

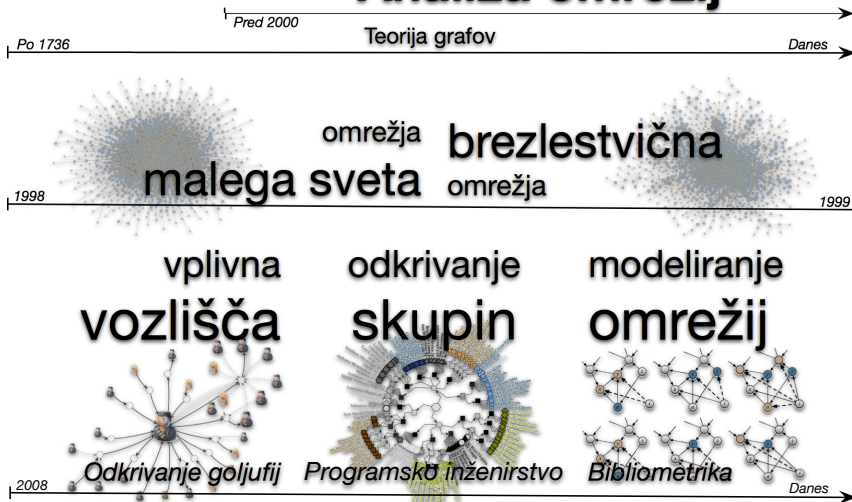
Analiza omrežij

Lovro Šubelj

Univerza v Ljubljani,
Fakulteta za računalništvo in informatiko

25.3.2014

Analiza omrežij



Analiza omrežij



Teorija grafov

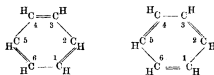
1736 *Problem königsberških mostov* (Euler)



1800s *Problem trgovskega potnika* (Hamilton)

1845 *Zakoni električnih vezij* (Kirchhoff)

1857 *Kemična struktura* (Kekulé)

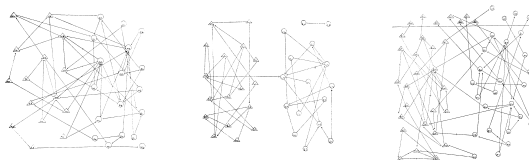


1950s *Operacijske raziskave* (Dijkstra, Kruskal, Ford)

1959 *Naključni grafi* (Erdős, Rényi)

Sociometrija & bibliometrika

1934 *Sociogrami* Moreno (1934)



1970s *Socialne mreže* Granovetter (1973); Zachary (1977); Freeman (1977, 1979)

1965 *Znanstveni prispevki* Price (1965)



SCIENCE CITATION INDEX

Analiza omrežij

< 2000 Manjši grafi (10^2 - 10^3 vozlišč)

$\approx$ 2000 *Komunikacijska omrežja* (10^5 - 10^8 vozlišč)

$\approx$ 2005 *Spletna družabna omrežja* (10^8 vozlišč)

$\approx$ 2010 *Svetovni splet* (10^9 vozlišč)

analiza omrežij $\rightarrow$ struktura & funkcija omrežij (*teorija omrežij*)

analiza povezav $\rightarrow$ rudarjenje vozlišč (*rudarjenje omrežij*)

GBDM $\rightarrow$ manjši grafi (*rudarjenje grafov*)



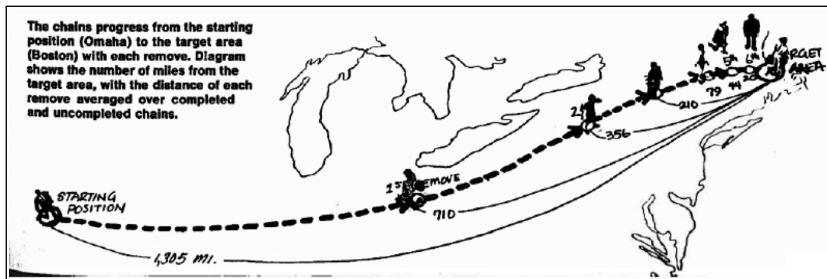
Analiza omrežij



Milgramov eksperiment

Pošiljanje pisma osebi v Bostonu prek prijateljev:

mali svet omrežij ali 6 stopenj ločenosti. Milgram (1967)



Vir: Milgram (1967)

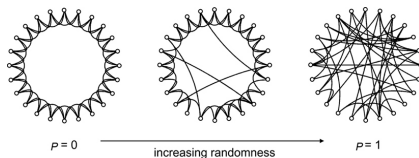
Moč šibkih povezav & šibkost krepih povezav. Granovetter (1973)

Model malega sveta

Naj bo k_i stopnja vozlišča i , Δ_i # povezanih sosedov & d_{ij} razdalja med vozliščema i, j .

$$L = \frac{1}{\binom{n}{2}} \sum_{ij} d_{ij} \quad C = \frac{1}{n} \sum_i \frac{\Delta_i}{\binom{k_i}{2}}$$

Naključno prevezovanje: Watts & Strogatz (1998)



Vir: <http://www.frontiersin.org/>

Regularni grafi *visok* L & C , naključni grafi *nizek* L & C .
Realna omrežja nizek L & *visok* C !

Omrežja malega sveta

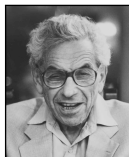
Lastnosti omrežij malega sveta:

6 stopenj ločenosti Milgram (1967)

7 (4) stopenj ločenosti za e-mail Dodds et al. (2003)

4 stopenj ločenosti na Facebooku Backstrom et al. (2012)

Majhna *Erdösova & Baconova* števila...



Realna omrežja (npr. P2P) omogočajo usmerjanje? Kleinberg (2001)

Decentralizirano iskanje v času polinomskem v $\mathcal{O}(\log n)$.

Analiza omrežij

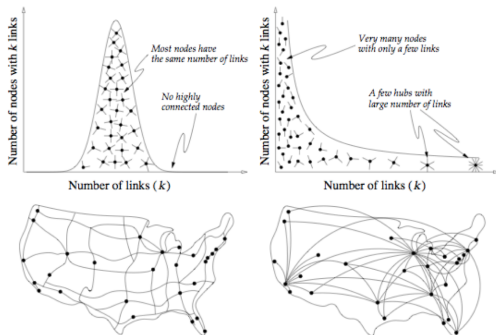


Normalna & potenčna porazdelitev

Citiranja, internet & splet imajo *potenčno* $P(k)$: Price (1965); Faloutsos et al. (1999)

$$P(k) \sim k^{-\alpha}$$

k_i je stopnja vozlišča i & α je potenčni eksponent, $\alpha > 1$.



Vir: Barabási (2002)

Model brezlestvičnih omrežij

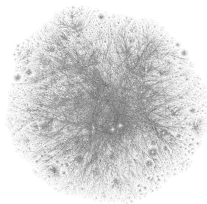
Omrežja s potenčnim repom $P(k)$ so *brezlestvična*.

Prednostno povezovanje: Barabási & Albert (1999)

- (1) tvori naključno povezavo z verjetnostjo p
- (2) sicer tvori povezavo z verjetnostjo $\frac{k_j}{\sum k_j}$

$$P(k) = k^{-\alpha} \quad \alpha = 1 + \frac{1}{1 - p}$$

Pojav, da *bogati bogatijo* (*kazala*).



Brezlestvična omrežja

Naj bo n & m # vozlišč ali povezav, zaporedoma, & k povprečna stopnja.

Lastnosti brezlestvičnih omrežij:

redka z $m \approx n$ Del Genio et al. (2011)

$1 < \alpha < 2$, k narašča z n (brez lestvice)

$2 < \alpha < 3$, *infekcija v epidemijo* Sinha (2011)

ranljiva na napade Albert et al. (2000)



Internet po odstranitvi 10% kazal in 10% naključnih vozlišč.

Analiza omrežij



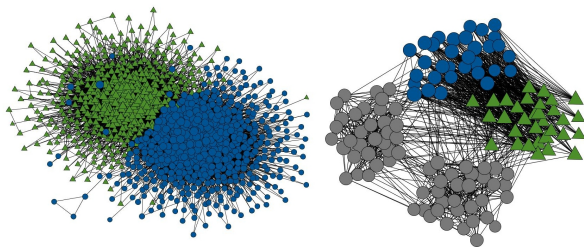
Skupine v realnih omrežjih

skupnosti tesno povezane skupine, ki so dobro ločene

(gosti podgrafi redkih grafov) Girvan & Newman (2002)

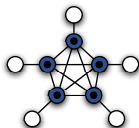
module skupine povezane s podobnimi vozlišči Newman & Leicht (2007)

(skupine s podobnim vzorcem povezovanja) Šubelj & Bajec (2012)

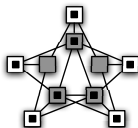


Formalizem skupin vozlišč

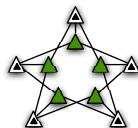
Naj bo S skupina vozlišč (polna) & T njen vzorec (označena). Šubelj et al. (2013a)



Skupnost ($S = T$)



Mešanica ($S \approx T$)



Modul ($S \neq T$)



$\tau_{S,T}$ je parameter skupine S & vzorca T .

$$\tau_{S,T} = \frac{|S \cap T|}{|S \cup T|}$$

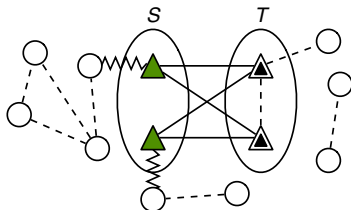
$\tau = 1$ za skupnosti, $\tau \approx \frac{1}{2}$ za mešanice & $\tau = 0$ za module.

Kriterij skupin vozlišč

Naj bo $L_{S,T}$ število povezav med S & T .

$$W_{S,T} = \dots \left(\frac{L_{S,T}}{|S||T|} - \frac{L_{S,T^C}}{|S||T^C|} \right) \text{ Šubelj et al. (2013a)}$$

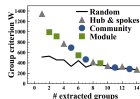
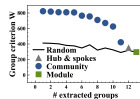
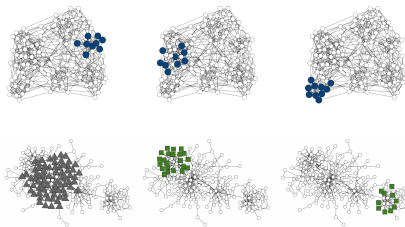
Lokalni asimetrični kriterij, ki *podpira povezave* v (S, T) & *kaznuje povezave* v (S, T^C) . Skladen z različnimi modeli za $S = T$. Zhao et al. (2011)



Odkrivanje skupin vozlišč

Zaporedno odstranjevanje: Šubelj et al. (2013a) & Zhao et al. (2011)

- (1) S & T , ki maksimizirata W
- (2) Odstrani *povezave med S & T*
- (–) Dokler je W večji kot naključno



<http://github.com/gw0/nodegroups>

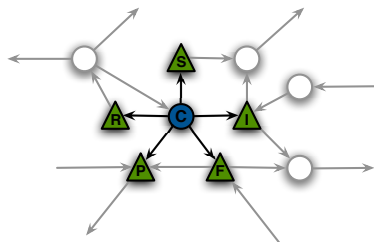
Programska omrežja

Omrežja programskih odvisnosti: Šubelj & Bajec (2011)

vozišča → *razredi* objektno-usmerjenega programa

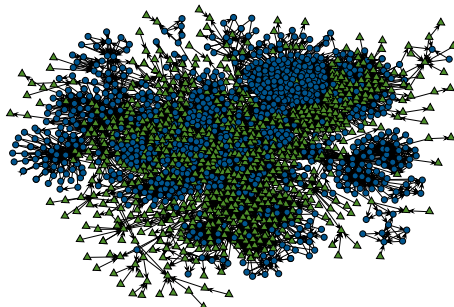
povezave → *odvisnosti* med razredi (npr. dedovanje)

```
class C extends S implements I {  
    F field;  
    public C() { ... }  
    void foo(P parameter) { ... }  
    private R bar() { ... }  
}
```



Zgradba programskih omrežij

Programska omrežja podobna drugim omrežjem. Valverde et al. (2002)

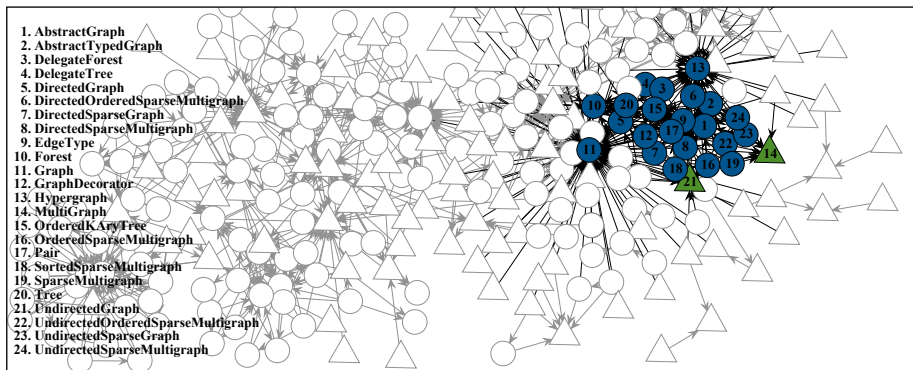


programska omrežja = Šubelj et al. (2013b)

= gosta družbena omrežja + redka topologija interneta

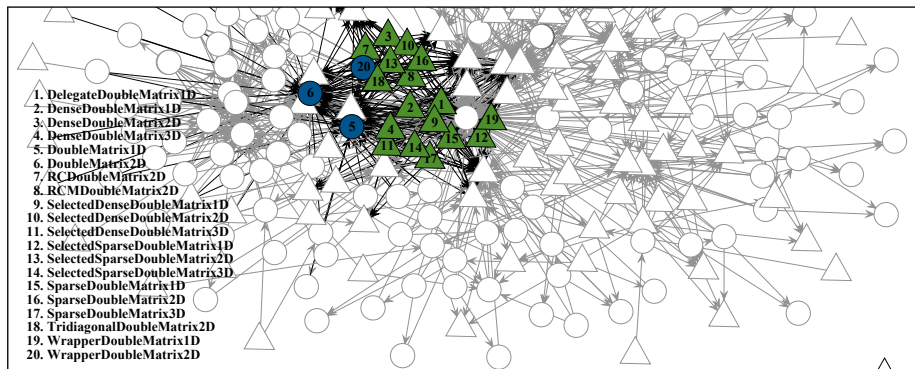
Skupnosti v programskih omrežjih

Skupnosti so razredi v *jedru programja*. Šubelj & Bajec (2011)



Moduli v programskih omrežjih

Moduli so razredi z *enako funkcionalnostjo*. Šubelj & Bajec (2012)



Programsko inženirstvo

CA pri napovedovanju *paketov razredov*: Šubelj et al. (2013b)

Programje	# Razredov	# Paketov	Sosedi Γ	Skupine S	Omrežje N	Privzeto	Naključno
<i>JBullet</i>	107	11	72.0%	75.7%	64.5%	28.0%	8.6%
<i>colt</i>	154	16	58.4%	73.4%	55.2%	22.7%	5.9%
<i>JUNG</i>	237	31	72.2%	74.2%	65.0%	11.4%	3.3%
<i>Lucene</i>	1335	178	47.1%	49.2%	43.7%	6.4%	0.6%

CA pri napovedovanju *visoko-nivojskih paketov razredov*:

Programje	# Razredov	# Paketov	Sosedi Γ	Skupine S	Omrežje N	Privzeto	Naključno
<i>JBullet</i>	107	5	84.6%	85.0%	78.5%	64.5%	20.4%
<i>colt</i>	154	10	86.4%	83.8%	69.5%	39.0%	9.7%
<i>JUNG</i>	237	5	89.1%	90.5%	91.1%	44.3%	20.3%
<i>Lucene</i>	1335	15	85.5%	90.8%	85.0%	28.2%	6.6%

CA pri napovedovanju *vrste, verzije & avtorja razreda*:

Problem	# Kategorij	Sosedi Γ	Skupine S	Omrežje N	Privzeto	Naključno
Vrsta razreda	2	65.0%	85.2%	84.8%	84.4%	49.9%
Verzija razreda	9	67.7%	72.8%	66.2%	44.3%	11.2%
Avtor razreda	11	71.6%	71.0%	70.9%	44.3%	9.2%

Analiza omrežij

Teorija grafov

omrežja **brezlestvična**
malega sveta omrežja

vplivna
vozlišča

odkrivanje skupin

modeliranje omrežij

Ödkrivande goljufij

Programsko inženirstvo

Bibliometrika

2008

Danes

Družbena & ostala omrežja

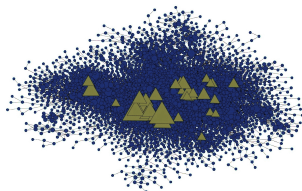
Omrežja so brezlestvična, malega sveta, se gostijo & krčijo.

Nakopičenost C & mešanje r (korelacije stopenj): Newman (2002)

Družbena $\rightarrow$ *visok C & asortativnost $r > 0$* Newman & Park (2003)

Ostala $\rightarrow$ *nižji C & disasortativnost $r < 0$* Šubelj & Bajec (2012)

Citiranja $\rightarrow$ *nižji C & mešanje $r \approx 0$*

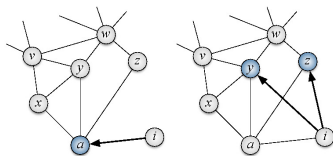


Model gorečega gozda

Zaporedno dodajanje: Leskovec et al. (2007)

- (1) vozlišče izbere ambasadorja a & tvori povezavo z a
- (2) vozlišče izbere sosede $a_1, \dots, a_{x_p}$ & tvori povezave z a_i
- (3) $a_1, \dots, a_{x_p}$ so novi ambasadorji

p je verjetnost gorenja & $x_p \sim G(\frac{p}{1-p})$.

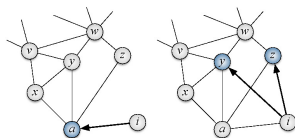


Omrežja so brezlestvična, malega sveta & *asortativna* $r > 0$.

Model omrežij citiranja

Dinamika avtorjev:

- (1) avtor prebere prispevek a & ga citira
- (2) avtor izbere reference $a_1, \dots, a_{x_p}$ & jih citira
- (3) $a_1, \dots, a_{x_p}$ izbrane za nadaljnje branje



Avtorji preberejo vse citirane prispevke...

$\approx 80\%$ citiranih prispevkov ni prebranih. Simkin & Roychowdhury (2003)

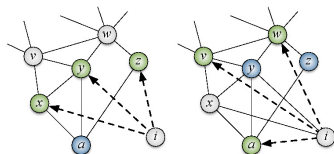
Branje & citiranje prispevkov sta dva procesa!

Model citiranja

Zaporedno dodajanje: Šubelj & Bajec (2013)

- (1) vozlišče izbere ambasadorja a
- (2) vozlišče izbere sosede $a_1, \dots, a_{x_p}$
vozlišče izbere sosede $l_1, \dots, l_{x_q}$ & tvori povezave z l_i
- (3) $a_1, \dots, a_{x_p}$ so novi ambasadorji

q je verjetnost povezovanja & $x_q \sim G(\frac{q}{1-q})$.



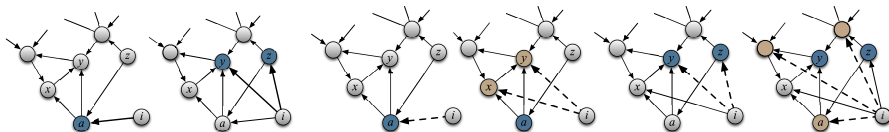
Omrežja so brezlestvična, malega sveta & (dis)asortativna $r \approx 0$.

Model citiranj (II)

Razširitev na usmerjena omrežja:

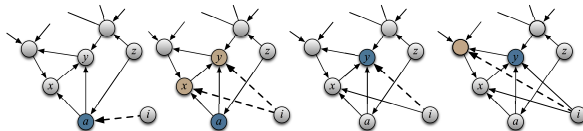
gorenje $\rightarrow$ verjetnosti p_{fwd} & p_{bck}

povezovanje $\rightarrow$ verjetnosti q_{amb} & q_{lnk}



Goreči gozd Leskovec et al. (2007)

Citiranja (naš)



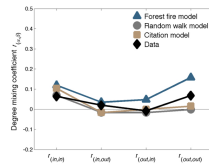
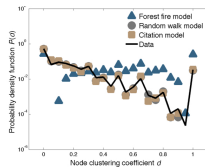
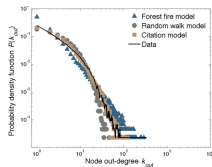
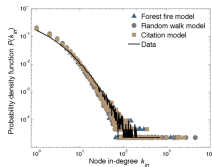
Naključni sprehod Vázquez (2003)

Primerjava modelov citiranj

Dobljena omrežja so brezlestvična & malega sveta.

Model	k	C	D	α_{in}	α_{out}	$r_{(in,in)}$	$r_{(in,out)}$	$r_{(out,in)}$	$r_{(out,out)}$	$\mathcal{F}$
WoS ILS (1945–)	8.05	0.11	0.14	2.47	3.35	0.06	0.02	−0.01	0.07	1.000
Citiranja (naš)	7.97	0.11	0.13	2.58	3.31	0.12	−0.01	0.01	0.01	0.863
Naključni sprehod	7.91	0.11	0.13	2.61	4.68	0.07	−0.01	−0.02	0.00	0.695
Naključni avtor	7.48	0.21	0.26	2.36	3.26	0.16	−0.02	0.08	0.15	0.400
Goreči gozd	7.88	0.40	0.52	2.51	2.45	0.13	0.04	0.06	0.17	0.087

Le omrežja modela citiranj primerljiva omrežjem citiranj!



Bibliometrika

Citiranih prispevkov: $k/2$.

Prebranih prispevkov: $s = \left(1 - \frac{p_{fwd}}{1-p_{fwd}} - \frac{p_{bck}}{1-p_{bck}}\right)^{-1}$.

% Prebran članek tudi citiran: q_{amb} .

Citatov kopiranih iz drugih prispevkov: $s \cdot \frac{q_{lnk}}{1-q_{lnk}}$.

Razmerje med citati & storitvami: $\frac{p_{fwd}}{1-p_{fwd}} : \frac{p_{bck}}{1-p_{bck}}$.

WoS baza	Citiranje		Branje		Pridobivanje		
	# Citati	% Kopiran	# Branje	% Citiran	% Citati	% Storitve	% Drugo
ILS (1945–)	3.98	86.1%	2.14	27.9%	29.2%	41.0%	29.8%
TM (1954–)	2.93	79.7%	1.47	45.2%	74.7%	0.5%	24.9%
AI (1964–)	4.52	87.3%	1.47	40.9%	25.8%	47.6%	26.6%
SE (1954–)	2.78	81.5%	1.58	36.4%	68.8%	2.0%	29.2%
CYB (1961–)	2.18	69.6%	1.59	43.2%	24.5%	37.8%	37.6%

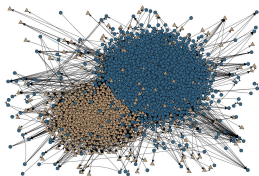
Bibliometrika (II)

WoS ILS	Citiranje		Branje		Pridobivanje		
	# Citati	% Kopiran	# Branje	% Citiran	% Citati	% Storitve	% Drugo
1945–2013	3.98	86.1%	2.14	27.9%	29.2%	41.0%	29.8%
1970–1980	2.23	52.1%	3.39	33.5%	41.4%	0.0%	58.5%
1980–1990	2.62	65.1%	2.96	33.0%	48.3%	1.1%	50.6%
1990–2000	3.42	81.6%	2.38	29.0%	40.3%	23.2%	36.5%
2000–2010	5.06	83.6%	2.90	32.2%	40.7%	27.5%	31.7%

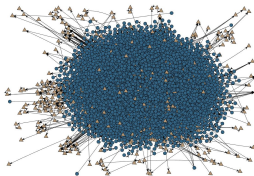
WoS TM	Citiranje		Branje		Pridobivanje		
	# Citati	% Kopiran	# Branje	% Citiran	% Citati	% Storitve	% Drugo
1954–2013	2.93	79.7%	1.47	45.2%	74.7%	0.5%	24.9%
1970–1980	1.99	33.4%	2.27	60.8%	35.4%	24.7%	39.9%
1980–1990	2.52	63.8%	1.76	56.7%	55.1%	10.9%	34.0%
1990–2000	2.93	76.5%	1.61	47.7%	58.8%	12.0%	29.2%
2000–2010	2.96	81.0%	1.45	43.0%	70.8%	2.9%	26.3%

Bibliometrika (III)

WoS revija	Citiranje		Branje		Pridobivanje		
	# Citati	% Kopiran	# Branje	% Citiran	% Citati	% Storitve	% Drugo
<i>Scientometrics</i>	3.83	87.6%	1.40	38.5%	54.3%	23.4%	22.3%
<i>Expert Syst. Appl.</i>	2.12	81.0%	1.67	25.9%	34.7%	27.8%	37.6%
<i>Commun. ACM</i>	1.32	46.5%	1.70	42.0%	8.7%	38.5%	52.8%



Scientometrics

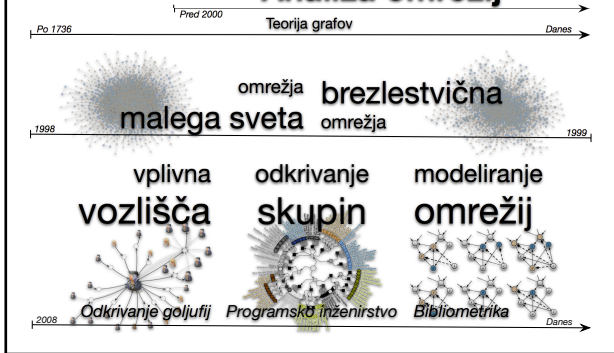


Expert Syst. Appl.



Commun. ACM

Analiza omrežij



<http://lovro.lpt.fri.uni-lj.si/>

lovro.subelj@fri.uni-lj.si

- R. Albert, H. Jeong, & A. L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
- L. Backstrom, P. Boldi, M. Rosa, J. Ugander, & S. Vigna. Four degrees of separation. In *Proceedings of the ACM International Conference on Web Science*, pages 45–54, Evanston, IL, USA, 2012.
- A. L. Barabási. *Linked: The new science of networks*. Perseus, 2002.
- A. L. Barabási & R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- H. Burch & B. Cheswick. Network of autonomous systems of internet by internet mapping project.
- C. I. Del Genio, T. Gross, & K. E. Bassler. All scale-free networks are sparse. *Phys. Rev. Lett.*, 107(17):178701, 2011.
- P. S. Dodds, R. Muhamad, & D. J. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.
- M. Faloutsos, P. Faloutsos, & C. Faloutsos. On power-law relationships of the internet topology. *Comput. Commun. Rev.*, 29(4):251–262, 1999.
- L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.
- L. C. Freeman. Centrality in social networks: Conceptual clarification. *Soc. Networks*, 1(3):215–239, 1979.
- M. Girvan & M. E. J. Newman. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.
- M. S. Granovetter. The strength of weak ties. *Am. J. Sociol.*, 78(6):1360–1380, 1973.

- J. Kleinberg. Small-world phenomena and the dynamics of information. In *Proceedings of the International Conference in Advances in Neural Information Processing Systems*, 2001.
- J. Leskovec, J. Kleinberg, & C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):1–41, 2007.
- S. Milgram. The small world problem. *Psychol. Today*, 1(1):60–67, 1967.
- J. L. Moreno. *Who Shall Survive?* Beacon House, Beacon, 1934.
- M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.
- M. E. J. Newman & E. A. Leicht. Mixture models and exploratory analysis in networks. *P. Natl. Acad. Sci. USA*, 104(23):9564, 2007.
- M. E. J. Newman & J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, 2003.
- D. J. d. S. Price. Networks of scientific papers. *Science*, 149:510–515, 1965.
- M. V. Simkin & V. P. Roychowdhury. Read before you cite! *Compl. Syst.*, 14:269–274, 2003.
- S. Sinha. Few and far between. *Physics*, 4:81, 2011.
- L. Šubelj & M. Bajec. Community structure of complex software systems: Analysis and applications. *Physica A*, 390(16):2968–2975, 2011.
- L. Šubelj & M. Bajec. Ubiquitousness of link-density and link-pattern communities in real-world networks. *Eur. Phys. J. B*, 85(1):32, 2012.
- L. Šubelj & M. Bajec. Model of complex networks based on citation dynamics. In *Proceedings of the WWW Workshop on Large Scale Network Analysis*, pages 527–530, Rio de Janeiro, Brazil, 2013.
- L. Šubelj, N. Blagus, & M. Bajec. Group extraction for real-world networks: The case of communities, modules, and hubs and spokes. In *Proceedings of the International Conference on Network Science*, pages 152–153, Copenhagen, Denmark, 2013a.

- L. Šubelj, S. Žitnik, N. Blagus, & M. Bajec. Node mixing and group structure of complex software networks. *sub. to Adv. Complex Syst.*, page 23, 2013b.
- S. Valverde, R. F. Cancho, & R. V. Solé. Scale-free networks from optimal design. *Europhys. Lett.*, 60(4):512–517, 2002.
- A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E*, 67(5):056104, 2003.
- D. J. Watts & S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.
- W. W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33(4):452–473, 1977.
- Y. Zhao, E. Levina, & J. Zhu. Community extraction for social networks. *P. Natl. Acad. Sci. USA*, 108(18):7321–7326, 2011.