

Utilizing scale-free networks to support the search for scientific publications

Claudia Hauff, Andreas Nürnberger

14.03.2006

Overview

- Introduction
 - Motivation
 - Idea
- Network Theory
 - Scale-free networks
- Experiments and Results
 - Citebase data set
 - Visual Paper Finder (ViPF)
- Conclusion and Future Work

Motivation

- Search for scientific publications:
 - Standard tools: Google (Scholar), Citeseer, Citebase

User knows research area or wanted publication



User knows very few or nothing about the research area



- Search engines offer only inadequate indicators for the importance or influence of a publication
- Absolute number of incoming references as indicator disadvantages recently published papers

Idea

- Take the age of a publication into account during the calculation of the importance indicator
- Applying a network model, the expected number of incoming references is calculated
- Comparison between the actual *act* and expected number *exp* of incoming references
 - $act > exp \rightarrow$ high importance
 - $act < exp \rightarrow$ low importance
- Problem: determination of the parameters of the network model

Overview

- Introduction
 - Motivation
 - Idea
- Network Theory
 - Scale-free networks
- Experiments and Results
 - Citebase data set
 - Visual Paper Finder (ViPF)
- Conclusion and Future Work

Networks

- Nodes represent objects and edges the relations between them
 - Citation network: publications (nodes) and references (edges)
- Properties of networks
 - Degree distribution $P(k)$: probability that a node is attached to exactly k edges
 - Average shortest path λ
- Traditional Network model
 - „Random graph“: fixed number of nodes, random edge insertion

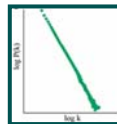
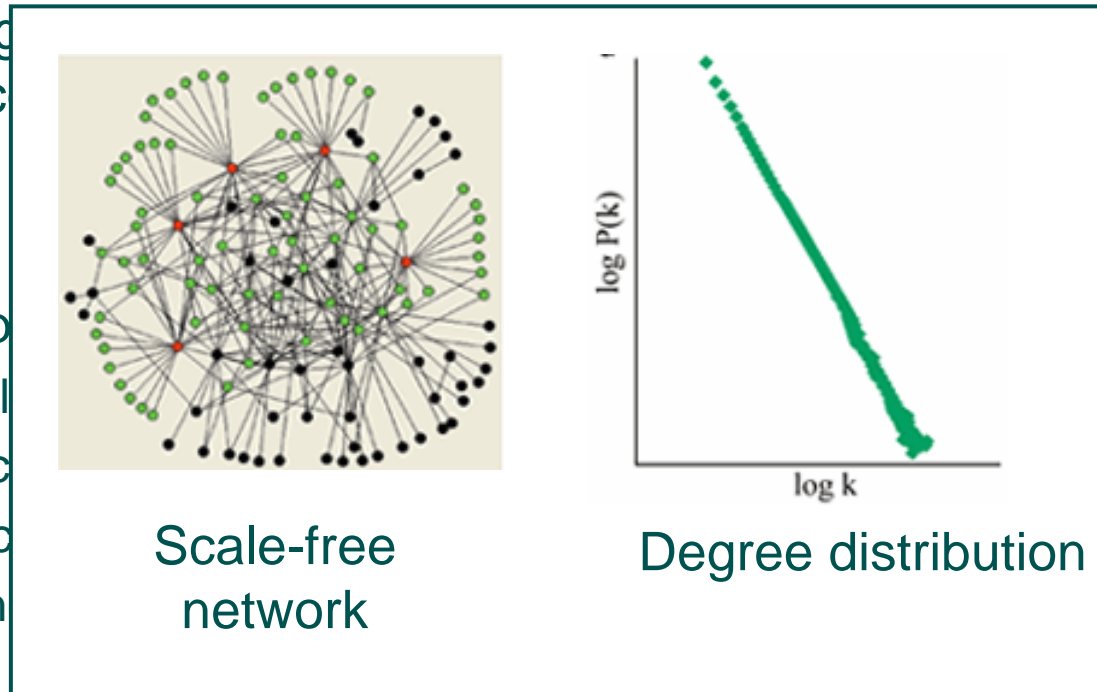
Scale-Free Networks

Barabási et al., 1998

- In the late 1990s: Analysis of large data sets became possible

- Finding
lowly c

- Examp
– Biol
– Tec
– Soc
– Oth



Barabási-Albert Model

- Dynamic network model based on growth and preferential attachment
- Algorithm:
 - Starting with m_0 nodes, at each time step one node and $m \leq m_0$ undirected edges are added to the network
 - The probability of a node s receiving one of the new edges at time t linearly depends on its degree $k(s,t)$

$$P(k(s,t)) = \frac{k(s,t)}{\sum_u k(u,t)}$$

- Numerical simulations and analytical approaches resulted in a power law distribution with $\gamma=3$

Continuum Theory

- Analytical approach by Barabási and Albert
 - Calculation of the number of edges $k(s,t)$ of node s at time t (the node was born at time s)
 - Corresponds to the expected number exp of edges

$$\frac{\partial k(s,t)}{\partial t} = m \frac{k(s,t)}{\sum_u k(u,t)}$$

$$k(s,t) = m \left(\frac{s}{t} \right)^{-\frac{1}{2}}$$

- Necessary network statistics for the Barabási-Albert model
 - Number of newly introduced edges (m)
 - Sorting of nodes according to their age (s)
 - Collection size (t)

- Model accounts for directed networks
- The number of edges m , that enter the network at time t varies (c_0 and b are constants)

$$m(t) = c_0 t^b$$

- Utilizing the continuum approach for networks with $\gamma > 2$ yields (B is a constant)

$$\frac{\partial k_{in}(s, t)}{\partial t} = \frac{1+b}{1+B} \frac{k_{in}(s, t) + Bc_0 t^b / (1+b)}{t}$$
$$k_{in}(s, t) = \left(\frac{Bc_0 s^b}{1-Bb} \right) \left(\frac{s}{t} \right)^{-(1+b)/(1+B)} - \frac{Bc_0 t^b}{1-Bb}$$

Overview

- Introduction
 - Motivation
 - Idea
- Network Theory
 - Scale-free networks
- Experiments and Results
 - Citebase data set
 - Visual Paper Finder (ViPF)
- Conclusion and Future Work

Citebase I

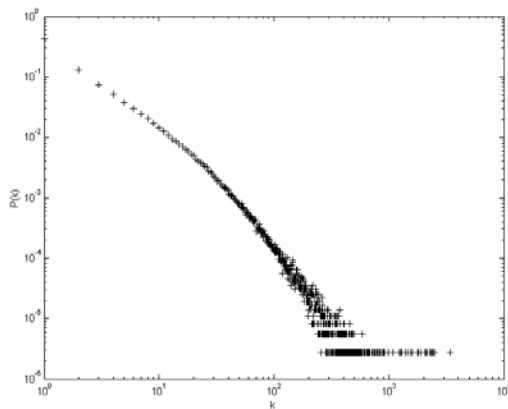
- Web service for scientific publications in the fields of physics, mathematics and computer science
 - ~363000 publications, ~2.5 million references
 - 25% of the publications do not contain citations
 - Meta data: titel, authors, creation date, abstract, topic
 - 39% of the publications belong to one or more topics
- Average shortest path length: $\lambda_{\text{undir}}=6.47$

Citebase II

- Accelerated growth of the average number of citations:
 - Publications until december 1998: $m=4.8$ (~96000 Publikationen)
 - Publications until december 2002: $m=6.5$ (~154000 Publikationen)
 - Publications until december 2004: $m=6.9$ (~340000 Publikationen)

$$m(t) = 0.1043t^{0.3439}$$

- Degree distribution of the incoming citations is approximated by the accelerated growth model with $\gamma=2.16$



Visual Paper Finder (ViPF) I

scale-free Search

Citation network

Top topics

- statistical mechanics (54)
- disordered systems neural networks (29)
- citation (3)
- collaborations (3)
- condensed matter (2)

Result list

16. Scale-Free Magnetic Networks: Comparing Observational Data with a Self-Organizing Model of the Coronal Field (2003-09-08)
Paczuski, Maya and Hughes, David
We propose that the coronal magnetic field, linking concentrations on the photosphere through an interwoven web of flux, embodies a scale-free network. It arises from a self-organized critical dynamics ...
<http://arxiv.org/abs/astro-ph/0309230>

17. Critical threshold and dynamics of a general rumor model on complex networks (2005-08-01)
Nekovee, Maziar and Moreno, Yamir and Bianconi, Ginestra and Marsili, Matteo
We derive mean-field equations that describe the dynamics of a general model of rumor spreading on complex networks, and use analytical and numerical solutions of these equations to examine the threshold ...
<http://arxiv.org/abs/physics/0508008>

18. Epidemic dynamics in finite size scale-free networks (2002-02-18)
Pastor-Satorras, Romualdo and Vespignani, Alessandro
Many real networks present a bounded scale-free behavior with a connectivity cut-off due to physical constraints or a finite network size. We study epidemic dynamics in bounded scale-free networks with ...
<http://arxiv.org/abs/cond-mat/0202298>

19. Scaling properties of scale-free evolving networks: Continuous approach (2000-12-01)
Mendes, J. F. F. and Dorogovtsev, S. N.
Scaling behavior of scale-free evolving networks arising in communications, citations, collaborations, etc. areas is studied. We derive universal scaling relations describing properties of such networks ...
<http://arxiv.org/abs/cond-mat/0012009>

20. Free Energy of QCD at High Temperature (1995-10-25)
Braaten, Eric and Nieto, Agustin

Graph retrieved min max Page 2 out of 7

Visual Paper Finder (ViPF) II

scale-free Search

Topic

Info Panel

Hierarchical Organization in Complex Networks (2002-06-08)
Ravasz, Erzsébet Barabasi, Albert-Laszlo
and society : they are high degree of se two features are the consequence of a hierarchical ...
<http://arxiv.org/abs/cond-mat/0206130>

Information panel

statistical mechanics (54)
disordered systems neural networks (29)
condensed matter (4)
architecture (4)
networking internet architecture (4)

31. Generation of uncorrelated random scale-free networks (2004-08-05)
Pastor-Satorras, Romualdo and Boguna, Marian and Catanzaro, Michele
Uncorrelated random scale-free networks are useful null models to check the accuracy on the analytical solutions of dynamical processes defined on complex networks. We propose and analyze a model capable ...
<http://arxiv.org/abs/cond-mat/0408110>

32. Minimal models of weighted scale-free networks (2004-09-04)
Mendes, J. F. F. and Dorogovtsev, S. N.
We consider a class of simple, non-trivial models of evolving weighted scale-free networks. The network evolution in these models is determined by attachment of new vertices to ends of preferentially chosen ...
<http://arxiv.org/abs/cond-mat/0408343>

33. Scale Free Subnetworks by Design and Dynamics (2005-02-06)
Costa, Luciano da Fontoura
This article addresses the degree distribution of subnetworks, namely the number of links between the nodes in each subnetwork and the remainder of the structure (cond-mat/0408076). The transformation ...
<http://arxiv.org/abs/cond-mat/0502156>

34. Clustering and Synchronization of Oscillator Networks (2005-06-09)
Menzinger, Michael and McGraw, Patrick N.
Using a recently described technique for manipulating the clustering coefficient of a network without changing its degree distribution, we examine the effect of clustering on the synchronization of phase ...
<http://arxiv.org/abs/cond-mat/0501663>

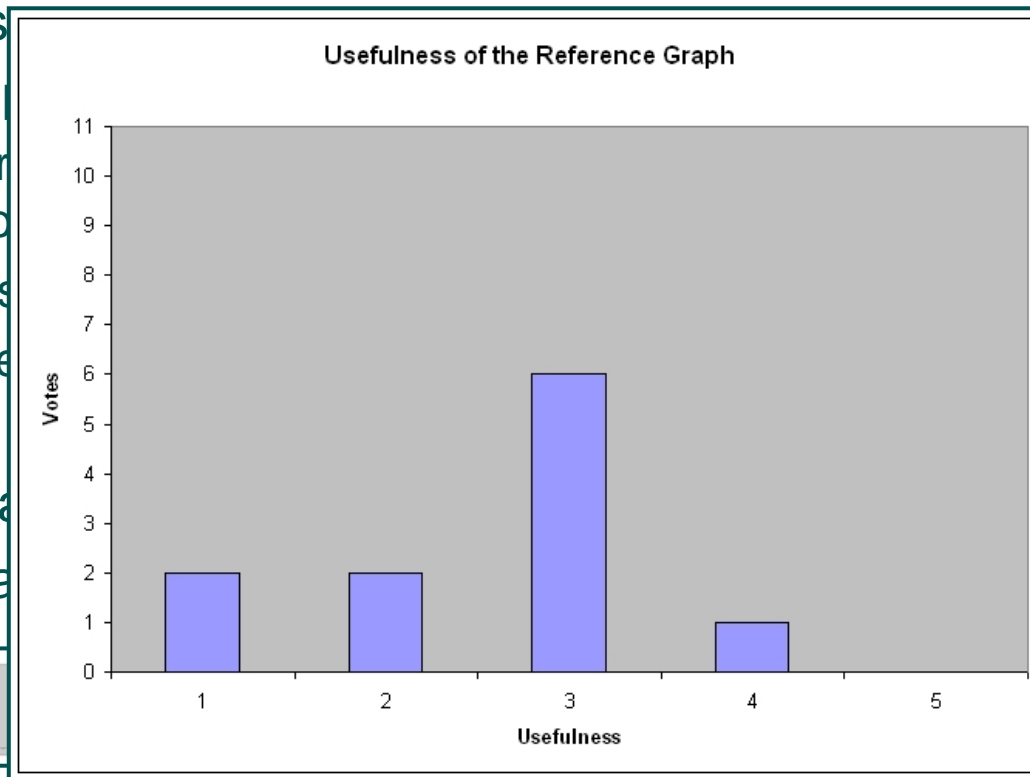
35. Statistical ensemble of scale-free random graphs (2001-11-26)
Burda, Z. and Correia, J. D. and Krzywicki, A.
A thorough discussion of the statistical ensemble of scale-free connected random free graphs is presented. Methods borrowed from field theory are used to define the ensemble and to study analytically its ...
<http://arxiv.org/abs/cond-mat/0104155>

36. Scale-free network on a vertical plane (2003-07-07)

Graph retrieved min max Page 3 out of 7

User Study

- 11 users
 - 1 male
 - Br
 - co
 - Us
 - Se

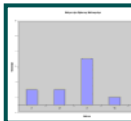


Retrieval (1),

“e learning“,

approved of
useful

- Display
- Age a



- Results of user study show a positive tendency, but do not allow a clear conclusion

Overview

- Introduction
 - Motivation
 - Idea
- Network Theory
 - Scale-free networks
- Experiments and Results
 - Citebase data set
 - Visual Paper Finder (ViPF)
- Conclusion and Future Work

Conclusion

- Approach, based on scale-free networks, which allows the principal inclusion of age in the calculation of a fitness indicator for network nodes
- Citation networks can be approximated by a scale-free network model
- Implementation of a tool that aims to support the user in his search for scientific publications

Future Work

- ViPF contains only a very simple visualization
 - Cluster visualization
 - Overview of entire network
- Accelerated growth does not capture all actions in a citation network
 - Usage of a cluster model for the different topics
 - Introduction of a weighted model to discount citations between papers written by one and the same author
- Larger user study

Thank You For Your Attention!