553-node network of email exchange at an university [Kossinets-Watts, Science ‘06]
  - 4.4-million-node network of declared friendships on a blogging community [Liben-Nowell et al., PNAS ‘05]
  - 240-million-node network of communication on Microsoft Messenger [Leskovec-Horvitz, WWW ’08]
  - 800-million-node Facebook network [Backstrom et al. ‘11]

Networks Really Matter

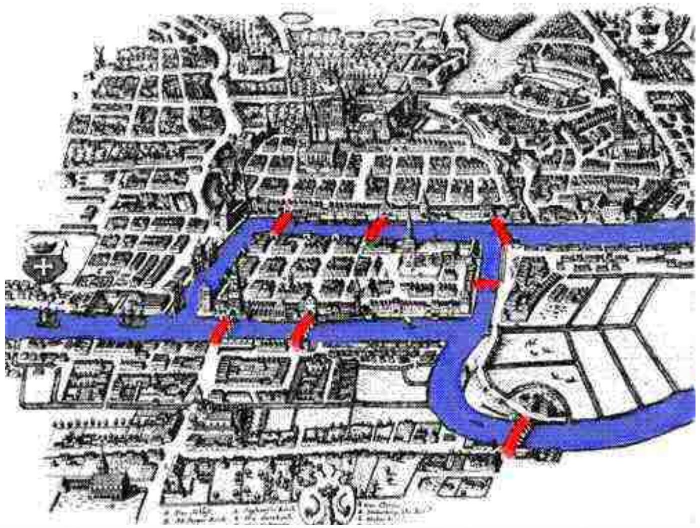
- If you were to understand the spread of diseases, can you do it without social networks?
- If you were to understand the WWW structure and information, hopeless without invoking the Web's topology.
- If you want to understand dissemination of news or evolution of science, it is hopeless without considering the information networks

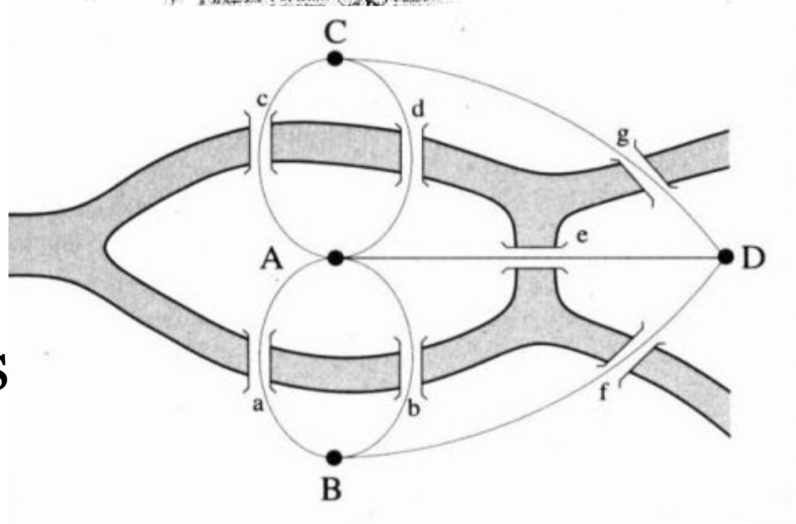
Networks – Social and Technological

- Social network analysis: sociologists and computer scientists – influence goes both ways
  - Large-scale network data in "traditional" sociological domains
    - Friendship and informal contacts among people
    - Collaboration/influence in companies, organizations, professional communities, political movements, markets, ...
  - Emerge of rich social structure in computing applications
    - Content creation, on-line communication, blogging, social networks, social media, electronic markets, ...
    - People seeking information from other people vs. more formal channels: MySpace, del.icio.us, Flickr, LinkedIn, Yahoo Answers, Facebook, ...

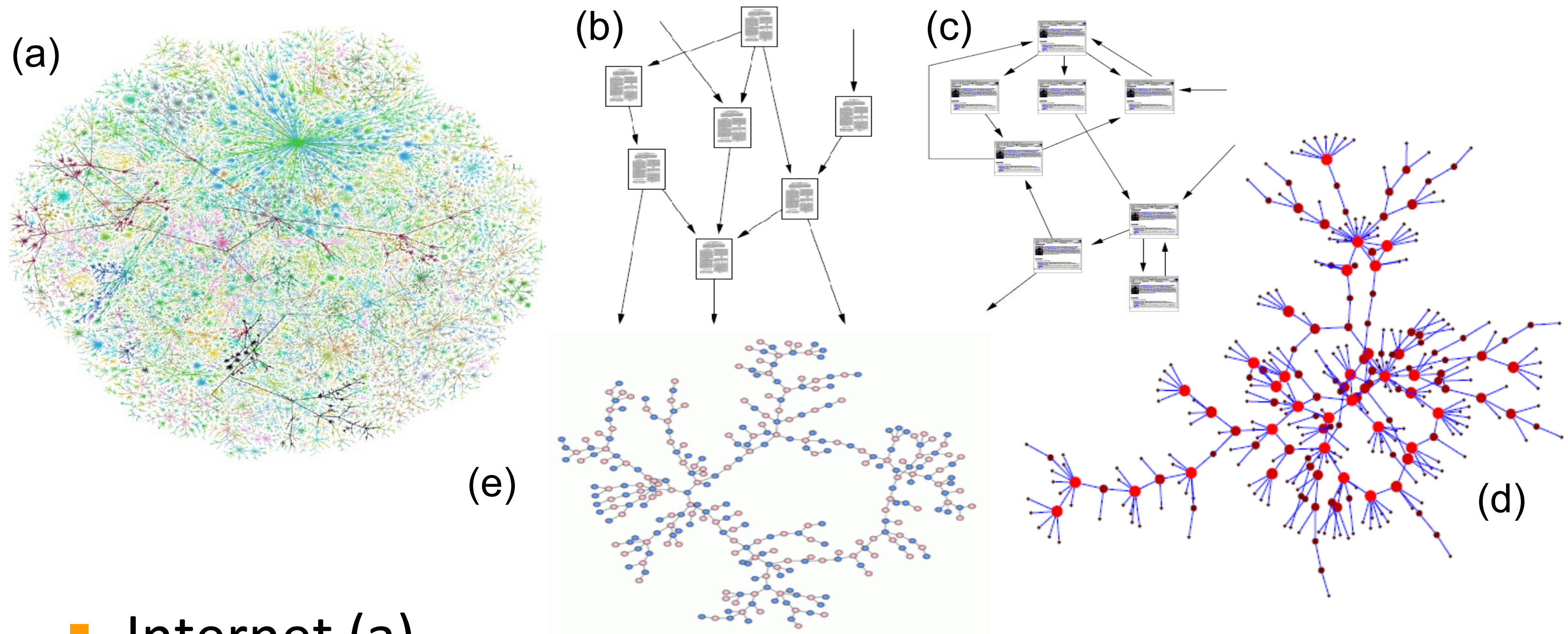
Como tudo começou...

- Leonhard Euler, 1875
- As pontes de Königsberg:
  - “Pode alguém caminhar pelas 7 pontes sem nunca cruzar a mesma ponte duas vezes?”
- A resposta: não é possível, pois o grafo precisa ter no máximo dois nós com grau ímpar;
- Surgimento Teoria dos Grafos





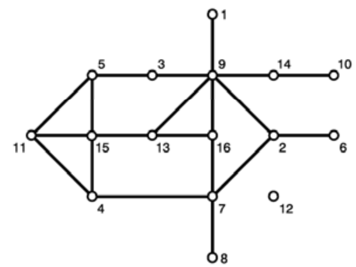
Examples of Networks



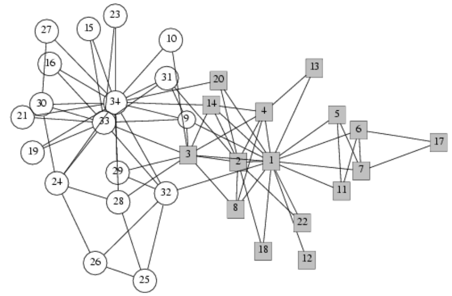
- Internet (a)
- Citation network (b)
- World Wide Web (c)
- Sexual network (d)
- Dating network(e)

Networks of the Real-world (1)

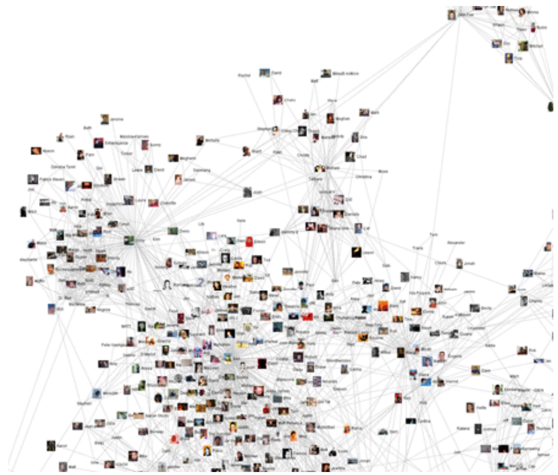
- Information networks:
  - World Wide Web: hyperlinks
  - Citation networks
  - Blog networks
- Social networks: people + interactions
  - Organizational networks
  - Communication networks
  - Collaboration networks
  - Sexual networks
- Technological networks:
  - Power grid
  - Airline, road, river networks
  - Telephone networks
  - Internet
  - Autonomous systems



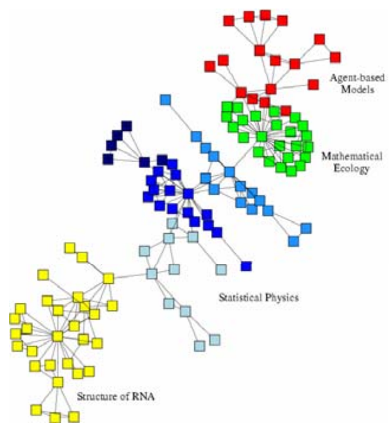
Florence families



Karate club network



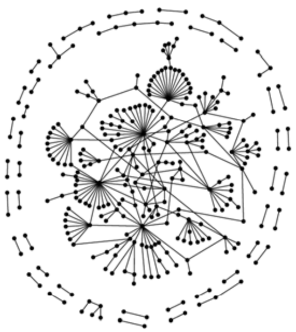
Friendship network



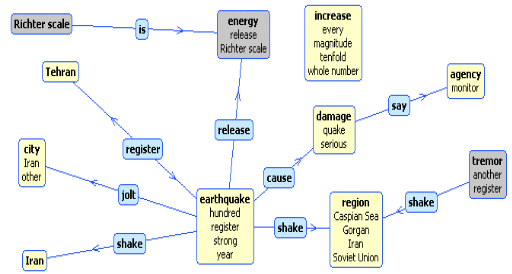
Collaboration network

Networks of the Real-world (2)

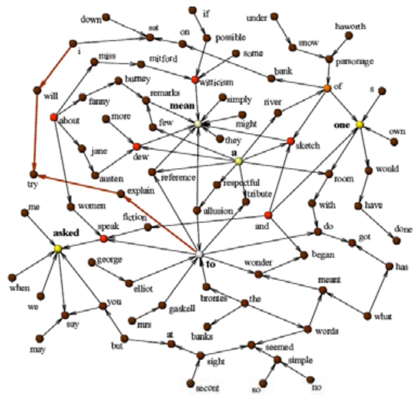
- Biological networks
  - metabolic networks
  - food web
  - neural networks
  - gene regulatory networks
- Language networks
  - Semantic networks
- Software networks



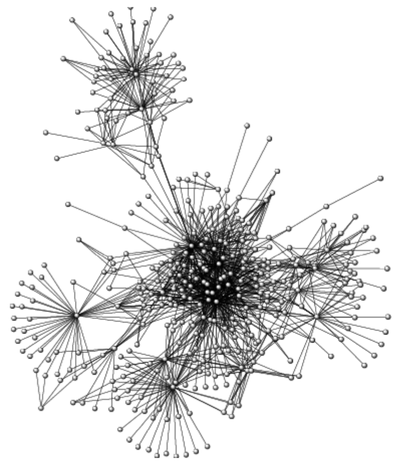
Yeast protein interactions



Semantic network



Language network



Software network

Networks as Phenomena

The emergence of ‘cyberspace’ and the World Wide Web is like the discovery of a new continent.

- Jim Gray, 1998 Turing Award address

- Complex networks as phenomena, not just designed artifacts
- What are the common patterns that emerge?

Models and Laws of Networks

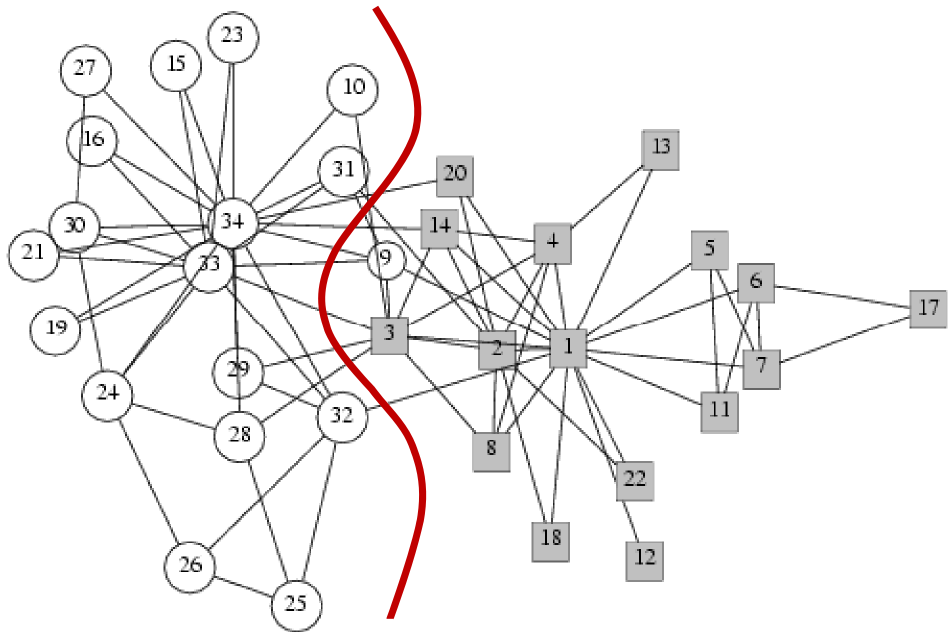
We want Kepler's Laws of Motion for the Web.

- Mike Steuerwalt, NSF KDI workshop, 1998

- Need statistical methods and tools to quantify large networks
- What do we hope to achieve from models of networks?
  - Patterns and statistical properties of network data
  - Design principles and models
  - Understand why networks are organized the way they are (predict behavior of networked systems)

Mining Social Network Data

- Mining social networks has a long history in social sciences:
  - Wayne Zachary's PhD work (1970-72): observe social ties and rivalries in a university karate club
  - During his observation, conflicts led the group to split
  - Split could be explained by a minimum cut in the social network



Networks: Rich Data

- Traditional obstacle:
- Can only choose 2 of 3:
  - Large-scale
  - Realistic
  - Completely mapped
- Now: large on-line systems leave detailed records of social activity
  - On-line communities: MyScace, Facebook, LiveJournal
  - Email, blogging, electronic markets, instant messaging
  - On-line publications repositories, arXiv, MedLine

Scale Matters

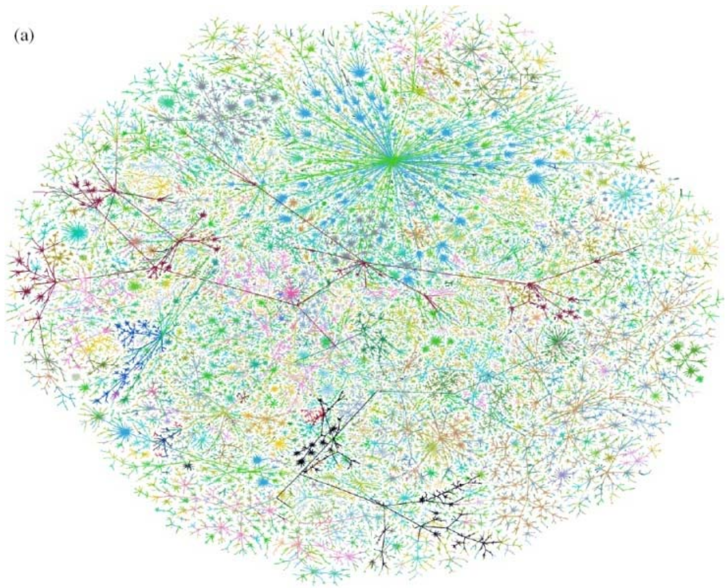
- How does massive network data compare to small-scale studies?
- Massive network datasets give you both more and less:
  - More: can observe global phenomena that are genuine, but literally invisible at smaller scales
  - Less: don’t really know what any node or link means. Easy to measure things, hard to pose right questions
  - Goal: Find the point where the lines of research converge

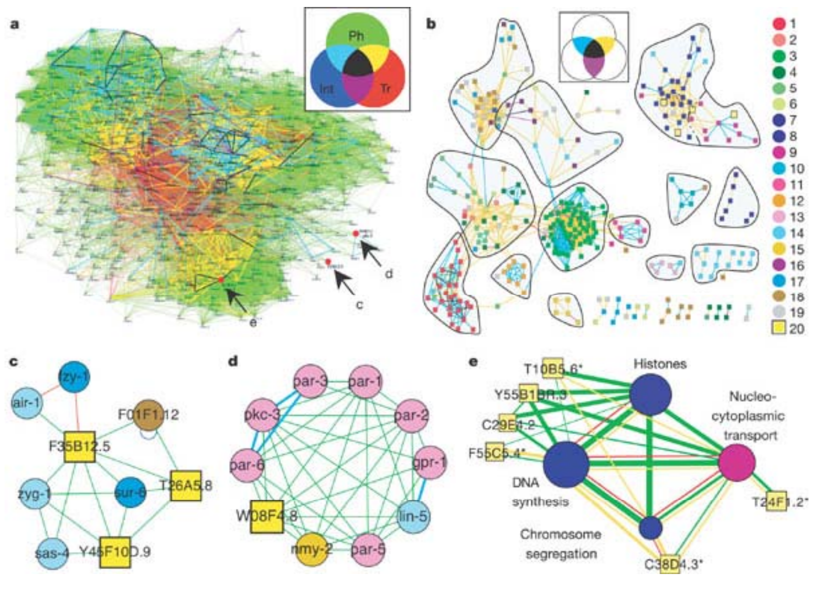
Structure vs. Process

- What have we learned about large networks?
- We know about the structure: Many recurring patterns
  - Scale-free, small-world, locally clustered, bow-tie, hubs and authorities, communities, bipartite cores, network motifs, highly optimized tolerance
- We know about the processes and dynamics
  - Cascades, epidemic threshold, viral marketing, virus propagation, threshold model

Structure of Networks

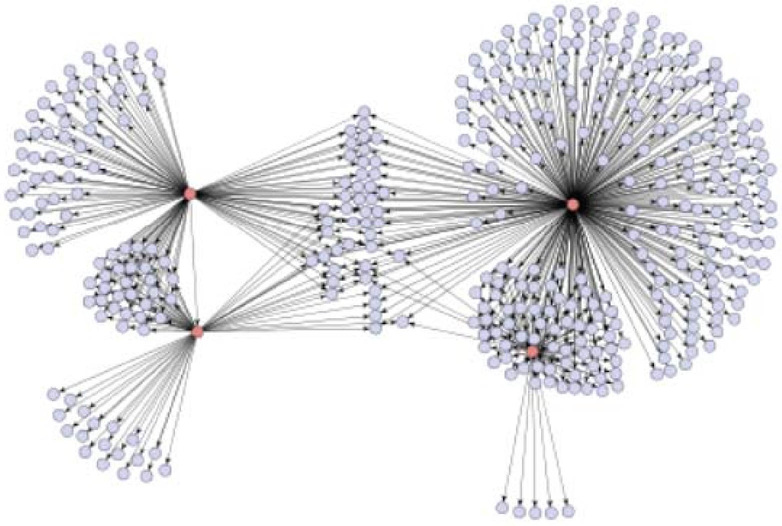
- What is the structure of a large network?
- Why and how did it became to have such structure?





Diffusion in Networks

- One of the networks is a spread of a disease, the other one is product recommendations
- Which is which? ☺



Traditional approach

- Sociologists were first to study networks:
  - Study of patterns of connections between people to understand functioning of the society
  - People are nodes, interactions are edges
  - Questionnaires are used to collect link data (hard to obtain, inaccurate, subjective)
  - Typical questions: Centrality and connectivity
- Limited to small graphs (~100 nodes) and properties of individual nodes and edges

Motivation: New approach (1)

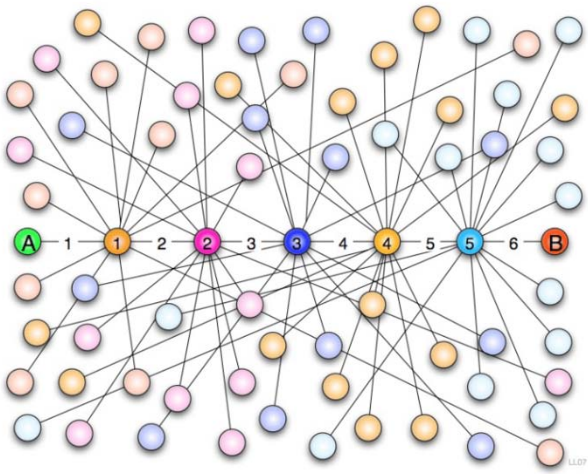
- **Large** networks (e.g., web, internet, on-line social networks) with millions of nodes
- Many traditional questions not useful anymore:
  - Traditional: What happens if a node u is removed?
  - Now: What percentage of nodes needs to be removed to affect network connectivity?
- Focus moves from a single node to study of **statistical** properties of the network as a whole

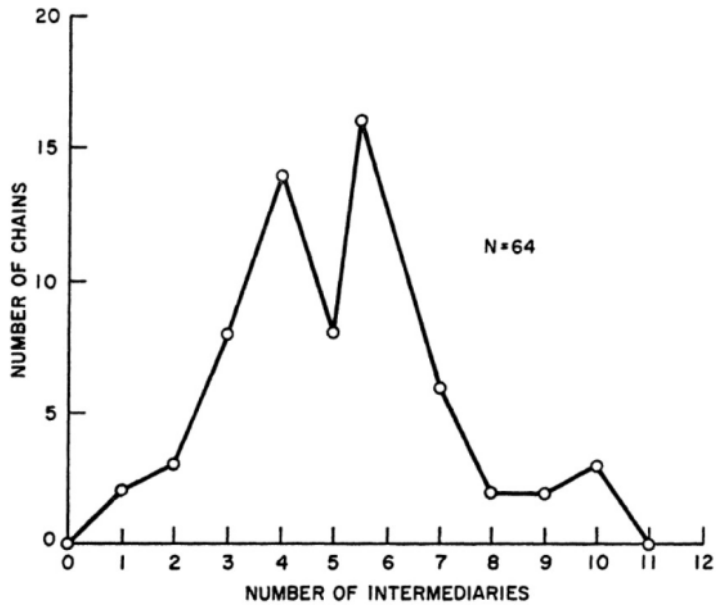
Motivation: New approach (2)

- How the network "looks like" even if I can't look at it?
- Need statistical methods and tools to quantify large networks
- 3 parts/goals:
  - Statistical properties of large networks
  - Models that help understand these properties
  - Predict behavior of networked systems based on measured structural properties and local rules governing individual nodes

Small-world effect (1)

- Six degrees of separation [Milgram 60s]
  - Random people in Nebraska were asked to send letters to stock brokers in Boston
  - Letters can only be passed to first-name acquaintances
  - Only 25% letters reached the goal
  - But they reached it in about **6 steps**



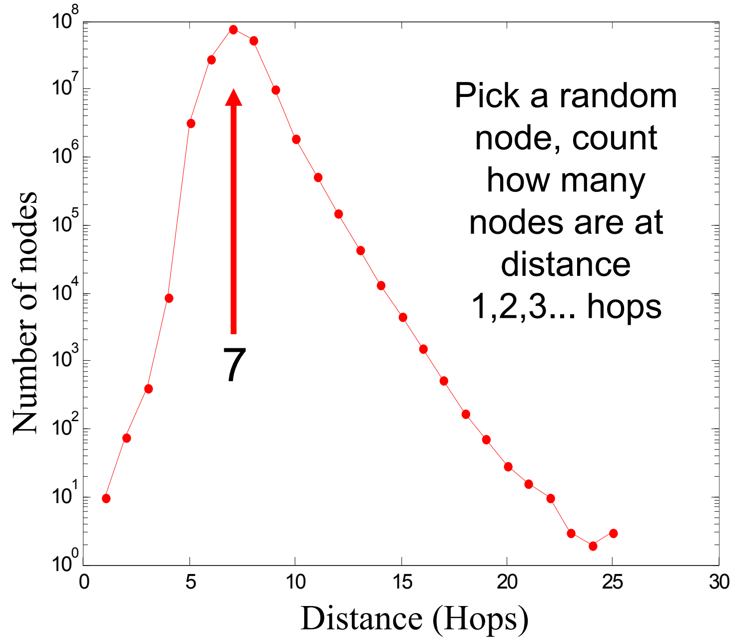


Small-world effect (2)

- Microsoft Messenger network
  - 180 million people
  - 1.3 billion edges
  - Edge if two people exchanged at least one message in one month period

[Leskovec&Horvitz,07]

Average path length is **6.6**
90% of nodes is reachable <8 steps



Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

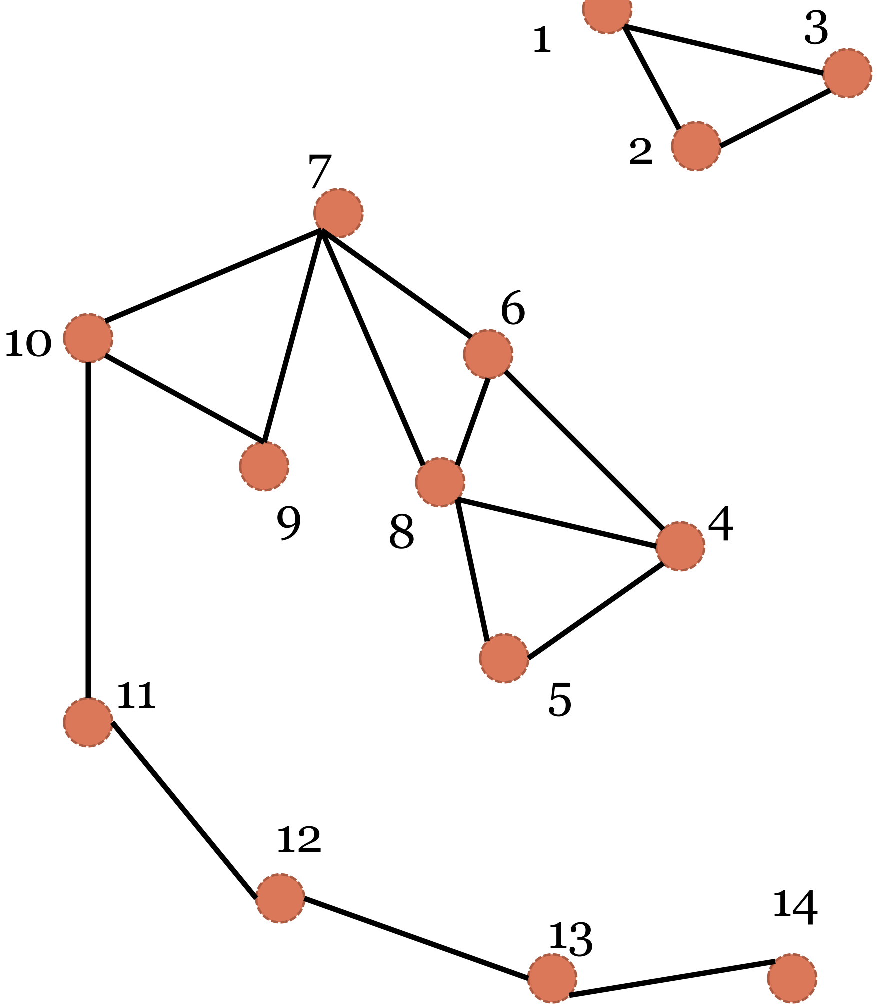
Measuring diameter

- **Measuring path lengths:**
  - Diameter (longest shortest path): *max dij*
  - Effective diameter: distance at which 90% of all connected pairs of nodes can be reached
  - Mean geodesic (shortest) distance *l*

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \quad \text{or} \quad \ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

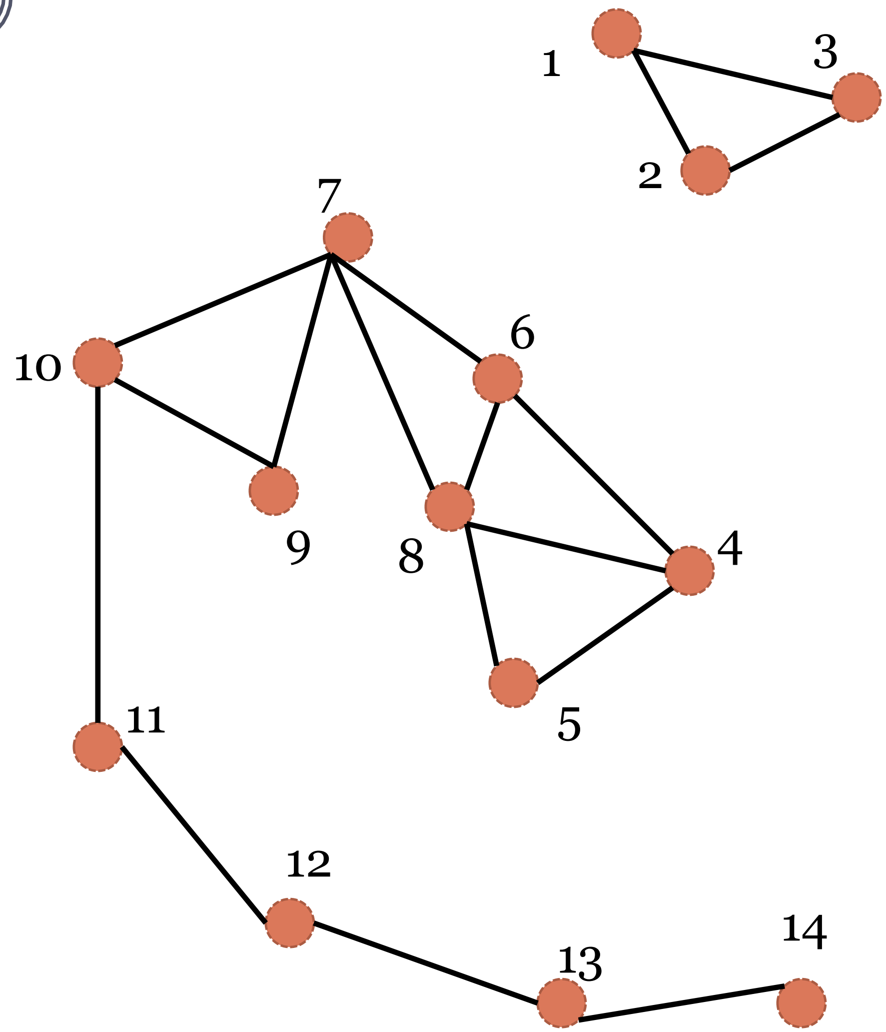
Diâmetro Efetivo

	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3			
5	2			
6	3			
7	4			
8	4			
9	2			
10	3			
11	2			
12	2			
13	2			
14	1			



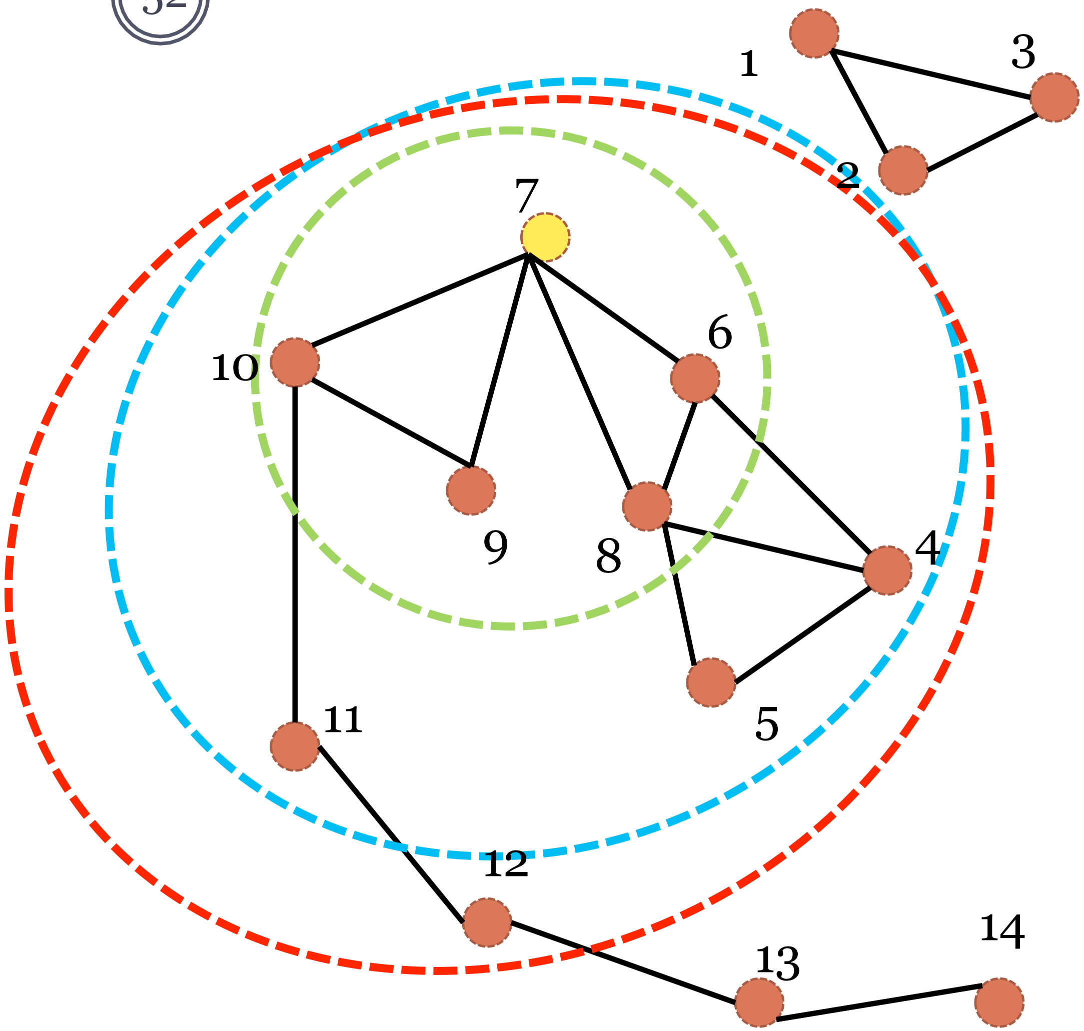
Diâmetro Efetivo

	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3	4	6	
5	2	3	4	
6	3	6	7	
7	4	7	8	
8	4	7	8	
9	2	6	9	
10	3	6	9	
11	2	5	8	
12	2	4	6	
13	2	3	4	
14	1	2	3	



Diâmetro Efetivo

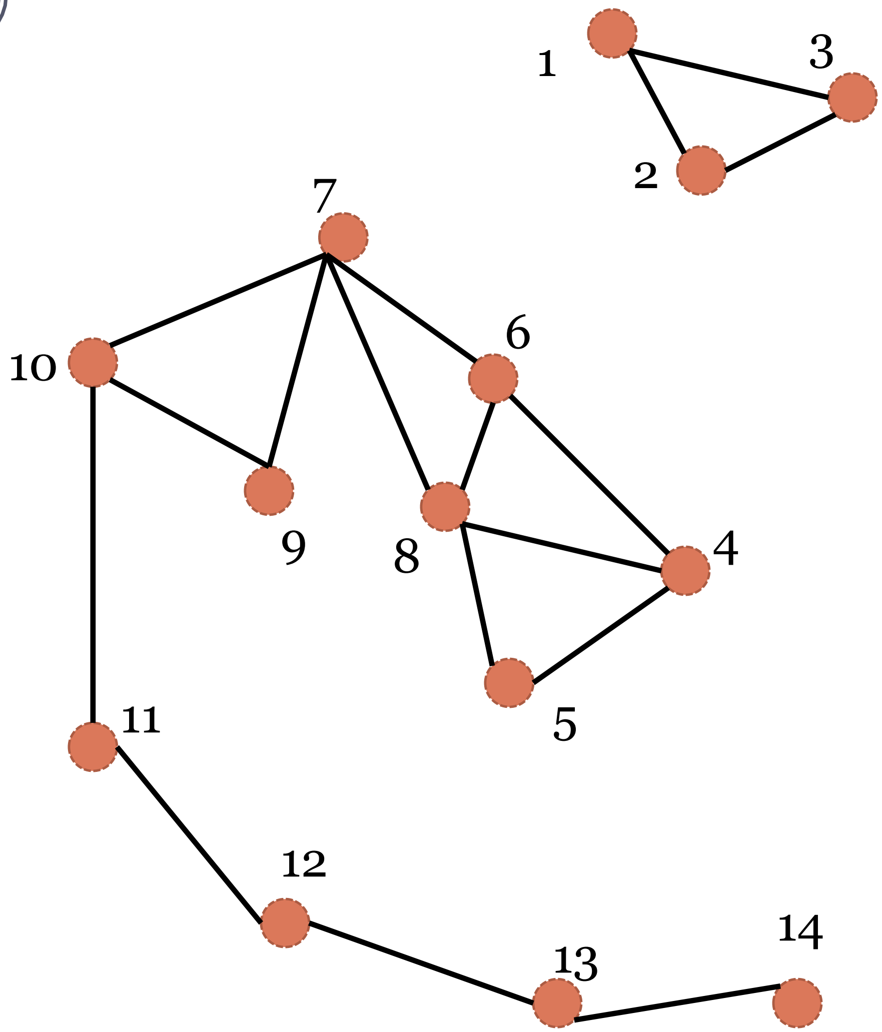
	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3	4	6	
5	2	3	4	
6	3	6	7	
7	4	7	8	
8	4	7	8	
9	2	6	9	
10	3	6	9	
11	2	5	8	
12	2	4	6	
13	2	3	4	
14	1	2	3	



Diâmetro Efetivo

	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3	4	6	
5	2	3	4	
6	3	6	7	
7	4	7	8	
8	4	7	8	
9	2	6	9	
10	3	6	9	
11	2	5	8	
12	2	4	6	
13	2	3	4	
14	1	2	3	

1h = 34

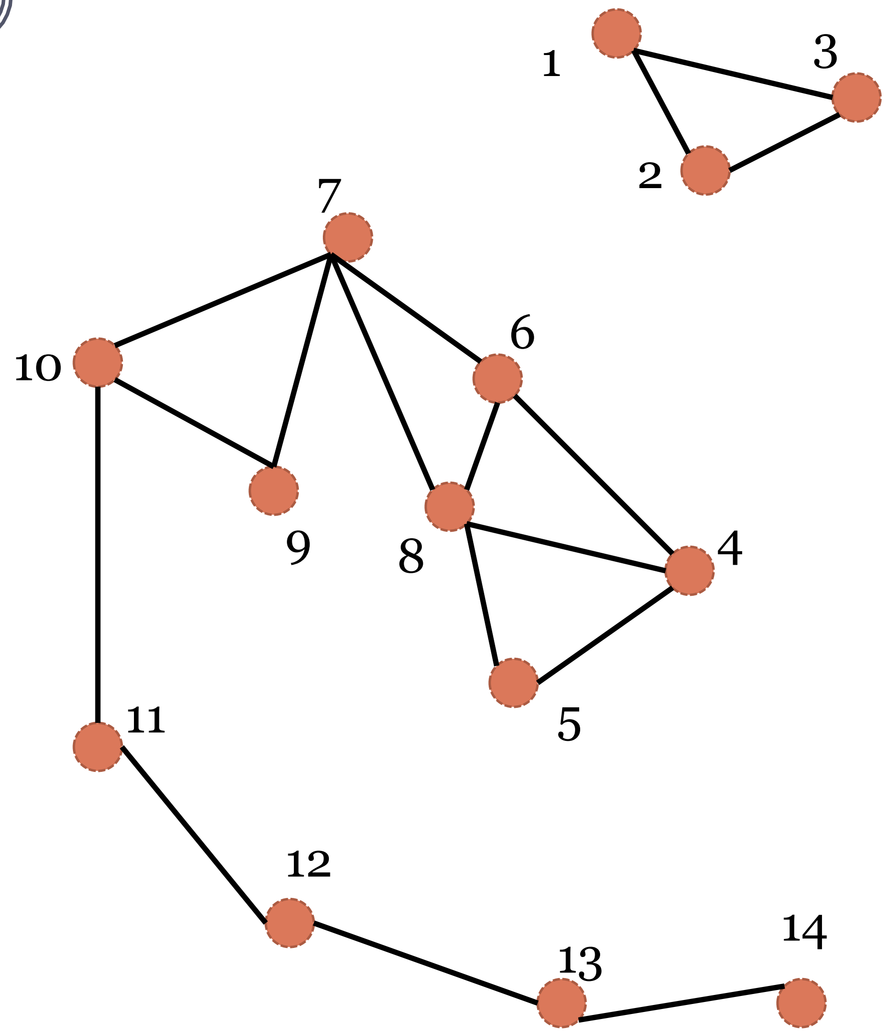


Diâmetro Efetivo

	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3	4	6	
5	2	3	4	
6	4	6	7	
7	4	7	8	
8	4	7	8	
9	2	6	9	
10	3	6	9	
11	2	5	8	
12	2	4	6	
13	2	3	4	
14	1	2	3	

1h = 34

2h = 59



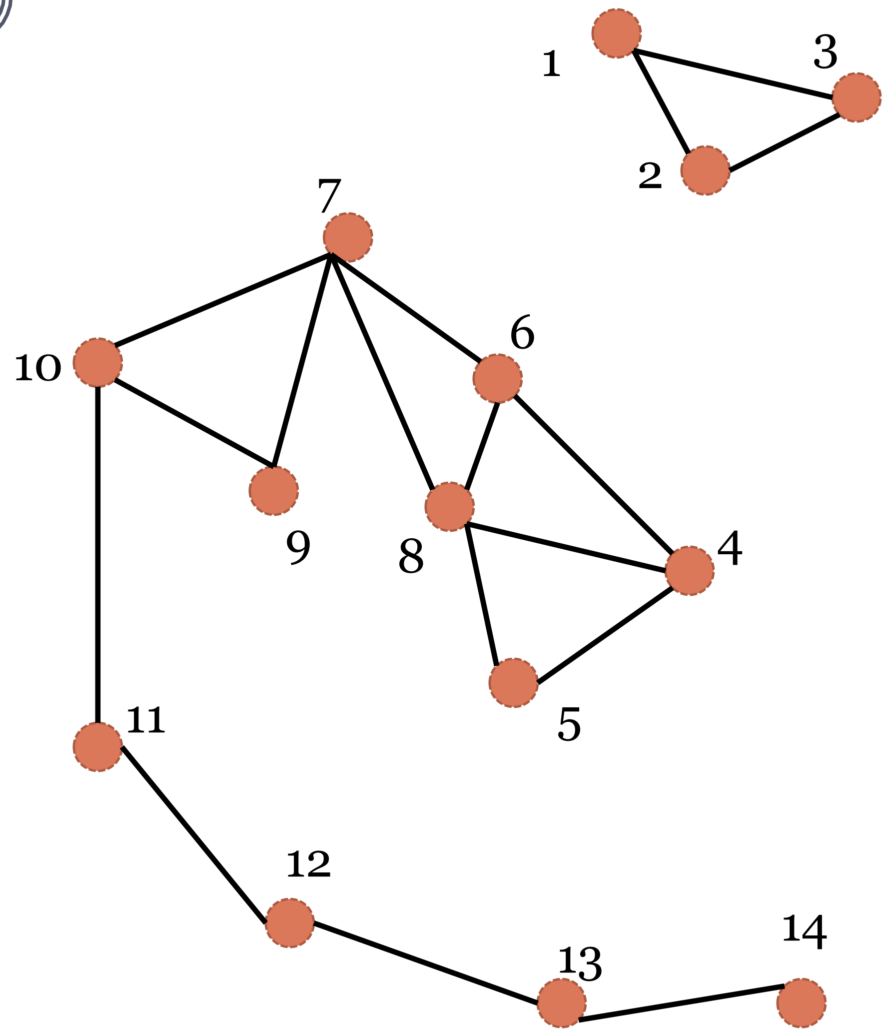
Diâmetro Efetivo

	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3	4	6	
5	2	3	4	
6	4	6	7	
7	4	7	8	
8	4	7	8	
9	2	6	9	
10	3	6	9	
11	2	5	8	
12	2	4	6	
13	2	3	4	
14	1	2	3	

1h = 34

2h = 59

3h = 78



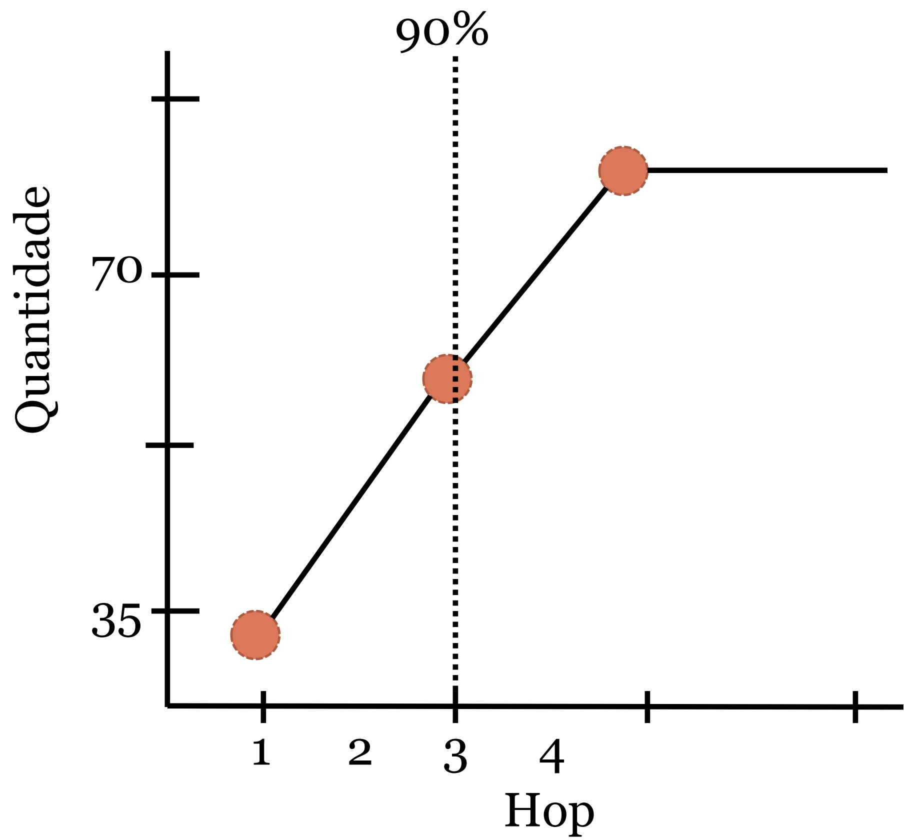
Diâmetro Efetivo

	1h	2h	3h	..
1	2	2	2	
2	2	2	2	
3	2	2	2	
4	3	4	6	
5	2	3	4	
6	4	6	7	
7	4	7	8	
8	4	7	8	
9	2	6	9	
10	3	6	9	
11	2	5	8	
12	2	4	6	
13	2	3	4	
14	1	2	3	

1h = 34

2h = 59

3h = 78

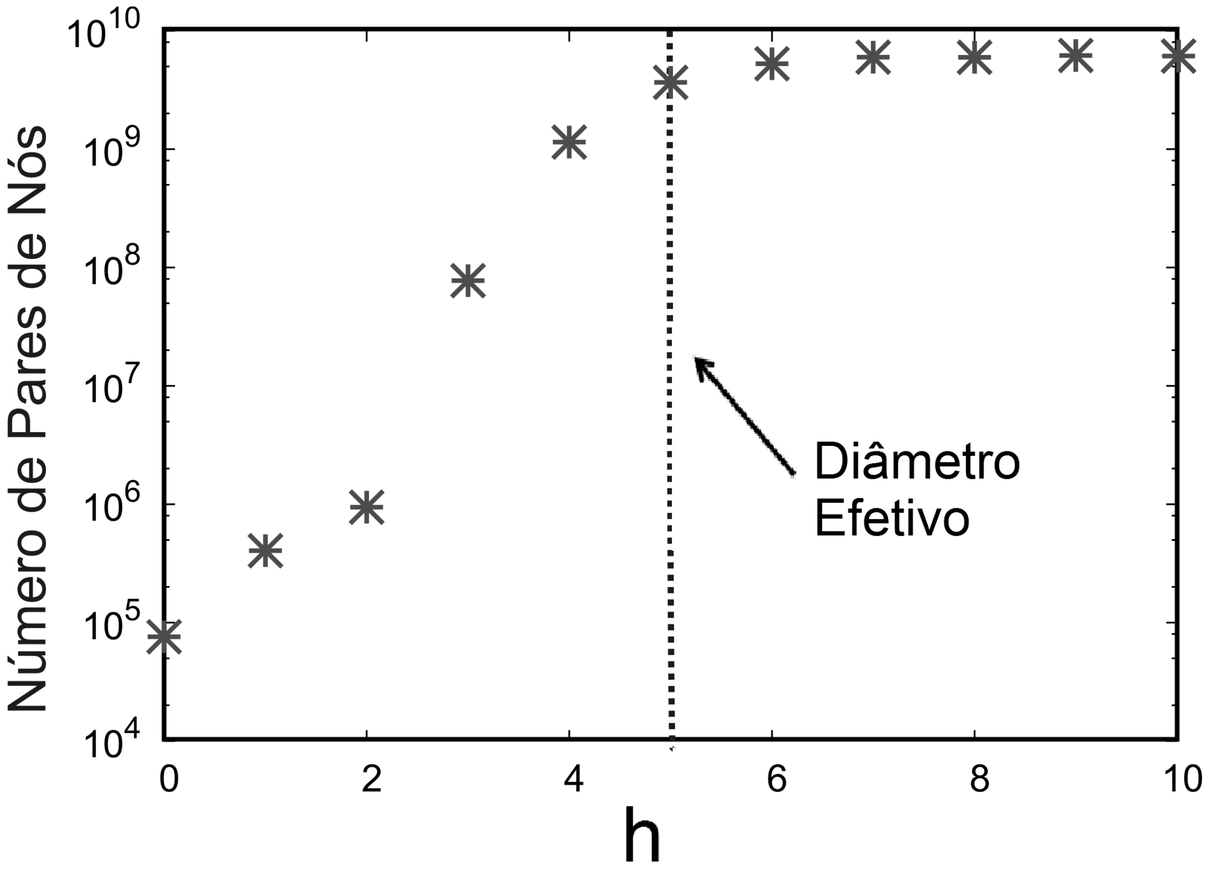


Diâmetro Efetivo

- Diâmetro Efetivo:
  - É o menor número de "arestas" em que no mínimo 90% de todos os nós da maior componente conexa do grafo podem ser alcançados entre si
  - É um valor mais robusto que o diâmetro tradicional
    - somente os pares de nós conexos são considerados
    - a direção das arestas (no caso de grafos direcionados) são ignoradas
    - experimentos mostram que o diâmetro efetivo exibe comportamento qualitativamente similar ao diâmetro tradicional
  - Principal algoritmo é o ANF que calcula o diâmetro efetivo em O(N)

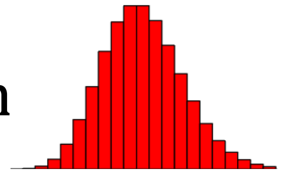
Palmer, C. R.; Gibbons, P. B. & Faloutsos, C. ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs *KDD* 2002*, 1*, 81-90

Gráfico Hop-Plot

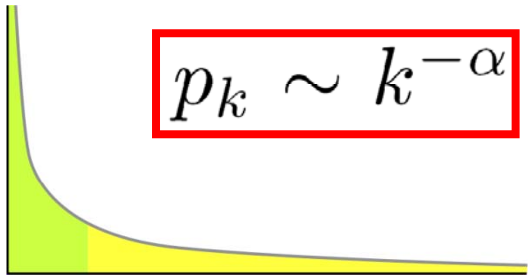


Degree distributions (1)

- Let *pk* denote a fraction of nodes with degree *k*
- We can plot a histogram of *pk* vs. *k*
- In a (Erdos-Renyi) random graph degree distribution follows Poisson distribution



- Degrees in real networks are heavily skewed to the right
- Distribution has a long tail of values that are far above the mean
- Power-law [Faloutsos et al], Zipf's law, Pareto's law, Long tail, Heavy-tail
- Many things follow Power-law:
  - Amazon sales,
  - word length distribution,
  - Wealth, Earthquakes, …



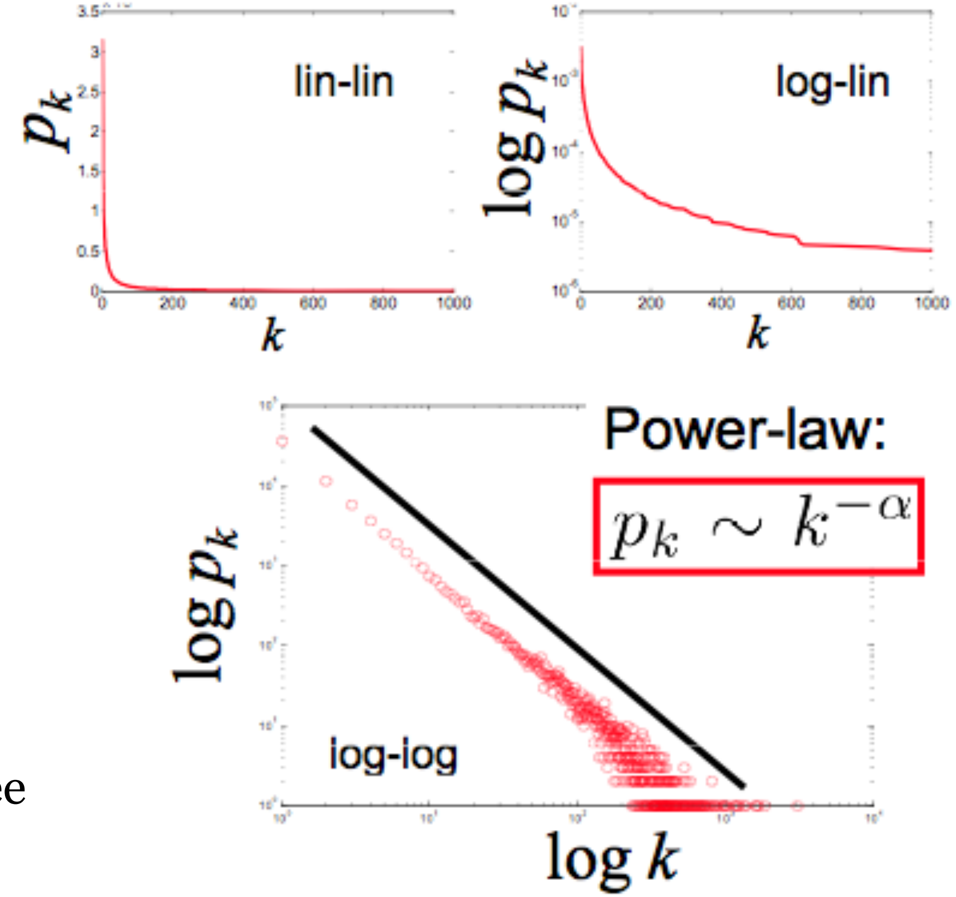
$$\boxed{p_k \sim k^{-\alpha}}$$

Degree distributions (2)

- Many real world networks contain hubs: highly connected nodes
- We can easily distinguish between exponential and power-law tail by plotting on log-lin and log-log axis
- Power-law is a line on log-log plot

For statistical tests and estimation see Clauset-Shalizi-Newman 2007

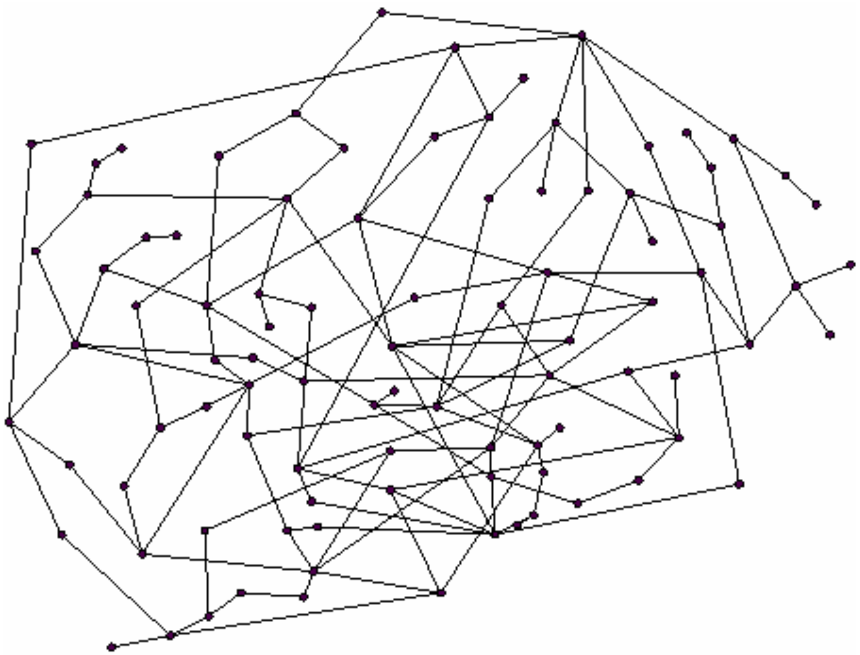
Degree distribution in a blog network
(plot the same data using different scales)



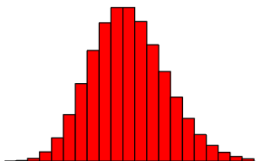
Power Law degree expoents

- Power law degree exponent is typically $2 < \alpha < 3$
  - Web graph [Broder et al. 00]:
    - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$
  - Autonomous systems [Faloutsos et al. 99]:
    - $\alpha = 2.4$
  - Actor collaborations [Barabasi- Albert 00]:
    - $\alpha = 2.3$
  - Citations to papers [Redner 98]:
    - $\alpha \approx 3$
  - Online social networks [Leskovec et al. 07]:
    - $\alpha \approx 2$

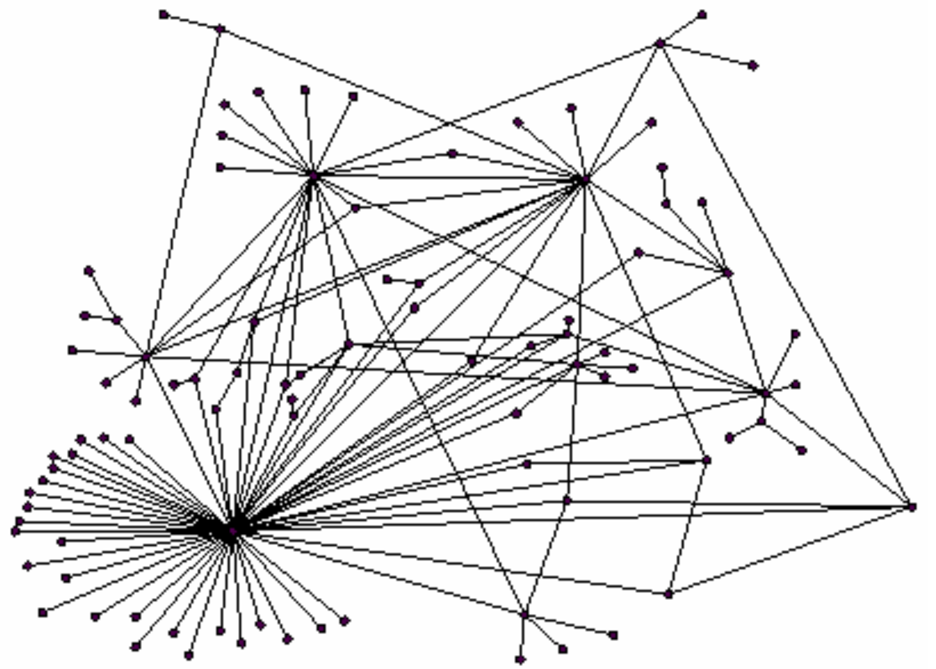
Poisson vs. Scale-free network



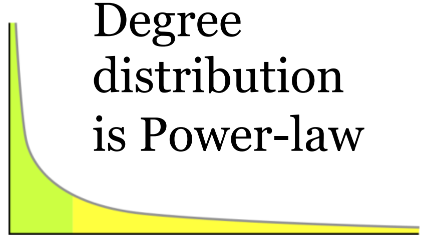
Poisson network
(Erdos-Renyi random graph)



Degree distribution is Poisson



Scale-free (power-law) network



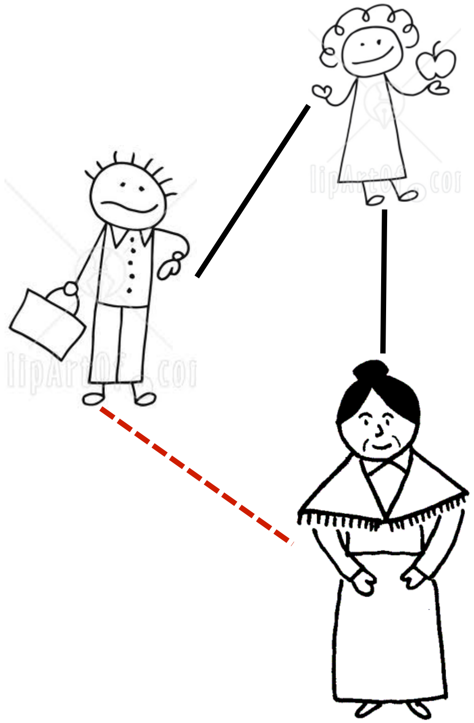
Degree distribution is Power-law

Function is scale free if:
$f(ax) = c\, f(x)$

- The basic role of triadic closure in social networks has motivated the formulation of simple social network measures to capture its prevalence.
- The clustering coefficient of a node A is defined as the probability that two randomly selected friends of A are friends with each other. In other words, it is the fraction of pairs of A's friends that are connected to each other by edges.

Triângulos

- Em uma rede social, nós são pessoas e as arestas são os relacionamentos;
- Sabe-se que se A é amigo de B que é amigo de C, há uma grande chance de A ser/se tornar amigo de C.
- A transitividade significa a presença de um alto número de triângulos (D(vi)) na rede.



Coeficiente de Clusterização

- indicar quão próximo o grafo está de ser um grafo completo
- Do nó

$$C(v_i) = \frac{2 * \Delta(v_i)}{d(v_i) * (d(v_i) - 1)}$$

- Da rede

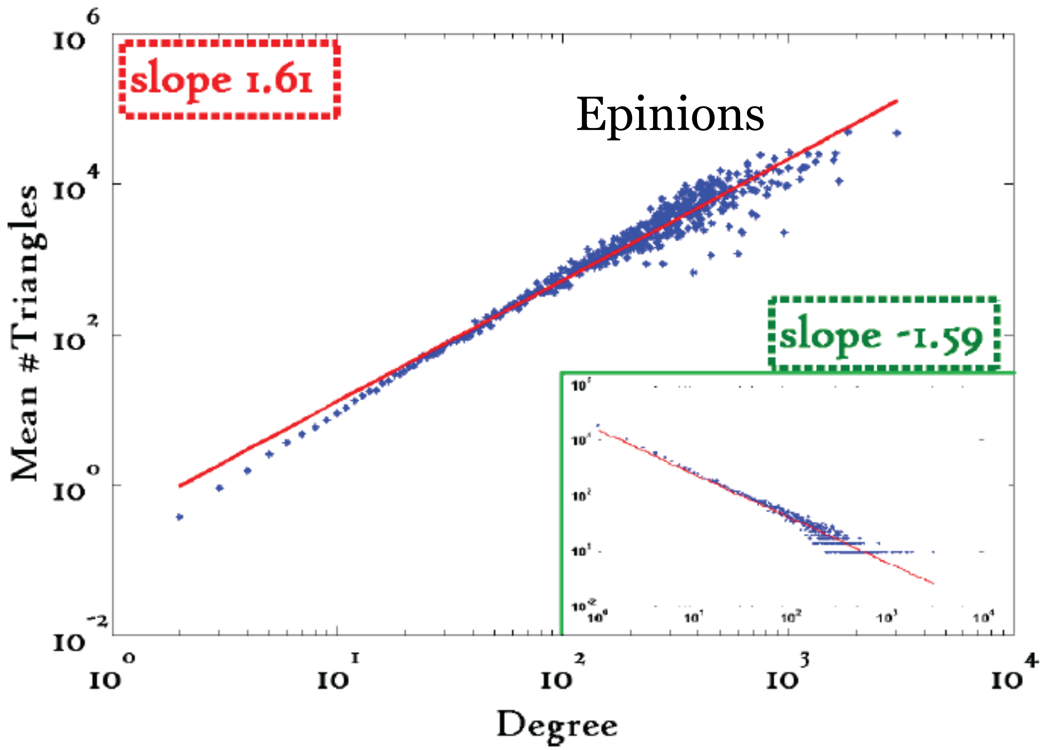
$$C(\mathscr{G}) = \frac{1}{N} * \sum_{i=1}^{N} C(v_i)$$

Triângulos

- Além do grau do nó os triângulos também seguem uma lei de potência

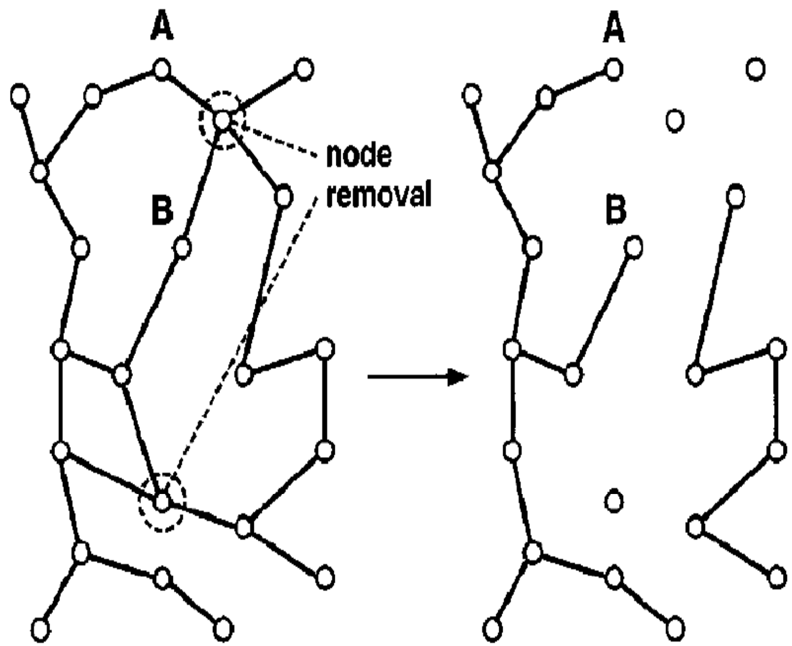
$$\Delta(G) = \frac{1}{6}\sum_{i=1}^{n}\lambda_i^3$$

Tsourakakis, C. E.
Fast Counting of Triangles in Large Real Networks without Counting: Algorithms and Laws
ICDM '08, IEEE Computer Society, 2008, 608-617



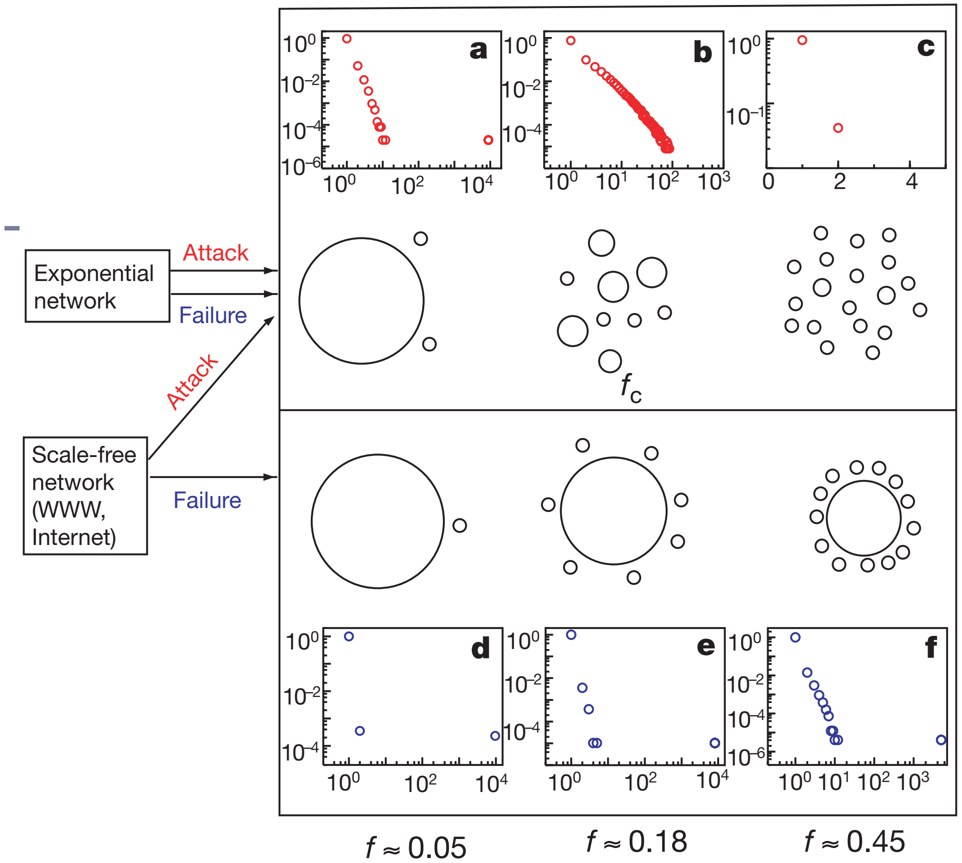
Network resilience (1)

- We observe how the connectivity (length of the paths) of the network changes as the vertices get removed [Albert et al. 00; Palmer et al. 01]
- Vertices can be removed:
  - Uniformly at random
  - In order of decreasing degree
- It is important for epidemiology
  - Removal of vertices corresponds to vaccination

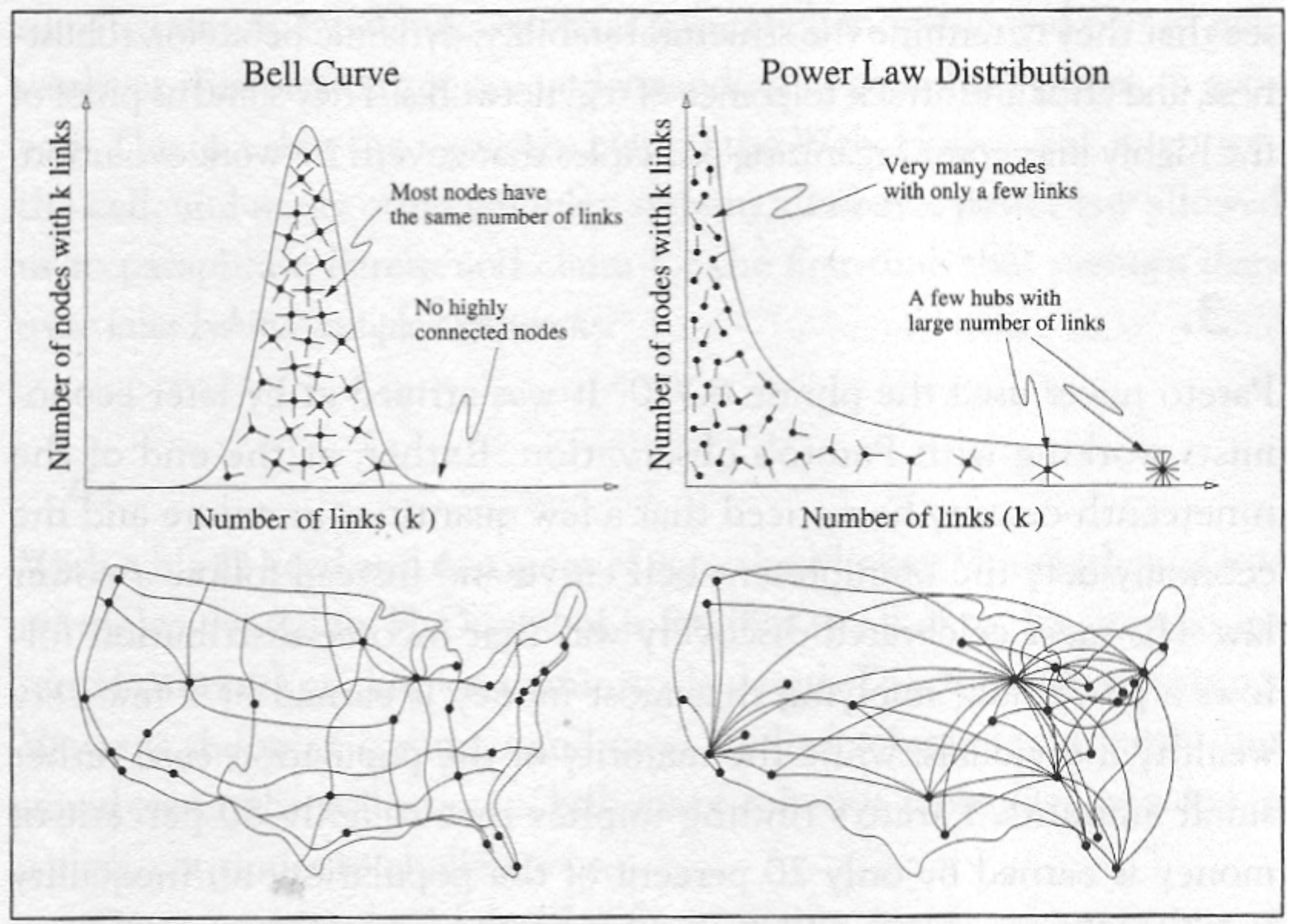


Network resilience (2)

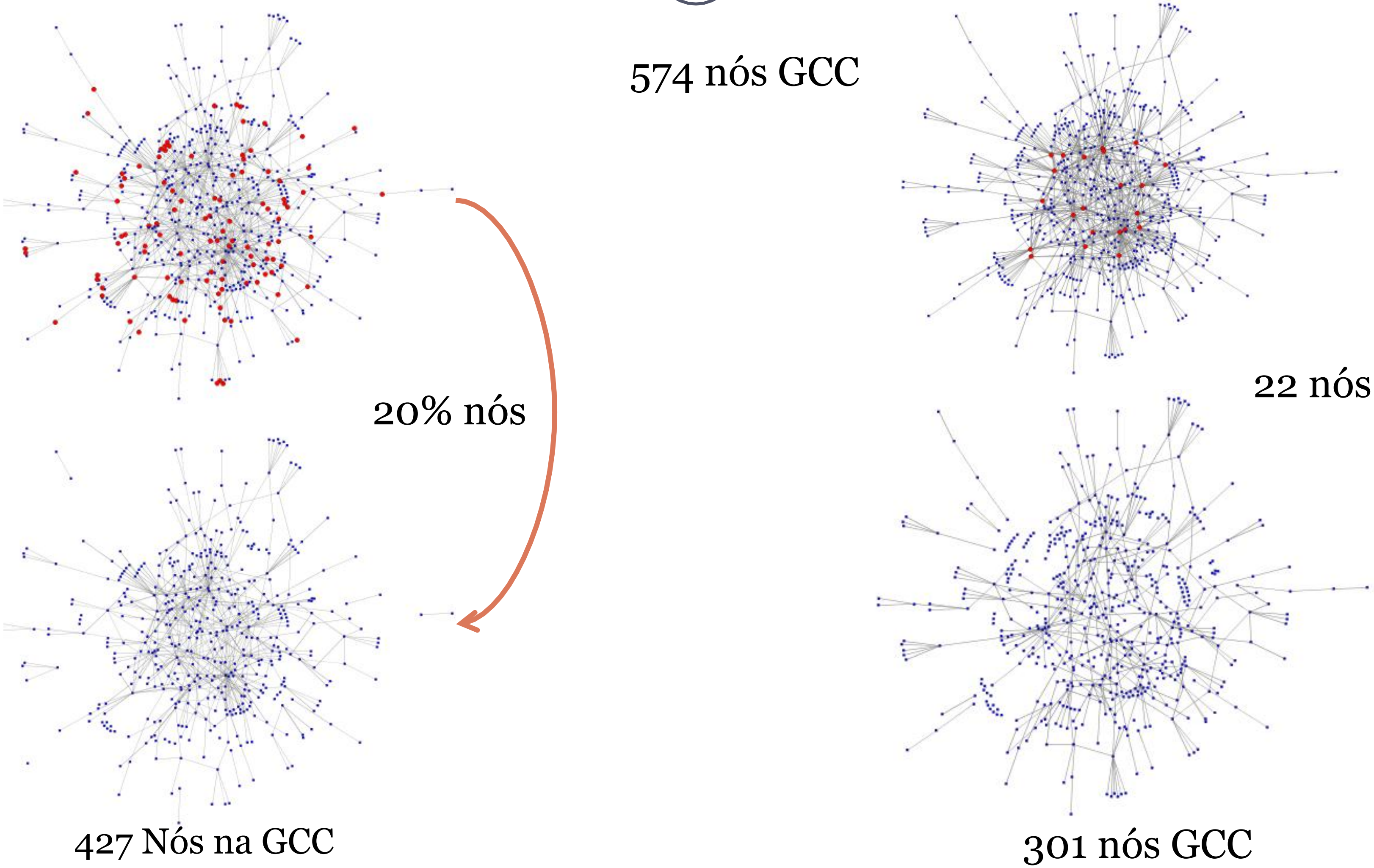
- Real-world networks are resilient to random attacks
  - One has to remove all web-pages of degree > 5 to disconnect the web
  - But this is a very small percentage of web pages
- Random network has better resilience to targeted attacks



Poisson vs. Scale-free network

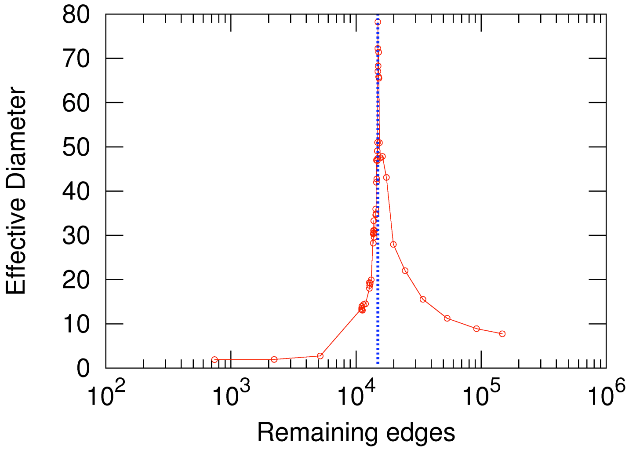


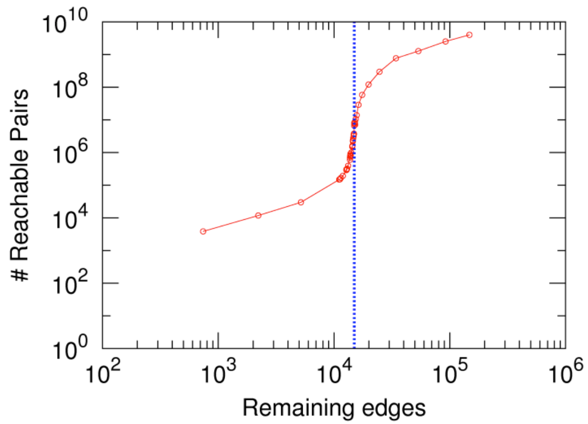
Exemplo

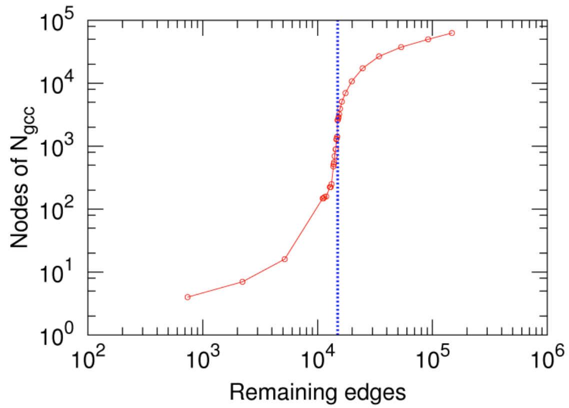


ShatterPlots (1)

- ShatterPlots
  - A simple and powerful algorithm to tease out patterns of real graphs, helping us to spot fake/masked graphs
  - Force a graph to reach a critical ("Shattering") point, randomly deleting edges, and study its properties at that point.

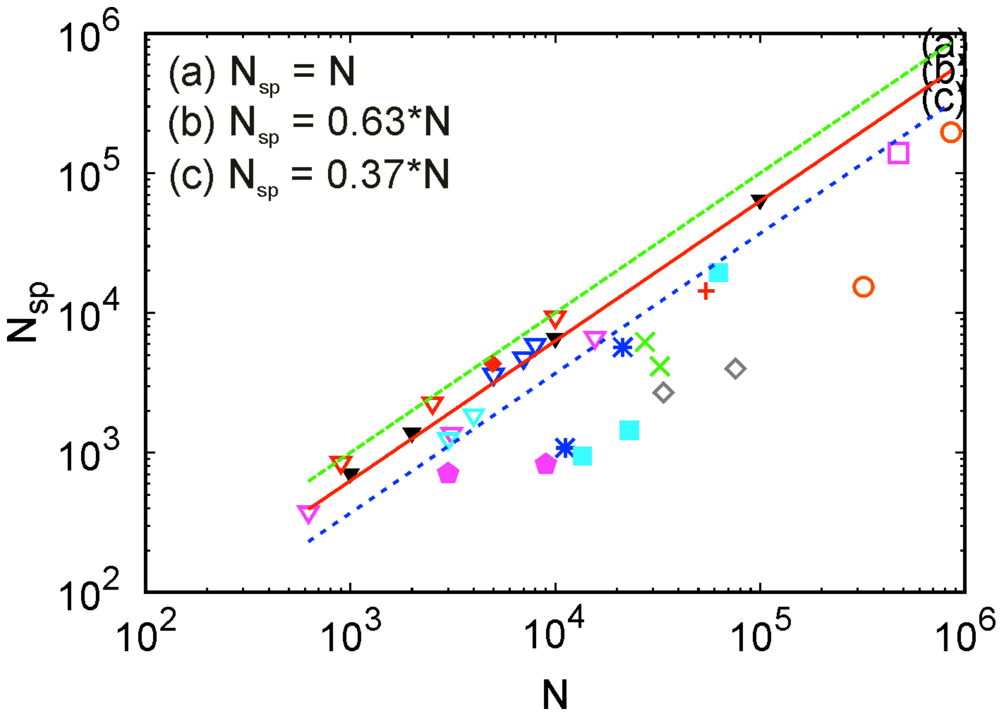






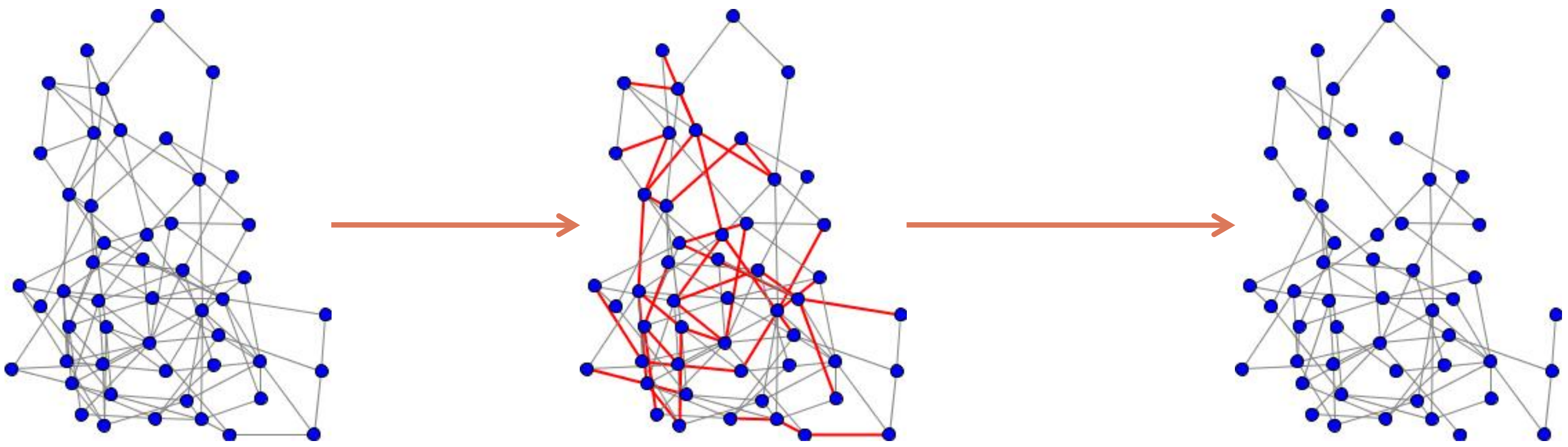
ShatterPlots (2)

- Node Shattering Ratio, which presents the relation of nodes at the Shattering point N_{sp} versus total number of nodes N of a graph.



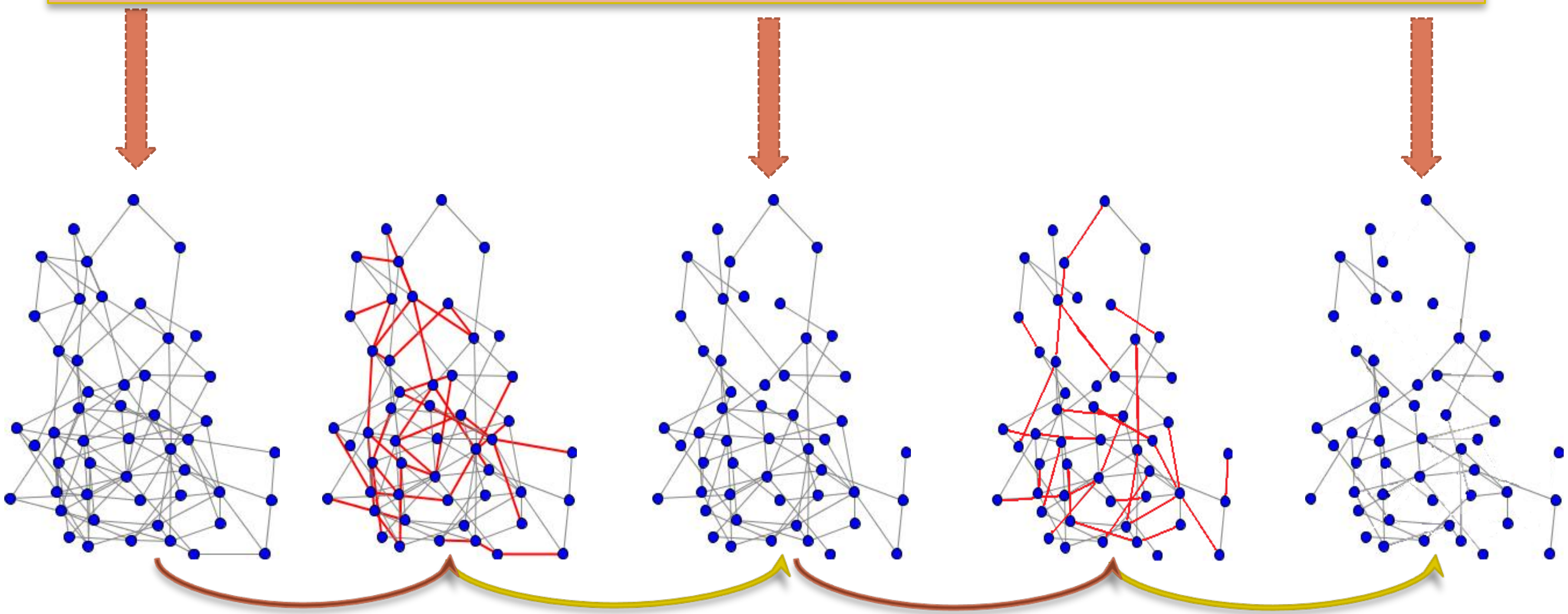
Arestas

- Remoção de arestas
  - bond percolation: cada aresta é removida com probabilidade p
    - Falhas aleatória dos links
  - Ataque: causa grandes danos na rede com a remoção de poucas arestas
    - Estratégias: remover arestas que são mais suscetíveis a quebrar a rede ou aumentar os menores caminhos →betweenness



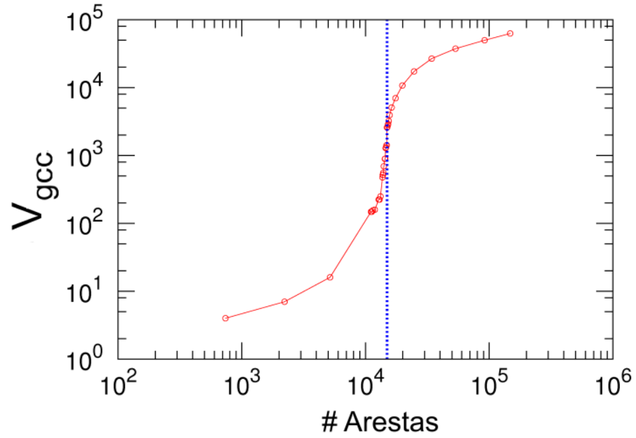
ShatterPlots

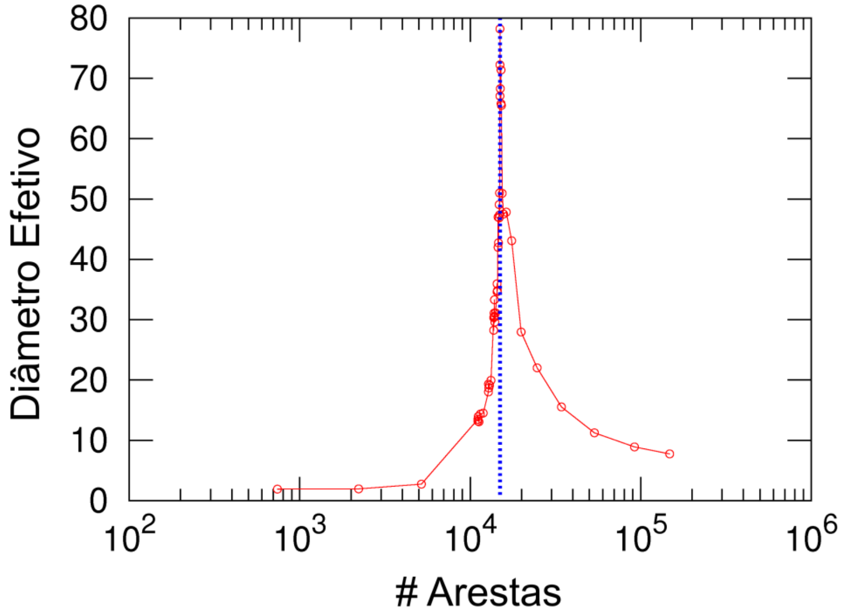
Medidas→ Diâmetro, autovalores, triângulos, componente conexas, pares alcançáveis

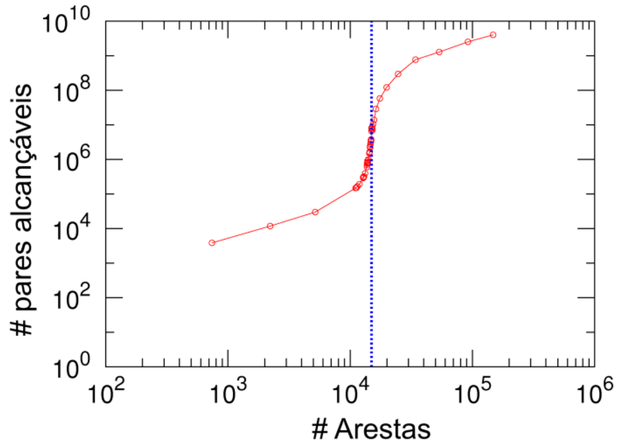


Shattering Point

Grande componente conexa e pares alcançáveis apresentam ponto critico mas APENAS o diâmetro tem um pico.







Experimentos

- 19 redes reais;
  - AS-Oregon, AS-Caida, Enron, AuthorToPaper, Gnutella, Web-Google, Berkley-Stanford, Epinions, etc.
- Redes sintéticas - triângulos;
  - Preferencial Attachment, Small-World, 2D Grid, Hierarchical;
  - ER → Validar resultados;
- Média de 10 Execuções;

Perguntas

- Todas as redes reais tem Shattering point?

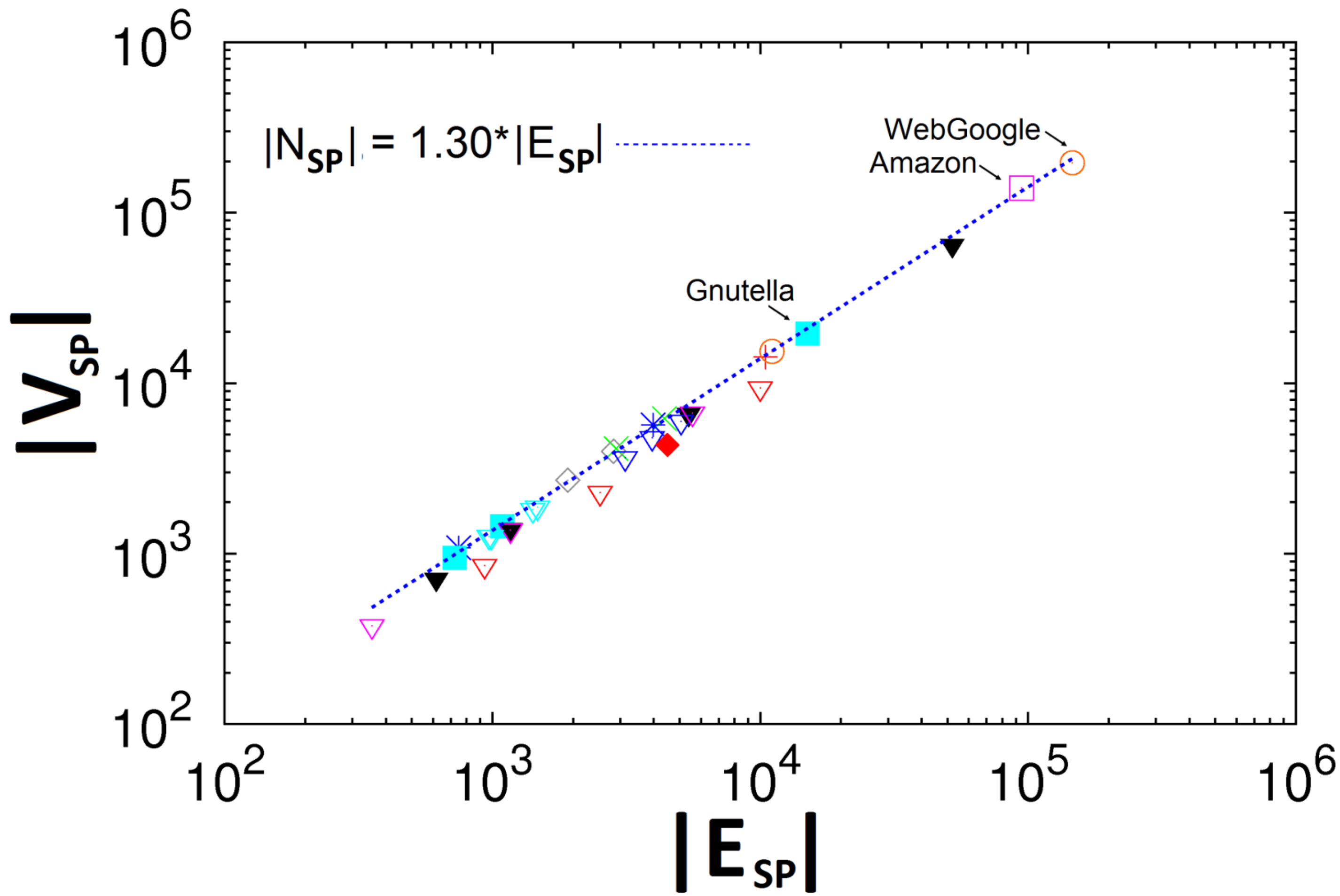
Todas as redes testadas possuem um Shattering Point.

- Quão próximas estão as redes reais do Shattering point?

As redes reais estão longe do Shattering Point

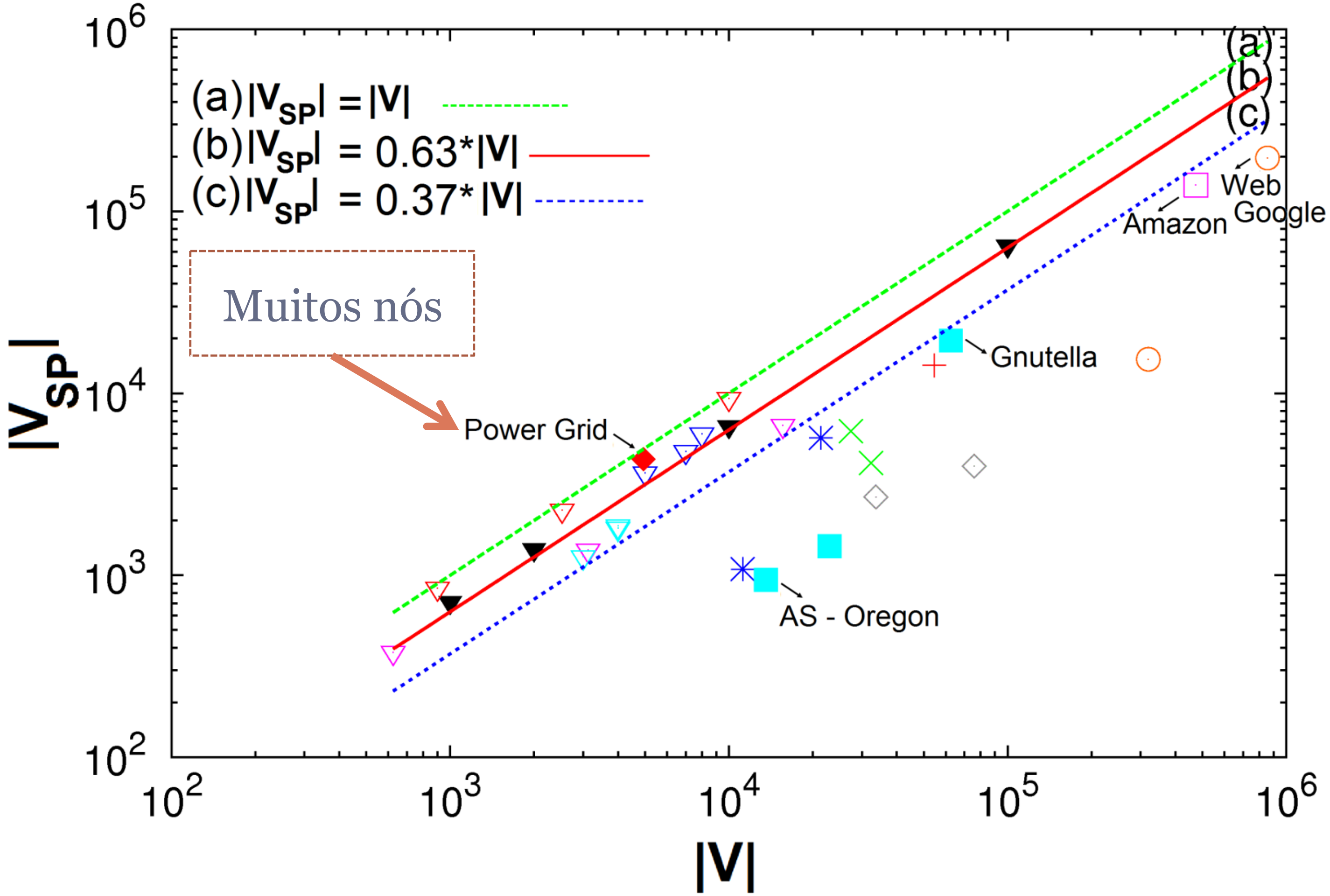
- As redes sintéticas tem comportamento parecido ou não com as redes reais quanto ao Shattering Point?

30%

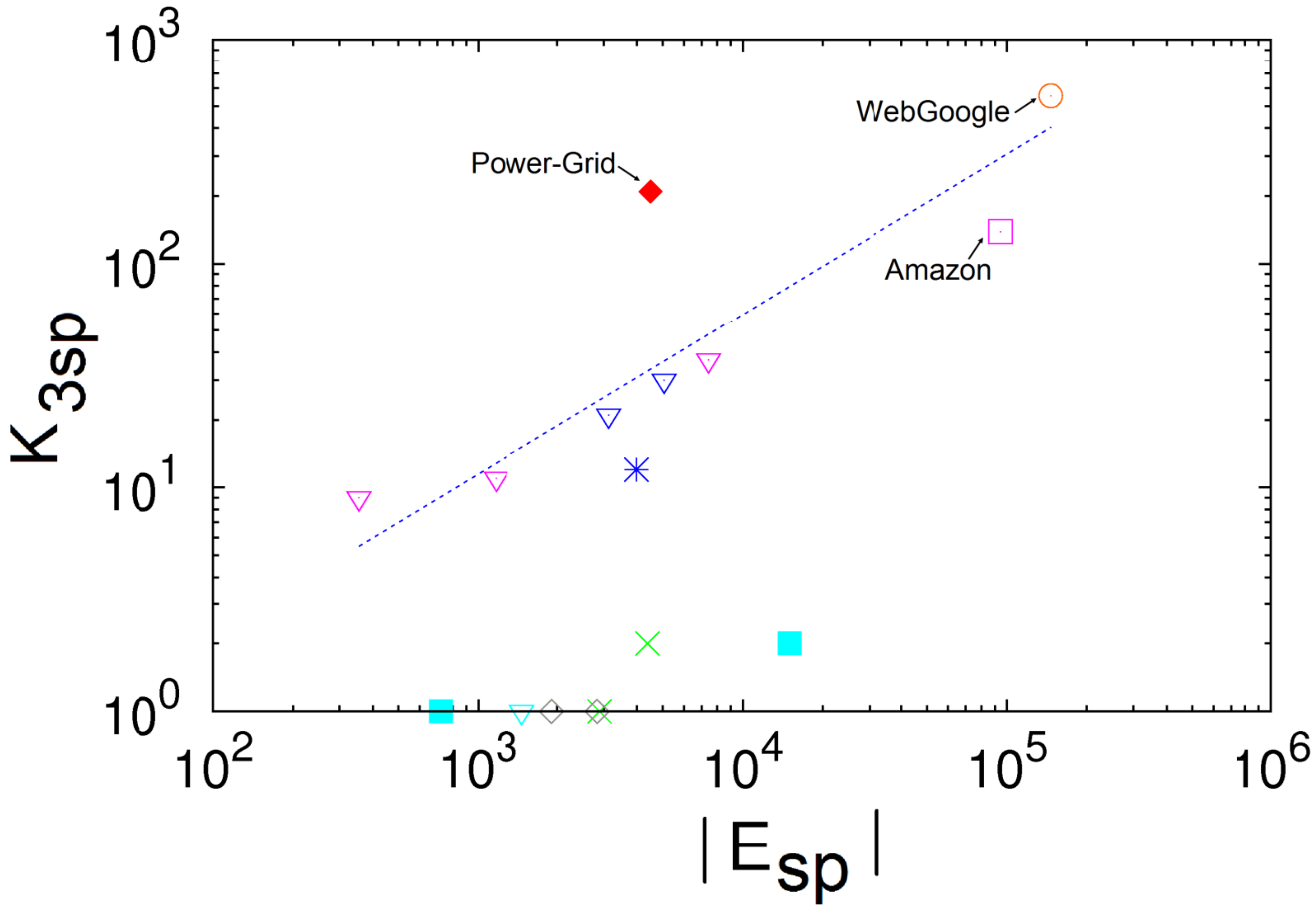


$|N_{SP}| = 1.30*|E_{SP}|$
WebGoogle
Amazon
Gnutella
10^2
10^3
10^4
10^5
10^6
$|V_{SP}|$
$|E_{SP}|$

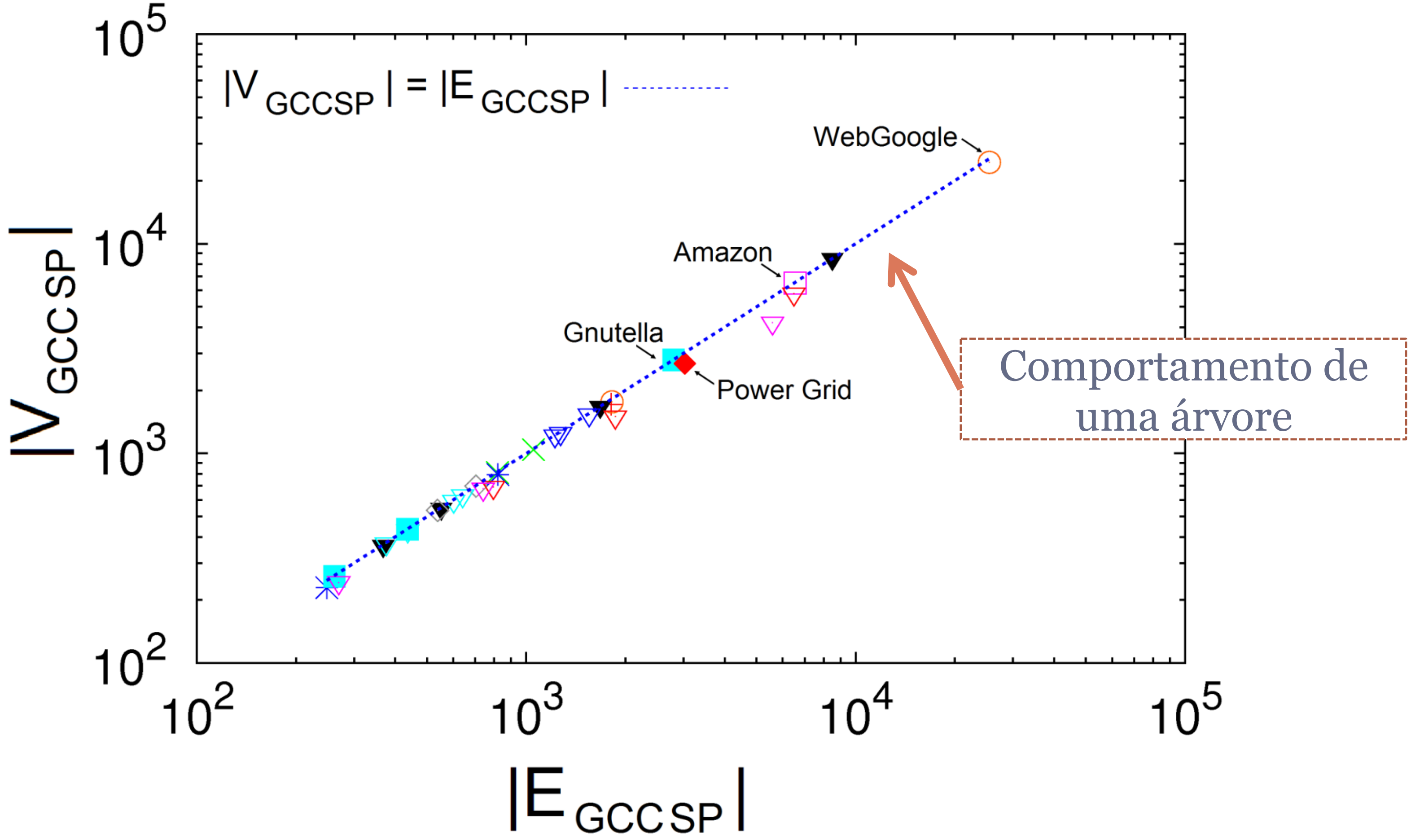
NodeShatteringRatio



TriangleRatio

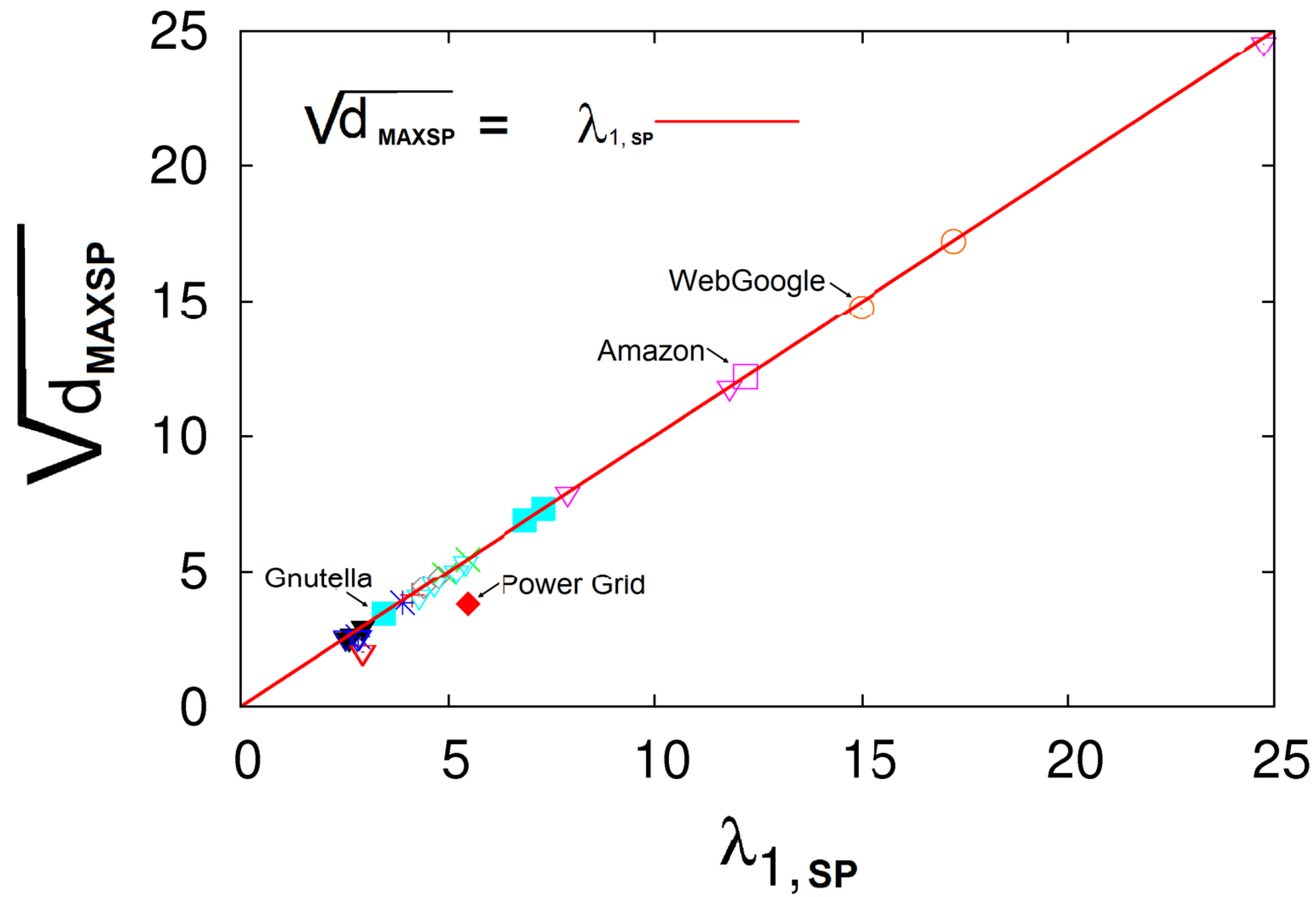


TreeGcc



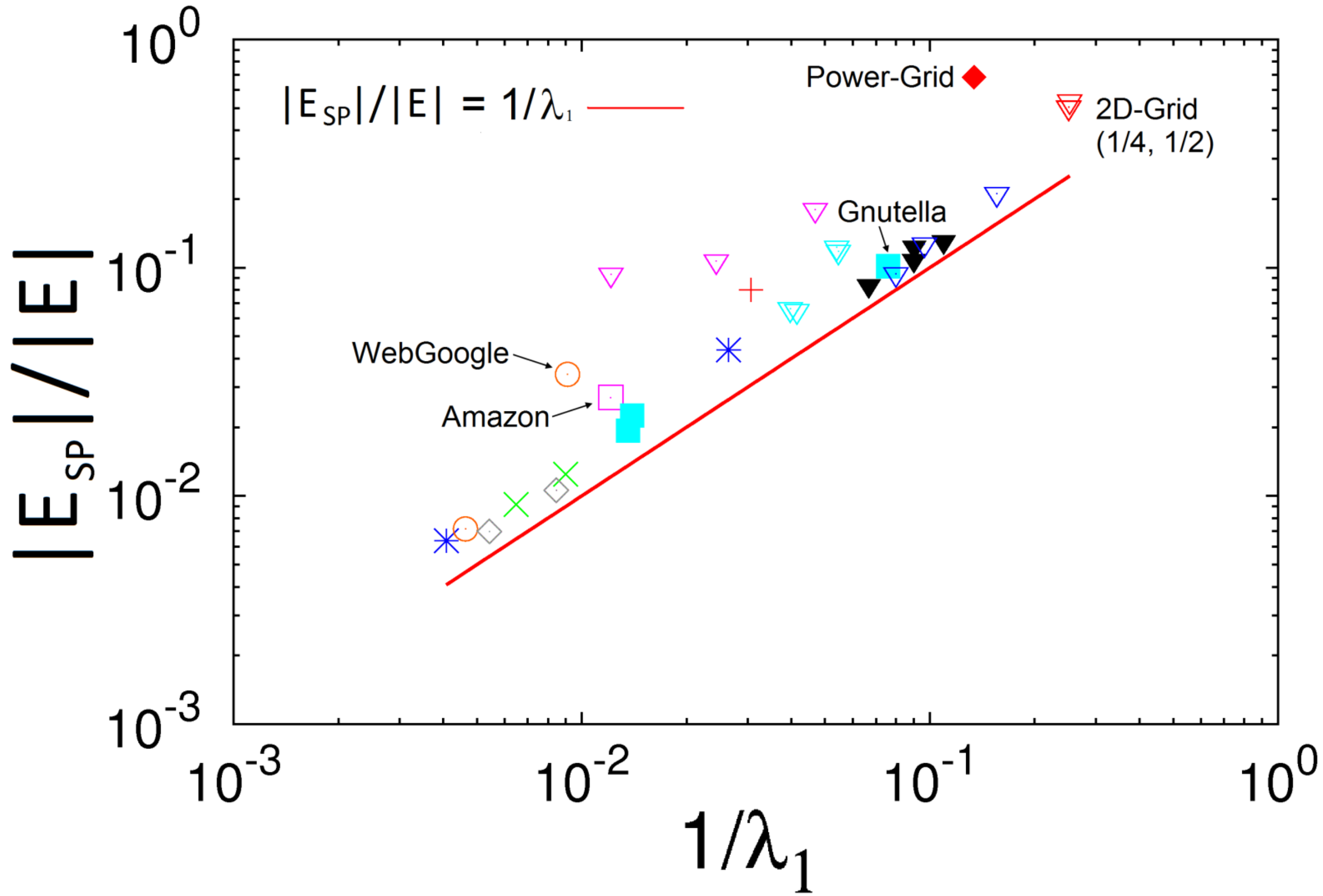
|V GCCSP | = |E GCCSP |
WebGoogle
Amazon
Gnutella
Power Grid
Comportamento de uma árvore
10^2
10^3
10^4
10^5
|E GCCSP |
|V GCCSP |

RootDegree



$\sqrt{d_{MAXSP}} = \lambda_{1,SP}$
WebGoogle
Amazon
Gnutella
Power Grid
$\sqrt{d_{MAXSP}}$
$\lambda_{1,SP}$
0
5
10
15
20
25

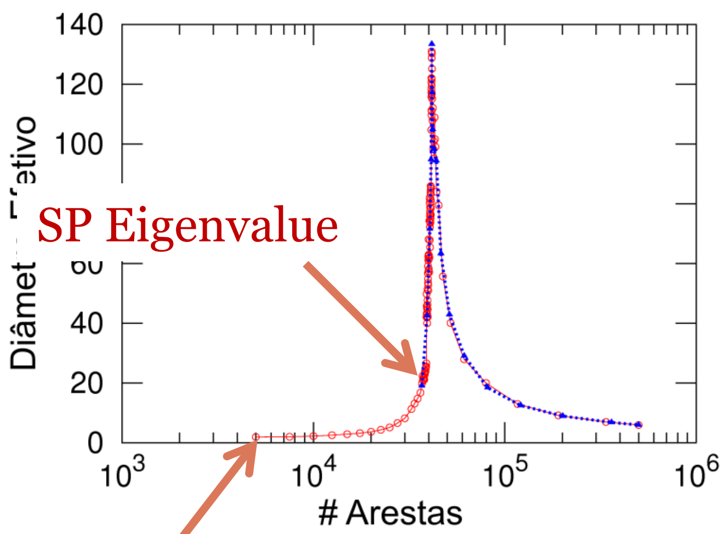
Eigenvalue



$|E_{SP}|/|E| = 1/\lambda_1$
Power-Grid
2D-Grid
(1/4, 1/2)
Gnutella
WebGoogle
Amazon
10^0
10^{-1}
10^{-2}
10^{-3}
$|E_{SP}|/|E|$
$1/\lambda_1$

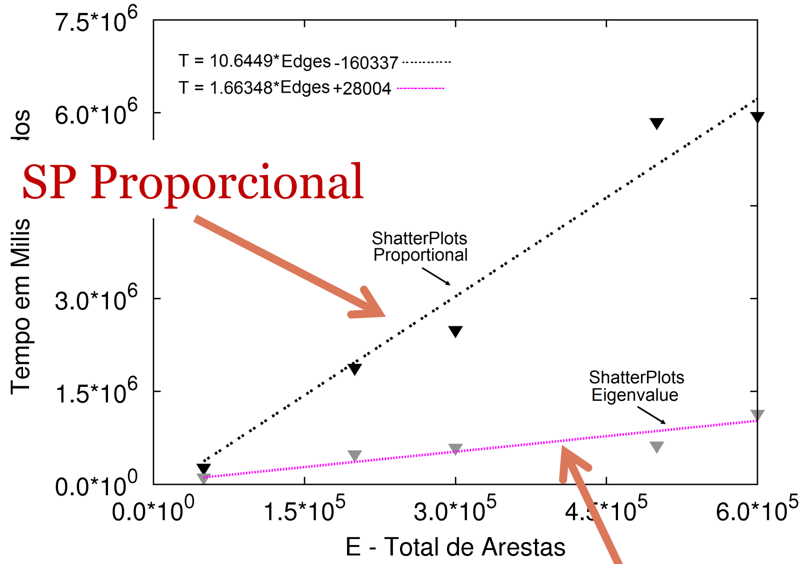
Escalabilidade

- ER → variar o tamanho e mesmo comportamento;



SP Eigenvalue

SP Proporcional



SP Proporcional

SP Eigenvalue

$$E_{sp} \geq E_t * 1/\lambda_1$$

What about evolving graphs?

- Conventional wisdom/intuition:
  - Constant average degree: the number of edges grows linearly with the number of nodes
- Slowly growing diameter: as the network grows the distances between nodes grow

Networks over time: Densification

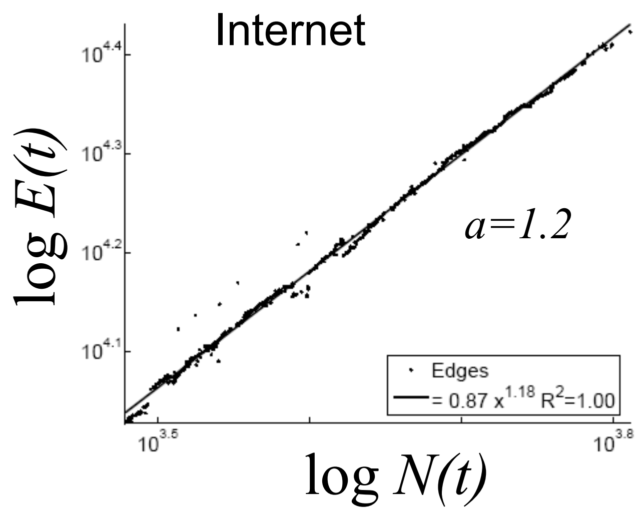
- A simple question: What is the relation between the number of nodes and the number of edges in a network over time?
- Let:
  - $N(t)$ ... nodes at time t
  - $E(t)$ ... edges at time t
- Suppose that:
  - $N(t+1) = 2 * N(t)$
- Q: what is your guess for
  - $E(t+1) =? \cancel{2} * E(t)$
- A: over-doubled!
  - But obeying the Densification Power Law [KDD05]

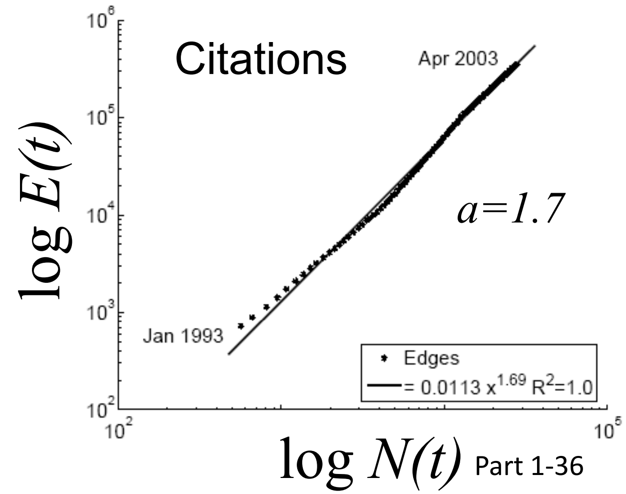
Networks over time: Densification

- Networks are denser over time
- The number of edges grows faster than the number of nodes – average degree is increasing

$$E(t) \propto N(t)^a$$

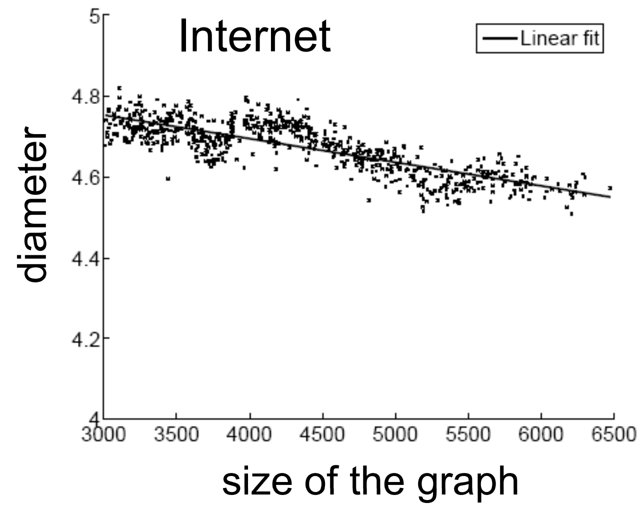
  - *a* … densification exponent
- ***1 ≤ a ≤ 2:***
  - *a=1:* linear growth – constant out- degree (assumed in the literature so far)
  - *a=2:* quadratic growth – clique

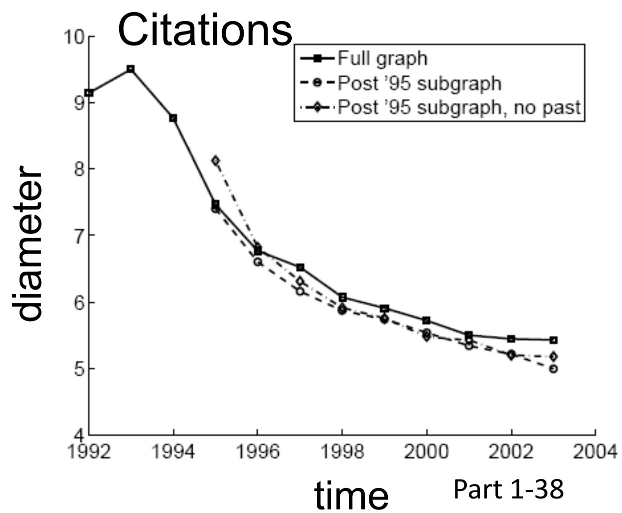




Shrinking diameters

- ~~Intuition and prior work say that distances between the nodes slowly grow as the network grows (like *log n*):~~
  - ~~*d ~ O(log N)*~~
  - ~~*d ~ O(log log N)*~~
- Diameter Shrinks/ Stabilizes over time
  - as the network grows the distances between nodes slowly decrease [KDD 05]





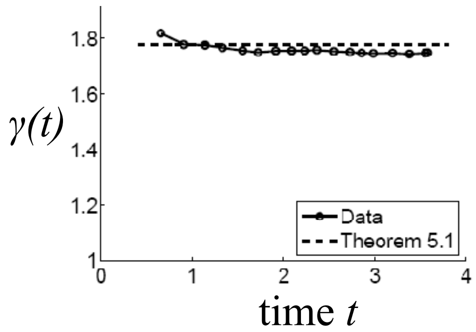
Densification & degree distribution

- How does densification affect degree distribution?
- Densification:
- Degree distribution: $p_k = k^{\gamma}$
- Given densification exponent a, the degree exponent is [TKDD '07]:
    - (a) For γ=*const* over time, we obtain densification only for $1<\gamma<2$, and then it holds: $\gamma = a/2$
    - (b) For $\gamma<2$ degree distribution evolves according to:

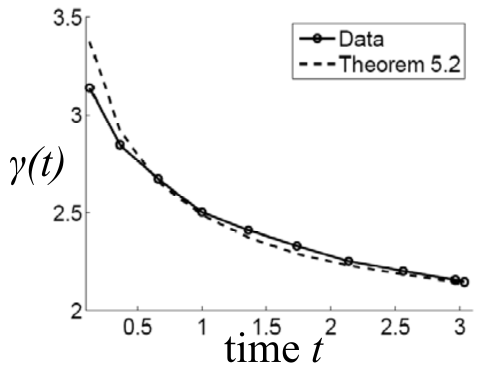
$$\gamma_n = \frac{4n^{a-1} - 1}{2n^{a-1} - 1}$$

Given: densification a, number of nodes n

Case (a): Degree exponent γ is constant over time. The network densifies, $a=1.2$



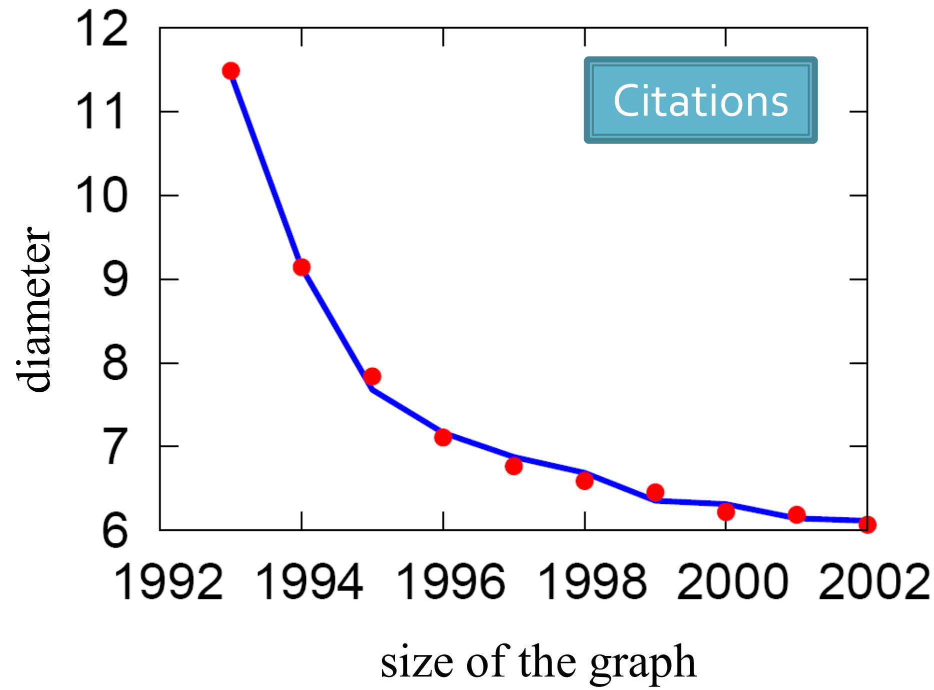
Case (b): Degree exponent γ evolves over time. The network densifies, $a=1.6$



Diameter of a rewired network

- Compare diameter of a:
  - True network (red)
  - Random network with the same degree distribution (blue)

Densification + degree sequence give shrinking diameter

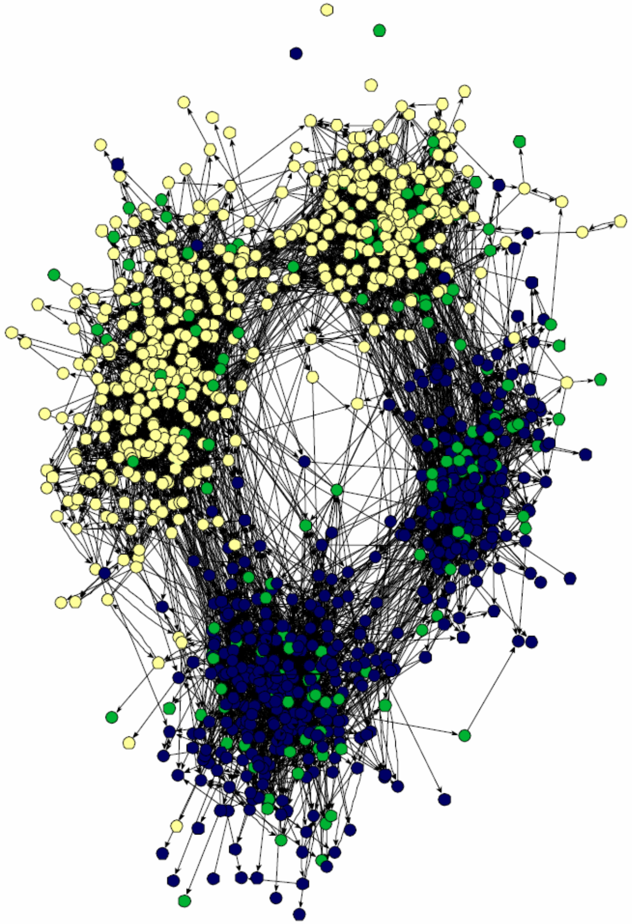


Properties hold in many graphs

- These patterns can be observed in many real world networks:
  - World wide web [Barabasi]
  - On-line communities [Holme, Edling, Liljeros]
  - Who call whom telephone networks [Cortes]
  - Internet backbone – routers [Faloutsos, Faloutsos, Faloutsos]
  - Movies to actors network [Barabasi]
  - Science citations [Leskovec, Kleinberg, Faloutsos]
  - Click-streams [Chakrabarti]
  - Autonomous systems [Faloutsos, Faloutsos, Faloutsos]
  - Co-authorship [Leskovec, Kleinberg, Faloutsos]
  - Sexual relationships [Liljeros]

Community structure

- Most social networks show community structure
  - groups have higher density of edges within than across groups
  - People naturally divide into groups based on interests, age, occupation, …
- How to find communities:
  - Spectral clustering (embedding into a low-dim space)
  - Hierarchical clustering based on connection strength
  - Combinatorial algorithms (min cut style formulations)
  - Block models
  - Diffusion methods

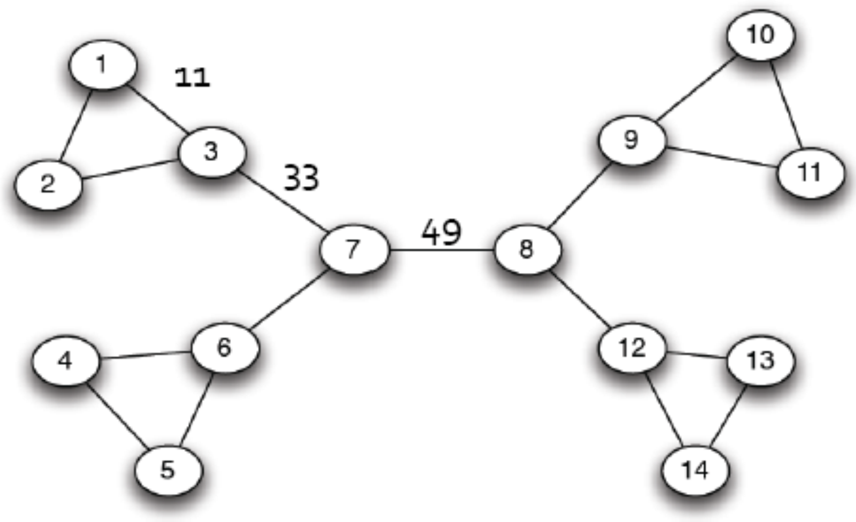


Friendship network of children in a school

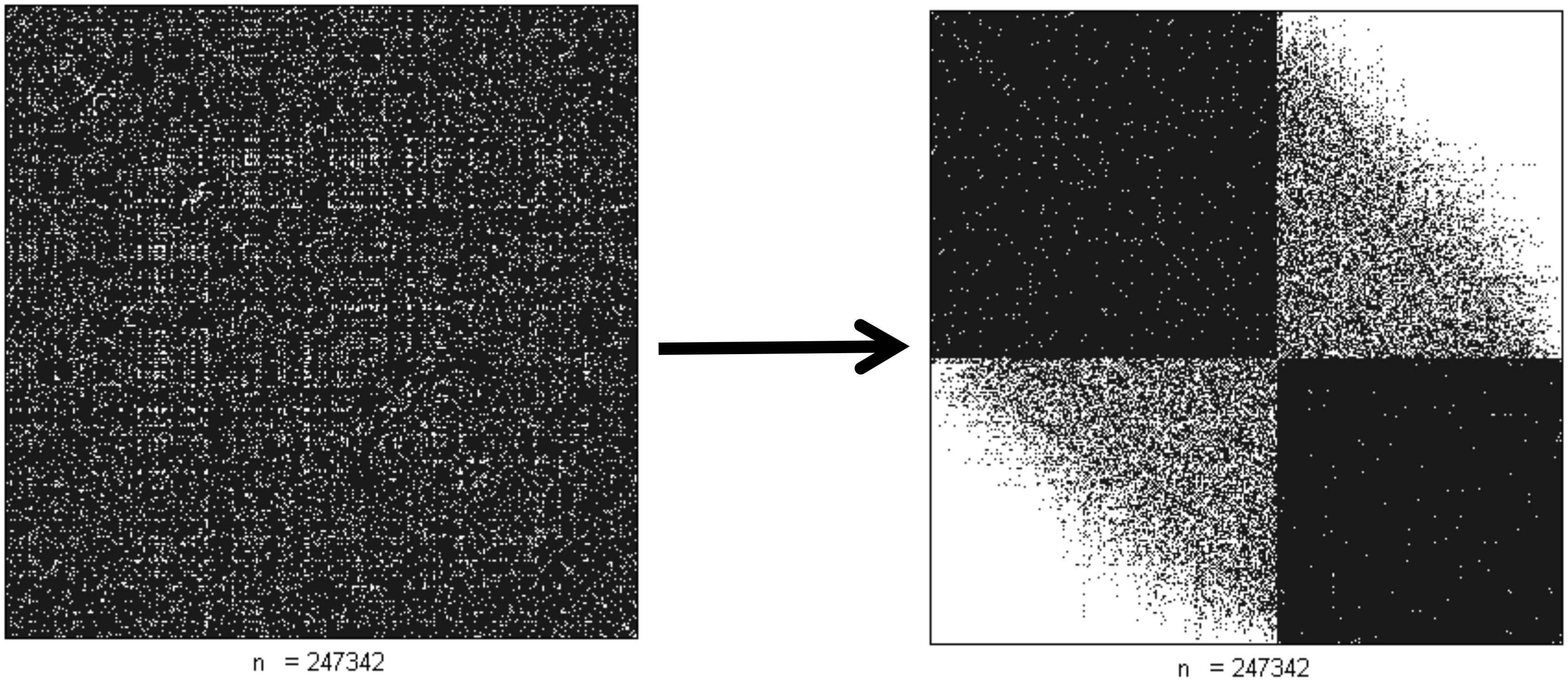
Girvan-Newman

- Detecção de cluster divisivo e hierarquico baseado na noção de betweenness:
- Número de caminhos mínimos que passam por cada aresta.
- Remover as aresta de modo decrescer o betweenness

Girvan, M. & Newman, M. E. J.
Community structure in social and biological networks
Proc. Natl. Acad. Sci. USA, 2002, *99*

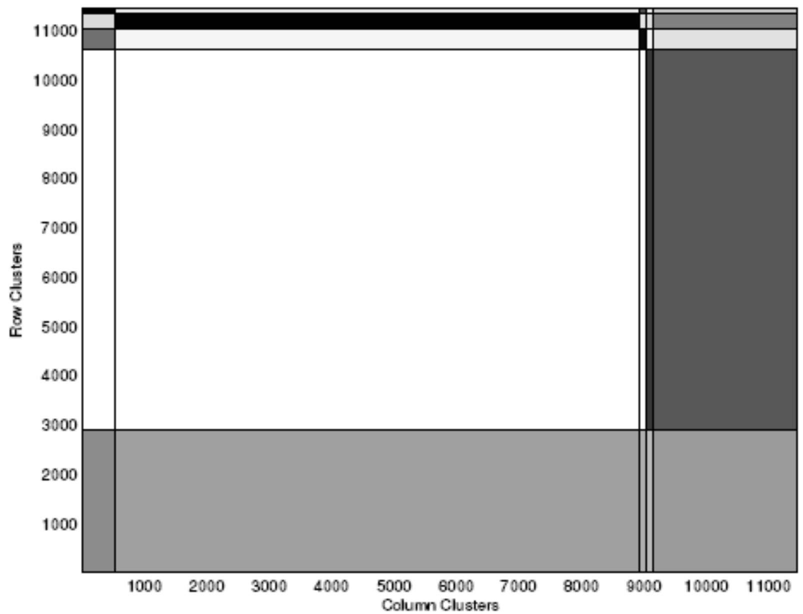


Spectral Partition

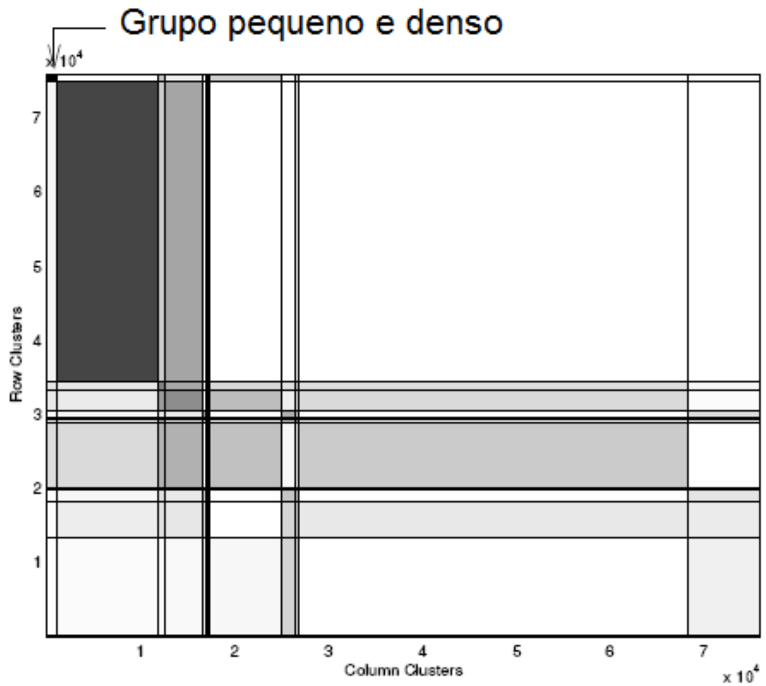


CrossAssociation

Chakrabarti, D.; Papadimitriou, S.; Modha, D. S. & Faloutsos, C.
Fully automatic cross-associations *KDD* 2004, 79-88



EPINIONS ($k^* = 18, \ell^* = 16$)

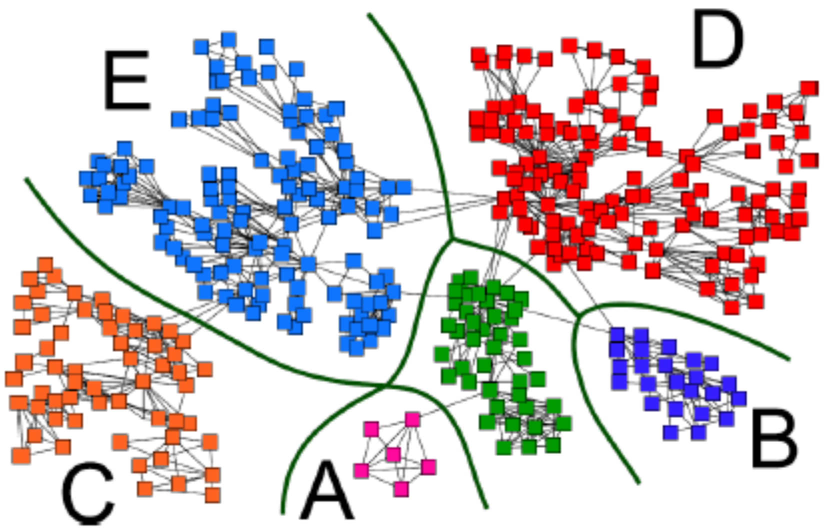


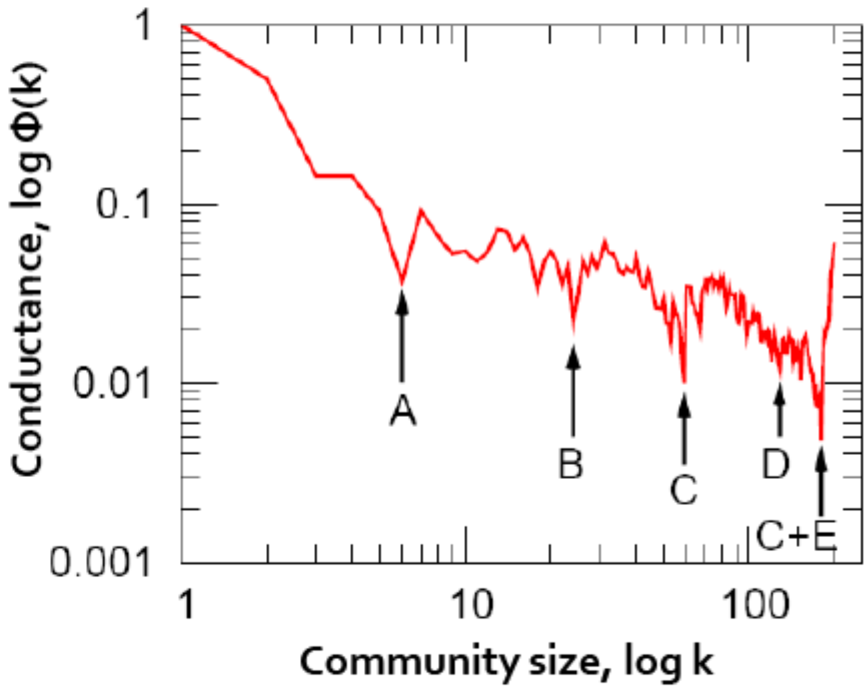
OREGON ($k^* = 9, \ell^* = 8$)

NCP - Plot

Collaborations between scientists in Networks

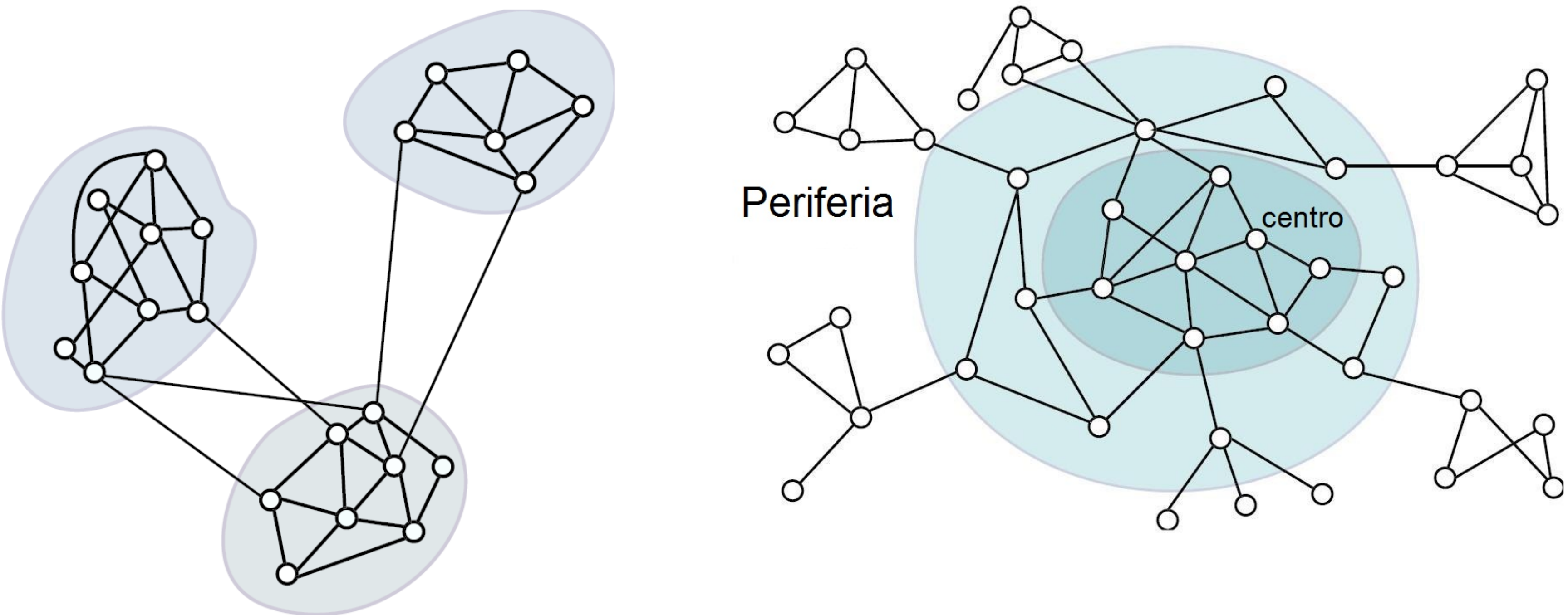
[Newman, 2005]





NCP - Plot

- Maioria das comunidade → cerca de 100 nós (número de Dunbar)
- Estrutura redes complexas grandes diferente das



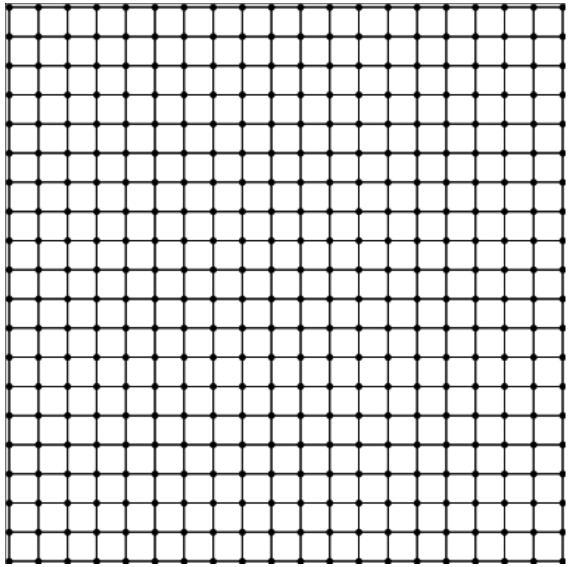
Models: Outline

- **The timeline of graph models:**
  - (Erdos-Renyi) Random graphs (1960s)
  - Exponential random graphs
  - Small-world model
  - Preferential attachment
  - Edge copying model
  - Community guided attachment
  - Forest fire
  - Kronecker graphs (today)

Graphs and networks

- What is the simplest way to generate a graph?
- Random graph model (Erdos-Renyi model, Poisson random graph model):
  - Given n vertices connect each pair i.i.d. with probability p
- How good ("realistic") is this graph generator?

Grafos vs. Redes Complexas

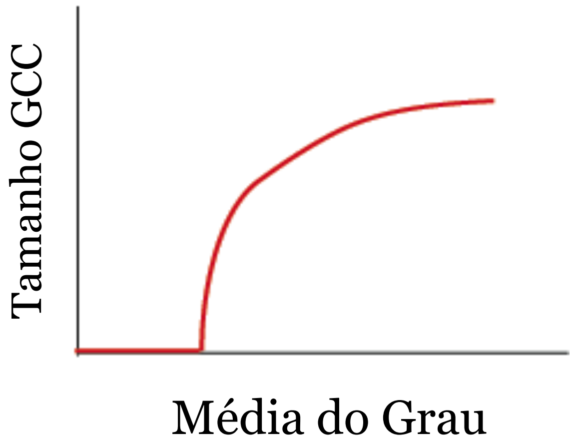


- Difere dos grafos tradicionais:
  - Grafos regulares (lattice)
- Novos grafos: Estrutura Complexa
- Grafos → Rede Complexa
- Modelo de grafo randômico:
  - Modelo Erdos-Renyi ou Poisson random graph model:
  - Dado *n* nós conectar cada par de nó com probabilidade p
- Não é um gerador muito realista.. Mais detalhes a seguir!!

(Erdos-Renyi) Random graph

- Also known as Poisson random graphs or Bernoulli graphs [Erdos&Renyi, 60s]
  - Given n vertices connect each pair i.i.d. with probability p
- Two variants:
  - $G_{n,p}$: graph with m edges appears with probability $p^m(1-p)^{M-m}$, where $M=0.5n(n-1)$ is the max number of edges
  - $G_{n,m}$: graphs with n nodes, m edges
- Does not mimic reality
- Very rich mathematical theory: many properties are exactly solvable

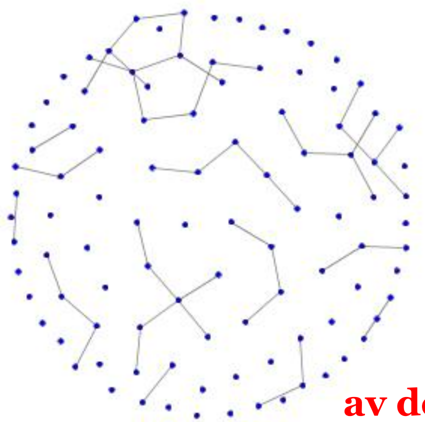
Fase de Transição

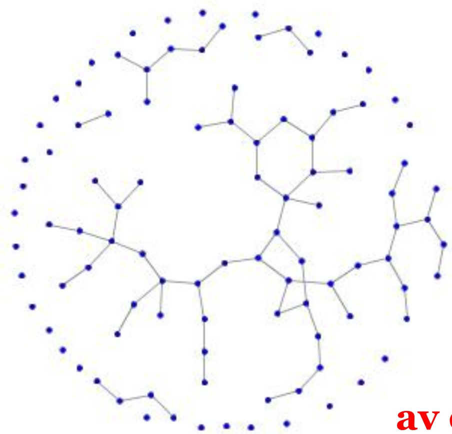


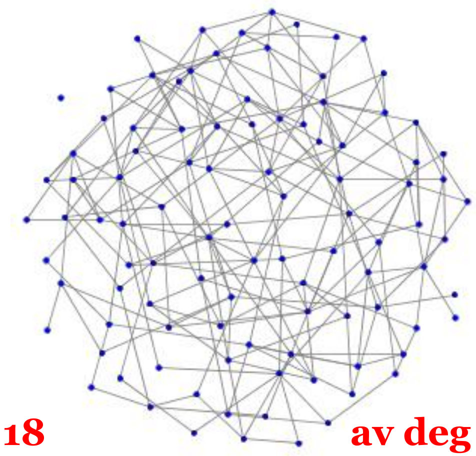
Fase de transição (Percolation threshold): Quantas arestas devem ser inseridas até a maior componente conexa aparecer?

Média do grau z = 1, a GCC aparece

z<1 rede desconexa e z> 1 rede fortemente conexa







Autovalores e autovetores

- Seja *A a matriz de adjacência do grafo*
- O autovalor λ é:
- $A\,v = \lambda\,v$, *na qual v é um vetor qualquer*
- Os autovalores são fortemente relacionados a topologia do grafo
- Por exemplo, ajudam a responder:
  - Quão importante é um nó?

Autovalores e autovetores

- Dependende se o grafo é representado com uma matriz de adjacência ou a Laplaciana os autovalores tem diferente significado.

- Laplaciana:
  - A multiplicidade do valor zero entre os autovalores de L(G) é igual ao número de componentes conexas.
  - O segundo menor autovetor é usado para detectar comunidades

$$L(u,v) = \begin{cases} d_v & \text{if } u = v, \\ -1 & \text{if } u \text{ and } v \text{ are adjacent}, \\ 0 & \text{otherwise.} \end{cases}$$

Propriedades

- Autovalores

$$\lambda_1 \geq \lambda_2 \geq \ldots . \geq \lambda_n$$

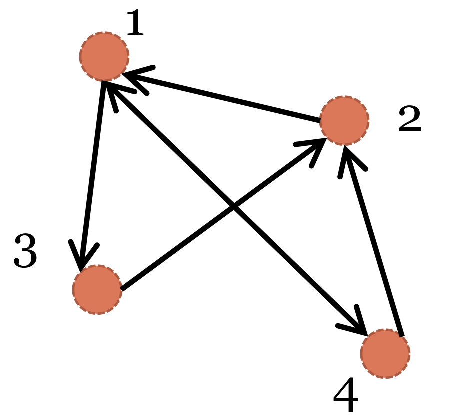
- Se a rede não possui ciclos

$$\lambda_1 = \sqrt{d}_{Max}$$

- A soma do quadrado dos autovalores é igual ao número de arestas da rede

$$\sum_{i=1}^{n} \lambda_i^2 = \sum_{i=1}^{n} d_i$$

Page Rank



PageRank é a distribuição de probabilidade usada para representar a verossimilhança que uma pessoa clica randomicamente em um link que vai para um determinada página

$$PR_{t+1} = (1\text{-}d)/n + d{*}A{*}PR_t$$

- PR é um vetor com o valor do PageRank da matriz A
- d é o fator de "pulo" e esta entre 0 < d < 1, usualmente é 0.85

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

Número de links saindo de uma página que aponta para a sua página. Quanto menos melhor

Número de links entrando. Quanto mais melhor.

Page, L.; Brin, S.; Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web *Stanford Digital Library,* 1998

Why should we care?

- Gives insight into the graph formation process:
  - *Anomaly detection* – abnormal behavior, evolution
  - *Predictions* – predicting future from the past
  - *Simulations* of new algorithms where real graphs are hard/ impossible to collect
  - *Graph sampling* – many real world graphs are too large to deal with
  - "What *if*" scenarios

Outline

- Part 1: Statistical properties of static and evolving networks.
  - Power law degree distributions found in static networks
  - Small world phenomena and six degrees of separation
  - Densification of time evolving networks
  - Shrinking diameters of growing networks
  - Communities and clusters in networks
- Part 2: Link predictions in complex networks.
  - Link Prediction
    - Link existence
    - Link weight
    - Link type
    - Link cardinality
  - Applications

Social Interaction on the Web

- Rich social structure in online computing applications
- Such structures are modeled by networks
- Most social network analyses view links as positive
  - Friends
  - Fans
  - Followers
- But generally links can convey either friendship or antagonism

Link Prediction via node distance

- Link prediction in a evolving network:
  - Task: Given $G[t_0,t_0']$ a graph on edges up to time t0' output a ranked list L of links (not in $G[t_0,t_0']$) that are predicted to appear in $G[t_1,t_1']$
  - Evaluation: $n=|E_{new}|$: # new edges that appear during the test period $[t_1,t_1']$
    Take top n elements of L and count correct edges

Link Prediction in Networks

- Network modeling is all about predicting links but so far we have not tackled this problem directly
- Task: predict missing links in a network
  - In a evolving network
  - In a static network
- 2 types of approaches:
  - Node distance approaches:
    - define a distance function, closer nodes are more likely to link
  - Statistical approaches:
    - Design a model of link creation and fit to data

Methods for Link Prediction

- Take the input graph during a training period [$G_0=(V,E)$]
- Pick a pair of nodes (u,v)
- Assign a connection weight score (u,v)
- Make a list in descending order of score
- Verify the prediction on the future graph [$G_1=(V,E_{new})$]

score is a measure of proximity / similarity

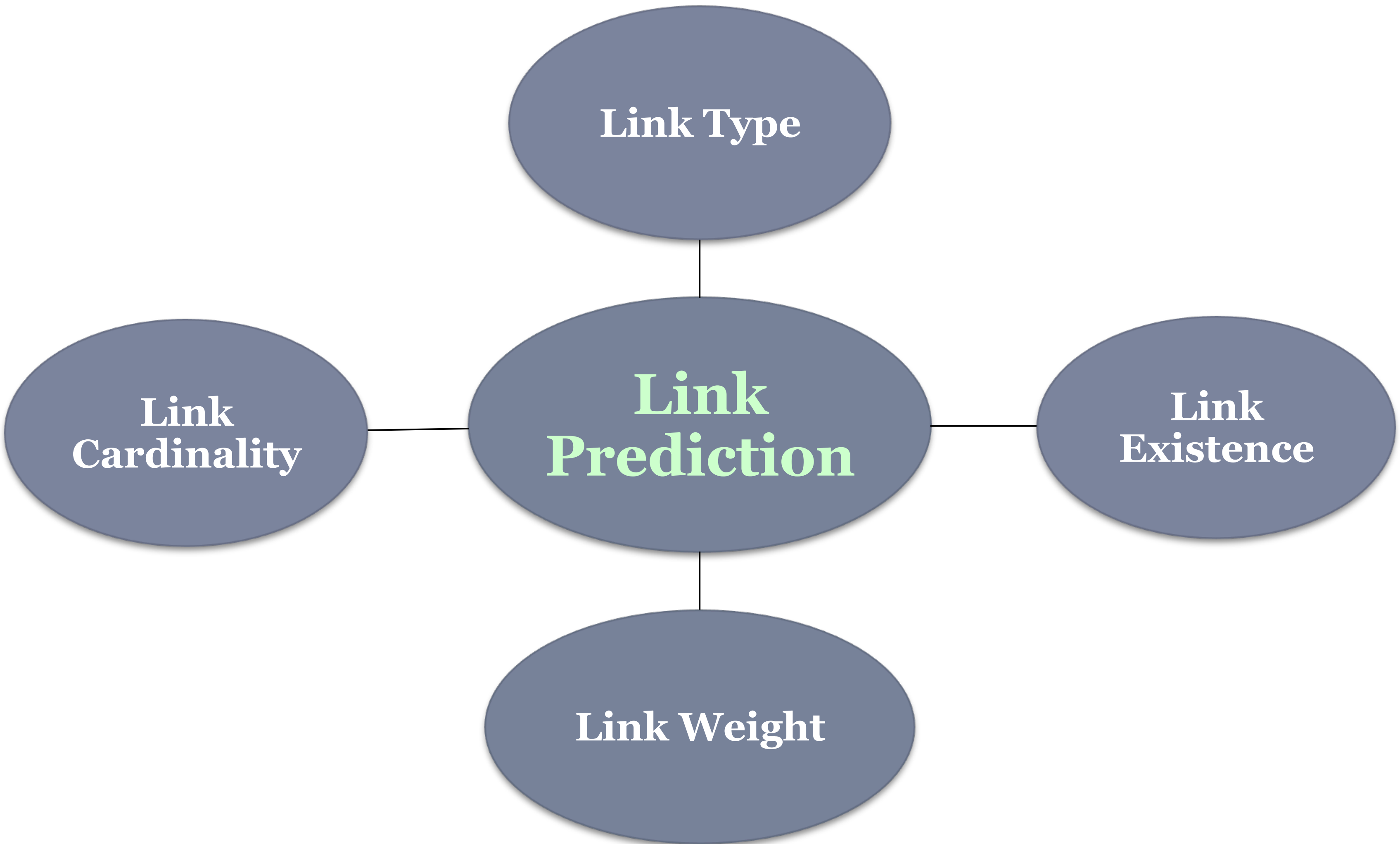
Methods for Link Prediction

- Node similarity can be defined by using the essential attributes of nodes:
  - two nodes are considered to be similar if they have many common features.
- The attributes of nodes are generally hidden
- Thus structural similarity is used, which is based solely on the network structure.

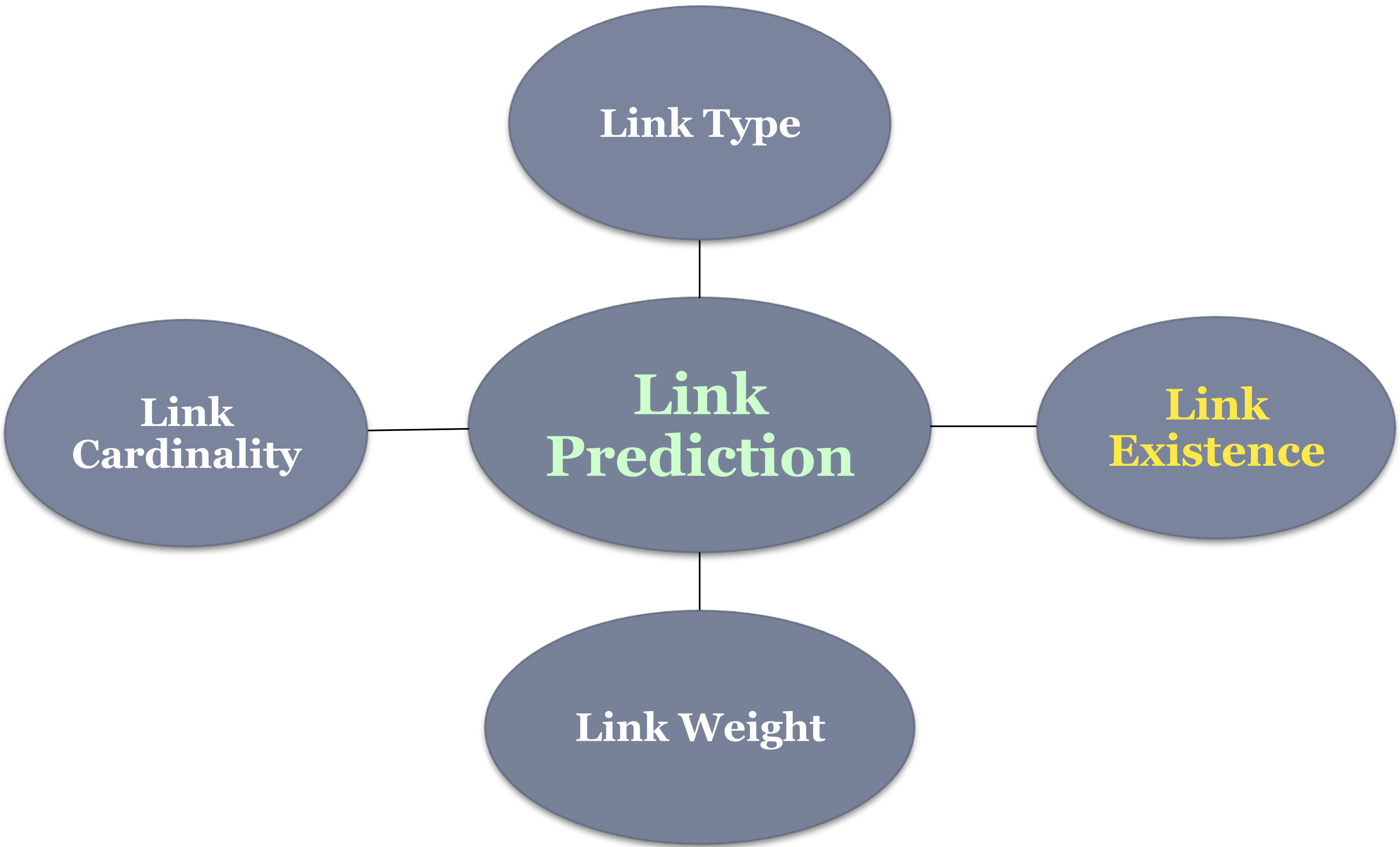
Reminder

- If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future

Link Prediction Task



Link Prediction Task



Unsupervised Link Prediction

- Unsupervised measurements could rely on different structural property:
- Neighborhood measures
  - Common Neighbors, Adamic Adar, Jaccard, Preferential Attachment
- Path-based measures
  - Graph distance, Katz
- Ranking
  - Sim Rank, Hitting time, Page Rank

Neighborhood Measures

- "How many friends we have to share in order to become friends?"
- Common Neighbors: the more friends we share, the more likely that we will become friends

$$\mathsf{score}(x, y) := |\Gamma(x) \cap \Gamma(y)|$$

Neighborhood Measures

- Jaccard: the more similar our friends circles are, the more likely that we will become friends

$$\mathsf{score}(x, y) := |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$$

Neighborhood Measures

- Adamic Adar: the more selective our mutual friends are, the more likely that we will become friends

$$\mathsf{score}(x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}.$$

Neighborhood Measures

- Preferential Attachment: more friends we have, the more likely that we will become friends

$$\text{score}(x, y) := |\Gamma(x)| \cdot |\Gamma(y)|.$$

Path-based Measures

- ”How distant we are?”
- Graph Distance: (negated) length of shortest path between u & v
- Katz$_\beta$: weighted sum over all the paths between u & v

$$score(u,v) = \sum_{l=1}^{\infty} \beta^l \left| paths_{u,v}^{\langle l \rangle} \right|$$

- where: $paths^{\langle l \rangle}_{u,v}$ ={paths of length exactly l from u to v}

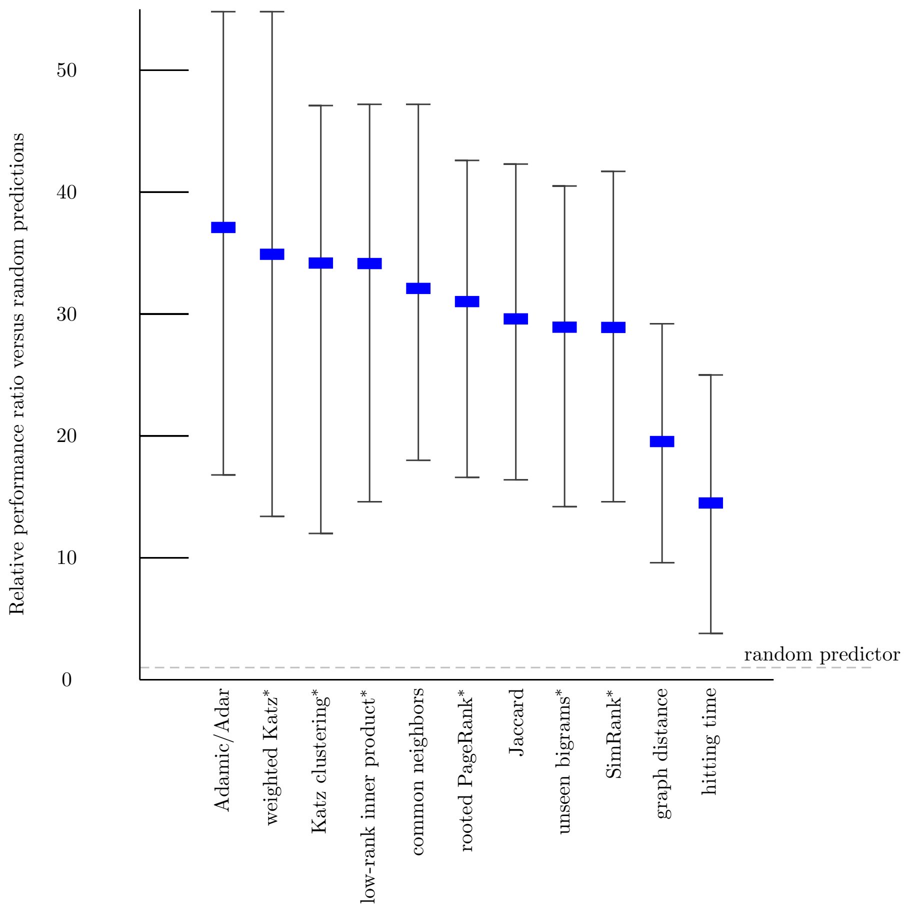
SimRank

- "Two nodes are similar to the extent that they are joined by similar neighbors"

$$similarity(u, v) = \gamma * \frac{\sum_{a \in \Gamma(u)} \sum_{n \in \Gamma(v)} similarity(a,b)}{|\Gamma(u)| * |\Gamma(v)|}$$

$$score(u, v) = similarity(u, v)$$

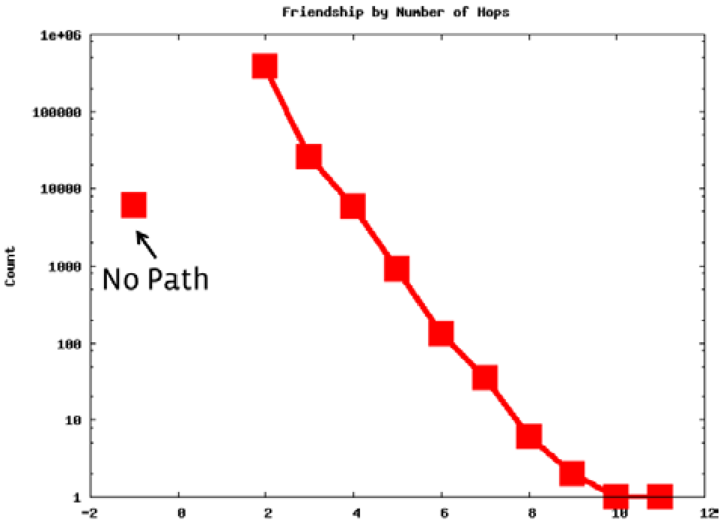
Comparison

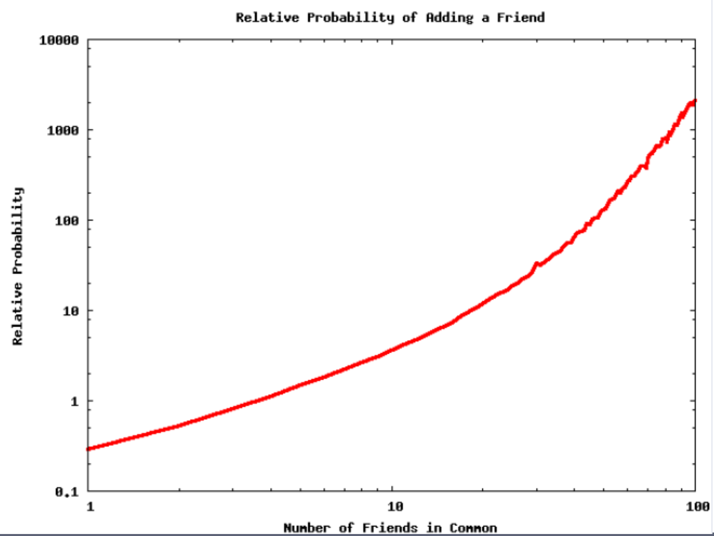


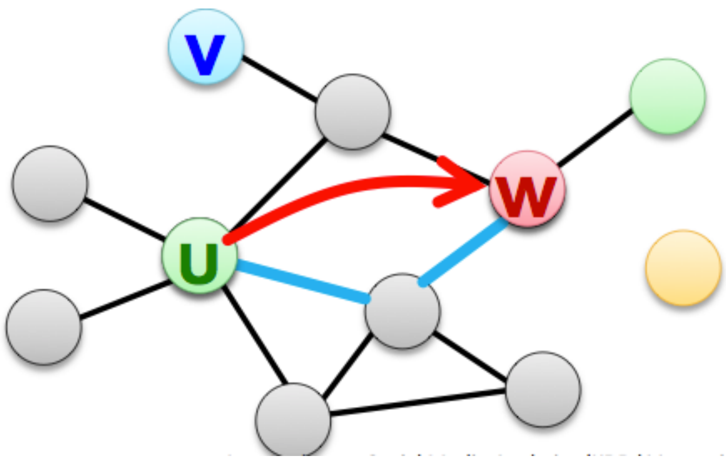
Relative performance ratio versus random predictions
50
40
30
20
10
0
random predictor
Adamic/Adar
weighted Katz*
Katz clustering*
low-rank inner product*
common neighbors
rooted PageRank*
Jaccard
unseen bigrams*
SimRank*
graph distance
hitting time

Supervised Link Prediction

- How to learn to predict new friends in networks?
- Facebook's People You May Know
- Looking at the data:
  - 92% of new friendships on FB are friend-of-a-friend
  - More common friends helps







Supervised Link Prediction

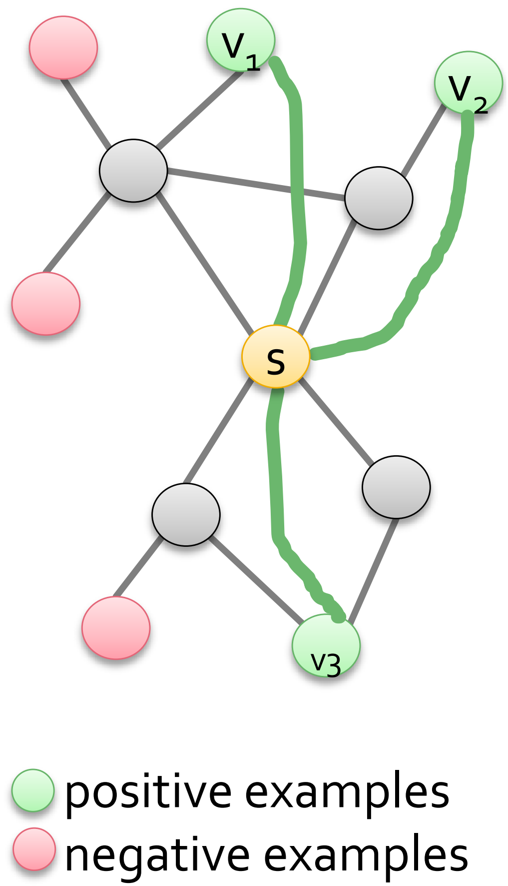
- How do characteristics of users (e.g., age, gender, home town) interact with the creation of new edges?
- In a social network, there can be many reasons exogenous to the network for two users to become connected:
  - it could be that they met at a party, and then connected on it.
    - Same age, Same town
  - this link might also be hinted at by the structure of the network:
    - two people are more likely to meet at the same party if they are "close"
- A pair of people likely has friends in common, and travel in similar social circles.
- Despite the exogenous event (i.e., a party) there are clues in social networks which suggest a high probability of a future friendship.

Supervised Link Prediction

- *Supervised Random Walks*
  - combines the network structure
  - the characteristics (attributes, features) of nodes
  - edges strengths of the network.
- *Supervised* way learns how to bias a PageRank-like random walk on the network
  - Visits given nodes (i.e., positive training examples) more often than the others.
  - Positive nodes are nodes to which new edges will be created in the future
  - Negative are all other nodes

Supervised Link Prediction

- Recommend a list of possible friends
- Supervised machine learning setting:
  - Training example:
    - For every node *s* have a list of nodes that will create links to *{v1, …, vk}*
  - Problem:
    - For a given node *s* learn to rank nodes *{v1, …, vk}* higher than other nodes in the network
- Supervised Random Walks based on word by Agarwal&Chakrabarti



Prophet + NELL

- Can computers learn to read? We think so.
- "Read the Web" is a CMU research project that attempts to create a computer system that learns over time to read the web.
- Since January 2010, the computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:
  - First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., playsInstrument(George_Harrison, guitar)).
  - Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

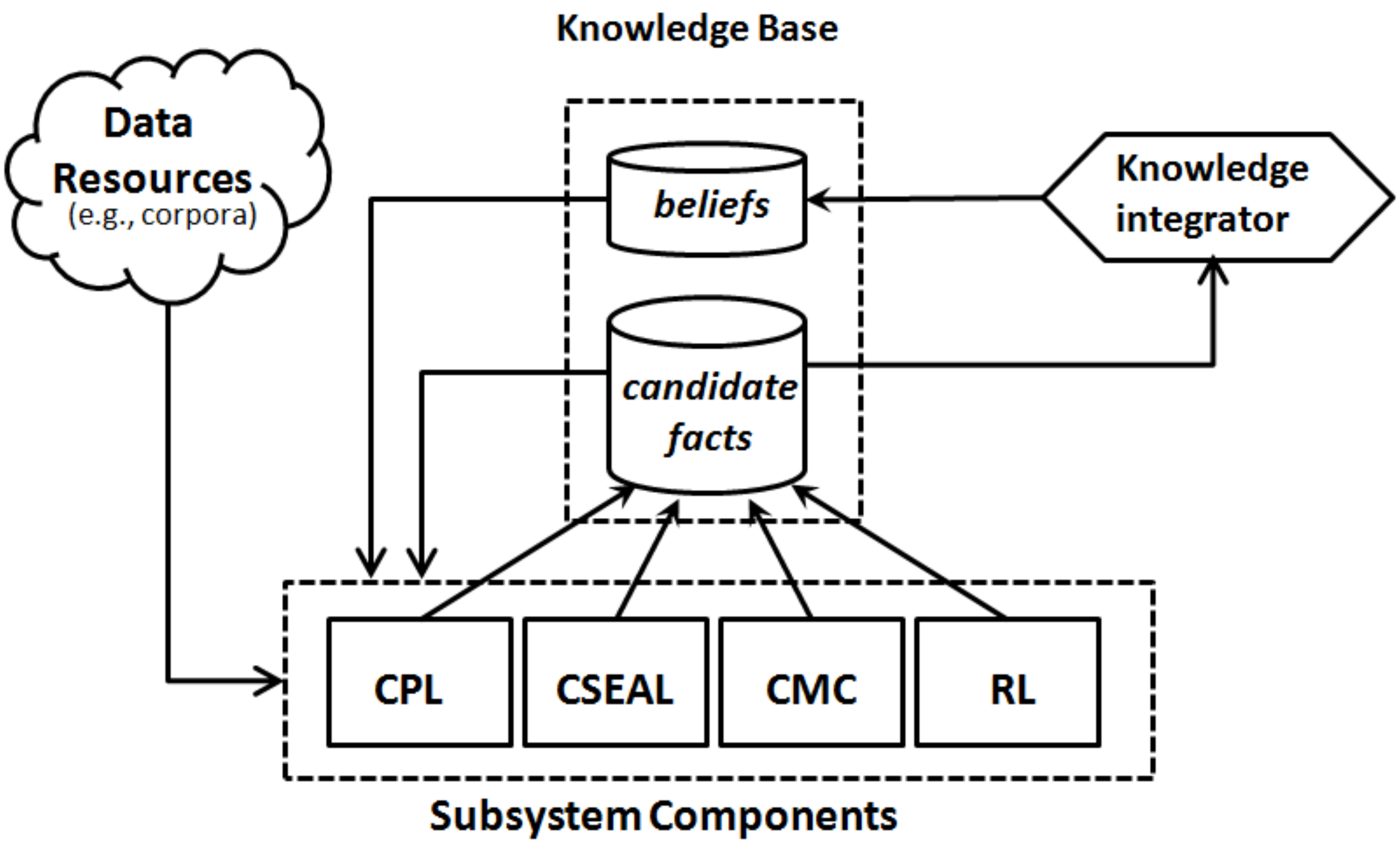
NELL: Never-Ending Language Learner

- Inputs:
  - initial ontology
  - handful of examples of each predicate in ontology
  - the web
  - occasional interaction with human trainers
- The task:
  - run 24x7, forever
  - each day:
    1. extract more facts from the web to populate the initial ontology
    2. learn to read (perform #1) better than yesterday

NELL: Never-Ending Language Learner

- Today…
- Running 24 x 7, since January, 2010
- Input:
  - ontology defining ~500 categories and relations
  - 10-20 seed examples of each
  - 500 million web pages (ClueWeb – Jamie Callan)
- Result:
  - continuously growing KB with ~440,000 extracted beliefs

NELL



Knowledge Base
Data Resources (e.g., corpora)
beliefs
candidate facts
Knowledge integrator
CPL
CSEAL
CMC
RL
Subsystem Components

Read The Web Project

- http://rtw.ml.cmu.edu

Recently-Learned Facts twitter

Refresh

instance	iteration	date learned	confidence
gasparilla island beach is a beach	427	27-sep-2011	100.0
abstract strategy games is a board game	430	07-oct-2011	98.6
visual thinking seminar is a cognitive action	430	07-oct-2011	100.0
senescent fish is a mollusk	431	08-oct-2011	96.6
andrew cockburn is a person	428	29-sep-2011	95.0
english is a language used in the university harvard college	430	07-oct-2011	99.2
dorothy chandler pavilion is a stadium or event venue located in the city los angeles	430	07-oct-2011	96.9
hitachi has acquired ibm	428	29-sep-2011	93.8
randy walker coaches the team northwestern oklahoma state university	431	08-oct-2011	93.8
kusf is a radio station in the city san francisco	427	27-sep-2011	96.9

wledge Base Browser
/eb Project

log in | preferences | help/instruct

s | relations

ributedtocreativework
rredinmovie
otebook
directedmovie
ysport
rkcontributedtobyagent
ter
aractor
rectedbydirector
trophy
conomicsector
rrency
ountry
nt
ectorcompany
tusedbysport

te
gamedate
irthdate
eathdate
existsatdate
tsatdate
ectedreleasedate
nyear
edwithagent

nba (sportsleague)

literal strings: NBA, nba, Nba

Help NELL Learn!

NELL wants to know if these beliefs are correct.
If they are or ever were, click thumbs-up. Otherwise, click thumbs-down.

- nba is a sports league
- chuck daly coaches in the league nba (sportsleague)
- doc sadler coaches in the league nba (sportsleague)
- jay triano coaches in the league nba (sportsleague)
- pat riley coaches in the league nba (sportsleague)

---

categories

- **sportsleague**(100.0%)
  - CPL @155 (100.0%) on 28-sep-2010 ["favorite player with _" "_ 's New Jersey Nets" "watch featuring _" "undesirable such as _" "_ 's Dallas Mavericks" "_ rebounding title" "_ 's Orlando Magic" "only winless team in _" "_ 's Eastern Conf shooting guards" "product is officially licensed by _" "_ 's Western Conference"] using nba
  - Seed
  - MBL @215 (75.0%) on 02-mar-2011 [Promotion of "sportsleague:nba" leaguestadiums "attraction:us_bank_arena"]
  - SEAL @172 (98.4%) on 05-dec-2010 [1 2 3 4 5 6] using nba

New Categories

biotechcompany	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bird	12	237	74	69	46	43	26	28	25	32	24	16	27	28	48	28	24	35	37	26	35	27	47	199	155	153	97	42	25	26	49	42	27	52	32	27	50	33	27	26
blog	9	92	28	28	23	17	18	6	25	13	11	7	2	2	10	8	4	14	13	7	8	3	6	5	2	2	2	6	3	3	5	8	5	21	13	13	17	2	20	13
boardgame	3	6	0	1	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bodypart	7	3	2	6	14	15	12	9	12	11	18	12	16	13	10	19	11	15	9	13	16	14	9	16	4	4	15	13	19	9	8	14	0	0	0	0	0	0	0	0
bone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
book	7	31	13	0	0	12	4	2	8	3	2	0	0	0	1	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
braintissue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
building	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
buildingfeature	13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
buildingmaterial	5	0	0	0	0	0	0	0	0	0	0	2	0	0	5	3	3	1	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
candy	10	16	22	2	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cardgame	2	4	1	1	2	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cave	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
celebrity	13	204	90	178	195	113	69	83	135	98	60	113	66	106	73	31	66	57	42	72	28	68	70	54	73	91	101	72	46	84	69	72	62	46	87	52	38	87	66	29
celltype	2	0	0	1	2	0	0	1	1	17	6	7	6	4	7	13	11	9	9	6	7	12	7	4	4	7	10	1	9	12	8	14	9	14	12	5	8	12	12	7
ceo	19	21	13	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
charactertrait	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cheese	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chef	12	112	9	12	8	19	10	3	7	24	4	0	0	0	3	0	4	18	11	4	0	0	1	4	3	4	7	1	0	1	0	0	0	0	0	1	0	0	0	0
chemical	10	131	107	48	33	30	36	37	40	28	40	30	32	27	34	27	29	27	28	26	31	27	26	25	31	46	30	27	29	26	28	28	25	25	30	28	25	126	71	26
city	21	117	151	164	176	191	196	209	198	197	227	172	182	110	130	118	147	174	95	108	106	119	111	158	87	122	165	100	199	89	147	121	149	184	135	204	136	31	56	104
clothing	15	53	56	67	44	37	39	30	28	30	26	27	26	26	23	25	27	24	29	25	27	30	30	34	30	29	29	25	26	33	31	28	27	4	3	30	38	33	43	27
coach	0	12	5	6	12	11	4	26	13	5	2	1	2	0	0	1	2	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0

New Relations

y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eofanimal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ators	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
son	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hasanimal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
city	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nownas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
estadium	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sforteam	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sinleague	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ssport	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ssportsteamposition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
city	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
roducesmodel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
roducedbymaker	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ntainsbodypart	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
thinbodypart	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	6	4	8	1	7	1	5	4	9	4	2	1	2	1	0	2	3	0	0	2	0	0	0	0	2	4	0	0	0	0	0	0	0	2	0	1	0	0	0	1	0	0	3	0	1	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
atedincity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ypeofchemical	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

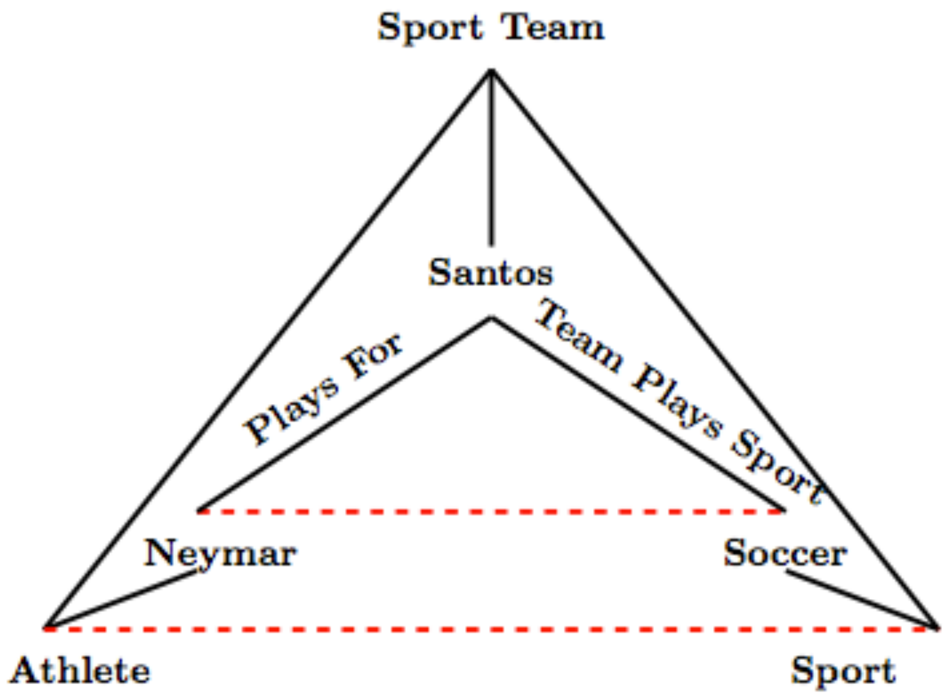
Discover New Coupling Constraints

- first order, probabilistic horn clause constraints
  - connects previously uncoupled relation predicates
  - infers new beliefs for KB

$$0.93 \quad \text{athletePlaysSport}(?x,?y) \leftarrow \text{athletePlaysForTeam}(?x,?z) \; \text{teamPlaysSport}(?z,?y)$$

Problem

- How can NELL learn new relations?
- Specially the hidden ones ?



Solution

- NELL knowledge base is an ontology
- A ontology can be mapped as a graph (rtwgraph)
- Thus we can apply graph mining techniques

Prophet (DaMNet 2011)

- A link prediction component coupled to NELL to help the automatic ontology extension that predicts new rules and relations with a higher accuracy.
- The goal is to extend the traditional link prediction task to be applied in complex network data that represents knowledge extracted from the Web and thus predicts (infer) new relations and rules that are presented by edges.
- The results show that the use of a common neighboring measure with some heuristics helps NELL learn more and better.

Motivation

- During the extraction phase there are some knowledge that NELL is not be able to learn.
- “Milwaukee Bucks is a basketball sport team which plays for NBA league.”
- NELL will be able of extract only
  - rules SportTeam and TeamPlaysInLeague in its beliefs.
    - `SportTeam(Basketball,Milwaukee Bucks)`
    - `TeamPlaysInLeague(Milwaukee Bucks, NBA)`

Data

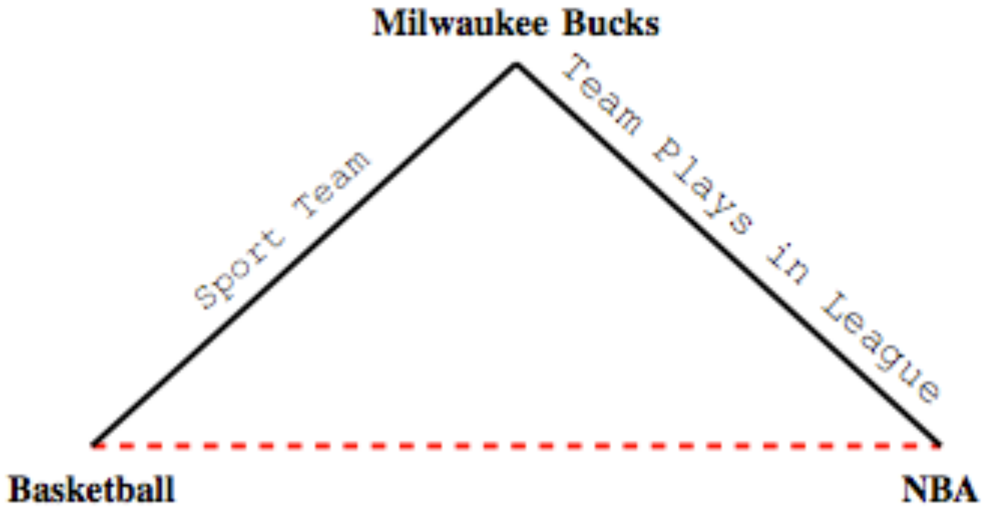
- Two Graphs
  - RTWGRAPH → instanced graph
  - Rule graph → rules
- Which one should we use? → Both
  - RTWGRAPH → redundancy
  - Rules → few information

Goals

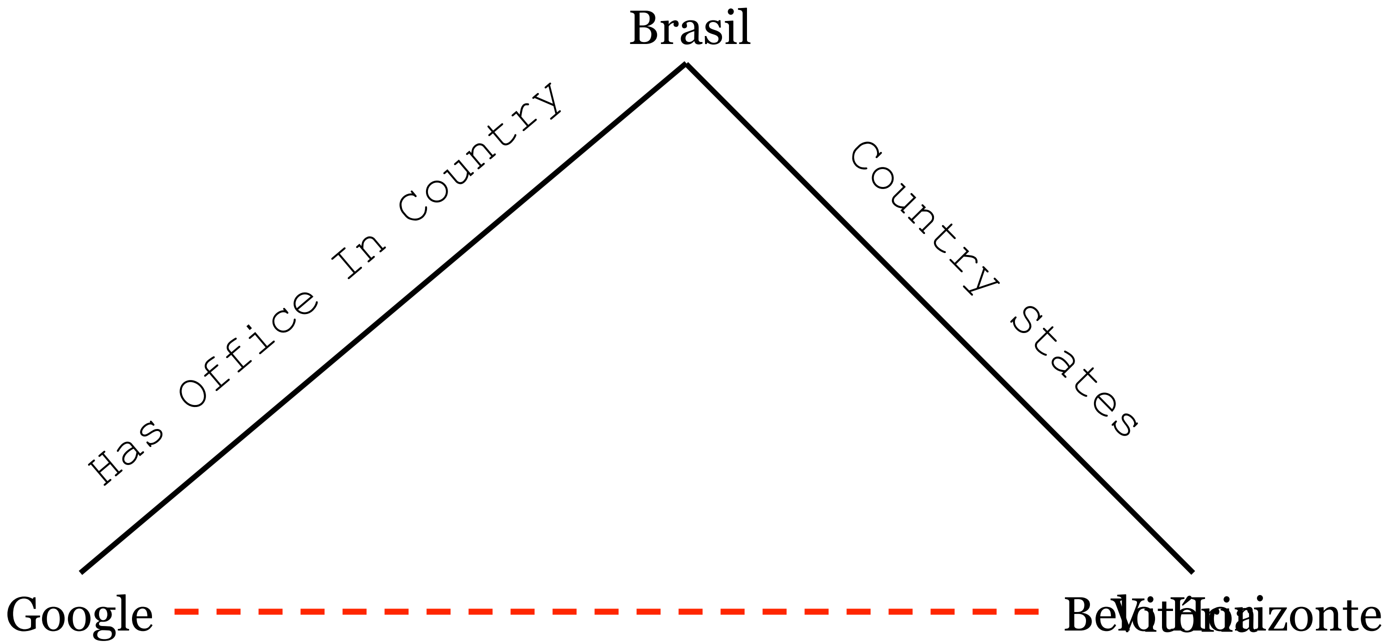
1. Extend the KB by predicting new relations (edges)that might exist between pairs of nodes;
2. Predict new rules that might help NELL learn more and better;
3. Identify misplaced edges which can be used by NELL as hints to identify wrong connections between nodes (wrong knowledge);

New Relations

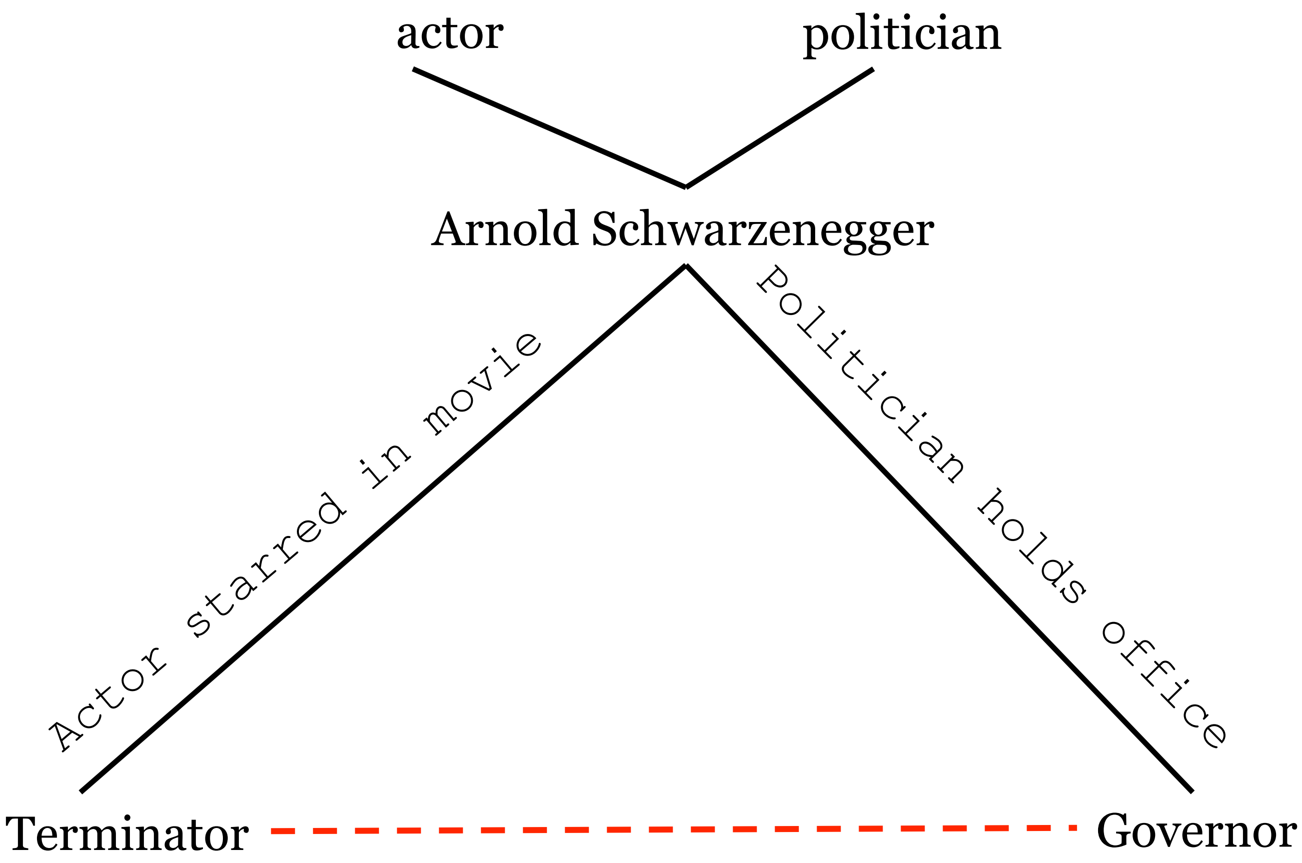
- New relations → just close triangles ???



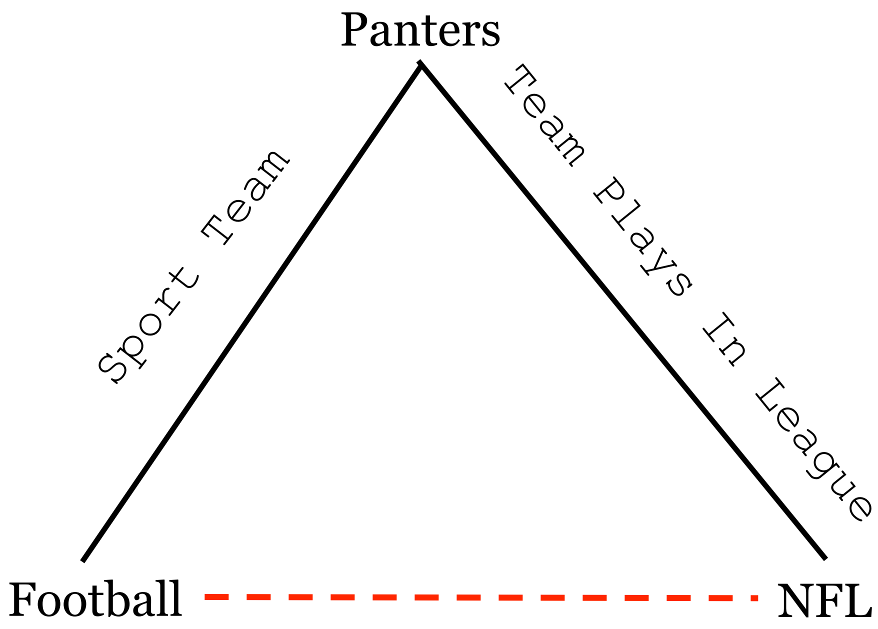
Wrong combinations

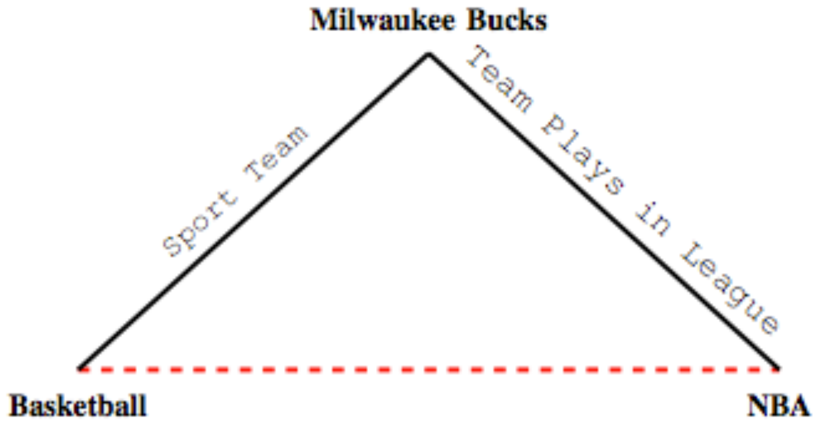


Different Categories



Redundancy


Panters
Sport Team
Team Plays In League
Football
NFL

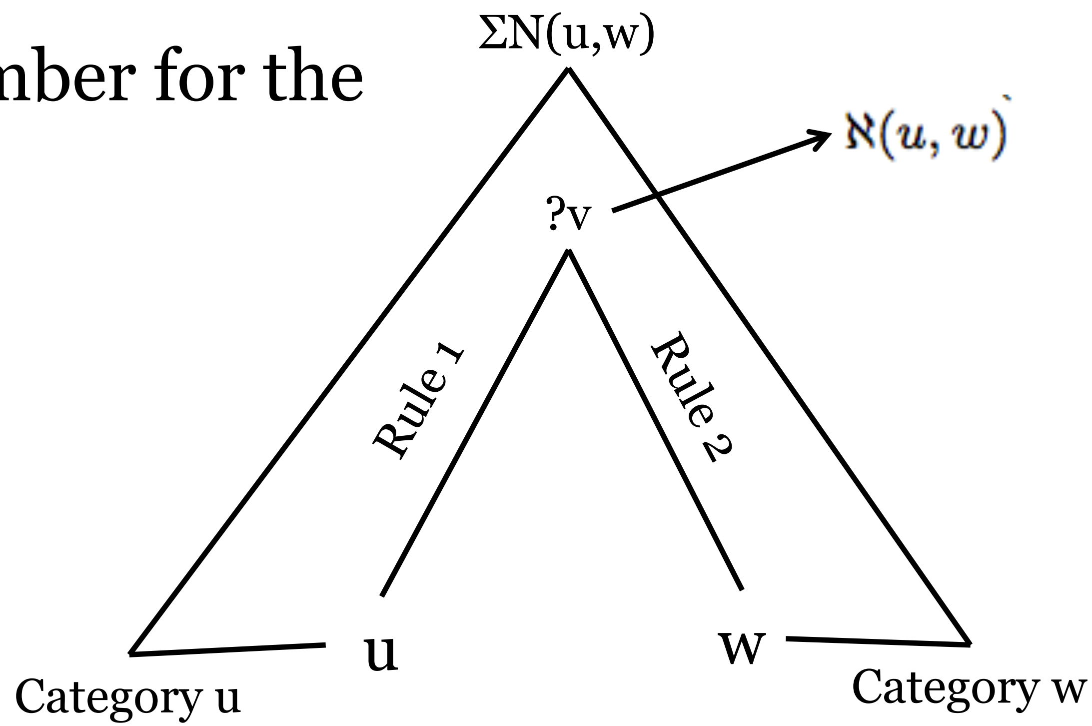

Milwaukee Bucks
Sport Team
Team Plays in League
Basketball
NBA

Prophet

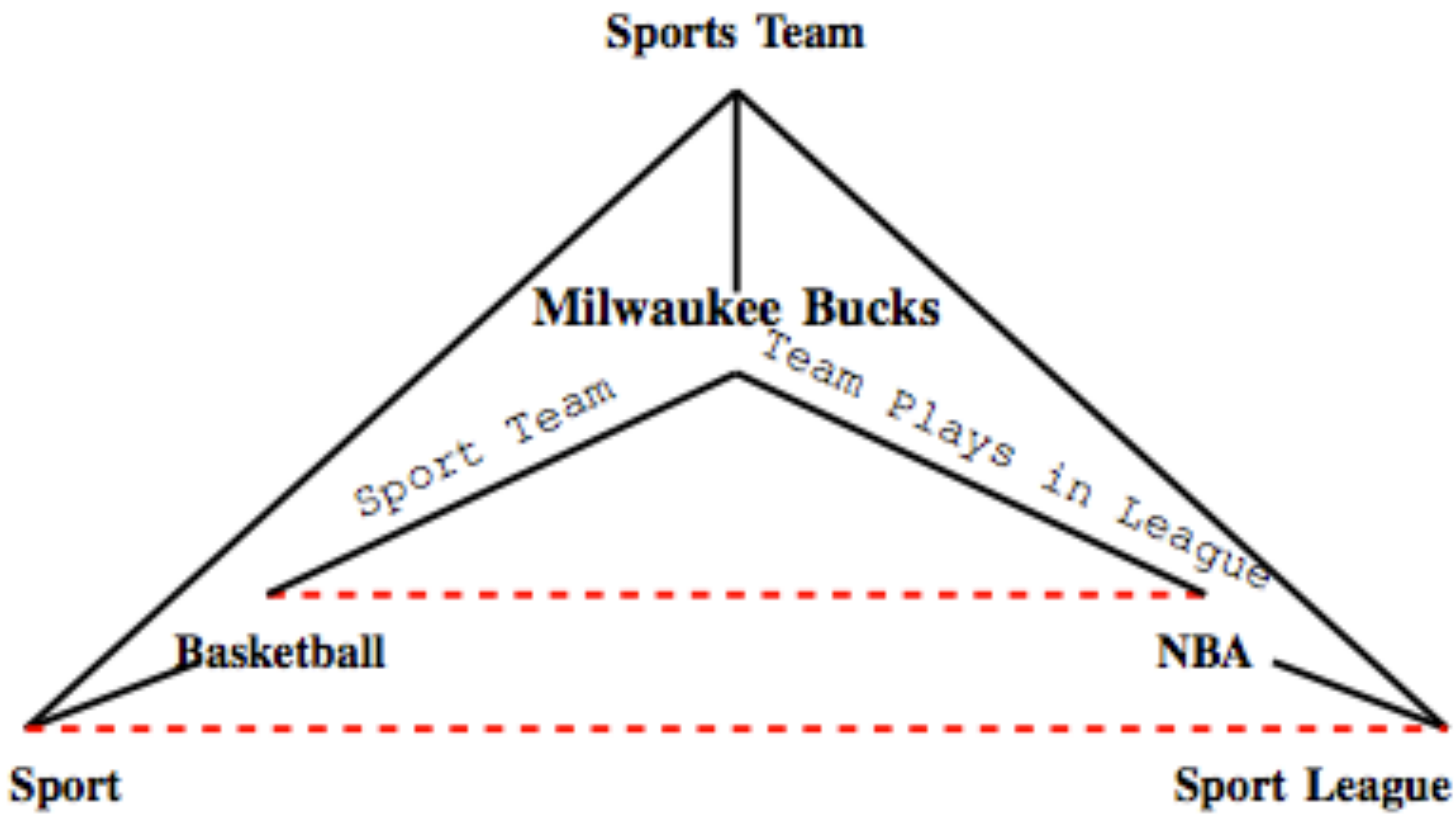
- First all open triangles are found
  - Combining both graphs RTWGRAPH+Rules
    - Avoid combine instances from different categories

Prophet

- Compute the number of common neighbors
- For instanced nodes u and v and
- The cumulative number for the categories nodes



Example


Sports Team
Milwaukee Bucks
Sport Team
Team Plays in League
Basketball
NBA
Sport
Sport League

Prophet

- Problem
  - Rules with more instances have high probability of have more common neighbors

Prophet

U	V	W	vizinhos	total
awardtrophytournament	coach sportsteam	athlete	5 1190	5 1162
awardtrophytournament	sportsteam	sport	217	47
awardtrophytournament	coach sportsteam	sportsleague	4 236	3 53
awardtrophytournament	sportsteam	stadiumoreventvenue	164	122
city	company	economicsector	205	178
company	city	newspaper	2225	2212
company	city country	stateorprovince	738 233	669 233
currency	country	stateorprovince	201	138
economicsector	company	city	190	165
sport	sportsteam	awardtrophytournament	234	55
sport	athlete sportsteam	coach	12 127	12 116
sport	athlete sportsteam stadiumoreventvenue	sportsleague	716 249 5	12 17 4
sportsleague	coach sportsteam	awardtrophytournament	4 243	3 58
sportsleague	athlete sportsteam stadiumoreventvenue	sport	716 244 5	12 13 4
stadiumoreventvenue	sportsteam	awardtrophytournament	170	127
stateorprovince	city country	company	859 193	780 193

Prophet

- Normalize the cumulative number of neighbors

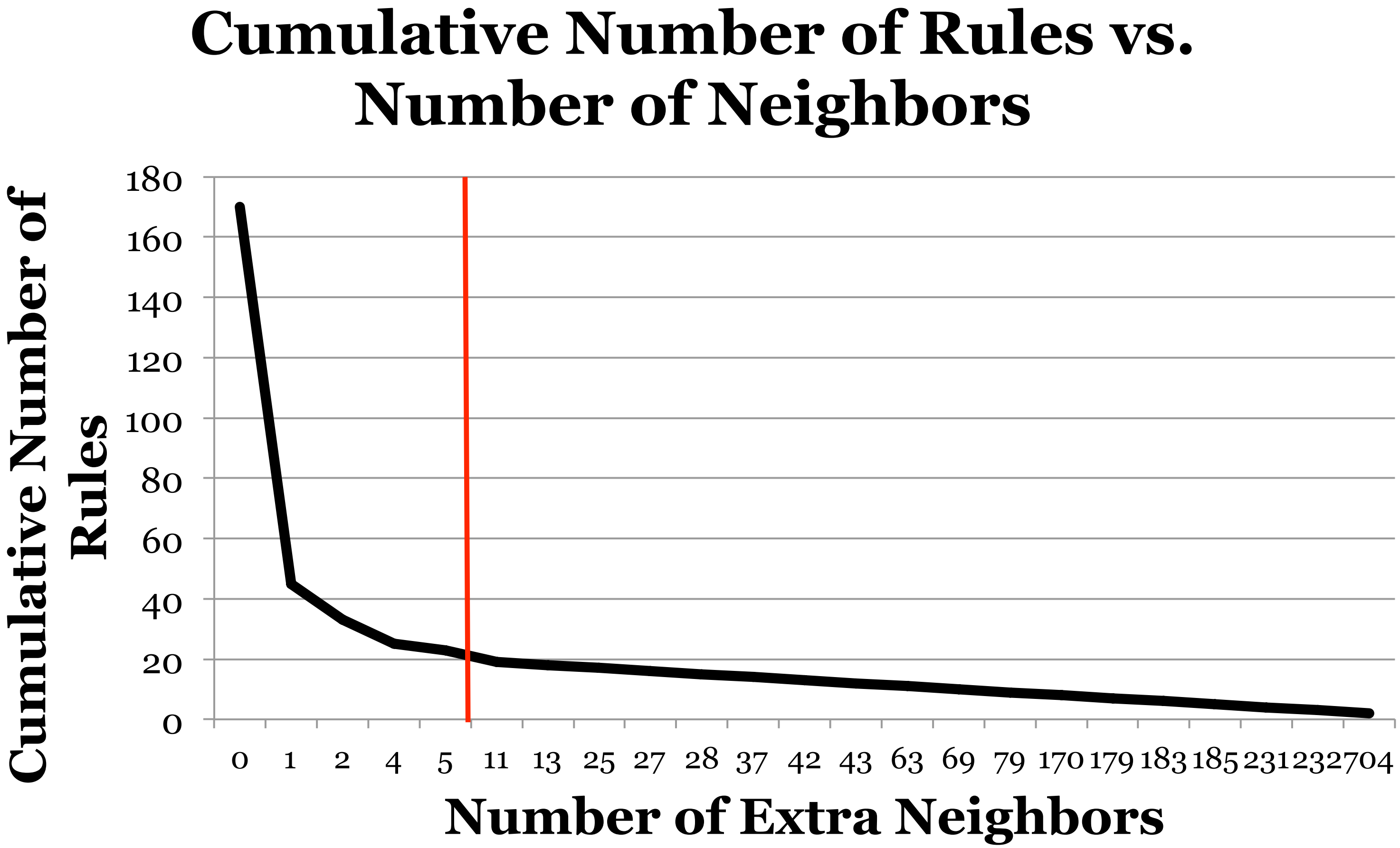
$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$

- $N_c(u_c,w_c)=0$ → all instanced rule only one neighbor
- $N_c(u_c,w_c)>\xi$ → select rules

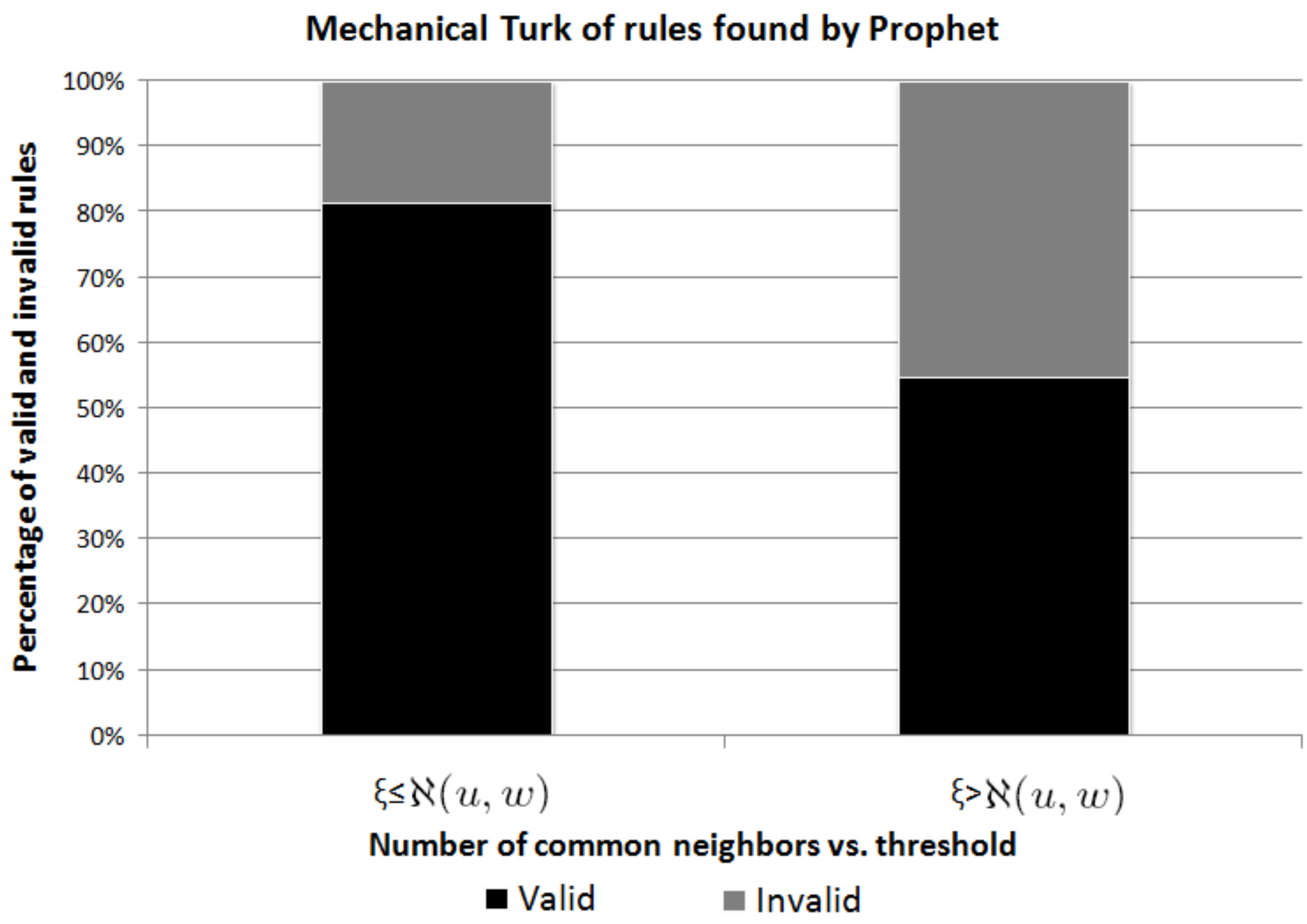
Prophet

U	V	W	vizinhos	total	diferença vizinhos
awardtrophytournament	coach	athlete	5	5	0
	sportsteam		1190	1162	28
awardtrophytournament	sportsteam	sport	217	47	170
awardtrophytournament	coach	sportsleague	4	3	1
	sportsteam		236	53	183
awardtrophytournament	sportsteam	stadiumoreventvenue	164	122	42
city	company	economicsector	205	178	27
company	city	newspaper	2225	2212	13
company	city	stateorprovince	738	669	69
	country		233	233	0
currency	country	stateorprovince	201	138	63
economicsector	company	city	190	165	25
sport	sportsteam	awardtrophytournament	234	55	179
sport	athlete	coach	12	12	0
	sportsteam		127	116	11
sport	athlete	sportsleague	716	12	704
	sportsteam		249	17	232
	stadiumoreventvenue		5	4	1
sportsleague	coach	awardtrophytournament	4	3	1
	sportsteam		243	58	185
sportsleague	athlete	sport	716	12	704
	sportsteam		244	13	231
	stadiumoreventvenue		5	4	1
stadiumoreventvenue	sportsteam	awardtrophytournament	170	127	43
stateorprovince	city	company	859	780	79
	country		193	193	0

Prophet


Cumulative Number of Rules vs. Number of Neighbors
Cumulative Number of Rules
180
160
140
120
100
80
60
40
20
0
0 1 2 4 5 11 13 25 27 28 37 42 43 63 69 79 170 179 183 185 231 232 704
Number of Extra Neighbors

Mechanical Turkey

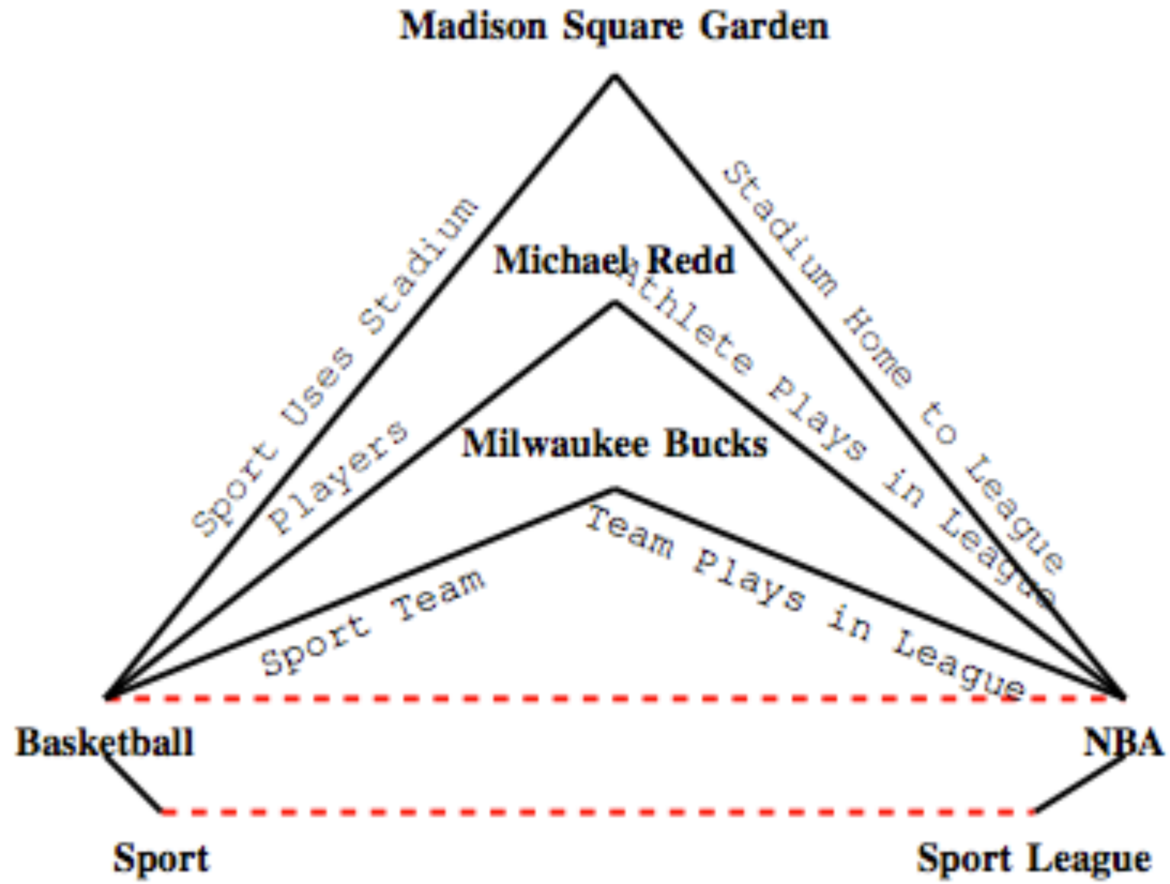


Prophet

- Another restriction to create the instances
  - Number of independent paths

- If the number of independent path is less than the original number of paths → the number of common neighbors (>ξ) is taken into account

Example



Madison Square Garden
Sport Uses Stadium
Stadium Home to League
Michael Redd
Players
Athlete Plays in League
Milwaukee Bucks
Sport Team
Team Plays in League
Basketball
NBA
Sport
Sport League

Example

Entity	U	relation	V	relation	Entidty	W	Neighbors
baseball	sport	players	athlete	athleteplaysinleague	major_league_baseball	sportsleague	13
baseball	sport	players	athlete	athleteplaysinleague	mlb	sportsleague	429
baseball	sport	sportteam	sportsteam	teamplaysinleague	mlb	sportsleague	40
baseball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	mlb	sportsleague	1
baseball	sport	players	athlete	athleteplaysinleague	nfl	sportsleague	1
baseball	sport	sportteam	sportsteam	teamplaysinleague	nfl	sportsleague	2
baseball	sport	sportteam	sportsteam	teamplaysinleague	nhl	sportsleague	1
soccer	sport	players	athlete	athleteplaysinleague	nba	sportsleague	1
basketball	sport	players	athlete	athleteplaysinleague	nba	sportsleague	44
basketball	sport	sportteam	sportsteam	teamplaysinleague	nba	sportsleague	58
basketball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	nba	sportsleague	1

Example

Entity	U	relation	V	relation	Entidty	W	Neighbors
baseball	sport	players	athlete	athleteplaysinleague	major_league_baseball	sportsleague	13
baseball	sport	players	athlete	athleteplaysinleague	mlb	sportsleague	429
baseball	sport	sportteam	sportsteam	teamplaysinleague	mlb	sportsleague	40
baseball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	mlb	sportsleague	1
baseball	sport	players	athlete	athleteplaysinleague	nfl	sportsleague	1
baseball	sport	sportteam	sportsteam	teamplaysinleague	nfl	sportsleague	2
baseball	sport	sportteam	sportsteam	teamplaysinleague	nhl	sportsleague	1
soccer	sport	players	athlete	athleteplaysinleague	nba	sportsleague	1
basketball	sport	players	athlete	athleteplaysinleague	nba	sportsleague	44
basketball	sport	sportteam	sportsteam	teamplaysinleague	nba	sportsleague	58
basketball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	nba	sportsleague	1

Example

Entity	U	relation	V	relation	Entidty	W	Neighbors
baseball	sport	players	athlete	athleteplaysinleague	major_league_baseball	sportsleague	13
baseball	sport	players	athlete	athleteplaysinleague	mlb	sportsleague	429
baseball	sport	sportteam	sportsteam	teamplaysinleague	mlb	sportsleague	40
baseball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	mlb	sportsleague	1
baseball	sport	players	athlete	athleteplaysinleague	nfl	sportsleague	1
baseball	sport	sportteam	sportsteam	teamplaysinleague	nfl	sportsleague	2
baseball	sport	sportteam	sportsteam	teamplaysinleague	nhl	sportsleague	1
soccer	sport	players	athlete	athleteplaysinleague	nba	sportsleague	1
basketball	sport	players	athlete	athleteplaysinleague	nba	sportsleague	44
basketball	sport	sportteam	sportsteam	teamplaysinleague	nba	sportsleague	58
basketball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	nba	sportsleague	1

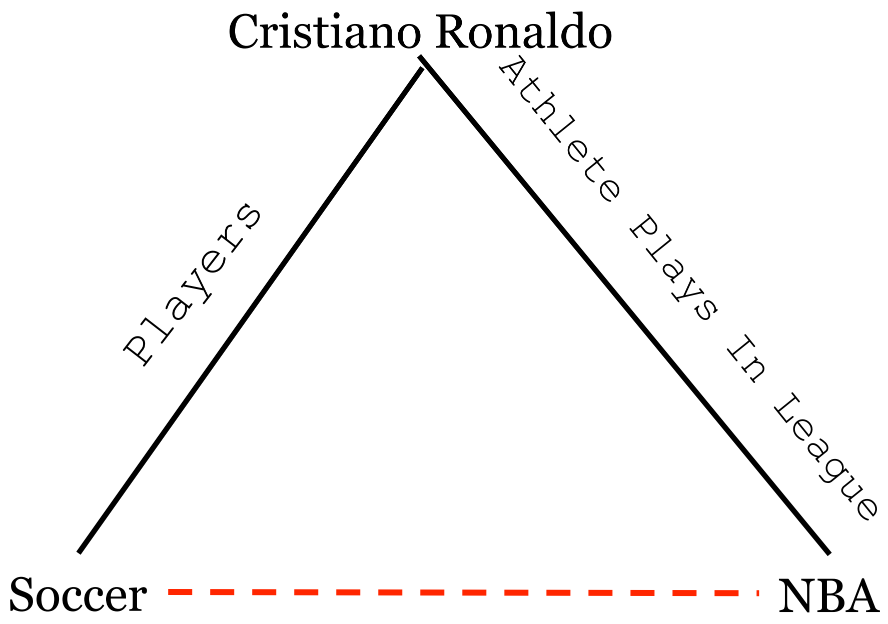
Example

Entity	U	relation	V	relation	Entidty	W	Neighbors
baseball	sport	players	athlete	athleteplaysinleague	major_league_baseball	sportsleague	13
baseball	sport	players	athlete	athleteplaysinleague	mlb	sportsleague	429
baseball	sport	sportteam	sportsteam	teamplaysinleague	mlb	sportsleague	40
baseball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	mlb	sportsleague	1
baseball	sport	players	athlete	athleteplaysinleague	nfl	sportsleague	1
baseball	sport	sportteam	sportsteam	teamplaysinleague	nfl	sportsleague	2
baseball	sport	sportteam	sportsteam	teamplaysinleague	nhl	sportsleague	1
soccer	sport	players	athlete	athleteplaysinleague	nba	sportsleague	1
basketball	sport	players	athlete	athleteplaysinleague	nba	sportsleague	44
basketball	sport	sportteam	sportsteam	teamplaysinleague	nba	sportsleague	58
basketball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	nba	sportsleague	1

Example

Entity	U	relation	V	relation	Entidty	W	Neighbors
baseball	sport	players	athlete	athleteplaysinleague	major_league_baseball	sportsleague	13
baseball	sport	players	athlete	athleteplaysinleague	mlb	sportsleague	429
baseball	sport	sportteam	sportsteam	teamplaysinleague	mlb	sportsleague	40
baseball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	mlb	sportsleague	1
baseball	sport	players	athlete	athleteplaysinleague	nfl	sportsleague	1
baseball	sport	sportteam	sportsteam	teamplaysinleague	nfl	sportsleague	2
baseball	sport	sportteam	sportsteam	teamplaysinleague	nhl	sportsleague	1
soccer	sport	players	athlete	athleteplaysinleague	nba	sportsleague	1
basketball	sport	players	athlete	athleteplaysinleague	nba	sportsleague	44
basketball	sport	sportteam	sportsteam	teamplaysinleague	nba	sportsleague	58
basketball	sport	sportusesstadium	stadiumoreventvenue	stadiumhometoleague	nba	sportsleague	1

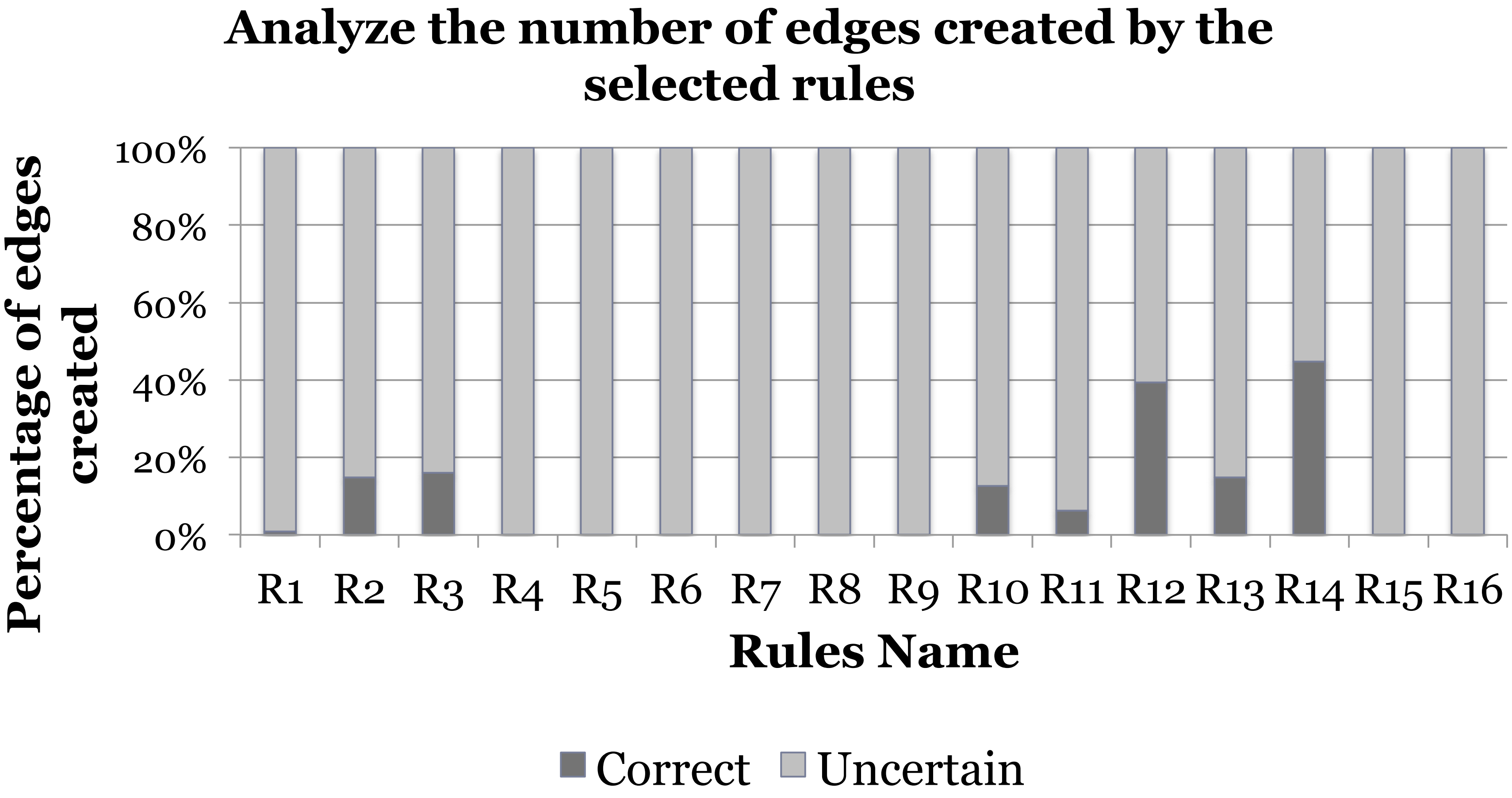
Outlier


Cristiano Ronaldo
Players
Athlete Plays In League
Soccer
NBA

Rules

- R1a(AwardTrophytournament, Athlete):- trophywonbycoaches(AwardTrophytournament, Coach), coachesathlete(Coach, Athlete), numberof(Coach) ≥ 10;
- R1b(AwardTrophytournament, Athlete):- trophywonbyteam(AwardTrophytournament, SportsTeam), teammember(SportsTeam, Athlete), numberof(SportsTeam) ≥ 10;
- R1c(AwardTrophytournament, Athlete):- trophywonbycoaches(AwardTrophytournament, Coach), coachesathlete(Coach, Athlete), trophywonbyteam (AwardTrophytournament, SportsTeam), teammember(SportsTeam, Athlete)

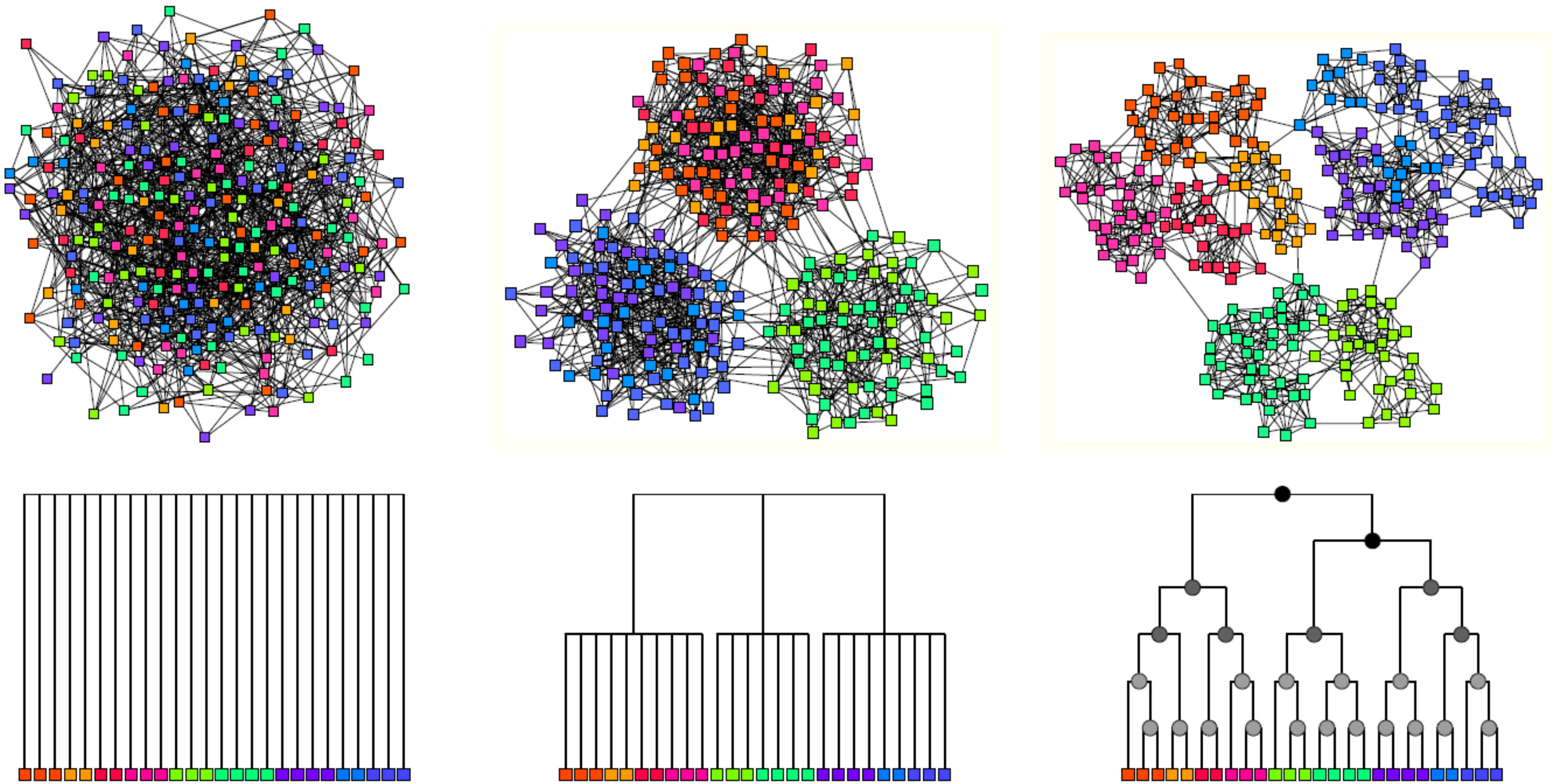
Results



Analyze the number of edges created by the selected rules
100%
80%
60%
40%
20%
0%
Percentage of edges created
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
Rules Name
Correct
Uncertain

Hierarchical Link Prediction (1)

- Graphs and corresponding hierarchies:



Hierarchical Link Prediction (2)

1. Given a network generate a set of hierarchical random graphs that fit its structure.
2. Evaluate pairs of vertices with a high probability of connection within the sampled hierarchical random graphs.
3. Rank the results by sorting based on the probability of their occurrence.

- Link prediction is an estimate of the likelihood or probability of the future occurrence of a link in a graph.
  - A maximum likelihood approach is used in missing link prediction based on a model of how links are organized in a network.
- This model considers all the possible arrangements of a given network and the distribution of such arrangements across a range of possible network structures [17].
- A maximum likelihood approach can also be used to predict false positives, which are links that are present but should not be present in a network.
- This is accomplished by looking at the minimum likelihood (lowest probability) of a link in a graph.
- A defining element of link prediction (as in [38], [17] and [13]) is that prediction methods are based purely on graph structure and focus on network evolution.

Hierarchical Link Prediction (3)

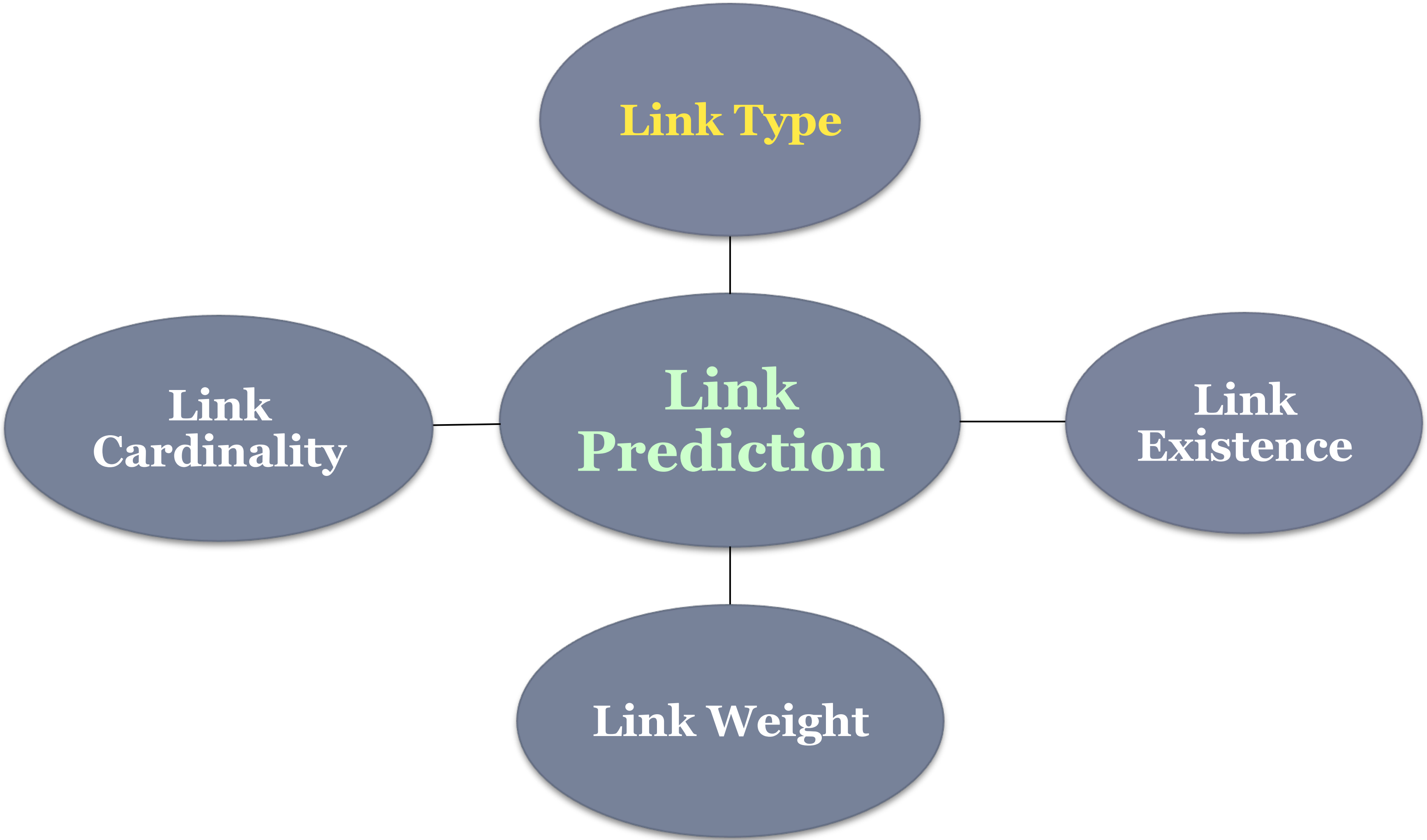
- Inferring hierarchical structure from network data that can be used in the prediction of missing links.
- Hierarchical structure is represented by a tree or *dendrogram* in which closely related pairs of vertices have lowest common ancestors that are lower in the tree than those of more distantly related pairs
- The prediction of missing links is then calculated as the probability that two nodes are connected over all the sampled dendograms.

Identify suicide ideation in social network

- The number of user communities to which a user belongs to
- The transitivity → number of triangles
- Fraction of suicidal neighbors in the social network, contributed the most to suicide ideation

Naoki Masuda, Issei Kurahashi, Hiroko Onari Suicide ideation of individuals in online social networks

Link Prediction Task



Social Media: Interaction (1)

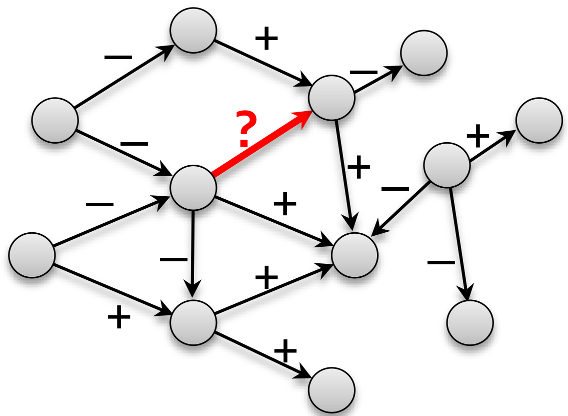
- In Social Media users **interact** with one another and the content they both crate and consume
- Traditional social network analysis only distinguishes between pairs of people that are **linked** vs. **not-linked**
- But, user interactions in social media are **much richer**

Social Media: Interaction (2)

- How to learn to recommend/predict links in social networks?
- User interactions in social media:
  - Strength: strong vs. weak ties
  - Friends vs. Foes
  - Trust vs. Distrust
  - Predict the directions

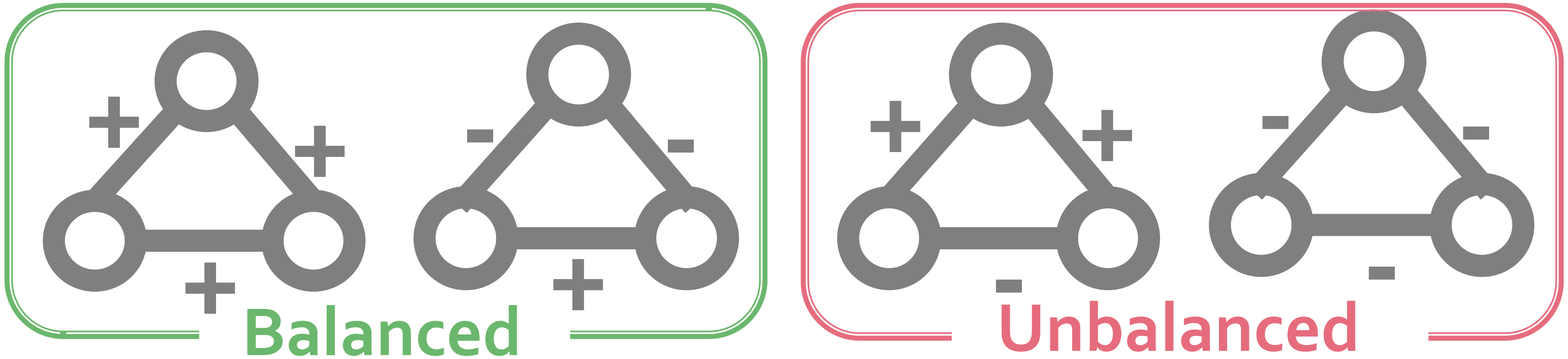
Friends vs. Foes

- So far we viewed links as positive but links can also be negative
- Question:
  - How do edge signs and network interact?
  - How to model and predict edge signs?
- Applications:
  - Friend recommendation
    - Not just whether you know someone but what do you think of them



Theory of Structural Balance

- Consider edges as undirected
- Start with intuition [Heider '46]:
  - Friend of my friend is my friend
  - Enemy of enemy is my friend
  - Enemy of friend is my enemy
- Look at connected triples of nodes:



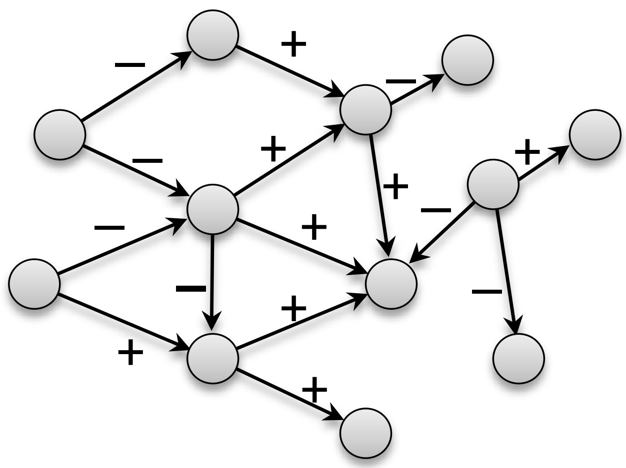
Consistent with "friend of a friend" or "enemy of the enemy" intuition

Inconsistent with the "friend of a friend" or "enemy of the enemy" intuition

Networks with Explicit Signs

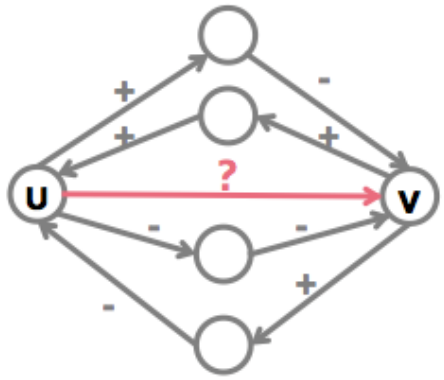
- Each link A→B is **explicitly** tagged with a sign:
  - **Epinions:** Trust/Distrust
    - Does A trust B's product reviews? (only positive links are visible)
  - **Wikipedia:** Support/Oppose
    - Does A support B to become Wikipedia administrator?
  - **Slashdot:** Friend/Foe
    - Does A like B's comments?

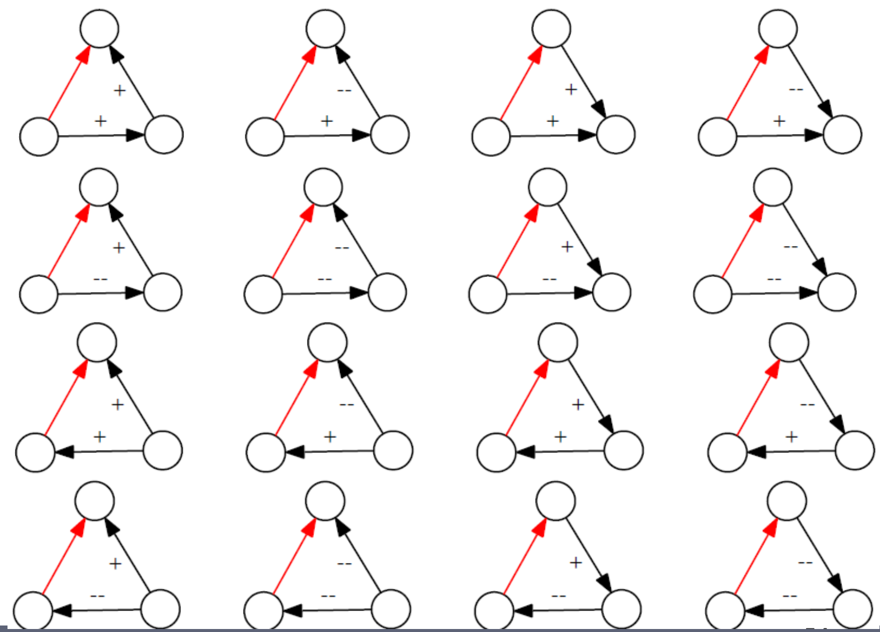
	Epinions	Slashdot	Wikipedia
Nodes	119,217	82,144	7,118
Edges	841,200	549,202	103,747
+ edges	85.0%	77.4%	78.7%
− edges	15.0%	22.6%	21.2%



Networks with Explicit Signs

- For each edge (u,v) create features:
- Triad counts (16):
  - Counts of signed triads edge u→v takes part in
- Degree (7 features):
  - Signed degree:
    - $d^{+}_{out}(u)$, $d^{-}_{out}(u)$, $d^{+}_{in}(v)$, $d^{-}_{in}(v)$
  - Total degree:
    - $d_{out}(u)$, $d_{in}(v)$
  - Embeddedness of edge (u,v)



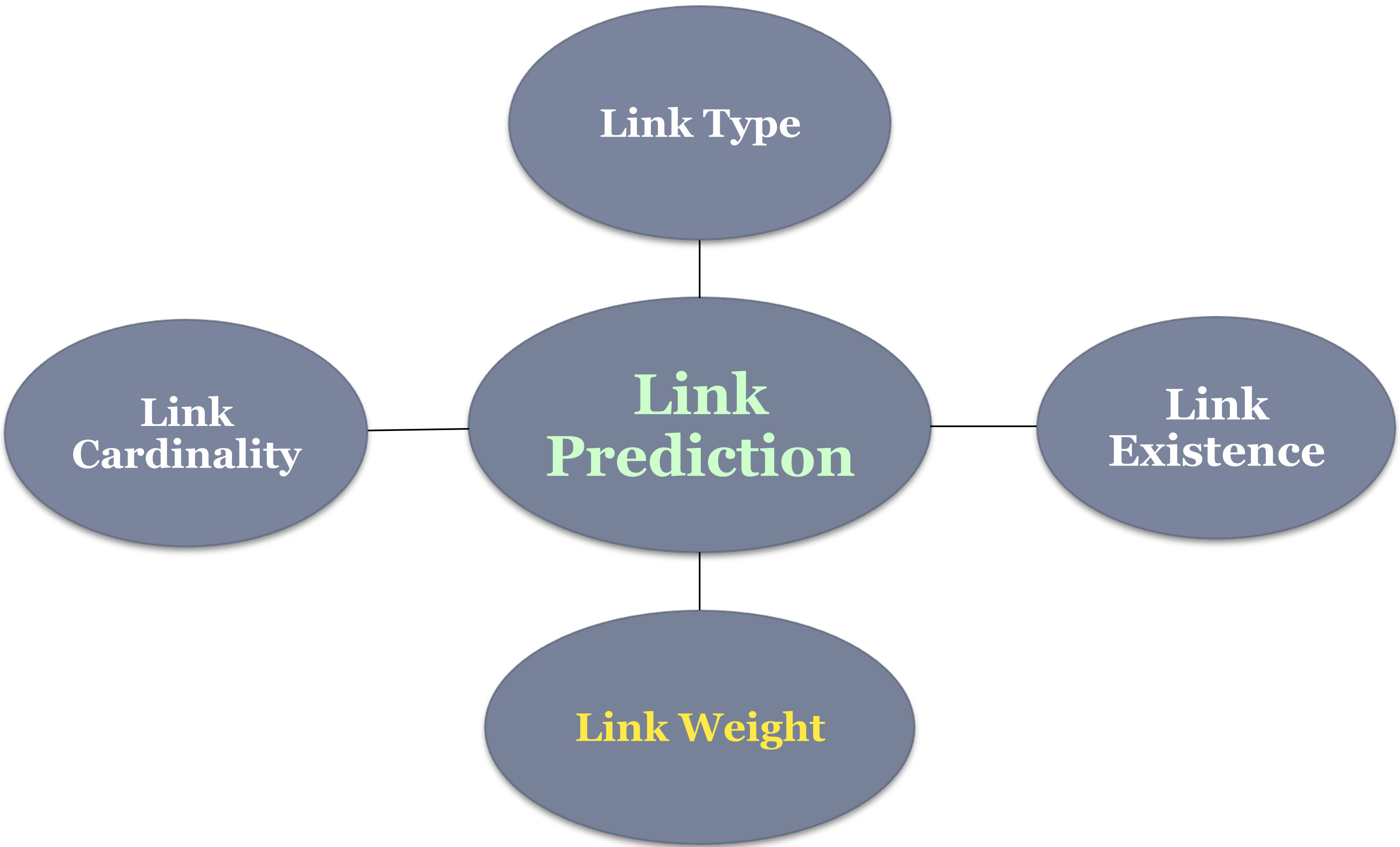


Networks with Explicit Signs

- Edge sign prediction problem
  - Given a network and signs on all but one edge, predict the missing sign
- Machine Learning formulation:
  - Predict sign of edge (u,v)
  - Class label:
    - +1: positive edge
    - -1: negative edge
  - Learning method:
    - Logistic regression

$$P(+|x) = \frac{1}{1 + e^{-(b_0 + \sum_i^n b_i x_i)}}$$

Link Prediction Task



Weighted Link Prediction

- Weighted common neighbors

$$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{2}$$

- Weighted Adamic/Adar

$$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{2} \times \frac{1}{\log(\sum_{z' \in \Gamma(z)} w(z', z))}$$

- Weighted Preferential Attachment

$$score(x, y) = \sum_{x' \in \Gamma(x)} w(x', x) \times \sum_{y' \in \Gamma(y)} w(y', y)$$

Weighted Link Prediction

- The fundamental task of link prediction in weighted networks, namely to predict the existence of links with the help of not only the observed links but also their weights
- How to properly exploit the information of weights to improve the prediction accuracy is still an unsolved problem.

- T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measure, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, ACM Press, New York, 2007.
- L Lu, T. Zhou, Link prediction in weighted networks: The role of weak ties, EPL 89 (2010) 18001.

Weighted Link Prediction

- Weighted common neighbors

$$score(x, y) = \sum \frac{w(x,z) + w(y,z)}{2}$$

- We

Not predict the weight of new edge!!!

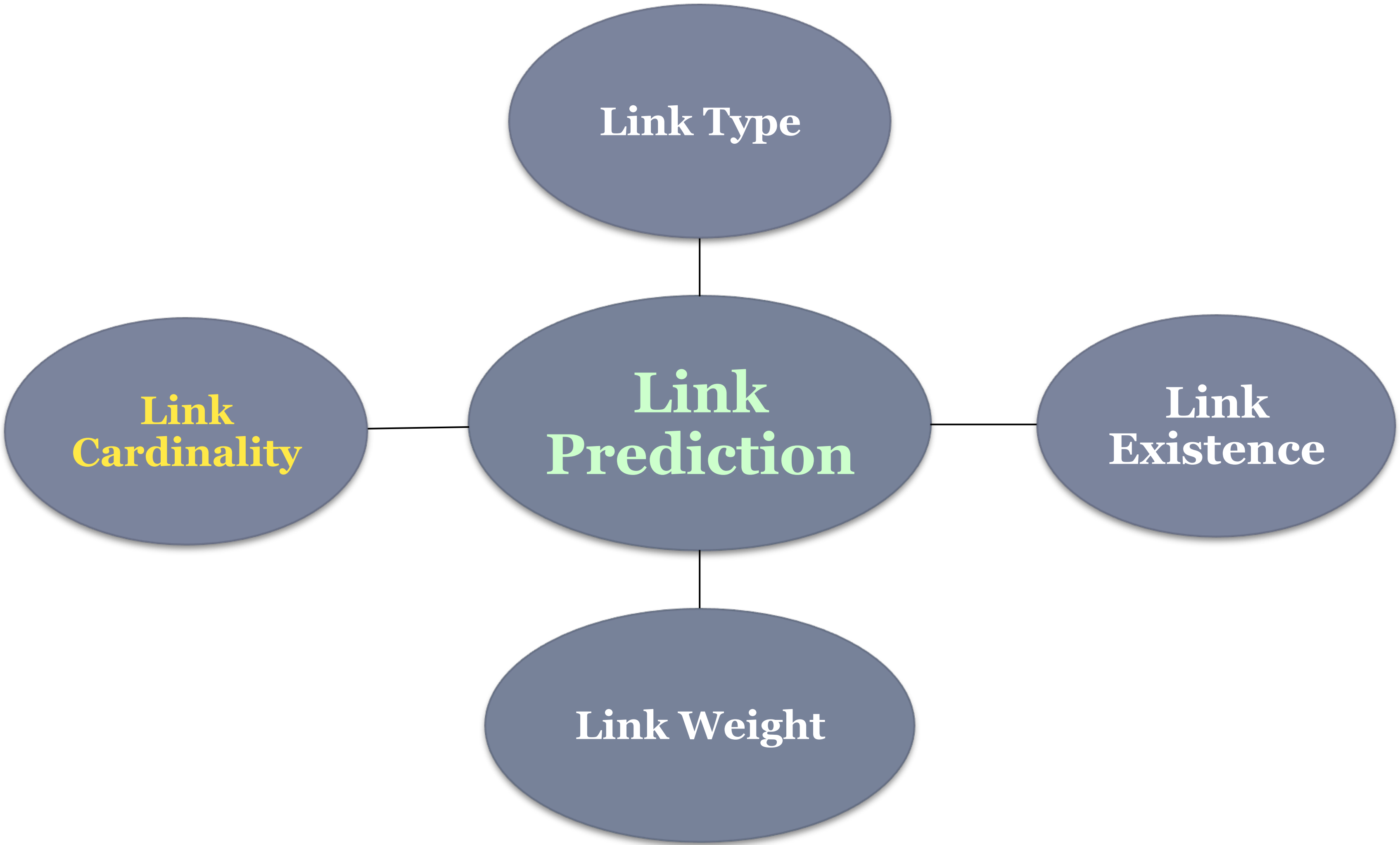
- We

$$score(x, y) = \sum_{x' \in \Gamma(x)} w(x', x) \times \sum_{y' \in \Gamma(y)} w(y', y)$$

Weighted Link Prediction

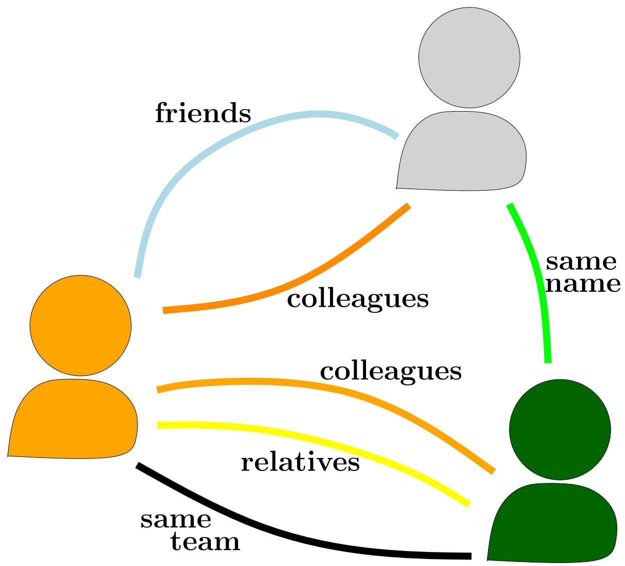
- A harder problem is to predict the weights of links
- Which is relevant to the traffic prediction for urban transportation and air transportation systems

Link Prediction Task



Multidimensional Network

- Each edge has a different meaning:
  - Social interaction
    - E-mail
    - Phone calls
    - Co-author
- Not only predict new link for disconnected node
  - New links for nodes connected
    - Different interaction



Multi-edge or multigraph

Multidimensional Predictors

- Multidimensional Common Neighbors

Consider only one dimension at time
NOT the relation among them!!

- M

$$Multidimensional\ Adamic\ Adar(u,v,d) = \sum_{z \in \{N(u,d) \cap N(v,d)\}} \frac{1}{log(|N(z,d)|)}$$

Reflections

- Open problems
  - Predict new edges and their direction or weight
  - Predict new edges for nodes already linked
- How weighted can improve accuracy is still not solved
  - Adapt or create new specific methods
- Bipartite graph
  - Change the graph to unipartite

Reflections

- Community structures can also help improving prediction accuracy
  - Same social circles
- In social networks, since one person may play different roles in different communities
  - The prediction in one domain can be inspired by the information in others.
    - Prediction the collaborations between authors can consider their affiliations to improve the accuracy.

Reflections

- Evolutions of link occurrences, which is more appropriate for dealing with the link prediction problem in evolving networks, such as online social networks
  - For now, it is impossible to predict whether and when two authors will collaborate again in co-authorship network
- Another way to involve time information is inspired by the fact that older events are less likely to be relevant to future links than recent ones.
  - For example, author's interests may change over time and thus old publications might be less relevant to his currents research area.

References

- *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations,* by J. Leskovec, J. Kleinberg, C. Faloutsos, ACM KDD 2005
- *Graph Evolution: Densification and Shrinking Diameters,* by J. Leskovec, J. Kleinberg and C. Faloutsos, *ACM TKDD 2007*
- *Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication,* by J. Leskovec, D. Chakrabarti, J. Kleinberg and C. Faloutsos, *PKDD 2005*
- *Scalable Modeling of Real Graphs using Kronecker Multiplication,* by J. Leskovec and C. Faloutsos, *ICML 2007*
- *The Dynamics of Viral Marketing, by* J. Leskovec, L. Adamic, B. Huberman, *ACM Electronic Commerce 2006*
- *Collective dynamics of 'small-world' networks,* by D. J. Watts and S. H. Strogatz, Nature 1998
- *Emergence of scaling in random networks,* by R. Albert and A.-L. Barabasi, Science 1999
- *On the evolution of random graphs,* by P. Erdos and A. Renyi, Publication of the
- Mathematical Institute of the Hungarian Acadamy of Science, 1960
- *Power-law distributions in empirical data,* by A. Clauset, C. Shalizi, M. Newman, 2007

References

- T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measure, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, ACM Press, New York, 2007.
- L Lu, T. Zhou, Link prediction in weighted networks: The role of weak ties, EPL 89 (2010) 18001.
- H. Yin, S. C. Wong, J. Xu, C. K. Wong, Urban traffic flow prediction using a fuzzy-neural approach, Transportation Res. C 10 (2002) 85.
- J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting Positive and Negative Links in Online Social Networks, In Proceedings of WWW'2010, ACM Press, New York, 2010.
- V. Leroy, B. B. Cambazoglu, F. Bonchi, Cold start link prediction, In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2010, p. 393.
- E. Zheleva, L. Getoor, J. Golbeck, Ugur Kuter, Using friendship ties and family circles for link prediction, In Proceedings of the 2nd Workshop on Social Network Mining and Analysis, ACM Press, New York, 2008.
- B. Cao, N. N. Liu, Q. Yang, Transfer learning for collective link prediction in multiple heterogenous domains, In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010.
- Z. Huang, D. K. J. Lin, The time-series link prediction problem with applications in communication surveillance, INFORMS J. Comput. 21 (2009) 286.
- T. Tylenda, R. Angelova, S. Bedathur, Towards time-aware link prediction in evolving social networks, In Proceedings of the 3rd Workshop on Social Network Mining and Analysis, ACM Press, New York, 2009.
- J. Kunegis, E. W. De Luca, S. Albayrak, The link predection problem in bipartite networks, arXiv: 1006.5367.
- B. Taskar, M.-F. Wong, P. Abbeel, D. Koller, Link prediction in relational data, In Proceedings of Neural Information Precessing Systems, MIT Press, Cambridge MA, 2004, p. 659.

References

- *The structure and function of complex networks,* M. Newman, SIAM Review 2003
- ·*Hierarchical Organization in Complex Networks,* Ravasz and Barabasi, Physical Review E 2003
- *A random graph model for massive graphs,* W. Aiello, F. Chung and L. Lu, STOC 2000
- *Community structure in social and biological networks,* by Girvan and Newman, PNAS 2002
- *On Power-law Relationships of the Internet Topology,* by Faloutsos, Faloutsos, and Faloutsos, SIGCOM 1999
- *Power laws, Pareto distributions and Zipf's law,* by M. Newman, Contemporary Physics 2005
- *Social Network Analysis : Methods and Applications,* by Wasserman, Cambridge University Press 1994
- *The web as a graph: Measurements, models and methods,* by J. Kleinberg and S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, COCOON 1998
- *Stochastic Kronecker Graphs,* by M. Mahdian and Y. Xu, Workshop on Algorithms and Models for the Web-Graph (WAW) 2007

Acknowledgements

Jure Leskovec