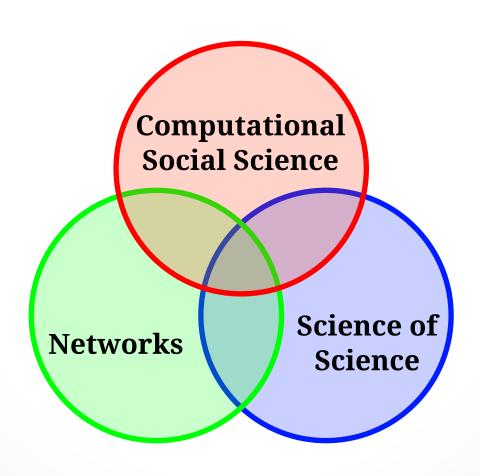
Community structure in complex networks

Santo Fortunato



Me in a nutshell...



Computational social science

REVIEWS OF MODERN PHYSICS, VOLUME 81, APRIL-JUNE 2009

Statistical physics of social dynamics

Claudio Castellano*

SMC. INFM-CNR and Dipartimento di Fisica, "Sapienza" Università di Roma, Plazzale A. Moro 2, 00185 Roma, Italy

Santo Fortunato

Complex Networks Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133 Torino, Italy

Vittorio Loreto

Dipartimento di Fisica, "Sapienza" Università di Roma and SMC, INFM-CNR Piazzale A. Moro 2, 00185 Roma, Italy and Complex Networks Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133 Torino, Italy

Statistical physics has proven to be a fruitful framework to describe phenomena outside the realm of traditional physics. Recent years have witnessed an attempt by physicists to study collective phenomena emerging from the interactions of individuals as elementary units in social structures. A wide list of topics are reviewed ranging from opinion and cultural and language dynamics to crowd behavior, hierarchy formation, human dynamics, and social spreading. The connections between these problems and other, more traditional, topics of statistical physics are highlighted. Comparison of model results with empirical data from social systems are also emphasized.

DOI: 10.1103/RevModPhys.81.591

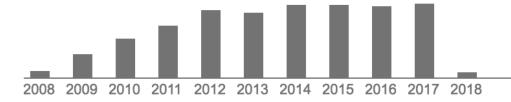
PACS number(s): 05.10.-a, 89.20.-a, 89.75.-k

CONTENTS

| I. Introduction | 59 |
|---|----|
| II. General Framework: Concepts and Tools | 59 |
| A. Order and disorder: The Ising paradigm | 59 |
| B. Role of topology | 59 |
| C. Dynamical systems approach | 59 |
| D. Agent-based modeling | 59 |
| II. Opinion Dynamics | 59 |
| | |

| C. Other multidimensional models | 616 | |
|--|-----|--|
| V. Language Dynamics | | |
| A. Evolutionary approaches | 617 | |
| Evolutionary language game | 617 | |
| 2. Quasispecies-like approach | 618 | |
| B. Semiotic dynamics approach | 619 | |
| The Naming Game | 619 | |
| 2. Symmetry breaking: A controlled case | 620 | |
| 3. The role of the interaction topology | 620 | |

Total citations Cited by 2540



ABOUT THOMSON REUTERS

ScienceWatch.com > Data & Rankings > New Hot Papers

Scholar articles Statistical physics of social dynamics C Castellano, S Fortunato, V Loreto - Reviews of modern physics, 2009

Cited by 2540 Related articles All 33 versions

NEW HOT PAPERS

Reputation and impact in academic careers

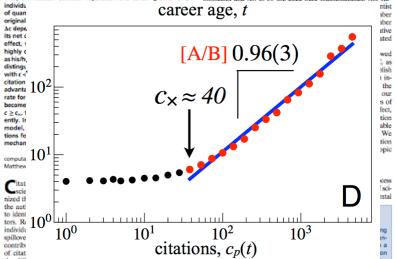
Alexander Michael Petersen^{a, 1}, Santo Fortunato^{b, 1}, Raj K. Pan^b, Kimmo Kaski^b, Orion Penner^c, Armando Rungi^a, Massimo Riccaboni^{c,d}, H. Eugene Stanley^{e,1}, and Fabio Pammolli^{a,e}

"Laboratory for the Analysis of Complex Economic Systems, Institutions, Markets, Technologies (MT) Lucca Institute for Advanced Studies, 55100 Lucca, Italy;
"Department of Biomedical Engineering and Computational Science, Aalto University Shool of Science, R-100076 Aalto, Finland," Laboratory of Honovation
Management and Economics, MT Lucca Institute for Advanced Studies, 55100 Lucca, Italy; "Department of Management all Economics, Strategy and Innovation." Katholieke Universiteit Leuven, 3000 Leuven, Belgium; and "Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215

Contributed by H. Eugene Stanley, December 17, 2013 (sent for review May 8, 2013)

informed quality assessments of both publications and careers of ever, the relation between reputation and career growth of an biologists, and [E] to 50 top-cited pure mathematicians (for fur-

Reputation is an important social construct in science, which enables 387,103 publication years. Dataset [A] refers to 100 top-cited physicists, [B] to another set of 100 highly prolific physicists, [C] scientists in the absence of complete systemic information. How- to 100 assistant professors in physics, [D] to 100 top-cited cell



data (10-13). As the measures are becoming increasingly common in evaluation scenarios throughout science, it is crucial to better understand what the citation measures actually represent in the context of scientists' careers. Moreover, how does reputation affect a scientist's access to key resources, the incentives to publish quality over quantity, and other key decisions along the career path (14-18)? In addition, what role does reputation play in the mentor-matching process within academic institutions, in the effectiveness of single/double blinding in peer review, and in the reward system of science (14, 15, 19)?

It is against this background that we have developed a quantitative framework with the goal of isolating the effect of author reputation upon citation dynamics. Specifically, by controlling for time- and author-specific factors, we quantify the role of author reputation on the citation life cycle of individual publications at the micro level. We use a longitudinal career dataset from Thomson Reuters Web of Science comprising 450 highly cited scientists, 83,693 articles, and 7,577,084 citations tracked over

www.pnas.org/cgi/doi/10.1073/pnas.1323111111

its citation life cycle, before a tipping point, after which reputation has much less influence relative to the paper's citation count. In science, perceived quality, and decisions made based on those perceptions, is increasingly linked to citation counts. Shedding light on the complex mechanisms driving these quantitative measures facilitates not only better evaluation of scientific outputs but also a more transparent evaluation of the scientists producing them.

tion plays a key role in driving a paper's citation count early in

AMP, SF, RKP, KK, OP, MR, HES, and FP, performed research: AMP, SF, RKP. A.R., and O.P. analyzed data; and A.M.P., S.F., R.K.P., K.K., O.P., A.R., M.R., H.E.S., and F.P.

The authors declare no conflict of interest.

¹To whom correspondence may be addressed. Email: hes@bu.edu, petersen.xander@ gmail.com, or santo.fortunato@aalto.fl.

This article contains supporting information online at www.pnas.org/lookup/supplidob10 1073/pnas.1323111111/-DCSupplemental.

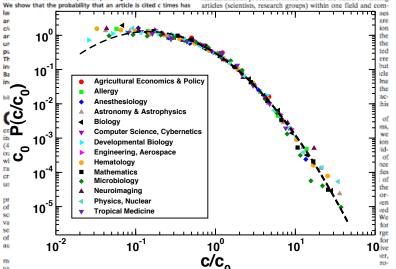
Universality of citation distributions: Toward an objective measure of scientific impact

Filippo Radicchi^a, Santo Fortunato^a, and Claudio Castellano^{b,1}

*Complex Networks Lagrange Laboratory, Institute for Scientific Interchanged Foundation, 10133 Torino, Italy; and *Centre for Statistical Mechanics and Complexity, National Institute for the Physics of Matter-Consiglio Nazionale delle Ricerche, and Dipartimento di Fisica, "Sapienza" Università di Roma. Piazzale A. Moro 2, 00185 Roma, Italy

Edited by Michael E. Fisher, University of Maryland, College Park, MD, and approved September 17, 2008 (received for review July 18, 2008)

We study the distributions of citations received by a single publi-possibilities exist also in the detailed implementation of the cation within several disciplines, spanning broad areas of science. standardization procedure. Some methods are based on ranking



much more or much less than in others. This may happen for several reasons, including uneven number of cited papers per article in different fields or unbalanced cross-discipline citations (11). A paradigmatic example is provided by mathematics: the highest 2006 impact factor (IF) (12) for journals in this category (Journal of the American Mathematical Society) is 2.55, whereas this figure is 10 times larger or more in other disciplines (for example, in 2006, New England Journal of Medicine had IF 51.30. Cell had IF 29.19, and Nature and Science had IF 26.68 and 30.03,

The existence of this bias is well-known (8, 10, 12) and it is widely recognized that comparing bare citation numbers is inappropriate. Many methods have been proposed to alleviate this problem (13-17). They are based on the general idea of normalizing citation numbers with respect to some properly chosen reference standard. The choice of a suitable reference standard, which can be a journal, all journals in a discipline, or a more complicated set (14), is a delicate issue (18). Many particular value of c_f is the same. Moreover, we show that c_f allows us to properly take into account the differences, within a single discipline, between articles published in different years. This provides a strong validation of the use of c_f as an unbiased relative indicator of scientific impact for comparison across fields

Variability of Citation Statistics in Different Disciplines

First, we show explicitly that the distribution of the number of articles published in some year and cited a certain number of

Author contributions: F.R., S.F., and C.C. designed research; F.R., S.F., and C.C. performed research; F.R. analyzed data; and C.C. wrote the paper The authors declare no conflict of interest.

This article is a PNAS Direct Submission

To whom correspondence should be addressed. E-mail: claudio.castellano@roma1.infn.it. © 2008 by The National Academy of Sciences of the USA

Science of science

Correspondence

Growing time lag threatens Nobels

The time lag between reporting a scientific discovery worthy of a Nobel prize and the awarding of the medal has increased, with waits of more than 20 years becoming common. If this trend continues, some candidates might not live long enough to attend their Nobel ceremonies.

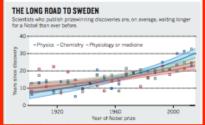
Before 1940, Nobels were awarded more than 20 years after the original discovery for only about 11% of physics, 15% of chemistry and 24% of physiology or medicine prizes, respectively. Since 1985, however, such lengthy delays have featured in 60%, 52% and 45% of these awards, respectively.

awards, respectively.
The increasing average interval between reporting discoveries and their formal recognition can be fitted to an exponential curve (see "The long road to Sweden'), with data points scattered about the mean value.

As this average interval becomes longer, so the average age at which laureates are awarded the prize goes up. By the end of this century, the prizewinners' predicted average age for receiving the award is likely to exceed his or her projected life expectancy (data not shown). Given that the Nobel prize cannot be awarded posthumously, this lag threatens to undermine science's most venerable institution. Santo Fortunato* Aalto University, Finland. santo.fortunato@gmail.com *On behalf of 6 co-authors; see go.nature.com/cmmxa5 for full list.

Livestock: tackle demand and yields

Among many otherwise laudable suggestions, Mark Eisler and colleagues propose limiting feedstuffs for livestock to fibrous fodder, such as grass and silage (see Nature 507, 32–34; 2014). However, we believe that any attempt to meet the rapid growth



in world demand for meat and dairy products by focusing on ruminant grazing systems would be damaging for biodiversity and for the global climate.

Although ruminants convert grass and silage into animal protein, they do so inefficiently; they therefore require much more land to produce a given amount of meat or milk than ruminants fed on diets that include grain. Growing enough fodder to satisfy demand would require the large-scale expansion of grazing lands (see go.nature.com/rmf63y) — a leading cause of biodiversity loss, tropical deforestation and carbon dioxide emissions.

dioxide crississons. The environmental impacts of meet and dairy production should instead be addressed by stringent efforts to decrease consumption, halt the expansion of grazing, and increase yields on land that is already used for livestock. Promoting extensive grazing without tackling demand would do more harm than good. Erasmus K. H. J. zu Ermgassen, David R. Williams, Andrew Balmford University of Cambridge, UK.

Livestock: limit red meat consumption

Mark Eisler and co-authors advocate eating only 300 grams of red meat a week (roughly the volume of three decks of playing cards) as a step towards producing sustainable livestock (Nature 507, 32–34; 2014). That amount corresponds to 3.5–7% of a 2,000–calorie-a-day diet, depending on the cut and type of meat. Such a move would also make for a more equitable global distribution of animal-product consumptions; these products comprise around 48% of the average diet in the United States, for example (S. Bonhommeau et al. Proc. Natl Acad. Sci. USA 110, 2061–20620; 2013).

Imposing a global dietary limit of 5% red meat as part of a 10% maximum for all animal-based products would enable more people to be fed using less land. For example, eliminating livestock and using existing agricultural lands to grow crops for direct human consumption instead of for livestock fodder could feed an extra 4 billion people (E. S. Cassidy et al. Environ. Res. Lett. 8, 034015; 2013), thereby reducing or eliminating the greenhouse-gas emissions and biodiversity loss associated with conversion of natural habitats. This would also reduce many other environmental impacts of agriculture that relate to the use of water, fertilizer and fossil fuels. Brian Machovina, Kenneth I. Feeley Florida International University, Miami; and The Fairchild Tropical Botanic Garden, Coral Gables, Florida, USA.

brianmachovina@gmail.com

Zoo visits boost biodiversity literacy

Zoos and aquaria worldwide attract more than 700 million visits every year. They are therefore well placed to make more people aware of the importance of biodiversity a prime target of the United Nations Strategic Plan for Biodiversity 2011–20.

We surveyed approximately 6,000 visitors to 30 zoos and aquaria in 19 countries (see go.nature.com/ws/8yf). More respondents showed improved understanding of biodiversity after their visit (75.1% compare with 69.8% before) and more could identify an individual action that would bolster biodiversity after their visit (58.8% compared with 50.5% before).

Regrettably, increased awareness does not necessarily change behaviour. The world's zoo and aquarium communities must also help to drive important behavioural and social changes to assist conservation.

Andrew Moss Chester Zoo, UK. Eric Jensen University of Warwick, Coventry, UK. Markus Gusset World Association of Zoos and Aquariums, Gland, Switzerland.markus gusset@waza.org

A protein that spells trouble

The gene CYLD is so named because one of its mutant forms is associated with cylindromatosis, which causes skin tumours.

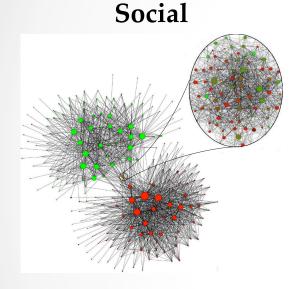
The CYLD protein is an enzyme; its active site in humans contains a cysteine residue at position 601 (denoted as C in the one-letter amino-acid code). The amino-acid sequence following this cysteine (C) is tyrosine (Y), leucine (L) and aspartate (D). What are the odds of that? David Boone Indiana University School of Medicine — South Bend, Indiana, USA.

SOUNDLE MOBILIFIED ONG VIVES ATYLIETT LETT UND THEFINST (2000)

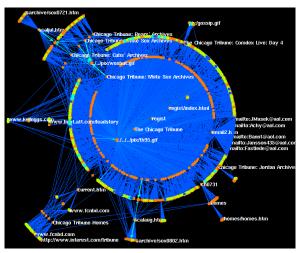
186 | NATURE | VOL 508 | 10 APRIL 2014

Network science

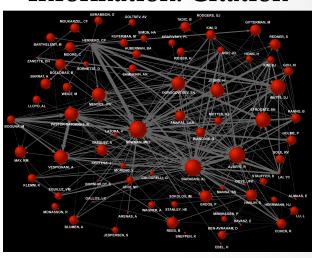
Analysis and modeling



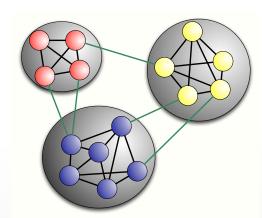
Information: WWW



Information: Citation



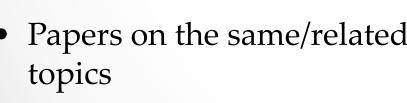
Community structure

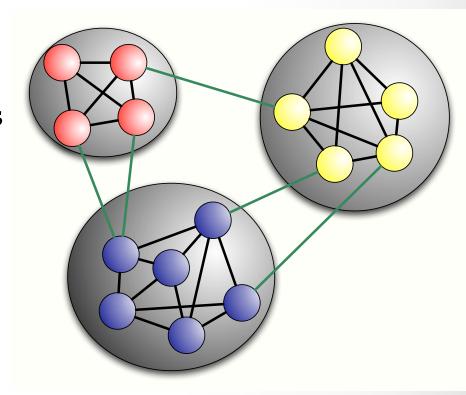


Community structure

Communities: sets of tightly connected nodes

- People with common interests
- Scholars working on the same field
- Proteins with equal/similar functions
- Papers on the same/related

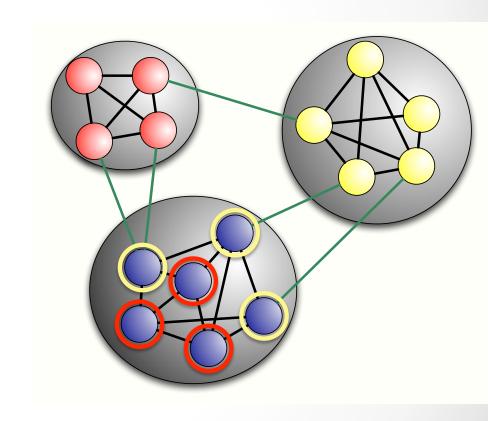




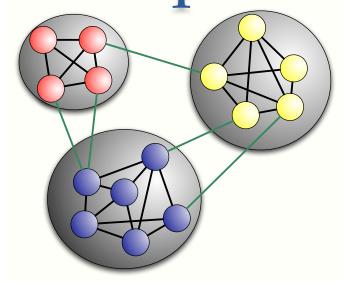
Community detection

What for?

- Organization
- Node classification
- Missing links
- Effect on dynamics
- •



Difficult problem!



```
      1
      2
      3
      4
      5
      6
      7
      8
      9
      10
      11
      12
      13
      14
      15

      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
```

```
      10
      1
      2
      9
      5
      13
      12
      14
      6
      11
      4
      3
      8
      7
      15

      10
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
      1
```

Difficult problem!

Ill-defined problem:

- What is a community/partition?
- What is a *good* community/partition?

Three basic questions

- 1) How to detect communities?
- 2) How to test community detection algorithms?
- 3) How to make partitions robust?

Acknowledgements

Alex Arenas Marc Barthelémy Ginestra Bianconi Richard Darst Alberto Fernández Sergio Gómez













Clara Granell











Vito Latora



Massimo Marchiori





José J. Ramasco Jari Saramäki





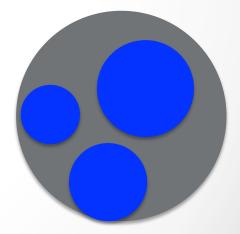
How to detect communities?

Global optimization

Principle:

- Function Q(P) that assigns a score to each partition
- Best partition of the network -> partition corresponding to the maximum/minimum of $Q(\mathcal{P})$

Problem: Answer depends on the whole graph -> it changes if one considers portions of it or if it is incomplete



How to detect communities?

Global optimization

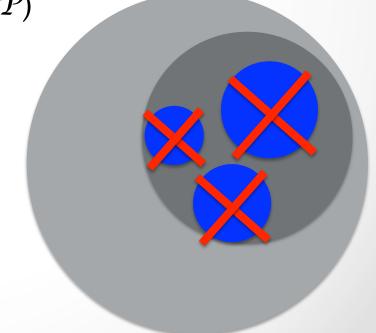
Principle:

• Function Q(P) that assigns a score to each partition

Best partition of the network -> partition corresponding

to the maximum/minimum of Q(P)

Problem: Answer depends on the whole graph -> it changes if one considers portions of it or if it is incomplete



$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left(l_c - \frac{d_c^2}{4m} \right)$$

$$E = m c^2$$

M. E. J. Newman, M. Girvan, Phys. Rev. E 69, 026113 (2004)

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left(l_c - \frac{d_c^2}{4m} \right)$$

$$E = m c^2$$

M. E. J. Newman, M. Girvan, Phys. Rev. E 69, 026113 (2004)

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left(l_c - \frac{d_c^2}{4m} \right)$$

$$E = m c^2$$

M. E. J. Newman, M. Girvan, Phys. Rev. E 69, 026113 (2004)

$$Q = \frac{1}{m} \sum_{\mathbf{c}=1}^{n_{\mathbf{c}}} \left(l_{\mathbf{c}} - \frac{d_{\mathbf{c}}^2}{4m} \right)$$

$$E = m c^2$$

M. E. J. Newman, M. Girvan, Phys. Rev. E 69, 026113 (2004)

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left(l_c - \frac{d_c^2}{4m} \right)$$

$$E = m c^2$$

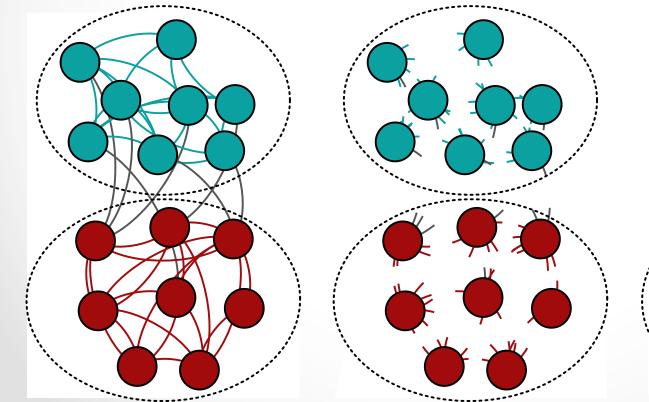
M. E. J. Newman, M. Girvan, Phys. Rev. E 69, 026113 (2004)

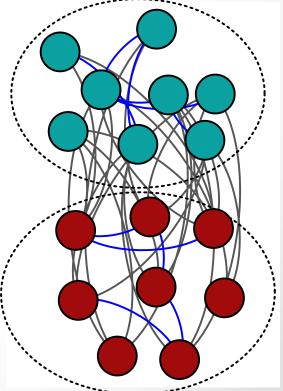
M. E. J. Newman, Phys. Rev. E 69, 066133 (2004)

Goal: find the maximum of Q over all possible network partitions

Problem: NP-complete (Brandes et al., 2007)!

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left(l_c - \frac{d_c^2}{4m} \right)$$

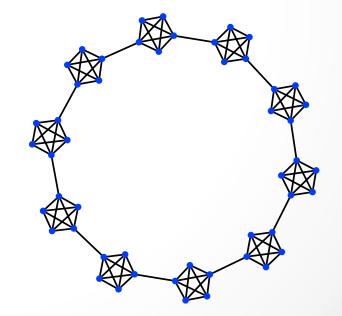




$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[l_c - \frac{1}{4} \left(\frac{d_c}{m} \right)^2 \right]$$
 modularity's scale

Result: clusters smaller than this scale cannot be resolved!

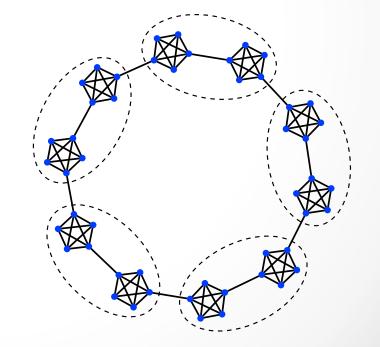
Consequences



$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[l_c - \frac{1}{4} \left(\frac{d_c}{m} \right)^2 \right]$$
 modularity's scale

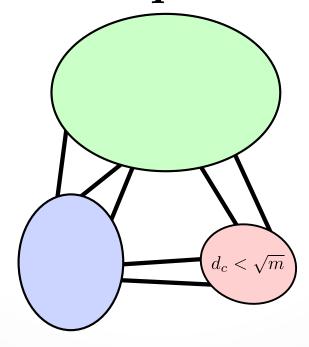
Result: clusters smaller than this scale cannot be resolved!

Consequences



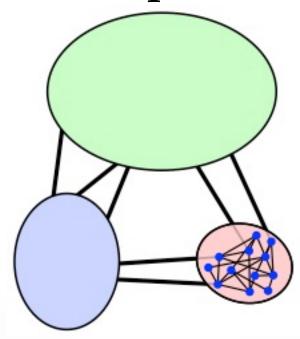
$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[l_c - \frac{1}{4} \left(\frac{d_c}{m} \right)^2 \right]$$
 modularity's scale

Consequences



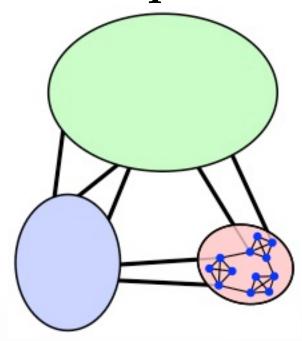
$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[l_c - \frac{1}{4} \left(\frac{d_c}{m} \right)^2 \right]$$
 modularity's scale

Consequences



$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left[l_c - \frac{1}{4} \left(\frac{d_c}{m} \right)^2 \right]$$
 modularity's scale

Consequences



Local optimization

Principle:

- Communities are local structures
- Local exploration of the network, involving the subgraph and its neighborhood

Advantages:

- Absence of global scales -> no resolution limit
- One can analyze only parts of the network

Local optimization

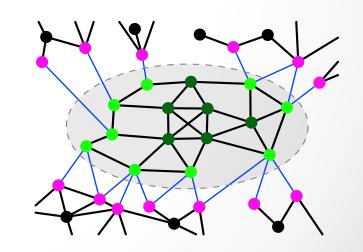
Implementation:

- Function Q(C) that assigns a score to each subgraph
- Best cluster -> cluster corresponding to the maximum/ minimum of Q(C) over the set of subgraphs including a seed node

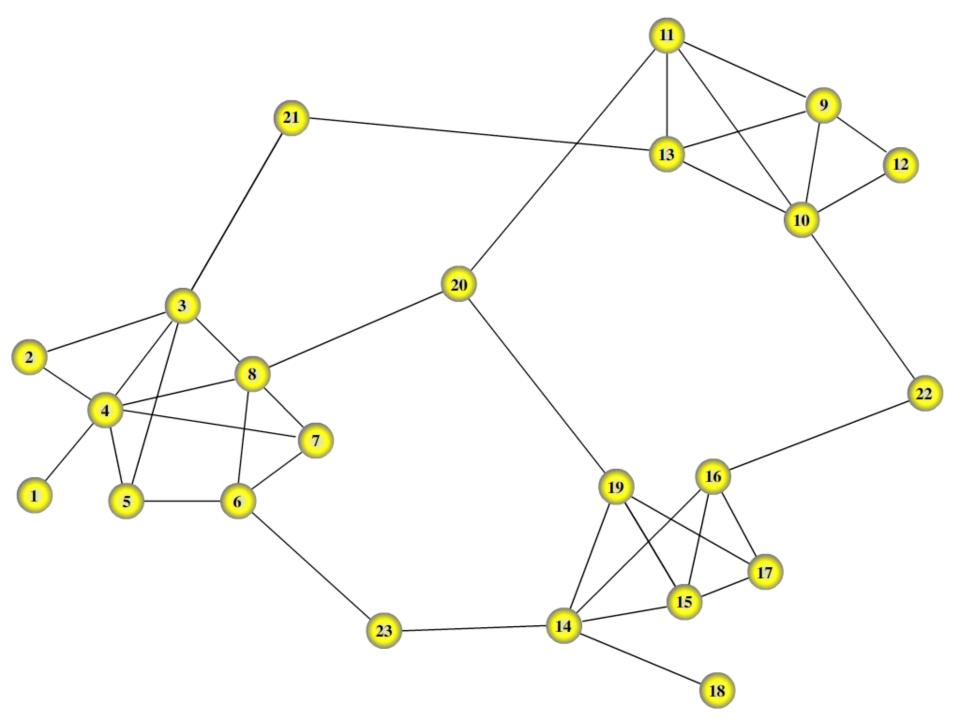
Example: Local Fitness Method (LFM)

Fitness of cluster *C* :

$$f_{\mathcal{C}} = \frac{k_{in}^{\mathcal{C}}}{(k_{in}^{\mathcal{C}} + k_{out}^{\mathcal{C}})^{\alpha}} = \frac{2l_{\mathcal{C}}}{d_{\mathcal{C}}^{\alpha}}$$



A. Lancichinetti, S. F., J. Kertész, New. J. Phys. 11, 033015 (2009)



Local optimization: OSLOM

Basics:

- LFM with fitness expressing the statistical significance of a cluster with respect to random fluctuations
- Statistical significance evaluated with Order Statistics

First multifunctional method:

- Link direction
- Link weight
- Overlapping clusters
- Hierarchy

A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. F., PLoS One 6, e18961 (2011)

Local optimization: OSLOM



Welcome to OSLOM's Web page

OSLOM means Order Statistics Local Optimization Method and it's a clustering algorithm designed for networks.

Download the code (beta version 2.4, last update: September, 2011)

The package contains the source code and the instructions to compile and run the program. You will also get a simple script which we implemented to visualize the clusters found by OSLOM. This script writes a pajek file which in turn can be processed by <u>pajek</u> or <u>gephi</u>.

This is a nice example of how the visualization looks like.



Home

Codes

Publications

Team

Contacts

http://www.oslom.org/

How to test community detection algorithms?

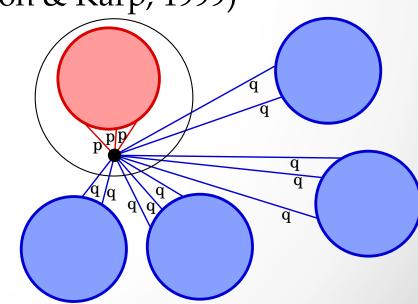
Question: how to test clustering algorithms?

Answer: checking whether they are able to recover the known community structure of benchmark graphs

Planted l-partition model (Condon & Karp, 1999)

Ingredients:

- 1) p=probability that vertices of the same cluster are joined
- 2) q=probability that vertices of different clusters are joined

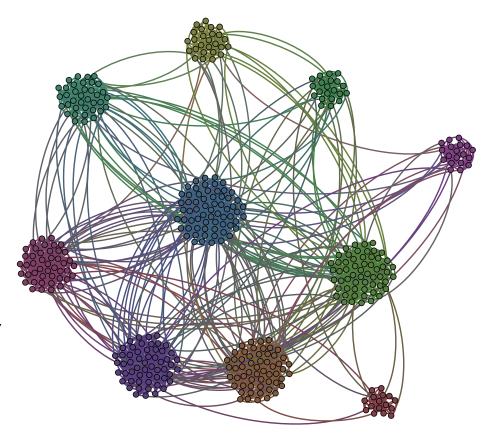


Principle: if p > q the groups are communities

The LFR benchmark

Realistic feature: power law distributions of degree and community size

A. Lancichinetti, S. F., F. Radicchi, Phys. Rev. E 78, 046110 (2008)



https://sites.google.com/site/andrealancichinetti/files/

https://github.com/networkx/networkx/blob/master/networkx/algorithms/community/community_generators.py

The LFR benchmark

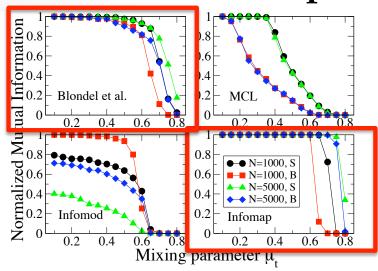
A comparative analysis

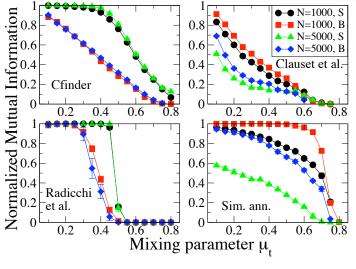
| Author | Label | Order |
|---------------------|----------------------|------------------------------------|
| Girvan & Newman | GN | $O(nm^2)$ |
| Clauset et al. | Clauset et al. | $O(n\log^2 n)$ |
| Blondel et al. | Blondel et al. | O(m) |
| Guimerà et al. | Sim. Ann. | parameter dependent |
| Radicchi et al. | Radicchi et al. | $O(m^4/n^2)$ |
| Palla et al. | Cfinder | $O(\exp(n))$ |
| Van Dongen | MCL | $O(nk^2), k < n \text{ parameter}$ |
| Rosvall & Bergstrom | Infomod | parameter dependent |
| Rosvall & Bergstrom | Infomap | O(m) |
| Donetti & Muñoz | DM | $O(n^3)$ |
| Newman & Leicht | EM | parameter dependent |
| Ronhovde & Nussinov | RN | $O(n^{\beta}), \beta \sim 1$ |

A. Lancichinetti, S. F., Phys. Rev. E 80, 056117 (2009)

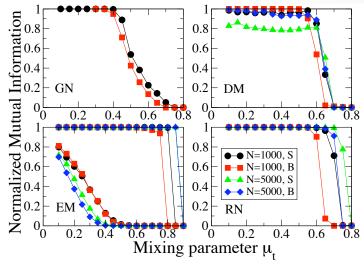
The LFR benchmark

A comparative analysis





- ... and the winner is:
- Infomap
- Louvain method *



Consensus clustering

Problem: Stochastic (non-deterministic) methods yield many result partitions: which one shall one choose?

Solution: Searching for the partition which is most similar, on average, to the input partitions (median or consensus partition)

Difficult combinatorial optimization task: greedy solution (consensus matrix)

A. Lancichinetti, S. F., Sci. Rep. 2, 336 (2012).

Consensus matrix

Definition

• Matrix **D** whose entry D_{ij} is the frequency that vertices i and j were in the same cluster in the input partitions

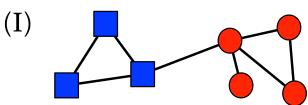
Starting point: network G with n vertices, clustering method A.

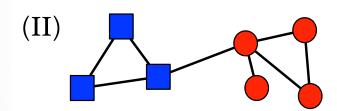
- Apply A on $G n_P$ times -> n_P partitions
- Compute the consensus matrix D: D_{ij} is the number of partitions in which vertices i and j of G are assigned to the same cluster, divided by n_p
- All entries of D below a chosen threshold t are set to zero
- Apply A on **D** n_P times -> n_P partitions
- If the partitions are all equal, stop (the consensus matrix would be block-diagonal). Otherwise go back to 2.

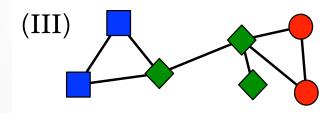
A simple example

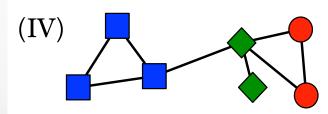
Original Graph

Consensus Matrix





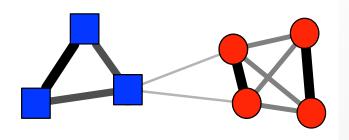






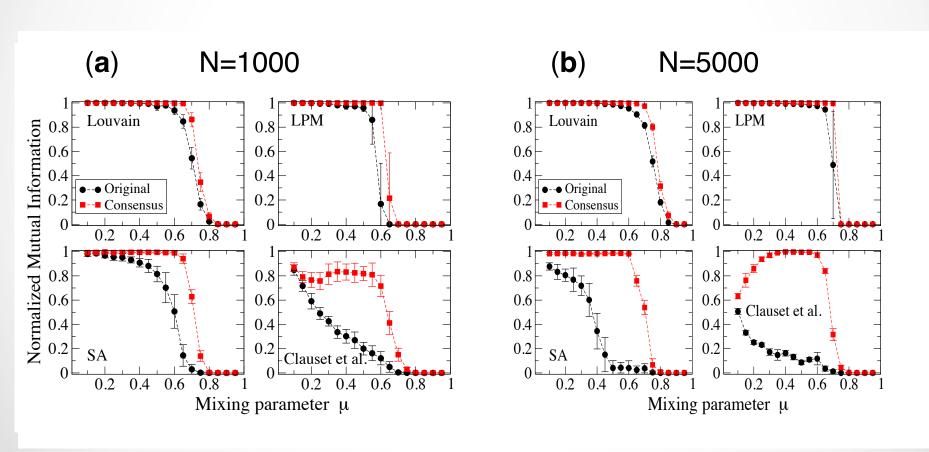
$$D_{ij} = 2/4$$

$$D_{ij} = 1/4$$



Results

LFR benchmarks



Consensus in dynamic networks

- Succession of snapshots, corresponding to overlapping time windows of size Δt : [t_0 , t_0 + Δt], [t_0 +1, t_0 +1+ Δt], [t_m - Δt , t_m]
- D_{ij} = number of times vertices i and j are clustered together, divided by the number of partitions corresponding to snapshots including both vertices

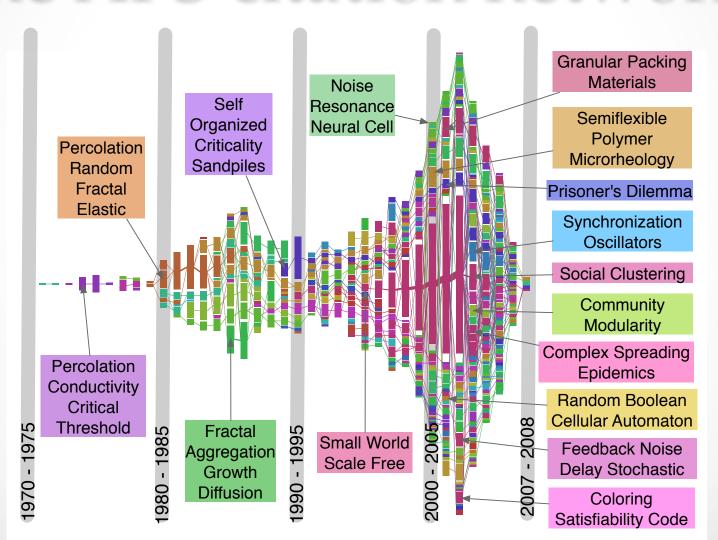
Tracking dynamic clusters: $C_t \longrightarrow C_{t+1}$?

Strategy: computing the Jaccard index of C_t with all clusters of the partition at time t+1, and pick the cluster with the highest value. Same procedure to find the "father" of cluster C_{t+1}

Criterion:

- A and B are each other's best match: A "survives" to time t +1
- A and B are not each other's best match: A "dies" at time t and B is considered as a new cluster.

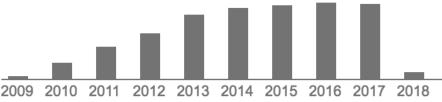
Tracking dynamic clusters: the APS citation network



Summary

- 1) What is a community? **No unique answer! Definition is system- and problem-dependent**
- 2) Magic method? No such thing! Domain dependent methods?
- 3) Global optimization methods have important limits: local optimization looks more natural and promising
- 4) Consensus clustering useful technique to find robust partitions
- 5) Attention on validation

Total citations Cited by 6256



Physics Reports 486 (2010) 75-174

ents lists available at ScienceDirect

Physics Reports

epage: www.elsevier.com/locate/physrep



Scholar articles

Community detection in graphs
S Fortunato - Physics reports, 2010
Cited by 6256 Related articles All 44 versions

Article history: Accepted 5 November 2009 Available online 4 December 2009 editor: 1, Procaccia phs

sdation, Viale S. Severo 65, 10133, Torino, I, Italy

STRACT

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly

S. F., Phys. Rep. 486, 75-174 (2010)

Top 25 Hottest Articles

Physics and Astronomy

Most Cited Physics Reports Articles

The most cited articles published since 2008, extracted from SciVerse Scopus.

Community detection in graphs

Volume 486, Issues 3-5, February 2010, Pages 75-174 Fortunato, S.

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks. © 2009 Elsevier B.V.

S. F., D. Hric, Phys. Rep. 659, 1-44 (2016)

Physics Reports 659 (2016) 1-44



Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep



Community detection in networks: A user guide



Santo Fortunato a,b,*, Darko Hricb

^a Center for Complex Networks and Systems Research, School of Informatics and Computing, and Indiana University Network Science Institute (IUNI), Indiana University, Bloomington, USA

ARTICLE INFO

Article history: Accepted 26 September 2016 Available online 6 October 2016 editor: I. Procaccia

Keywords: Networks Communities Clustering

ABSTRACT

Community detection in networks is one of the most popular topics of modern network science. Communities, or clusters, are usually groups of vertices having higher probability of being connected to each other than to members of other groups, though other patterns are possible. Identifying communities is an ill-defined problem. There are no universal protocols on the fundamental ingredients, like the definition of community itself, nor on other crucial issues, like the validation of algorithms and the comparison of their performances. This has generated a number of confusions and misconceptions, which undermine the progress in the field. We offer a guided tour through the main aspects of the problem. We also point out strengths and weaknesses of popular methods, and give directions to their use.

© 2016 Elsevier B.V. All rights reserved.

Contents

| 1. | Introd | uction | |
|----|--------|---|---|
| 2. | What | are communities? | |
| | 2.1. | Variables | |
| | 2.2. | Classic view | |
| | 2.3. | Modern view | |
| 3. | Valida | ntion | |
| | 3.1. | Artificial benchmarks | 1 |
| | 3.2. | Partition similarity measures | 1 |
| | 3.3. | Detectability | 1 |
| | 3.4. | Structure versus metadata | 1 |
| | 3.5. | Community structure in real networks | 1 |
| 4. | Metho | ods | 2 |
| | 4.1. | How many clusters? | 2 |
| | 4.2. | Consensus clustering | 2 |
| | 4.3. | Spectral methods | |
| | 4.4. | Overlapping communities: vertex or edge clustering? | 2 |
| | 4.5. | Methods based on statistical inference | 2 |
| | 4.6. | Methods based on optimisation | 2 |
| | 4.7. | Methods based on dynamics | 3 |
| | | | |

E-mail address: santo@indiana.edu (S. Fortunato).

http://dx.doi.org/10.1016/j.physrep.2016.09.002 0370-1573/© 2016 Elsevier B.V. All rights reserved.

b Department of Computer Science, Aalto University School of Science, P.O. Box 15400, FI-00076, Finland

^{*} Corresponding author at: Center for Complex Networks and Systems Research, School of Informatics and Computing, and Indiana University Network Science Institute (IUNI), Indiana University, Bloomington, USA.