Complex networks: from data through models to knowledge

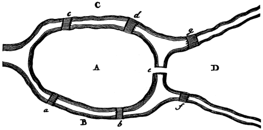
Matúš Medo

15 January 2018, Fribourg

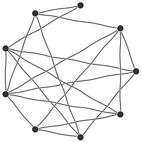
UESTC, Chengdu
Inselspital, Bern
University of Fribourg, Fribourg

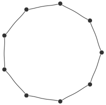
Complex networks

- 1736, Euler: bridges of Königsberg



- 1959, Erdős & Rényi: random graphs

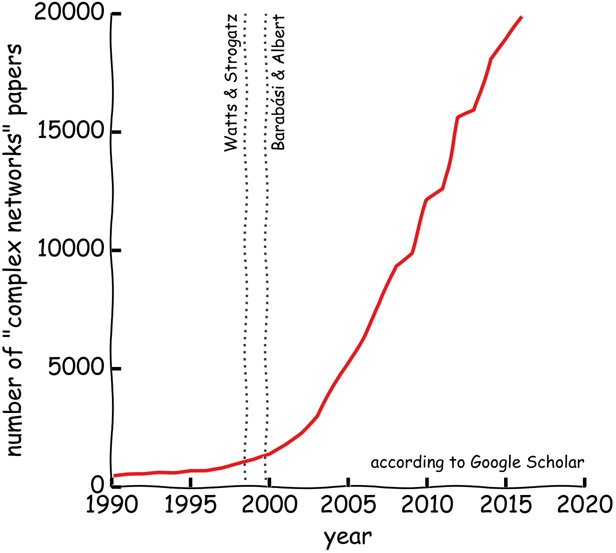




- 1998, Watts & Strogatz: disorder in regular networks
- 1999, Barabási & Albert: preferential attachment

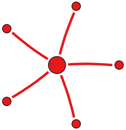
$$\boxed{\text{Complex network} = \text{graph} + \text{context}}$$

Historical milestones



Network glossary

undirected network



node with degree 5

directed network

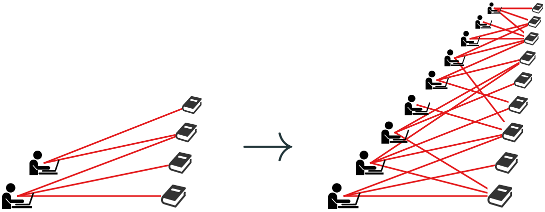
node with indegree 2

Information networks around us

- E-commerce systems: users and purchased items
- The World Wide Web: hyperlinked web pages
- Academia: citations among scientific papers

Information networks around us

- E-commerce systems: users and purchased items
- The World Wide Web: hyperlinked web pages
- Academia: citations among scientific papers

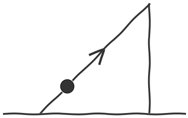


Why physicists study complex networks

- Many metrics, models, and algorithms can be introduced…
  The tough part is to decide which are useful

Why physicists study complex networks

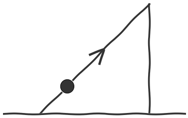
- Many metrics, models, and algorithms can be introduced… The tough part is to decide which are useful
- Historical note:
    - In Aristotelian physics, a projectile moves along a straight line until its “force” is exhausted and the projectile falls straight down



    - It took almost 2000 years and good measurements (Copernicus, Brahe, Galileo,…) to discredit the theory

Why physicists study complex networks

- Many metrics, models, and algorithms can be introduced... The tough part is to decide which are useful
- Historical note:
    - In Aristotelian physics, a projectile moves along a straight line until its “force” is exhausted and the projectile falls straight down



    - It took almost 2000 years and good measurements (Copernicus, Brahe, Galileo,...) to discredit the theory
- Proposing and testing models is how physicists can contribute

From data to models

Preferential attachment

- Re-discovered many times: Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)

Preferential attachment

- Re-discovered many times: Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
- Probability that node *i* attracts a new link at time *t*:

$$P(i,t) \sim \underbrace{k_i(t)}_{\text{node degree}}$$

Preferential attachment

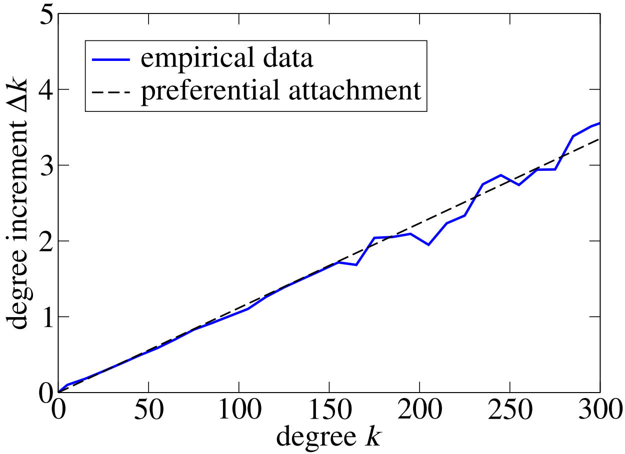
- Re-discovered many times: Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
- Probability that node i attracts a new link at time t:

$$P(i,t) \sim \underbrace{k_i(t)}_{\text{node degree}}$$

- Resulting growing networks have a power-law degree distribution similar to real systems

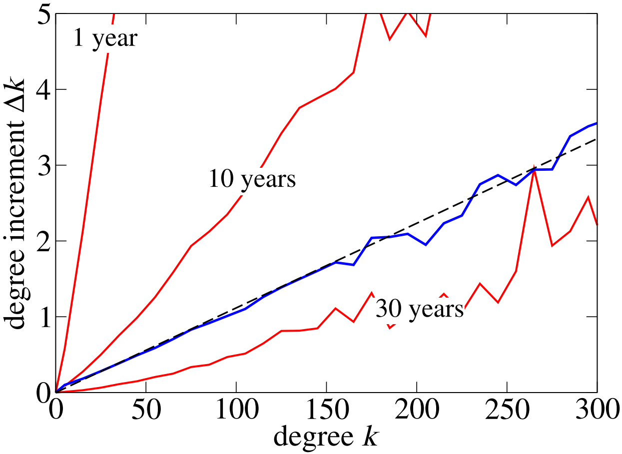
The missing element

American Physical Society papers, 1893–2009



The missing element

American Physical Society papers, 1893–2009



"All the News That's Fit to Print."

The New York Times.

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG;
866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH;
ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

WOMEN AND CHILDREN FIRST

Aging is fundamental

A better model (PRL 107, 238701, 2011)

- Probability that node *i* attracts a new link

$$P(i,t) \sim \underbrace{k_i(t)}_{\text{degree}} \times \underbrace{D_R(t)}_{\text{aging}} \times \underbrace{f_i}_{\text{fitness}}$$

  - $D_R(t)$ is a function that decreases with time
  - f_i is node parameter

A better model (PRL 107, 238701, 2011)

- Probability that node *i* attracts a new link

$$P(i,t) \sim \underbrace{k_i(t)}_{\text{degree}} \times \underbrace{D_R(t)}_{\text{aging}} \times \underbrace{f_i}_{\text{fitness}}$$

  - $D_R(t)$ is a function that decreases with time
  - f_i is node parameter

- This model:
  - Produces various realistic degree distributions
  - Explains data better than other models (likelihood maximization in PRE 89, 032801, 2014)
  - Obviously, it does not capture all effects (see paper by Golosovsky and Solomon in PRE, 2017)

The model's implication

- The expected final node degree is

$$\overline{k_i}(\infty) \propto \exp(\alpha f_i)$$

The model's implication

- The expected final node degree is

$$\overline{k_i}(\infty) \propto \exp(\alpha f_i)$$

- Hence paper 1 with 1000 citations is not 100-times better than a paper 2 with 10 citations
- Instead, the papers' fitness ratio is

$$f_1/f_2 = \ln k_1/\ln k_2 = 3$$

The model's implication

- The expected final node degree is

$$\overline{k_i}(\infty) \propto \exp(\alpha f_i)$$

- Hence paper 1 with 1000 citations is not 100-times better than a paper 2 with 10 citations
- Instead, the papers' fitness ratio is

$$f_1/f_2 = \ln k_1/\ln k_2 = 3$$

- A case for modesty
    - Citations counts magnify the qualitative differences between papers/researchers
    - Besides numbers, we should look at individuals' contribution in terms of ideas, service to community, etc.

Application 1:
Ranking network nodes

PageRank: A classical network centrality metric

- Centrality metrics quantify the importance of nodes
- Simplest centrality metric: in-degree
- PageRank weights links from important nodes more

PageRank: A classical network centrality metric

- Centrality metrics quantify the importance of nodes
- Simplest centrality metric: in-degree
- PageRank weights links from important nodes more

- PageRank score p_i of node i is

$$p_i = \underbrace{c \sum_{j \to i} \frac{p_j}{k_j^{out}}}_{\text{network contribution}} + \underbrace{1 - c}_{\textit{teleportation}}$$

  - $c = 0.85$ (WWW) or $c = 0.5$ (citation networks)

Evaluation on model networks

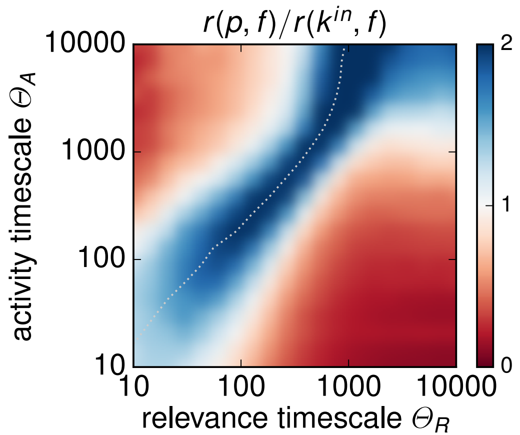
- Three key elements of the model:
    1. Node-specific fitness f_i
    2. Decay of relevance (attractiveness to incoming links): $D_R(t)$
    3. Decay of activity (activity to create outgoing links): $D_A(t)$
- Timescales of the two decays: Θ_R and Θ_A

Evaluation on model networks

- Three key elements of the model:
    1. Node-specific fitness f_i
    2. Decay of relevance (attractiveness to incoming links): $D_R(t)$
    3. Decay of activity (activity to create outgoing links): $D_A(t)$
- Timescales of the two decays: Θ_R and Θ_A

The key question:
Can PageRank uncover node fitness f_i?

When PageRank fails (Sci. Rep. 5, 16181, 2015)



PageRank vs indegree in a little more complicated model

Correcting PageRank (Journal of Informetrics 10, 1207, 2016)

- Citation data fall in a very wrong part of the (Θ_R, Θ_A) plane, yet PageRank is still commonly applied there...

Correcting PageRank (Journal of Informetrics 10, 1207, 2016)

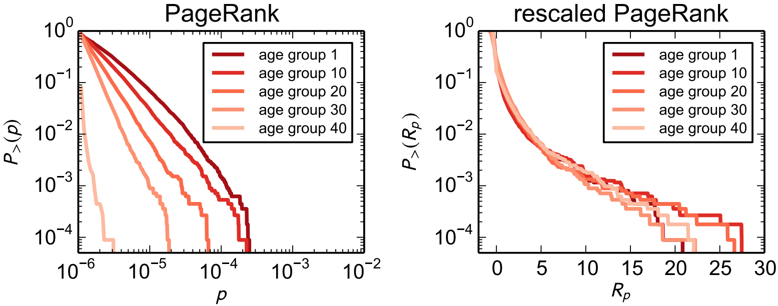
- Citation data fall in a very wrong part of the (Θ_R, Θ_A) plane, yet PageRank is still commonly applied there...

- We introduce rescaled PageRank of paper i as

$$R_i(p) = \frac{p_i - \mu_i}{\sigma_i}$$

  - p_i is PageRank score of paper i
  - μ_i and σ_i are the mean and standard deviation of PageRank score for papers published "close" to paper i

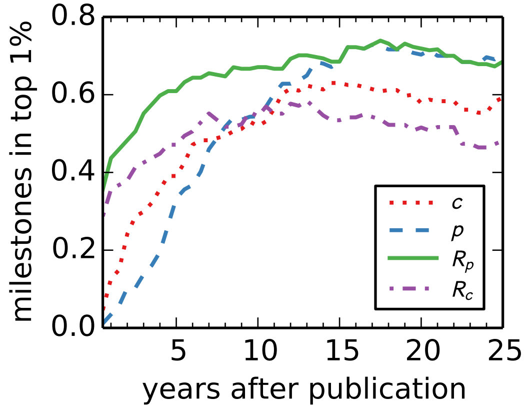
Divide the APS papers by age in 40 equally large groups



↓

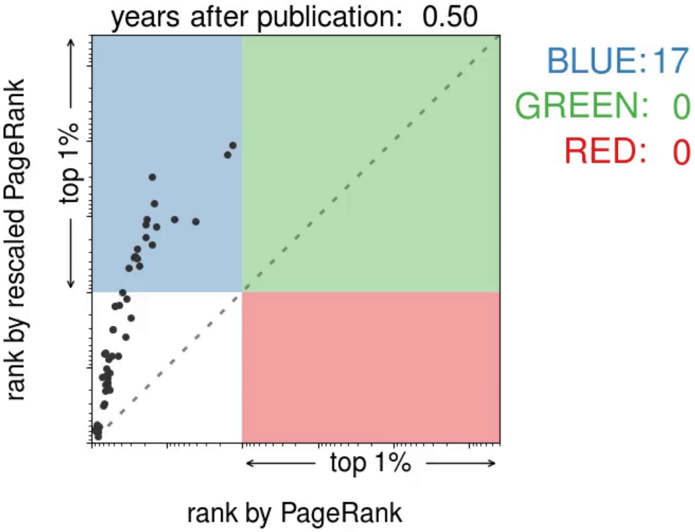
Allows us to fairly compare all papers!

Evaluation based on “milestone letters” announced by PRL



Note: CiteRank (Walker et al, 2007) is competitive with R_p in some aspects

PageRank vs. rescaled PageRank



From research to applications: www.sciencenow.info

ScienceNow Trending Blog About Leave a message

"Discover both old and recent significant research"

Here on ScienceNow, you can browse research papers published by the American Physical Society and see their *rescaled PageRank* score, $R(p)$. This new metric removes the time bias from Google's famous PageRank centrality. Since it is not biased by paper age, old seminal papers and new influential works have the same chance to appear at the top of the ranking by $R(p)$. Visit our blog to learn more.

You can:

- Search the papers by title and author (e.g., gravitational waves, topological insulators, Feynman) – see the search box at the top
- View the ranking history of papers (e.g., Einstein-Podolsky-Rosen paper on the completeness of quantum mechanics)
- See the publication record of individual researchers (e.g., Edward Witten)

Application 2:
Community detection

Introduction to community detection

- Many networks have community structure:
  - Some nodes are densely connected with each other (community)
  - Communities in social networks can be due to language, age, race, ...

Introduction to community detection

- Many networks have community structure:
  - Some nodes are densely connected with each other (community)
  - Communities in social networks can be due to language, age, race, ...
- Importance:
  - Can help us understand how the system works
  - Communities often have properties that differ a lot from the average network properties

Introduction to community detection

- Many networks have community structure:
  - Some nodes are densely connected with each other (community)
  - Communities in social networks can be due to language, age, race, …
- Importance:
  - Can help us understand how the system works
  - Communities often have properties that differ a lot from the average network properties
- “As long as there will be networks, there will be people looking for communities in them.” (Fortunato and Hric, 2016)
  - How best to find the communities?

Network modularity

- Popular approach to community detection: maximize the modularity function (Girvan & Newman, 2002)

$$Q = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \frac{k_i^{out} k_j^{in}}{m} \right) \delta(c_i, c_j)$$

Network modularity

- Popular approach to community detection: maximize the modularity function (Girvan & Newman, 2002)

in the same community

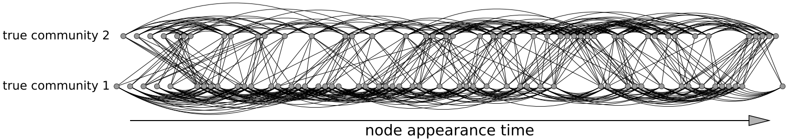
$$Q = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \frac{k_i^{out} k_j^{in}}{m} \right) \delta(c_i, c_j)$$

number of links

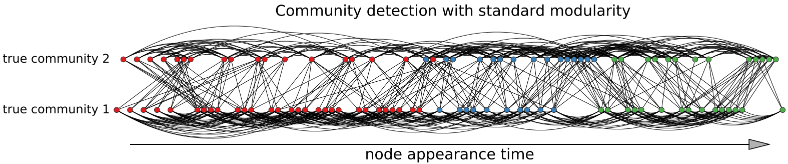
connected or not

link expectation

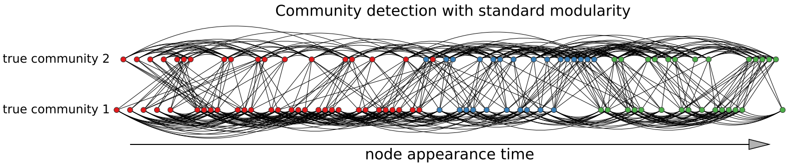
The problem in growing networks



The problem in growing networks

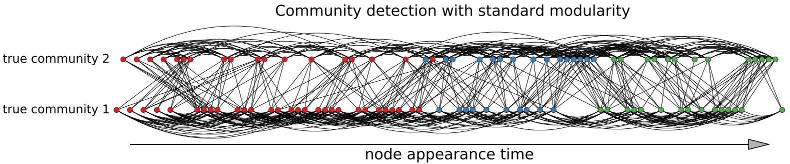


The problem in growing networks



- Standard modularity fails even if the true communities are disconnected (when $N \gtrsim 4\Theta_R$)

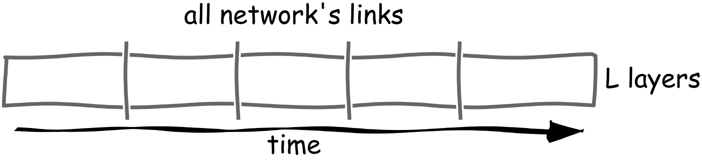
The problem in growing networks



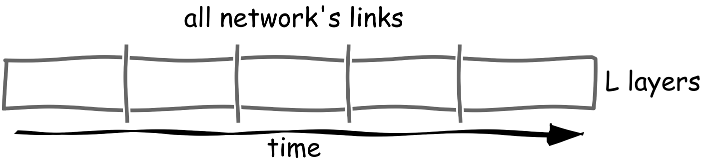
- Standard modularity fails even if the true communities are disconnected (when $N \gtrsim 4\Theta_R$)
- Reason of failure:
  If time matters, the link expectation term is wrong

$$Q = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \boxed{\frac{k_i^{out} k_j^{in}}{m}} \right) \delta(c_i, c_j)$$

Modularity for growing networks (to be submitted)



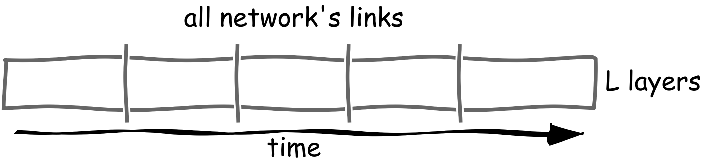
Modularity for growing networks (to be submitted)



Modularity with link expectation combined from all L layers

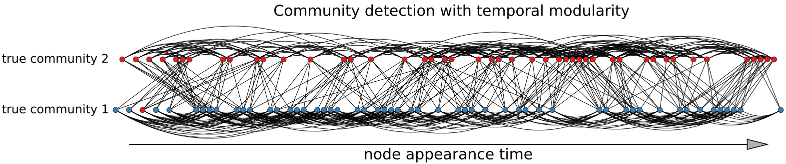
$$Q_T(L) = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \sum_{l=1}^{L} \frac{\Delta k_{i,l}^{out} \Delta k_{j,l}^{in}}{m_l} \right) \delta(c_i, c_j)$$

Modularity for growing networks (to be submitted)

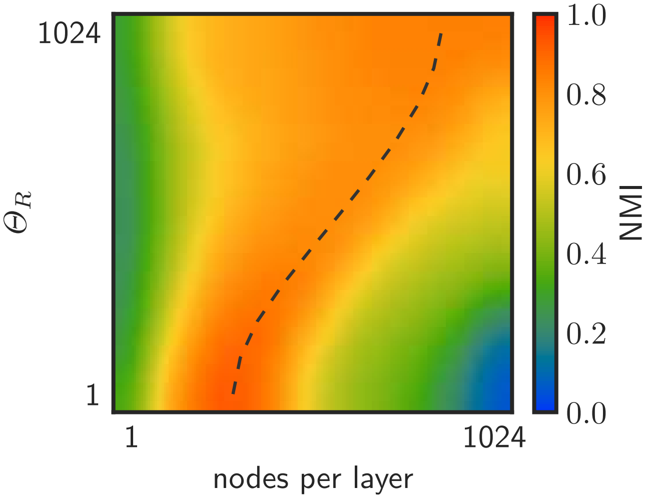


Modularity with link expectation combined from all L layers

$$Q_T(L) = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \sum_{l=1}^{L} \frac{\Delta k_{i,l}^{out} \Delta k_{j,l}^{in}}{m_l} \right) \delta(c_i, c_j)$$

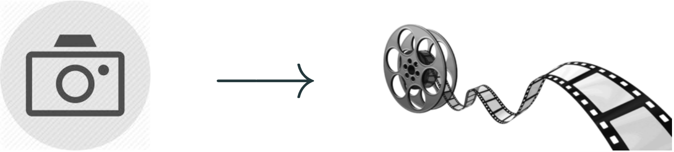


When it works and when it does not



Dashed line corresponds to median link timespan

Take-home message



1. We know a lot about the evolution of complex systems
2. Let the data drive you
3. Beware the application range of “good old” metrics and algorithms
4. By taking time into account, you can do better

Further related work:

1. H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, Physics Reports 689, 1-54, 2017
2. G. Vaccario, M. Medo, N. Wider, M. S. Mariani, Quantifying and suppressing ranking bias in a large citation network, Journal of Informetrics 11, 766-782, 2017
3. M. Medo, G. Cimini, Model-based evaluation of scientific impact indicators, Physical Review E 94, 032312, 2016
4. A. Vidmer, M. Medo, The essential role of time in network-based recommendation, EPL 116, 30007, 2016
5. M. Medo, M. S. Mariani, A. Zeng, Y.-C. Zhang, Identification and modeling of discoverers in online social systems, Scientific Reports 6, 34218, 2016

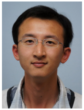
Web site: www.ddp.fmph.uniba.sk/~medo/physics/

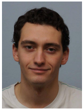

Yi-Cheng Zhang


Giulio Cimini


Stanislao Gualdi


Alex Vidmer


An Zeng


Manuel Mariani

Thank you for your attention!

Questions?