Odkrivanje skupin vozlišč v velikih realnih omrežjih na osnovi izmenjave oznak

Lovro Šubelj

Doktorska disertacija

Mentor: izr. prof. dr. Marko Bajec
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

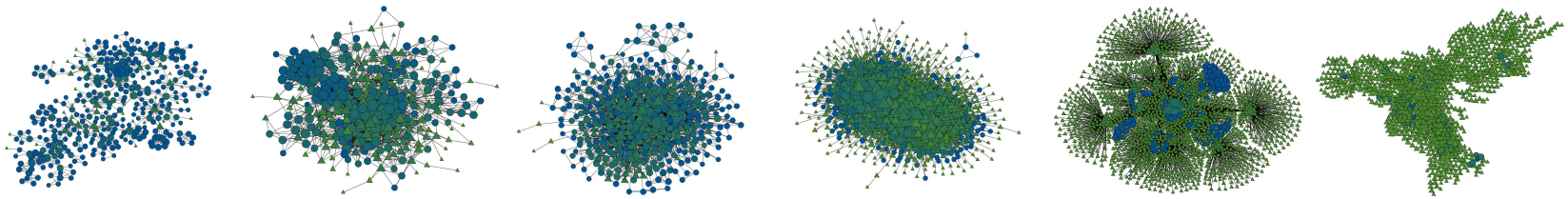
27. junij, 2013

Vsebina

Analiza omrežij

Različne vrste realnih omrežij (grafov).

socialna, informacijska, tehnološka, biološka itd.



Analiza omrežij: Newman (2008)

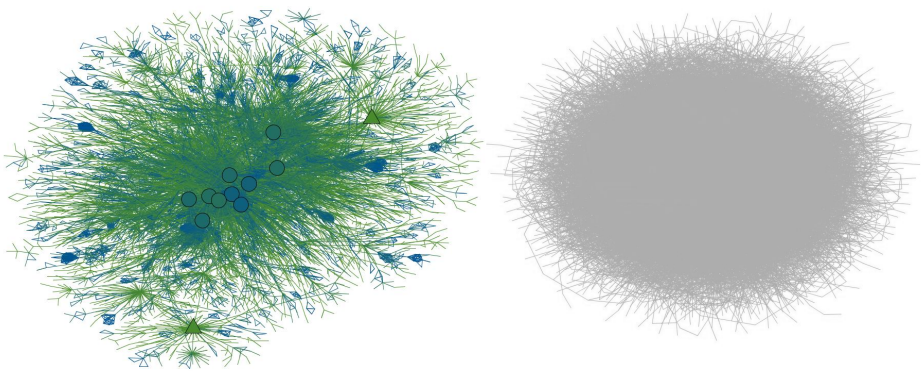
- preučevanje zgradbe omrežij
- razvoj potrebnih pristopov
- praktični primeri uporabe

Področje izjemno aktivno v številnih znanostih.

matematika, fizika, računalništvo, družboslovje, biologija itd.

Zgradba realnih omrežij

- globalne lastnosti omrežij (npr. mali svet)
- karakteristične skupine vozlišč (npr. skupnosti)
- pogosti vzorci vozlišč (npr. motifi, grafleti)
- lastnosti posameznih vozlišč (npr. zvezdišča)
- procesi nad omrežji (npr. širjenje)



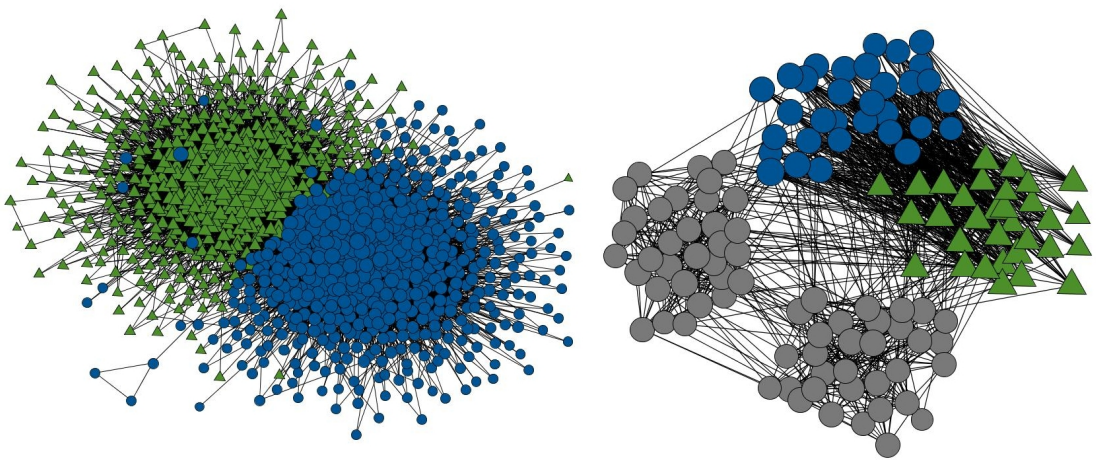
Omejimo se na neusmerjena omrežja.

Karakteristične skupine vozlišč

- skupnosti (*community*) Girvan and Newman (2002)

  (povezane) skupine tesno povezanih vozlišč
- moduli (*module*) Newman and Leicht (2007)

  (nepovezane) skupine podobno povezanih vozlišč



Omejimo se na neprekrivajoče skupine vozlišč.

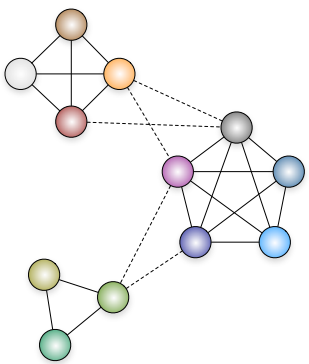
Vsebina

Osnovna izmenjava oznak

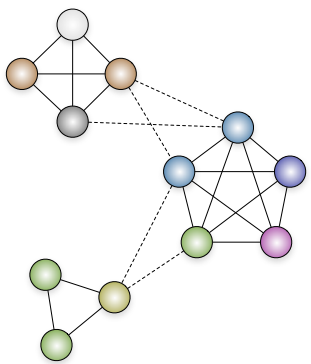
Osnovna izmenjava oznak (*label propagation*): Raghavan et al. (2007)

$$g_i = \underset{g}{\operatorname{argmax}} \sum_{v_j \in N_i} w_{ij} \delta(g_j, g)$$

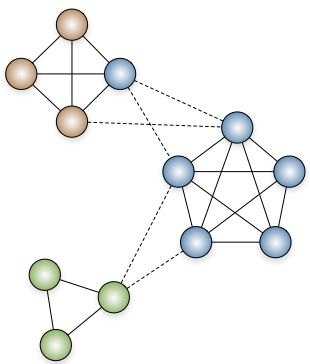
g_i posodabljamo zaporedno v naključnem vrstnem redu.



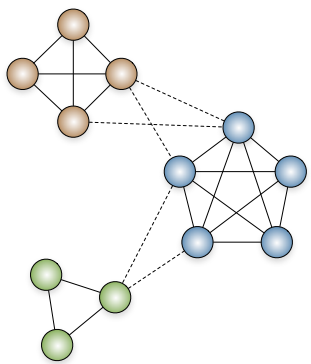
Začetno stanje



1. korak



2. korak



Končno stanje

g_i je oznaka skupine vozlišča v_i in w_{ij} utež na povezavi med v_i in v_j.

Analiza izmenjave oznak

Prednosti:

- brez predhodnega znanja (npr. število skupin)
- skoraj linearna časovna zahtevnost
- enostavna implementacija

Slabosti:

robustnost več razvrstitev že v manjših omrežjih Tibély and Kertész (2008)

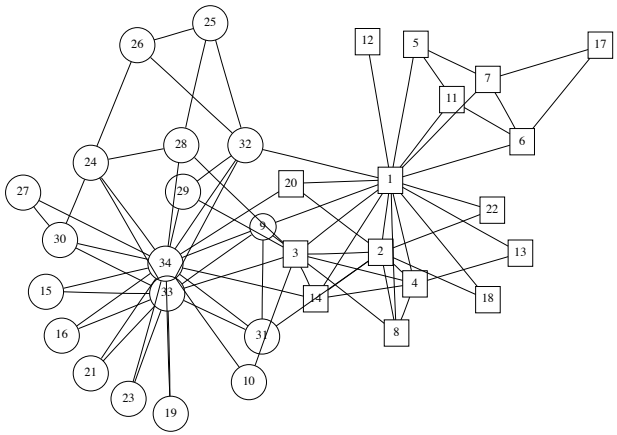
natančnost slabša natančnost v omrežjih z nejasno zgradbo Leung et al. (2009)

splošnost pristop omejen na odkrivanje skupnosti Šubelj and Bajec (2012c)

Vsebina

Robustnost izmenjave oznak

Osnovna izmenjava vrne $>$ 500 razvrstitev v skupine. Tibély and Kertész (2008)

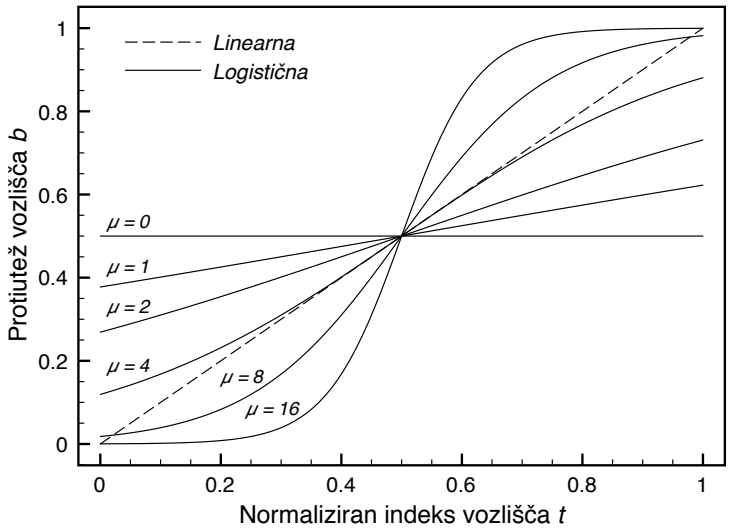


Vrstni red obravnave g_i se odraža kot preference vozlišč f_i. Šubelj and Bajec (2011a)

$$g_i = \operatorname*{argmax}_g \sum_{v_j \in N_i} f_j \cdot w_{ij} \delta(g_j, g)$$

f_i je "moč" širjenja oznake vozlišča v_i. Leung et al. (2009)

Uravnotežena izmenjava oznak



Uravnotežena izmenjava (*balanced propagation*): Šubelj and Bajec (2011a)

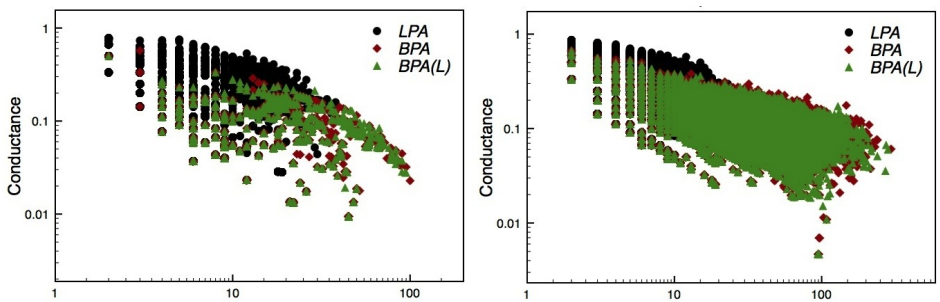
$$g_i = \operatorname*{argmax}_g \sum_{v_j \in N_i} b_j \cdot w_{ij} \delta(g_j, g)$$

b_i je protiutež in $t_i \in (0, 1]$ normaliziran indeks vozlišča v_i.

Eksperimentalni rezultati

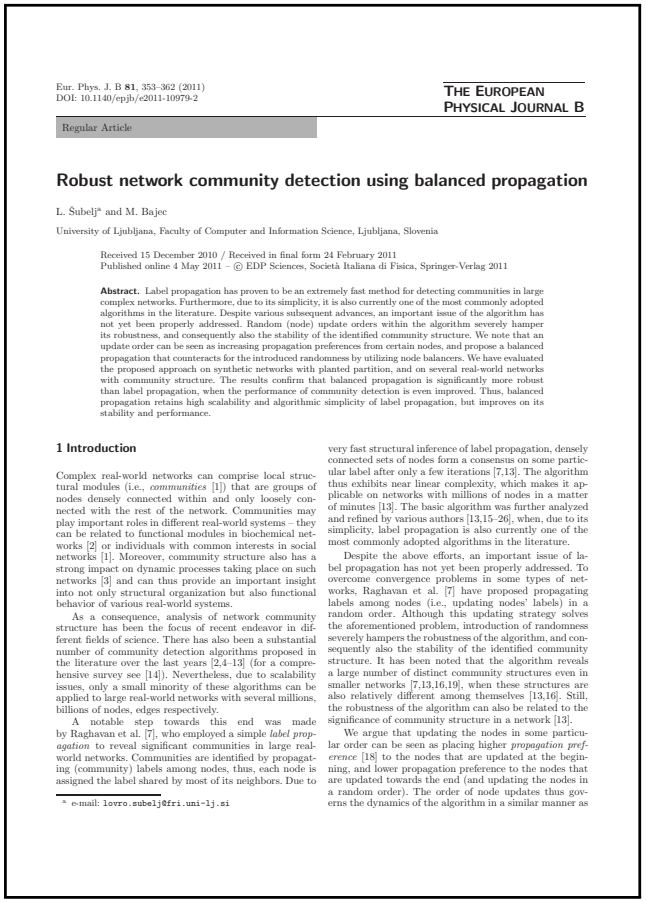
različnih razvrstitev v 1000 ponovitvah: Šubelj and Bajec (2011a)

	karate	*dolphins*	*books*	*football*	*jazz*	*elegans*
Osnovna izmenjava	184	525	269	414	63	707
Uravnotežena izmenjava	**19**	**36**	**29**	**154**	**20**	**75**



Robustnost se izboljša na račun časovne zahtevnosti. Šubelj and Bajec (2011c)

Znanstveni doprinos



Eur. Phys. J. B **81**, 353–362 (2011)
DOI: 10.1140/epjb/e2011-10979-2

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Robust network community detection using balanced propagation

L. Šubelj* and M. Bajec

University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

Received 15 December 2010 / Received in final form 24 February 2011
Published online 4 May 2011 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2011

Abstract. Label propagation has proven to be an extremely fast method for detecting communities in large complex networks. Furthermore, due to its simplicity, it is also currently one of the most commonly adopted algorithms in the literature. Despite various subsequent advances, an important issue of the algorithm has not yet been properly addressed. Random (node) update orders within the algorithm severely hamper its robustness, and consequently also the stability of the identified community structure. We note that an update order can be seen as increasing propagation preferences from certain nodes, and propose a balanced propagation that counteracts for the introduced randomness by utilizing node balancers. We have evaluated the proposed approach on synthetic networks with planted partition, and on several real-world networks with community structure. The results confirm that balanced propagation is significantly more robust than label propagation, when the performance of community detection is even improved. Thus, balanced propagation retains high scalability and algorithmic simplicity of label propagation, but improves on its stability and performance.

1 Introduction

Complex real-world networks can comprise local structural modules (i.e., *communities* [1]) that are groups of nodes densely connected within and only loosely connected with the rest of the network. Communities may play important roles in different real-world systems – they can be related to functional modules in biochemical networks [2] or individuals with common interests in social networks [1]. Moreover, community structure also has a strong impact on dynamic processes taking place on such networks [3] and can thus provide an important insight into not only structural organization but also functional behavior of various real-world systems.

As a consequence, analysis of network community structure has been the focus of recent endeavor in different fields of science. There has also been a substantial number of community detection algorithms proposed in the literature over the last years [2,4–13] (for a comprehensive survey see [14]). Nevertheless, due to scalability issues, only a small minority of these algorithms can be applied to large real-world networks with several millions, billions of nodes, edges respectively.

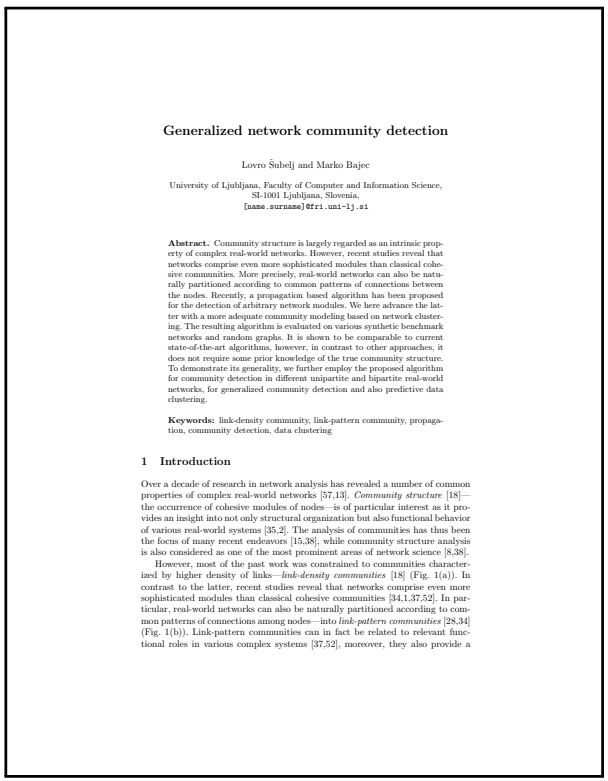
A notable step towards this end was made by Raghavan et al. [7], who employed a simple *label propagation* to reveal significant communities in large real-world networks. Communities are identified by propagating (community) labels among nodes, thus, each node is assigned the label shared by most of its neighbors. Due to very fast structural inference of label propagation, densely connected sets of nodes form a consensus on some particular label after only a few iterations [7,13]. The algorithm thus exhibits near linear complexity, which makes it applicable on networks with millions of nodes in a matter of minutes [13]. The basic algorithm was further analyzed and refined by various authors [13,15–26], when, due to its simplicity, label propagation is also currently one of the most commonly adopted algorithms in the literature.

Despite the above efforts, an important issue of label propagation has not yet been properly addressed. To overcome convergence problems in some types of networks, Raghavan et al. [7] have proposed propagating labels among nodes (i.e., updating nodes' labels) in a random order. Although this updating strategy solves the aforementioned problem, introduction of randomness severely hampers the robustness of the algorithm, and consequently also the stability of the identified community structure. It has been noted that the algorithm reveals a large number of distinct community structures even in smaller networks [7,13,16,19], when these structures are also relatively different among themselves [13,16]. Still, the robustness of the algorithm can also be related to the significance of community structure in a network [13].

We argue that updating the nodes in some particular order can be seen as placing higher *propagation preference* [18] to the nodes that are updated at the beginning, and lower propagation preference to the nodes that are updated towards the end (and updating the nodes in a random order). The order of node updates thus governs the dynamics of the algorithm in a similar manner as

* e-mail: lovro.subelj@fri.uni-lj.si

Eur. Phys. J. B, 2011 (12 cit.)



Generalized network community detection

Lovro Šubelj and Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science,
SI-1001 Ljubljana, Slovenia,
[name.surname]@fri.uni-lj.si

Abstract. Community structure is largely regarded as an intrinsic property of complex real-world networks. However, recent studies reveal that networks comprise even more sophisticated modules than classical cohesive communities. More precisely, real-world networks can also be naturally partitioned according to common patterns of connections between the nodes. Recently, a propagation based algorithm has been proposed for the detection of arbitrary network modules. We here advance the latter with a more adequate community modeling based on network clustering. The resulting algorithm is evaluated on various synthetic benchmark networks and random graphs. It is shown to be comparable to current state-of-the-art algorithms, however, in contrast to other approaches, it does not require some prior knowledge of the true community structure. To demonstrate its generality, we further employ the proposed algorithm for community detection in different unipartite and bipartite real-world networks, for generalized community detection and also predictive data clustering.

Keywords: link-density community, link-pattern community, propagation, community detection, data clustering

1 Introduction

Over a decade of research in network analysis has revealed a number of common properties of complex real-world networks [57,13]. *Community structure* [18]—the occurrence of cohesive modules of nodes—is of particular interest as it provides an insight into not only structural organization but also functional behavior of various real-world systems [35,2]. The analysis of communities has thus been the focus of many recent endeavors [15,38], while community structure analysis is also considered as one of the most prominent areas of network science [8,38].

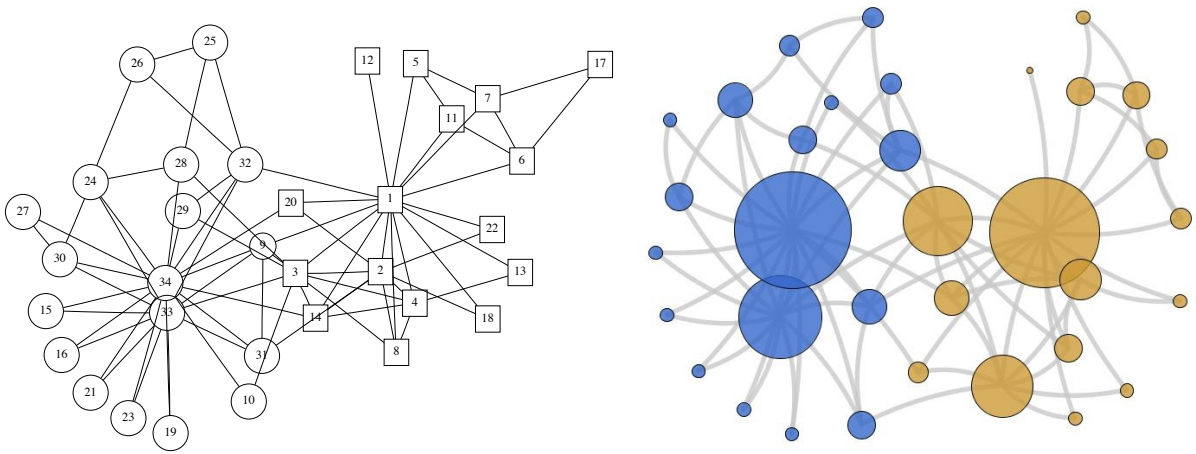
However, most of the past work was constrained to communities characterized by higher density of links—*link-density communities* [18] (Fig. 1(a)). In contrast to the latter, recent studies reveal that networks comprise even more sophisticated modules than classical cohesive communities [34,1,37,52]. In particular, real-world networks can also be naturally partitioned according to common patterns of connections among nodes—into *link-pattern communities* [28,34] (Fig. 1(b)). Link-pattern communities can in fact be related to relevant functional roles in various complex systems [37,52], moreover, they also provide a

NEMO (*ECML PKDD* '11)

Vsebina

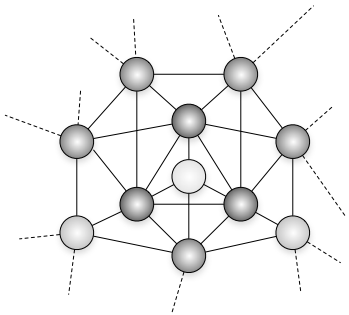
Natančnost izmenjave oznak

Natančnost izboljšamo z uporabo preferenc vozlišč f_i. Leung et al. (2009)

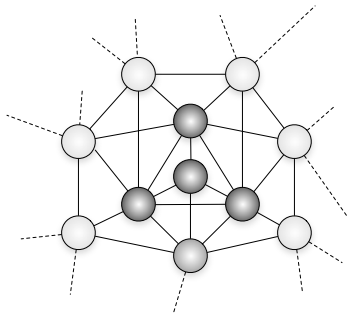


Osnovne lastnosti vozlišč niso primerne za f_i. Šubelj and Bajec (2011b)

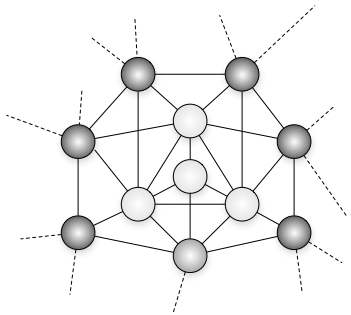
Zadržana in napadalna izmenjava oznak



Stopnje vozlišč



Zadržana



Napadalna

Zadržana izmenjava (*defensive propagation*): Šubelj and Bajec (2011b)

$$g_i = \operatorname*{argmax}_g \sum_{v_j \in N_i} p_j \cdot w_{ij} \delta(g_j, g)$$

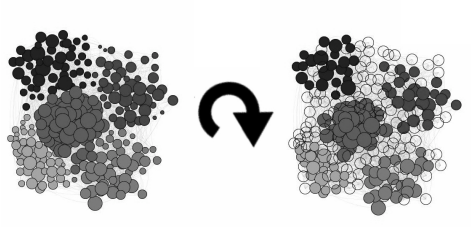
Napadalna izmenjava (*offensive propagation*):

$$g_i = \operatorname*{argmax}_g \sum_{v_j \in N_i} (1 - p_j) \cdot w_{ij} \delta(g_j, g)$$

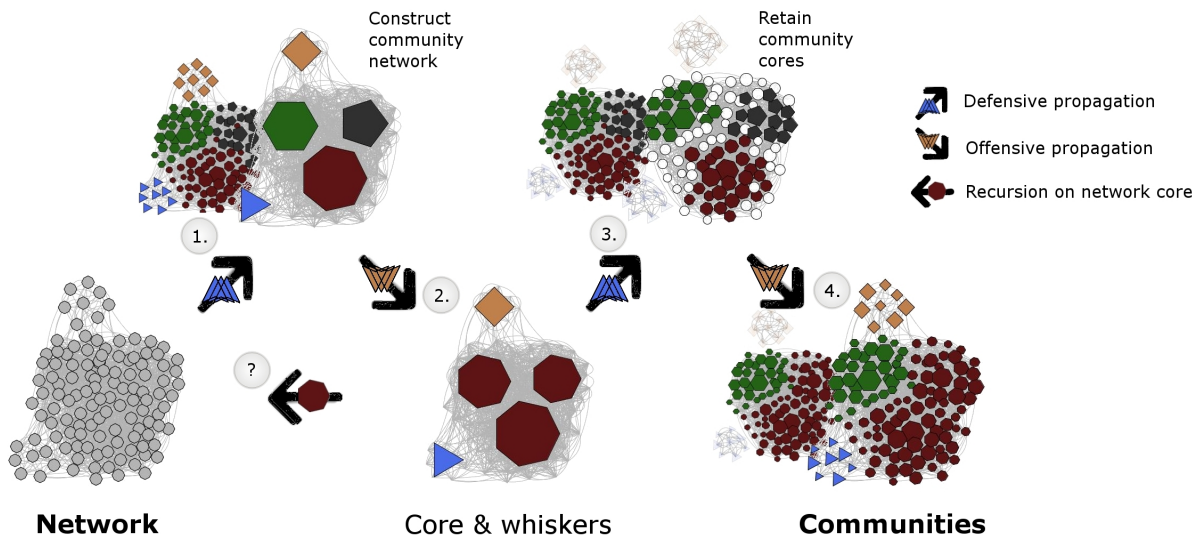
p_i je verjetnost pri naključnem sprehodu znotraj skupine g_i.

Napredna izmenjava oznak

Zadržana (napadalna) izmenjava doseže visok priklic (natančnost).



Napredna izmenjava (*diffusion propagation*): Šubelj and Bajec (2011b)



Eksperimentalni rezultati

Natančnost je primerljiva z najboljšimi pristopi. Šubelj and Bajec (2010)

Network	Description	*Nodes*	*Edges*	GMO	LPA	LPAD	LPAQ	LPAM	BDPA	DPA	No. CE[c]	T[c]
karate	Zachary's karate club [33]	34	78	0.381	0.416	0.402	0.399	**0.420**	**0.419**	**0.420**	0.02	
dolphins	Lusseau's bottlenose dolphins [38]	62	159		**0.529**	0.526	0.516	**0.529**	**0.528**	**0.529**	0.59	
books	Co-purchased political books [39]	105	441		**0.526**	0.519	0.522	**0.527**	**0.527**	**0.527**	0.46	
football	American football league [6]	115	616	0.556	**0.606**	**0.606**	0.604	**0.605**	**0.606**	**0.606**	0.37	
elegans	Metabolic network *C. elegans* [37]	453	2025	0.412	0.421	0.413	0.409	*0.452*	0.424	**0.427**[b]	0.17	
jazz	Jazz musicians [40]	198	2742	0.439	0.443	0.443	**0.445**	**0.445**	**0.444**	**0.444**	0.00	
netsci	Network scientists [9]	1589	2742		0.902	0.947			0.907	**0.960**	1.00	
yeast	Yeast protein interactions [41]	2114	4480		0.694	0.799			0.725	**0.824**	1.04	
emails	Emails within a university [42]	1133	5451	0.503	0.557	0.560	0.537	*0.582*	0.555	**0.562**	0.01	
power	Western US power grid [32]	4941	6594		0.612	0.804			0.668	**0.908**	1.14	
blogs	Weblogs on politics [43]	1490	16718		**0.426**	**0.426**			**0.426**	**0.426**	1.00	
pgp	*PGP* web of trust [44]	10680	24340	0.849	0.754	0.844	0.726	*0.884*	0.782	**0.869**	1.08	
asi	Autonomous syst. of Internet [25]	22963	48436		0.511	0.591			0.528	**0.600**[b]	1.02	0 s
codmat[3]	*Cond. Matt.* archive 2003[a] [45]	27519	116181	0.661	0.616	0.683	0.582	*0.755*	0.634	**0.735**	1.00	1.5 s
codmat[5]	*Cond. Matt.* archive 2005[a] [45]	36458	171736		0.586	0.643			0.608	**0.683**	1.00	
kdd[3]	*KDD-Cup* 2003 dataset [46]	27770	352285		0.624	**0.630**			0.619	0.617	1.00	3 s
nec	*nec* web overlay map [47]	75885	357317		0.693	0.738			0.703	**0.767**	1.03	
epinions	*Epinions* web of trust [48]	75879	508837		0.382	0.362			0.399	**0.402**	1.00	4.5 s
amazon[3]	*Amazon* co-purchasing 2003 [49]	262111	1.2M		0.682	0.749			0.701	**0.857**	1.01	20 s
ndedu	Webpages in nd.edu domain [50]	325729	1.5M		0.840	0.890			0.863	**0.903**	1.14	
google	Web graph of *Google* [3]	875713	4.3M		0.805	0.923			0.822	**0.968**	1.01	2.5 m
nber	*NBER* patents citations [51]	3.8M	16.5M		0.573	0.624			0.583	**0.759**	1.20	
live	*Live Journal* friendships [3]	4.8M	69.0M		0.538	0.539			0.557	**0.693**	1.00	44 m

Časovna zahtevnost blizu linearne $\mathcal{O}(m^{1,19})$. Šubelj and Bajec (2011b)

Znanstveni doprinos

Prispevek prejel fakultetno nagrado za raziskovalno delo.

PHYSICAL REVIEW E **83**, 036103 (2011)

Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction

Lovro Šubelj* and Marko Bajec†
University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
(Received 1 June 2010; revised manuscript received 6 November 2010; published 8 March 2011)

Label propagation has proven to be a fast method for detecting communities in large complex networks. Recent developments have also improved the accuracy of the approach; however, a general algorithm is still an open issue. We present an advanced label propagation algorithm that combines two unique strategies of community formation, namely, defensive preservation and offensive expansion of communities. The two strategies are combined in a hierarchical manner to recursively extract the core of the network and to identify whisker communities. The algorithm was evaluated on two classes of benchmark networks with planted partition and on 23 real-world networks ranging from networks with tens of nodes to networks with several tens of millions of edges. It is shown to be comparable to the current state-of-the-art community detection algorithms and superior to all previous label propagation algorithms, with comparable time complexity. In particular, analysis on real-world networks has proven that the algorithm has almost linear complexity, $O(m^{1.19})$, and scales even better than the basic label propagation algorithm (m is the number of edges in the network).

DOI: 10.1103/PhysRevE.83.036103 PACS number(s): 89.75.Fb, 89.75.Hc, 87.23.Ge, 89.20.Hh

I. INTRODUCTION

Large real-world networks can comprise local structural modules (*communities*) that are groups of nodes, densely connected within and only loosely connected with the rest of the network. Communities are believed to play important roles in different real-world systems (e.g., they may correspond to functional modules in metabolic networks [1]); moreover, they also provide a valuable insight into the structure and function of large complex networks [1–3]. Nevertheless, real-world networks can reveal even more complex modules than communities [4,5].

Over the last decade the research community has shown a considerable interest in detecting communities in real-world networks. Since the seminal paper of Girvan and Newman [6] a vast number of approaches have been presented in the literature—in particular, approaches optimizing modularity Q (the significance of communities due to a selected null model [7]) [8–12], graph partitioning [13,14] and spectral algorithms [9,15], statistical methods [4], algorithms based on dynamic processes [16–20], overlapping, hierarchical, and multiresolution methods [1,6,20], and others [21] (for an excellent survey see [22]).

The size of large real-world networks has forced the research community to develop scalable approaches that can be applied to networks with several millions of nodes and billions of edges. A promising effort was made by Raghavan *et al.* [18], who employed a simple label propagation to find significant communities in large real-world networks. Tibély and Kertész [23] have shown that label propagation is in fact equivalent to a large zero-temperature kinetic Potts model, while Barber and Clark [11] have further refined the approach into a modularity optimization algorithm. Just recently, Liu and Murata [12] have combined the modularity optimization version of the algorithm with a multistep greedy agglomeration [24] and derived an extremely accurate community detection algorithm.

Leung *et al.* [19] have investigated label propagation on large web networks, mainly focusing on scalability issues, and have shown that the performance can be significantly improved with label hop attenuation and by applying node preference (i.e., node propagation strength). We carry forward their work in developing two unique strategies of community formation, namely, defensive preservation of communities, where preference is given to the nodes in the core of each community, and offensive expansion of communities, where preference is given to the border nodes of each community. Cores and borders are estimated using random walks, formulating the diffusion over the network.

Furthermore, we propose an advanced label propagation algorithm—the diffusion and propagation algorithm—that combines the two strategies in a hierarchical manner: The algorithm first extracts the *core* of the network and identifies *whisker* communities [26] (Appendix A), and then recurses on the network core (Fig. 1). The performance of the algorithm has been analyzed on two classes of benchmark networks with planted partition and on 23 real-world networks ranging from networks with tens of nodes to networks with several tens of millions of edges. The algorithm is shown to be comparable to the current state-of-the-art community detection algorithms and superior to all previous label propagation algorithms, with comparable time complexity. In particular, the algorithm exhibits almost linear time complexity (in the number of edges of the network).

The rest of the paper is structured as follows. Section II gives a formal introduction to label propagation and reviews subsequent advances, relevant for this research. Section III presents the diffusion and propagation algorithm and discusses the main rationale behind it. Empirical evaluation with discussion is done in Secs. IV, and V is our conclusion.

*lovro.subelj@fri.uni-lj.si
†marko.bajec@fri.uni-lj.si

Phys. Rev. E, 2011 (10 cit.)

Unfolding network communities by combining defensive and offensive label propagation

Lovro Šubelj* and Marko Bajec

Laboratory for Data Technologies, University of Ljubljana, Ljubljana, Slovenia

Abstract. Label propagation has proven to be a fast method for detecting communities in complex networks. Recent work has also improved the accuracy and stability of the basic algorithm, however, a general approach is still an open issue. We propose different label propagation algorithms that convey two unique strategies of community formation, namely, defensive preservation and offensive expansion of communities. Furthermore, the strategies are combined in an advanced label propagation algorithm that retains the advantages of both approaches; and are enhanced with hierarchical community extraction, prominent for the use on larger networks. The proposed algorithms were empirically evaluated on different benchmarks networks with planted partition and on over 30 real-world networks of various types and sizes. The results confirm the adequacy of the propositions and give promising grounds for future analysis of (large) complex networks. Nevertheless, the main contribution of this work is in showing that different types of networks (with different topological properties) favor different strategies of community formation.

Keywords: Network communities, label propagation, defensive preservation, offensive expansion.

1 Introduction

Complex networks commonly comprise of local structural *modules* or *communities* that are groups of nodes strongly connected within and only weakly connected with the rest of the network. These modules play crucial roles in many real-world systems [15, 37], moreover, they provide an important insight into structure and function of (large) complex networks [37, 45, 27].

Over the last decade the research community has shown a considerable interest in detecting communities in real-world networks. Thus, a number of approaches has been presented in the literature. In particular, approaches optimizing *modularity*[1] Q [7, 6, 5], graph partitioning [14, 39, 38] and spectral [9, 33] algorithms, statistical methods [36], algorithms based on dynamic processes [40, 43, 38, 42], overlapping, hierarchical and multiresolution methods [37, 16, 42], and other [29, 30] (for a thorough review see [11]).

* Corresponding author: lovro.subelj@fri.uni-lj.si.
[1] Significance of communities due to a selected *null model* [35].

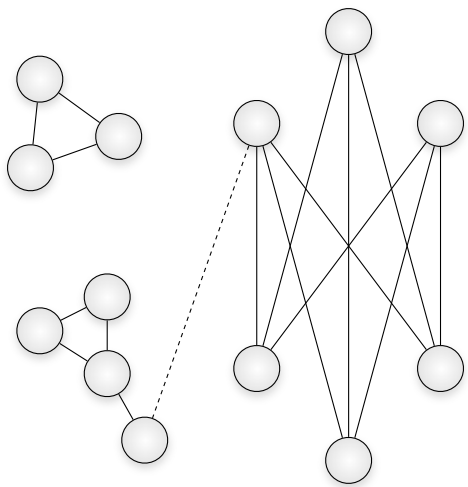
ACNE (*ECML PKDD* '10)

Vsebina

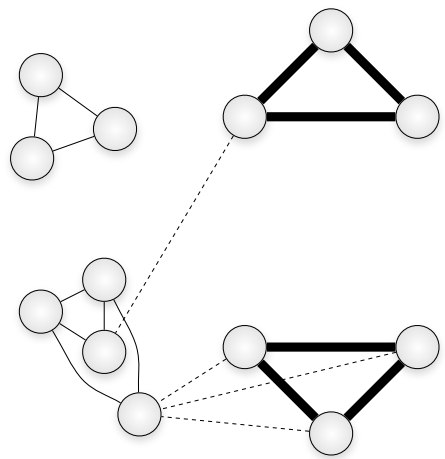
Splošnost izmenjave oznak

Pristop omejen na tesno povezana vozlišča (tj. skupnosti). Šubelj and Bajec (2012c)

Analogija med skupinami:



2 skupnosti, 2 modula



4 skupnosti

Oznake si izmenjujejo vozlišča na razdalji (največ) dva.

Posplošena izmenjava oznak

Posplošena izmenjava (*general propagation*): Šubelj and Bajec (2012c)

$$g_i = \underset{g}{\operatorname{argmax}} \left(\nu_g \sum_{v_j \in N_i} w_{ij} \delta(g_j, g) + (1 - \nu_g) \sum_{v_j \in N_i} w_{ij} / w_j \sum_{v_k \in N_j} w_{jk} \delta(g_k, g) \right)$$

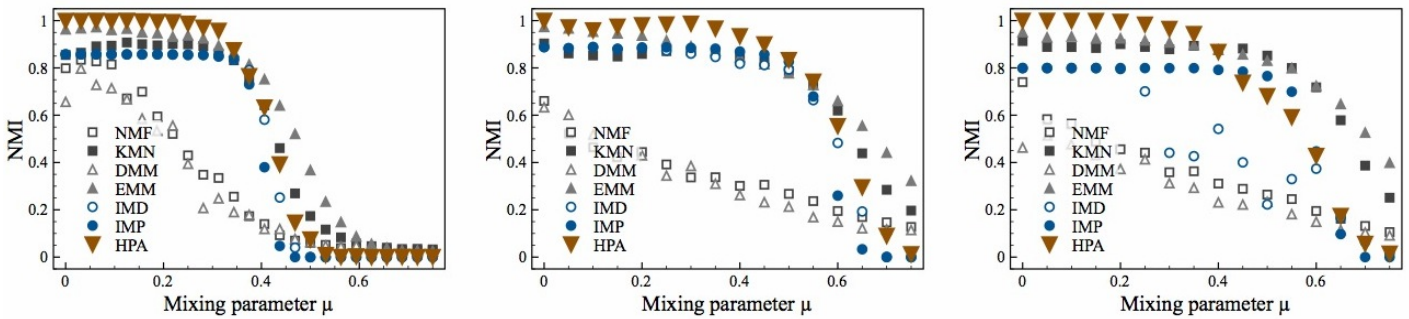
ν_g je blizu ena (nič) za skupnosti (module):

- lastnosti skupin Šubelj and Bajec (2012c)
- nakopičenost vozlišč Šubelj and Bajec (2011c)
- popravljena nakopičenost Šubelj and Bajec (2012a)

$\nu_g \in [0, 1]$ je parameter skupine vozlišč g.

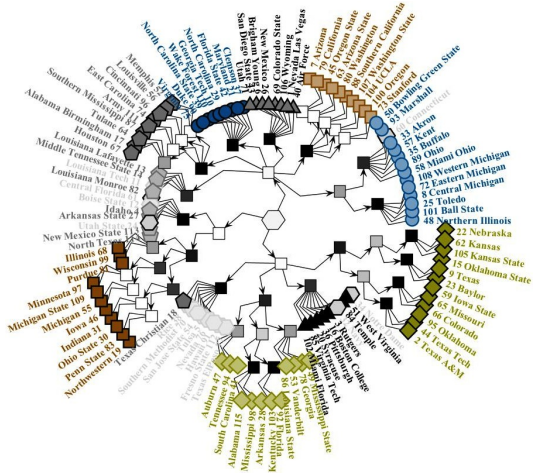
Eksperimentalni rezultati

Natančnost (vsaj) primerljiva z najboljšimi pristopi. Šubelj and Bajec (2012a)



AUC pri napovedovanju povezav:

	$\lvert N_i\rvert \cdot \lvert N_j\rvert$	$\lvert N_i \cap N_j\rvert$	Infomap	Izmenjava
football	0.222	*0.817*	**0.804**	**0.799**
politics	0.646	*0.862*	**0.763**	**0.762**
software	0.779	*0.826*	0.724	**0.766**
elegans	0.812	*0.920*	0.631	**0.641**
women	0.564	0.290	0.578	**0.699**
corporate	0.456	0.481	0.649	**0.748**



Znanstveni doprinos

Prispevek izpostavljen s strani uredništva ter na straneh *APS*, *ACM*, *IEEE* ipd.


Eur. Phys. J. B (2012) 85: 32
DOI: 10.1140/epjb/e2011-20448-7

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Ubiquitousness of link-density and link-pattern communities in real-world networks

L. Šubelj[a] and M. Bajec

University of Ljubljana, Faculty of Computer and Information Science, 1000 Ljubljana, Slovenia

Received 9 June 2011 / Received in final form 24 October 2011
Published online 18 January 2012 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2012

Abstract. Community structure appears to be an intrinsic property of many complex real-world networks. However, recent work shows that real-world networks reveal even more sophisticated modules than classical cohesive (link-density) communities. In particular, networks can also be naturally partitioned according to similar patterns of connectedness among the nodes, revealing link-pattern communities. We here propose a propagation based algorithm that can extract both link-density and link-pattern communities, without any prior knowledge of the true structure. The algorithm was first validated on different classes of synthetic benchmark networks with community structure, and also on random networks. We have further applied the algorithm to different social, information, technological and biological networks, where it indeed reveals meaningful (composites of) link-density and link-pattern communities. The results thus seem to imply that, similarly as link-density counterparts, link-pattern communities appear ubiquitous in nature and design.

1 Introduction

Complex real-world networks commonly reveal local cohesive modules of nodes denoted (*link-density*) *communities* [1]. These are most frequently observed as densely connected clusters of nodes that are only loosely connected between. Communities possibly play crucial roles in different real-world systems [2,3]; furthermore, community structure also has a strong impact on dynamic processes taking place on networks [4,5]. Thus, communities provide an insight into not only structural organization but also functional behavior of various real-world systems [3,6–8].

Consequently, analysis of community structure is currently considered one of the most prominent areas of network science [9–11], while it has also been the focus of recent efforts in a wide variety of other fields. Besides providing many significant theoretical grounds [8], a substantial number of different community detection algorithms has also been proposed in the literature [12–20] (for reviews see [10,11,21]). However, most of the past research was focused primarily on classical communities characterized by higher density of edges [22]. In contrast to the latter, some recent work demonstrates that real-world networks reveal even more sophisticated communities [23–26] that are indistinguishable under classical frameworks.

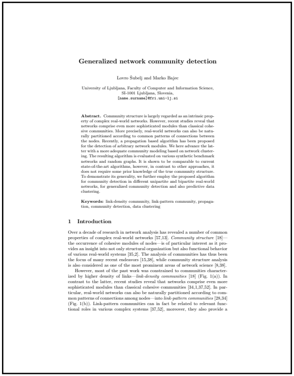
Networks can also be naturally partitioned according to similar patterns of connectedness among the nodes, revealing *link-pattern communities* [23,24]. (The term was formulated by Long et al. [24]). Loosely speaking, link-pattern communities correspond to clusters of nodes that are similarly connected with the rest of the network (i.e., share common neighborhoods). Note that link-density communities are in fact a special case of link-pattern communities (with some fundamental differences discussed later on). Thus, some of the research on the former also apply for the latter [13,14,28–30]. However, contrary to the flourish of the literature on classical communities in the last decade, a relatively small number of authors have considered more general link-pattern counterparts [23–26,31–37] (in the same sense as in this paper[1]). Although this could be attributed to a number of factors like increased complexity or lack of adequate generative models and algorithms, more importantly, existence of meaningful link-pattern communities has not been properly verified under the same framework in various different types of real-world networks that are commonly analyzed in the literature (still, some networks have been considered in the past). In this paper we try to address this issue. (Note that similar stance was also made by Newman and Leicht [23]).

We extend balanced propagation [38] with defensive preservation of communities [20] into a general approach that can extract arbitrary network modules ranging from link-density to link-pattern communities. To the best of our knowledge, this is the only such algorithm that does not require some prior knowledge of the true structure (e.g., the number of communities), or does not optimize

[a] e-mail: lovro.subelj@fri.uni-lj.si

[1] Link-pattern communities are known as blockmodels [27] in social networks literature. These were rigorously analyzed in the past, however, the main focus and employed formulation differs from ours.

Eur. Phys. J. B, 2012 (3 cit.)


Generalized network community detection

Lovro Šubelj and Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science,
SI-1001 Ljubljana, Slovenia,
[name.surname]@fri.uni-lj.si

Abstract. Community structure is largely regarded as an intrinsic property of complex real-world networks. However, recent studies reveal that networks comprise even more sophisticated modules than classical cohesive communities. More precisely, real-world networks can also be naturally partitioned according to common patterns of connections between the nodes. Recently, a propagation based algorithm has been proposed for the detection of arbitrary network modules. We here advance the latter with a more adequate community modeling based on network clustering. The resulting algorithm is evaluated on various synthetic benchmark networks and random graphs. It is shown to be comparable to current state-of-the-art algorithms, however, in contrast to other approaches, it does not require some prior knowledge of the true community structure. To demonstrate its generality, we further employ the proposed algorithm for community detection in different unipartite and bipartite real-world networks, for generalized community detection and also predictive data clustering.

Keywords: link-density community, link-pattern community, propagation, community detection, data clustering

1 Introduction

Over a decade of research in network analysis has revealed a number of common properties of complex real-world networks [57,13]. *Community structure* [18]—the occurrence of cohesive modules of nodes—is of particular interest as it provides an insight into not only structural organization but also functional behavior of various real-world systems [35,2]. The analysis of communities has thus been the focus of many recent endeavors [15,38], while community structure analysis is also considered as one of the most prominent areas of network science [8,38].

However, most of the past work was constrained to communities characterized by higher density of links—*link-density communities* [18] (Fig. 1(a)). In contrast to the latter, recent studies reveal that networks comprise even more sophisticated modules than classical cohesive communities [34,1,37,52]. In particular, real-world networks can also be naturally partitioned according to common patterns of connections among nodes—into *link-pattern communities* [28,34] (Fig. 1(b)). Link-pattern communities can in fact be related to relevant functional roles in various complex systems [37,52], moreover, they also provide a

NEMO (*ECML PKDD* '11)


Groups of nodes in complex real-world networks: An algorithm and comparison of the state-of-the-art

Lovro Šubelj
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Marko Bajec
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

ABSTRACT

Categories and Subject Descriptors

General Terms

Keywords

1. INTRODUCTION

2. BACKGROUND

2.1 Groups of nodes in networks

arXiv.org (2x)

Vsebina

Programska omrežja

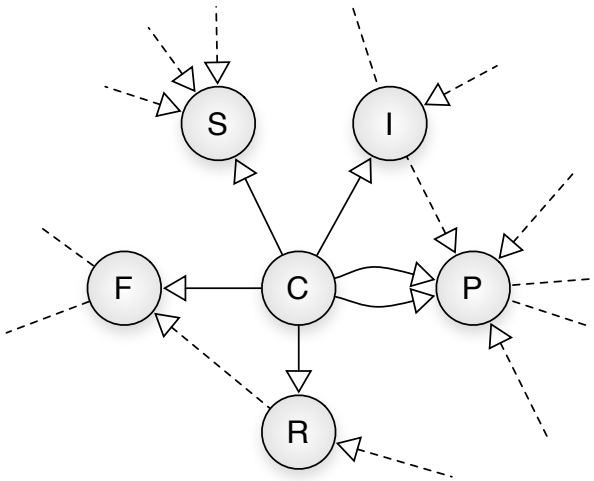
Omrežja odvisnosti med razredi (*class dependency*): Šubelj and Bajec (2011d)

```
class C extends S implements I {

  F field;

  public C(P parameter) {
    ...
  }

  public R function(P parameter) {
    ...
    return R;
  }
}
```

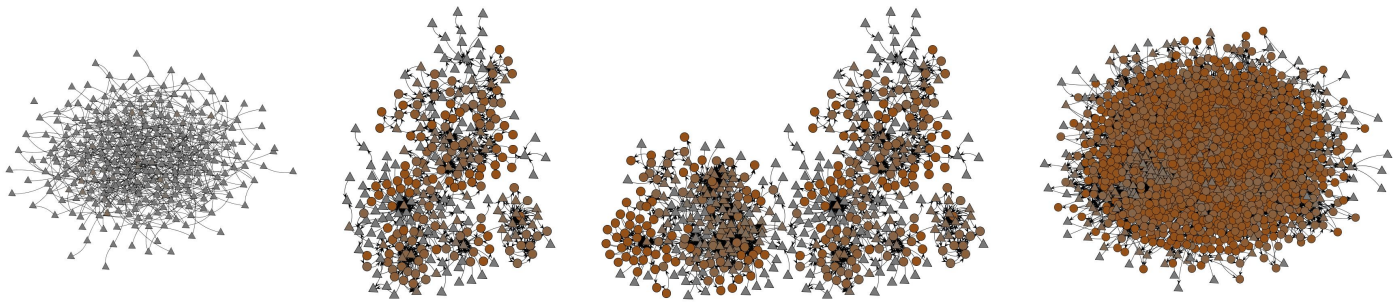


Podobne lastnosti kot druga realna omrežja. Valverde et al. (2002)

Skupine vozlišč še niso bile podrobno raziskane.

Analiza skupnosti vozlišč

Programska omrežja vsebujejo jasne skupnosti. Šubelj and Bajec (2011d)

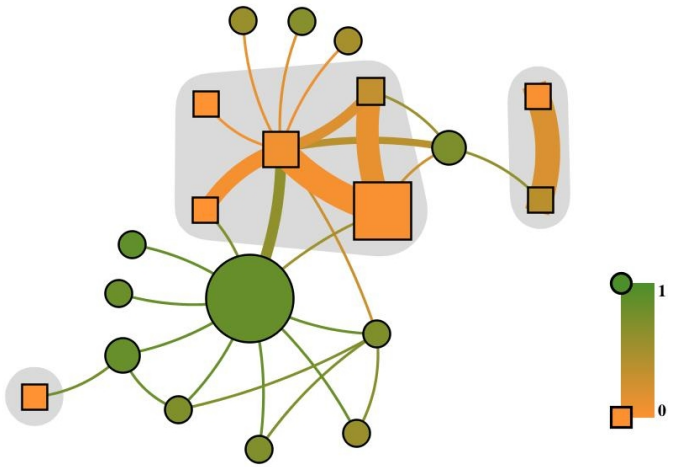


Skupnosti le delno sovpadajo s programskim paketi.

	Deljenje	Optimizacija	Izmenjava	Paketi*
flamingo	0.6466	**0.6823**	0.6485	*0.2511*
colt	-	**0.6025**	0.5599	*-0.0332*
jung	0.7210	**0.7324**	0.6874	*0.3212*
org	-	**0.5599**	0.5254	*0.1830*
javax	-	**0.7667**	0.7422	*0.2907*
java	-	**0.4664**	0.4132	*0.2206*

Analiza skupin vozlišč

Splošne skupine sovpadajo s programskim paketi. Šubelj and Bajec (2012c)



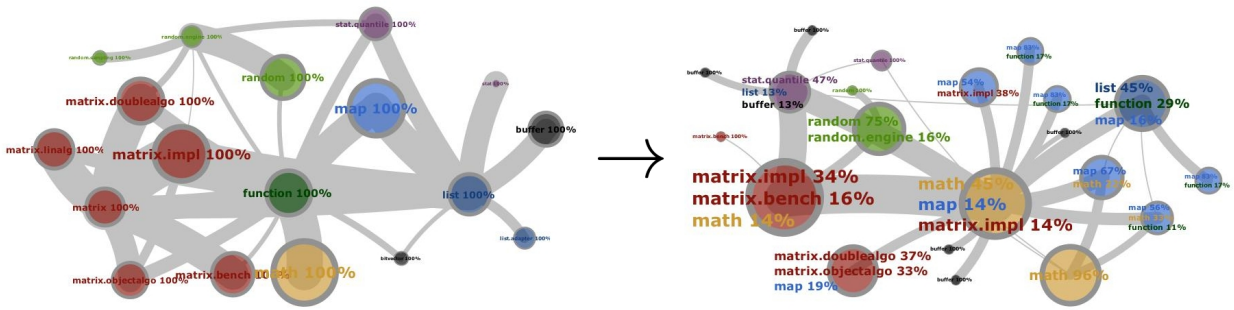
levo `jung.visualization.control.*Plugin` (100%)

sredina `jung.algorithms.layout.*Layout*` (54%), `....layout3d.*Layout` (23%);
`jung.graph.*(Graph|Multigraph|Tree)` (86%);
`jung.algorithms.filters.*Filter` (100%) itd.

desno `jung.io.graphml.parser.*Parser` (77%) in
`jung.io.graphml.*Metadata` (62%)

Uporaba skupin vozlišč

Reorganizacija paketov knjižnice (npr. modularno): Šubelj and Bajec (2011d)

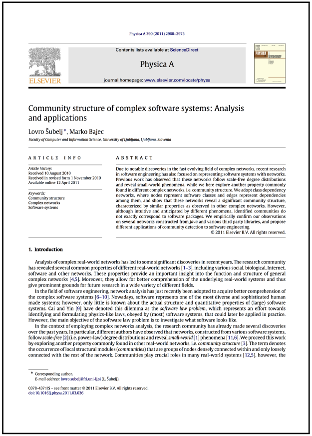


$Q = -0.0262$ $Q = 0.5020$

Točnost napovedovanja paketov razredov: Šubelj and Bajec (2012b)

	I	I_∞	P	P_4	P_3	P_2	P_1
flamingo	2.65	4	*0.566*	←	0.572	**0.793**	1.000
colt	3.35	4	*0.654*	←	**0.756**	0.942	1.000
jung	2.97	4	*0.617*	←	0.663	**0.857**	1.000
org	3.50	7	*0.616*	0.616	**0.714**	0.989	1.000
weka	3.02	6	*0.684*	0.692	**0.736**	0.871	1.000
javax	3.11	5	*0.626*	0.631	**0.816**	0.982	1.000

Znanstveni doprinos


Physica A 390 (2011) 2968–2975

Contents lists available at ScienceDirect

Physica A

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

Community structure of complex software systems: Analysis and applications

Lovro Šubelj *, Marko Bajec

Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 10 August 2010
Received in revised form 1 November 2010
Available online 12 April 2011

Keywords:
Community structure
Complex networks
Software systems

ABSTRACT

Due to notable discoveries in the fast evolving field of complex networks, recent research in software engineering has also focused on representing software systems with networks. Previous work has observed that these networks follow scale-free degree distributions and reveal small-world phenomena, while we here explore another property commonly found in different complex networks, i.e. community structure. We adopt class dependency networks, where nodes represent software classes and edges represent dependencies among them, and show that these networks reveal a significant community structure, characterized by similar properties as observed in other complex networks. However, although intuitive and anticipated by different phenomena, identified communities do not exactly correspond to software packages. We empirically confirm our observations on several networks constructed from Java and various third party libraries, and propose different applications of community detection to software engineering.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Analysis of complex real-world networks has led to some significant discoveries in recent years. The research community has revealed several common properties of different real-world networks [1–3], including various social, biological, Internet, software and other networks. These properties provide an important insight into the function and structure of general complex networks [4,5]. Moreover, they allow for better comprehension of the underlying real-world systems and thus give prominent grounds for future research in a wide variety of different fields.

In the field of software engineering, network analysis has just recently been adopted to acquire better comprehension of the complex software systems [6–10]. Nowadays, software represents one of the most diverse and sophisticated human made systems; however, only little is known about the actual structure and quantitative properties of (large) software systems. Cai and Yin [9] have denoted this dilemma as the software law problem, which represents an effort towards identifying and formulating physics-like laws, obeyed by (most) software systems, that could later be applied in practice. However, the main objective of the software law problem is to investigate what software looks like.

In the context of employing complex networks analysis, the research community has already made several discoveries over the past years. In particular, different authors have observed that networks, constructed from various software systems, follow scale-free [2] (i.e. power-law) degree distributions and reveal small-world [1] phenomena [11,6]. We proceed this work by exploring another property commonly found in other real-world networks, i.e. community structure [3]. The term denotes the occurrence of local structural modules (communities) that are groups of nodes densely connected within and only loosely connected with the rest of the network. Communities play crucial roles in many real-world systems [12,5], however, the

* Corresponding author.
E-mail address: lovro.subelj@fri.uni-lj.si (L. Šubelj).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.03.036

Physica A, 2011 (9 cit.)


Software Systems through Complex Networks Science: Review, Analysis and Applications

Lovro Šubelj
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Marko Bajec
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

ABSTRACT

Complex software systems are among most sophisticated human-made systems, yet only little is known about the actual structure of 'good' software. We here study different software systems developed in Java from the perspective of network science. The study reveals that network theory can provide a prominent set of techniques for the exploratory analysis of large complex software system. We further identify several applications in software engineering, and propose different network-based quality indicators that address software design, efficiency, reusability, vulnerability, controllability and other. We also highlight various interesting findings, e.g., software systems are highly vulnerable to processes like bug propagation, however, they are not easily controllable.

Categories and Subject Descriptors

D.2.8 [Software Engineering]: Metrics—complexity measures, performance measures, software science

General Terms

Theory, algorithms, experimentation.

Keywords

Software systems, Software engineering, Software networks, Network analysis.

1. INTRODUCTION

Complex software systems are among most sophisticated systems ever created by human. Nevertheless, only little is known about the actual structure and quantitative properties of large software systems [6]. For instance, in the context of software engineering, one is interested in how 'good' software looks like. Commonly adopted approaches and techniques fail to give a comprehensive answer [5, 7], moreover, there is also a lack of a simple but yet rigorous framework for

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SoftwareMining '12, August 12, Beijing, China
Copyright 2012 ACM 978-1-4503-1560-9/12/08 ...$15.00.

software analysis (to our knowledge). The above dilemma was denoted software law problem [6], which urges towards identifying (physical) laws obeyed by software systems that could be used in practical applications.

Networks possibly provide the most adequate framework for the analysis of the structure of complex systems like software projects[1]. Also, due to their simple and intelligible form, analysis of different networks has already provided several significant discoveries in the last decade [46, 3, 16, 23]. Note that the adoption of software networks is not novel [35, 27, 19, 39], however, network analysis is still only rarely used in software engineering. The main purpose of this study is thus to highlight different techniques developed in the field of network analysis, and to expose their use in software comprehension, development and engineering. We review most of the past work on different types of software networks, whereas we also include network analysis techniques proposed just recently [23, 44]. (Note that the main focus of the paper is merely a review, rather than a detailed comparison of network analysis techniques with other approaches.)

The study in the paper analyses software networks on different levels of granularity. First, we address the macroscopic properties of software networks like scale-free and small-world phenomena [46, 3] that are related to the structure and design of the entire project, or projects, represented by the network. Second, we analyze the microscopic properties of individual nodes, with special emphasis on different dynamical processes occurring on software networks like bug propagation [2, 30]. The above can be related to software quality, complexity, reusability, robustness, vulnerability and controllability. Third, we also identify mesoscopic structural modules within software networks [16, 44] and show their applicability in the context of software abstraction and refactoring. The paper thus exposes network analysis as a prominent set of techniques for software engineering.

The rest of the paper is structured as follows. Section 2 introduces software networks used in the study. Section 3 analyzes different characteristics of adopted networks and discusses their use in software engineering. Some applications of the presented techniques are given in Section 4, while Section 5 concludes the paper.

2. SOFTWARE NETWORKS

Various types of networks have been proposed for the analysis of the structure of complex software systems. For

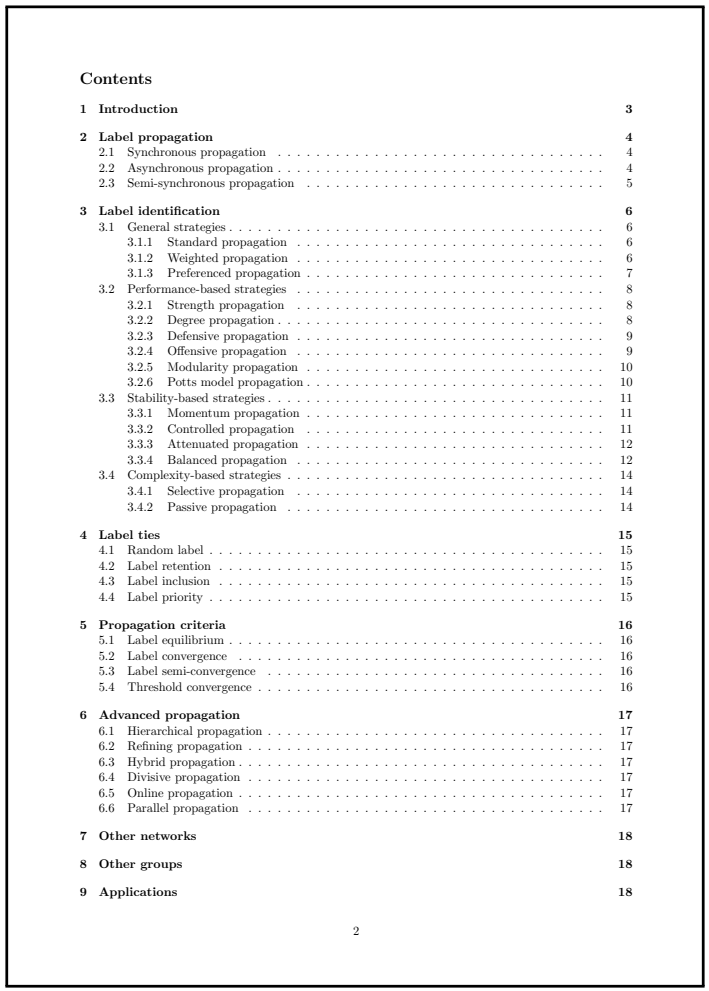
[1]Throughout the paper, the term project refers to a repository of software code.

SoftwareMining (*KDD* '12)

V pripravi

Vsebina

Pregled izmenjave oznak



Contents

2

V pripravi

prof. dr. Xin Liu

dr. Steve Gregory

Analiza skupin vozlišč

Group extraction for real-world networks

Lovro Šubelj[1], Neli Blagus & Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science, Slovenia

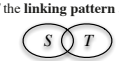
Background

Complex real-world networks contain characteristic groups of nodes with common linking pattern like densely linked **communities** [1]. These were the focus of most recent work and have diverse applications. However, many real-world networks also contain **other groups of nodes** that can be **overlapping** and other, whereas some parts of the networks reveal **no significant groups**.

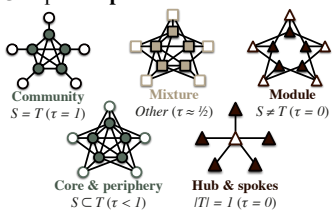
Group **formalism**

Let S be a **group of nodes**, T the **linking pattern** and τ the **group parameter**.

$$\tau(S,T) = \frac{|S \cap T|}{|S \cup T|}$$



Group **examples**



Group **criterion**

Let W be the **group criterion**, L the number of links and μ the (harmonic) mean size.

$$W(S,T) = \mu(S,T)\,(1-\mu(S,T))\left(\frac{L(S,T)}{|S||T|} - \frac{L(S,T^C)}{|S||T^C|}\right)$$

W is a **local asymmetric** criterion that **favors** the links between S and T, and **penalizes** for the links between S and T^C. (Note, however, that W **disregards** the links with both endpoints in S^C.) For $S = T$, W is consistent with a wide class of other models (e.g., *stochastic blockmodel*). [2]

Group **extraction**

A sequential extraction [2] of groups that can be **overlapping, nested etc.**

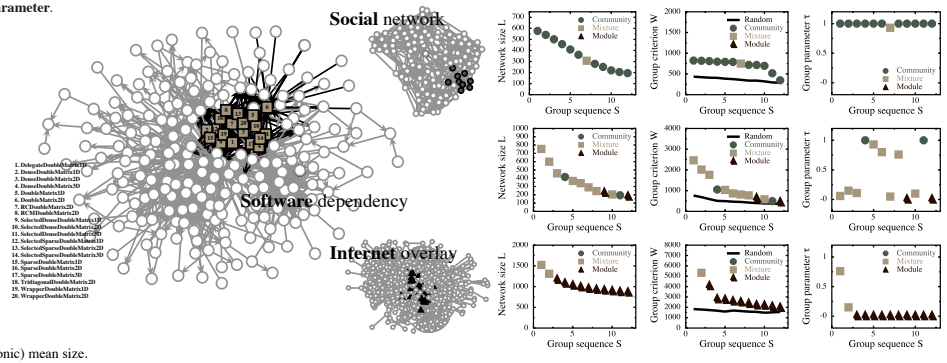
(1) **Find S and T** that optimize criterion W (e.g., *tabu search*).
(2) **Extract** only the **explained links** between S and T (and isolated nodes).
(3) **Repeat** until **W is larger than expected** in a random graph (by simulation).

Contributions

1. A simple **formalism and criterion for general groups** of nodes.
2. An **adequate extraction procedure** for statistically significant groups.
3. **Characterization of the group structure** of different real-world networks.

What are characteristic groups of nodes in real-world networks? **Network (type) dependent.**
What portion of network links is explained by the group structure? **Between 60% and 90%.**
What portion of network nodes is included in the group structure? **More than 50%.**

Groups in **real-world networks**



Network	Nodes	Links	Group #	Group |S|	Group τ	Community	Core	Mixture	Module	Background
						% Links (% nodes)				
Author collaborat. [3]	1589	2742	160	5.6	**0.94**	**71% (47%)**	0% (0%)	6% (5%)	1% (1%)	*22% (47%)*
American football [1]	115	613	13	8.6	**0.88**	**59% (83%)**	9% (11%)	3% (7%)	0% (0%)	*29% (98%)*
Lucene search engine	1657	6808	123	12.1	**0.55**	19% (25%)	1% (2%)	**30% (24%)**	**38% (34%)**	*11% (49%)*
Colt computing [4]	227	963	15	10.3	**0.41**	7% (11%)	5% (6%)	**69% (49%)**	4% (6%)	*15% (64%)*
Word adjacency [3]	112	425	4	11.2	**0.28**	0% (0%)	0% (0%)	**34% (33%)**	25% (15%)	*41% (99%)*
Internet overlay [5]	767	1857	33	10.6	**0.08**	0% (1%)	12% (4%)	13% (7%)	**34% (35%)**	*41% (80%)*
Southern women [6]	32	89	2	4.3	**0.00**	0% (0%)	0% (0%)	0% (0%)	**80% (41%)**	*20% (47%)*

All extracted groups are statistically significant at *1%* level.

References

[1] Girvan, M. & Newman, M.E.J.: Community structure in social and biological networks. **P. Natl. Acad. Sci. USA** 99(12), 7821–7826 (2002).
[2] Zhao, Y., Levina, E., & Zhu, J.: Community extraction for social networks. **P. Natl. Acad. Sci. USA** 108(18), 7321–7326 (2011).
[3] Newman, M. E. J.: Finding community structure in networks using the eigenvectors of matrices. **Phys. Rev. E** 74(3), 036104 (2006).
[4] Šubelj, L. & Bajec, M.: Community structure of complex software systems: Analysis and applications. **Physica A** 390(16), 2968-2975 (2011).
[5] Leskovec, J., Kleinberg, & J., Faloutsos, C.: Graphs over time: Densification laws, shrinking diameters and possible explanations. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, IL, USA, 2005), pp. 177–187.
[6] Davis, A., Gardner, B.B., & Gardner, M.R.: *Deep South* (Chicago University Press, Chicago, 1941).

[1]**Corresponding author:** lovro.subelj@fri.uni-lj.si

NetSci '13

Analiza različnih omrežij

Zgradba programskih omrežij:

May 6, 2013 18:50 WSPC/INSTRUCTION FILE sgs

Advances in Complex Systems
© World Scientific Publishing Company

NODE MIXING AND GROUP STRUCTURE OF SOFTWARE DEPENDENCY NETWORKS

LOVRO ŠUBELJ
Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

SLAVKO ŽITNIK
Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia
slavko.zitnik@fri.uni-lj.si

NELI BLAGUS
Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia
neli.blagus@fri.uni-lj.si

MARKO BAJEC
Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

Received (received date)
Revised (revised date)
Accepted (day month year)

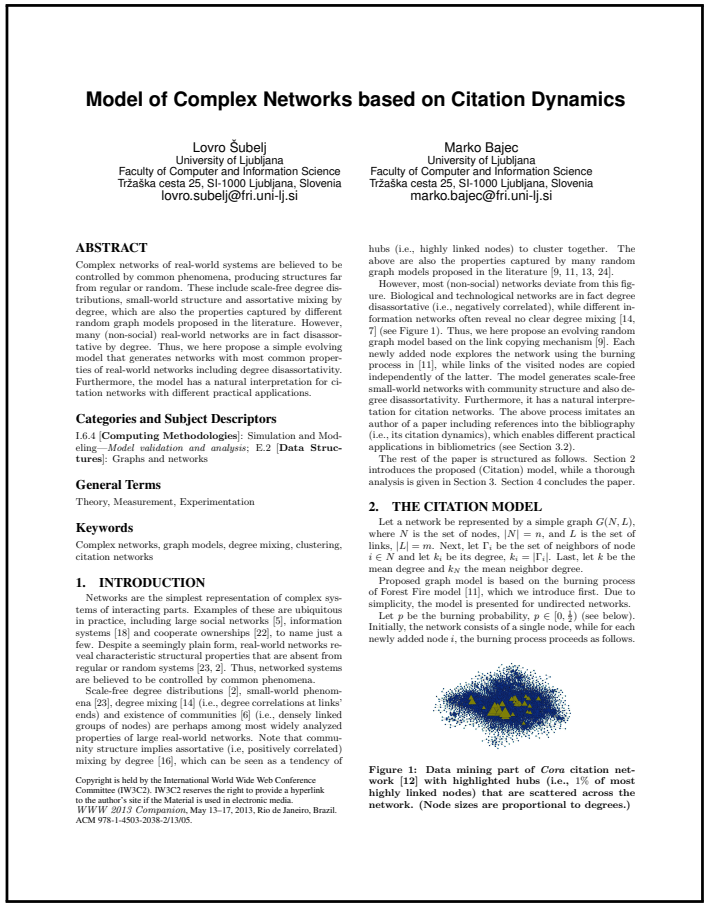
Communicated by (xxxxxxxxxx)

Large software projects are among most sophisticated human-made systems consisting of a network of interdependent parts. Past studies of software systems from the perspective of complex networks have already led to notable discoveries with different applications. Nevertheless, our comprehension of the structure of software networks remains to be only partial. We here investigate correlations or mixing between linked nodes and show that software networks reveal dichotomous degree mixing similar to that recently observed in biological networks. We further show that software networks also reveal characteristic degree-corrected clustering mixing. Hence, node mixing in software networks significantly differs from that in, e.g., Internet or social networks. We explain the observed mixing through the presence of groups of nodes with common linking pattern. More precisely, besides densely connected groups known as communities, software networks also consist of (disconnected) groups denoted modules, core/periphery structures, hubs and spokes, and other. Moreover, groups coincide with the intrinsic properties of the underlying software projects, which promotes practical applications in software engineering.

Keywords: Software networks; node mixing; node groups; software engineering.

V pripravi

Modeli omrežij citiranj:



Model of Complex Networks based on Citation Dynamics

Lovro Šubelj
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Marko Bajec
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

ABSTRACT

Complex networks of real-world systems are believed to be controlled by common phenomena, producing structures far from regular or random. These include scale-free degree distributions, small-world structure and assortative mixing by degree, which are also the properties captured by different random graph models proposed in the literature. However, many (non-social) real-world networks are in fact disassortative by degree. Thus, we here propose a simple evolving model that generates networks with most common properties of real-world networks including degree disassortativity. Furthermore, the model has a natural interpretation for citation networks with different practical applications.

Categories and Subject Descriptors

I.6.4 [**Computing Methodologies**]: Simulation and Modeling—*Model validation and analysis*; E.2 [**Data Structures**]: Graphs and networks

General Terms

Theory, Measurement, Experimentation

Keywords

Complex networks, graph models, degree mixing, clustering, citation networks

1. INTRODUCTION

Networks are the simplest representation of complex systems of interacting parts. Examples of these are ubiquitous in practice, including large social networks [5], information systems [18] and cooperate ownerships [22], to name just a few. Despite a seemingly plain form, real-world networks reveal characteristic structural properties that are absent from regular or random systems [23, 2]. Thus, networked systems are believed to be controlled by common phenomena.

Scale-free degree distributions [2], small-world phenomena [23], degree mixing [14] (i.e., degree correlations at links' ends) and existence of communities [6] (i.e., densely linked groups of nodes) are perhaps among most widely analyzed properties of large real-world networks. Note that community structure implies assortative (i.e, positively correlated) mixing by degree [16], which can be seen as a tendency of hubs (i.e., highly linked nodes) to cluster together. The above are also the properties captured by many random graph models proposed in the literature [9, 11, 13, 24].

However, most (non-social) networks deviate from this figure. Biological and technological networks are in fact degree disassortative (i.e., negatively correlated), while different information networks often reveal no clear degree mixing [14, 7] (see Figure 1). Thus, we here propose an evolving random graph model based on the link copying mechanism [9]. Each newly added node explores the network using the burning process in [11], while links of the visited nodes are copied independently of the latter. The model generates scale-free small-world networks with community structure and also degree disassortativity. Furthermore, it has a natural interpretation for citation networks. The above process imitates an author of a paper including references into the bibliography (i.e., its citation dynamics), which enables different practical applications in bibliometrics (see Section 3.2).

The rest of the paper is structured as follows. Section 2 introduces the proposed (Citation) model, while a thorough analysis is given in Section 3. Section 4 concludes the paper.

2. THE CITATION MODEL

Let a network be represented by a simple graph $G(N, L)$, where N is the set of nodes, $|N| = n$, and L is the set of links, $|L| = m$. Next, let Γ_i be the set of neighbors of node $i \in N$ and let k_i be its degree, $k_i = |\Gamma_i|$. Last, let k be the mean degree and k_N the mean neighbor degree.

Proposed graph model is based on the burning process of Forest Fire model [11], which we introduce first. Due to simplicity, the model is presented for undirected networks.

Let p be the burning probability, $p \in [0, 1)$ (see below). Initially, the network consists of a single node, while for each newly added node i, the burning process proceeds as follows.

Figure 1: Data mining part of *Cora* citation network [12] with highlighted hubs (i.e., 1% of most highly linked nodes) that are scattered across the network. (Node sizes are proportional to degrees.)

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

LSNA (*WWW* '13)

Vsebina

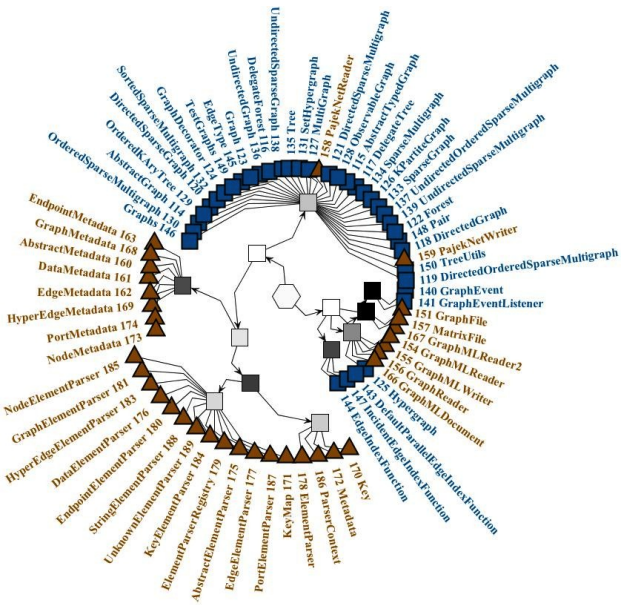
Zaključek

Znanstveni doprinos:

- pristopi za odkrivanje skupin vozlišč v omrežjih

  zahtevnost, splošnost, natančnost, robustnost, enostavnost, brez parametrov

- analiza in uporaba skupin v programskih omrežjih



Hvala za pozornost

Po zagovoru vabljeni v Laboratorij za podatkovne tehnologije.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of United States of America*, 99(12): 7821–7826, 2002.

I. X. Y. Leung, P. Hui, P. Liò, and J. Crowcroft. Towards real-time community detection in large networks. *Physical Review E*, 79(6):066107, 2009.

M. E. J. Newman. The physics of networks. *Physics Today*, 61(11):33–38, 2008.

M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of United States of America*, 104(23): 9564, 2007.

U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

L. Šubelj and M. Bajec. Unfolding network communities by combining defensive and offensive label propagation. In *Proceedings of the ECML PKDD Workshop on the Analysis of Complex Networks*, pages 87–104, Barcelona, Spain, 2010.

L. Šubelj and M. Bajec. Robust network community detection using balanced propagation. *European Physical Journal B*, 81(3):353–362, 2011a. doi: 10.1140/epjb/e2011-10979-2.

L. Šubelj and M. Bajec. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Physical Review E*, 83(3):036103, 2011b. doi: 10.1103/PhysRevE.83.036103.

L. Šubelj and M. Bajec. Generalized network community detection. In *Proceedings of the ECML PKDD Workshop on Finding Patterns of Human Behaviors in Network and Mobility Data*, pages 66–84, Athens, Greece, 2011c.

L. Šubelj and M. Bajec. Community structure of complex software systems: Analysis and applications. *Physica A: Statistical Mechanics and its Applications*, 390(16):2968–2975, 2011d. doi: 10.1016/j.physa.2011.03.036.

L. Šubelj and M. Bajec. Clustering assortativity, communities and functional modules in real-world networks. *e-print arXiv:12082518v1*, pages 1–21, 2012a.

L. Šubelj and M. Bajec. Software systems through complex networks science: Review, analysis and applications. In *Proceedings of the KDD Workshop on Software Mining*, pages 9–16, Beijing, China, 2012b. doi: 10.1145/2384416.2384418.

L. Šubelj and M. Bajec. Ubiquitousness of link-density and link-pattern communities in real-world networks. *European Physical Journal B*, 85(1):32, 2012c. doi: 10.1140/epjb/e2011-20448-7.

G. Tibély and J. Kertész. On the equivalence of the label propagation method of community detection and a potts model approach. *Physica A: Statistical Mechanics and its Applications*, 387(19-20):4982–4984, 2008.

S. Valverde, R. F. Cancho, and R. V. Solé. Scale-free networks from optimal design. *Europhysics Letters*, 60(4):512, 2002.