ANALIZA KOMPLEKSNIH OMREŽIJ: OSNOVNI POJMI IN PRIMERI UPORABE V PRAKSI

Lovro Šubelj, Neli Blagus, Štefan Furlan, Bojan Klemenc, Aleš Kumer, Dejan Lavbič, Aljaž Zrnec, Slavko Žitnik in Marko Bajec
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška cesta 25, Ljubljana
[ime.priimek]@fri.uni-lj.si

Povzetek

Predvsem z razmahom svetovnega spleta, družabnih omrežij in drugih oblik sporočanja so se v zadnjih letih začele pojavljati bogate podatkovne zbirke, ki poleg osnovnih lastnosti o posameznikih hranijo tudi različne vrste relacij med njimi. Relacije med entitetami navadno skrivajo veliko količino informacij, in tako razkrivajo kompleksne lastnosti, ki bi sicer pogosto ostale neodkrite. S preučevanjem relacij med entitetami se ukvarja *analiza omrežij*, ki trenutno velja za eno najbolj aktualnih področij v analizi podatkov. V prispevku najprej predstavimo osnovne pojme analize omrežij, ter izpostavimo vidnejše dosežke in odkritja iz preteklih let. V nadaljevanju nato predstavimo izbrane primere uporabe ter podamo nekatere možnosti uporabe v javni upravi.

Abstract

COMPLEX NETWORKS ANALYSIS: INTRODUCTION AND APPLICATIONS

With the recent expansion of *WWW*, social networks and other forms of communication there is an increase of rich datasets that include not only the basic properties of entities, but also various relations among them. Relations between entities usually hold a large amount of information, and thus reveal complex phenomena that would otherwise often remain unnoticed. Study of the relations among entities is called *network analysis*, which is currently considered one of the most prominent fields in the data analysis research. In this paper we first introduce basic concepts of network analysis, and highlight major discoveries in the recent years. We then present a selected set of practical applications and describe some possible uses in the public administration domain.

Ključne besede

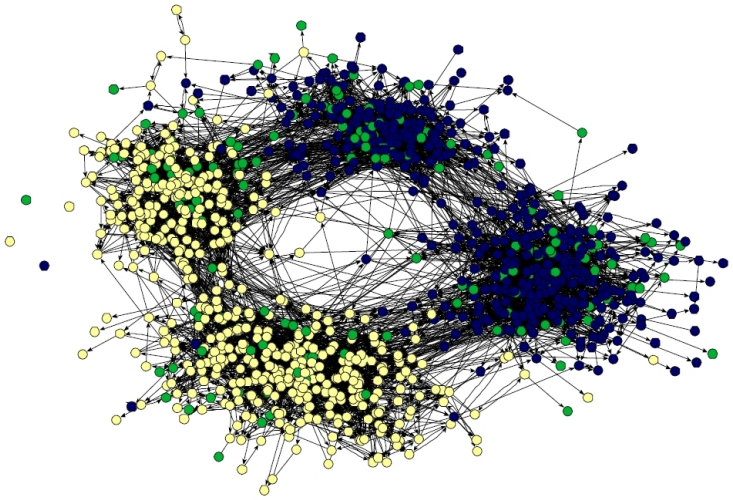
Analiza omrežij, teorija grafov, osnovni pojmi, primeri uporabe.

Key words

Network analysis, graph theory, introduction, applications.

1 UVOD

Predvsem z razmahom svetovnega spleta, pojavom družabnih omrežij ter drugih oblik sporočanja so se v zadnjih letih začele pojavljati bogate podatkovne zbirke, ki poleg osnovnih lastnosti o posameznikih hranijo tudi različne vrste relacij, odnosov ali interakcij med njimi. Relacije med poljubnimi entitetami navadno skrivajo veliko količino informacij, in



Slika 1: Socialno omrežje prijateljstev med otroci na ameriški šoli. Barve vozlišč predstavljajo različne rase, usmerjena povezava med vozlišči pa pomeni, da je prvi otrok prijatelj drugega. Otroci se očitno delijo v štiri množice vozlišč (skupnosti), ki se ločijo glede na raso (navpičen prerez) ter na nižje in višje letnike (vodoraven prerez). [10]

tako razkrivajo kompleksne lastnosti, ki bi sicer pogosto ostale neodkrite. S preučevanjem relacij med entitetami se ukvarja *analiza omrežij*.

Omrežje je sestavljeno iz množice vozlišč, ter množice povezav med njimi (Slika 1). Vozlišča lahko ponazarjajo različne entitete; v primeru, da vozlišča predstavljajo skupino oseb, omrežjem pravimo *socialna omrežja* (od tod ang. *social network analysis*). V preteklosti pa so bila veliko preučevana tudi *informacijska* (WWW, citati, itd.), *biološka* (proteinske interakcije, geni, itd.), *tehnološka* (Internet, električno omrežje, itd.) in mnoga druga omrežja.

V nadaljevanju najprej podamo osnovne pojme analize omrežij, ter izpostavimo vidnejše dosežke in odkritja iz preteklih let (razdelek 2). V razdelku 3 nato predstavimo izbrane primere uporabe v praksi, in podamo nekatere možnosti uporabe v javni upravi v razdelku 4. Na koncu sledijo še sklepne ugotovitve ter viri za nadaljnje branje (razdelek 5).

2 ANALIZA OMREŽIJ

Analiza omrežij je izjemno široko področje, ki vključuje (pod-)področja kot so *analiza socialnih omrežij* (ang. *social network analysis*), *analiza povezav* (ang. *link analysis*), *podatkovno rudarjenje nad grafi* (ang. *graph based data mining*), *odkrivanje skupnosti* (ang. *community detection*) in druga. Področje je prav tako tesno povezano z matematično *teorijo grafov* (ang. *graph theory*), *podatkovnim rudarjenjem* (ang. *data mining*), *strojnim učenjem* (ang. *machine learning*), *statistiko* (ang. *statistics*), *spektralno analizo* (ang. *spectral analysis*) in nekaterimi drugimi področji.

V nadaljevanju podamo osnovne termine analize omrežij (razdelek 2.1) in predstavimo nekatera pomembnejša odkritja o strukturi kompleksnih omrežij (razdelki 2.2-2.5).

2.1 Osnovni pojmi

Omrežje navadno predstavimo z matematičnim objektom, ki mu pravimo *graf* (Slika 1). Graf $G(N, E)$ je sestavljen iz množice *vozlišč* $N = \{e_1, e_2 \dots e_n\}$, in množice *povezav* med njimi $E = \{\{e_i, e_j\} |\ e_i, e_j \in N\}$.

V zgornji definiciji povezave niso usmerjene (graf je *neusmerjen*). V praksi običajno dopuščamo tudi več vzporednih povezav med vozlišči in pa *zanke* (povezave, ki povezujejo vozlišče s samim seboj). Grafom tedaj pravimo *multi*, *psevdo grafi* zaporedoma. Graf je lahko sestavljen iz več povezanih množic vozlišč, ki pa med seboj niso povezana. Takim množicam vozlišč pravimo *komponente* grafa. V nadaljevanju predpostavimo, da imamo opravka z enostavnim neusmerjenim grafom, v katerem so vsa vozlišča med seboj povezana (graf vsebuje le eno komponento).

Pogosto opazujemo število povezav, ki vodijo iz nekega vozlišča $e_i \in N$, kar imenujemo *stopnja* vozlišča. Poleg stopnje pa nas običajno zanima tudi razdalja med vozlišči. Naj bo $g(e_i, e_j)$ *geodetka* med vozlišči $e_i, e_j \in N$ – najkrajša pot med e_i in e_j, kjer štejemo število povezav na poti. Dolžina geodetke $g(e_i, e_j)$ predstavlja *razdaljo* med vozliščema e_i in e_j.

2.2 Potenčni zakon

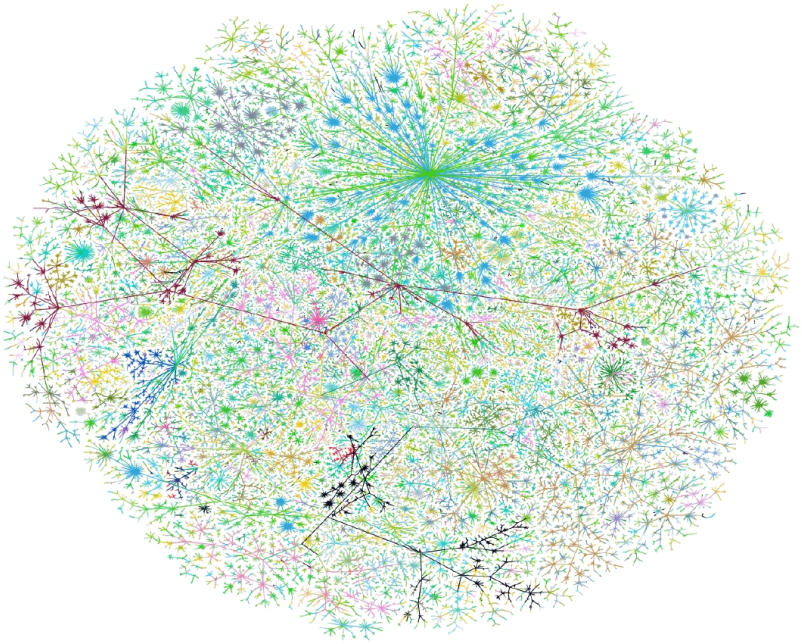
V preteklosti so raziskovalci veliko zanimanja posvetili preučevanju stopenj vozlišč. Izkaže se, da je porazdelitev stopenj v realnih omrežjih značilno drugačna od tiste v (enostavnih) naključnih omrežjih. Pri slednjih so stopnje vozlišč porazdeljene tesno okrog povprečja, kar pa ne velja za realna omrežja, kjer je porazdelitev navadno močno raztegnjena v desno. Povedano drugače, v realnih omrežjih obstajajo vozlišča z veliko večjo stopnjo, kot pa znaša povprečje. Izkaže se, da porazdelitev običajno sledi *potenčnemu zakonu* (ang. *power-law*); omrežjem, ki se podrejajo potenčnemu zakonu, pa pravimo ang. *scale-free* omrežja [2].

V primeru omrežja prijateljstev na Sliki 1 se potenčni zakon kaže v tem, da obstajajo osebe z ogromnim številom prijateljev (veliko večjim, kot pa znaša povprečje). Potenčni zakon je moč opaziti tudi v Internetnem omrežju (Slika 2), kjer obstajajo komponente strojne opreme, ki so povezane z zelo veliko drugimi komponentami.

Obstaja več razlag o izvoru *scale-free* omrežij. Ena od njih je načelo *prednostne povezanosti* [2] (ang. *preferential attachment*), ki pravi, da bogati bogatijo (ang. *rich get richer*).

2.3 Pojav majhnega sveta

Stanley Milgram je v 60. letih prejšnjega stoletja izvedel naslednji eksperiment. Posameznikom je zadal nalogo naj preko zaporedja pisem dosežejo neko poljubno izbrano osebo (na



Slika 2: Internetno omrežje na nivoju "avtonomnih sistemov" (vsako vozlišče predstavlja tudi do več tisoč posameznih računalnikov). [4]

drugem koncu sveta). Večina pisem se je tekom eksperimenta sicer izgubila, ostala pisma pa so prišla na cilj v presenetljivo majhnem številu korakov. V objavljenih primerih je bilo to število v povprečju enako 6, od koder izhaja izraz *šest stopenj separacije* (ang. *six degrees of separation*). Slednje velja za prvi prikaz *pojava majhnega sveta* [18] (ang. *small-world effect*), ki pravi, da sta poljubni dve vozlišči v omrežju povezani preko zelo kratke poti.

Pojav navadno merimo z opazovanjem povprečne razdalje med vozlišči l. Raziskave kažejo, da je pojav majhnega sveta prisoten v številnih omrežjih različnih vrst – povprečna razdalja l navadno ni večja od 10, tudi v zelo velikih omrežjih (Slika 1, 2). Zadnji rezultati raziskav nad največjim kdajkoli analiziranim omrežjem (omrežje *Yahoo!* s preko milijardo vozlišči) prav tako potrjujejo, da vrednost l verjetno ni večja od 7 [7].

2.4 Tranzitivnost

V realnih omrežjih pogosto opazimo pojav *tranzitivnosti* – v kolikor obstaja povezava med vozliščema $e_i, e_j \in N$, in med vozliščema $e_j, e_k \in N$, pogosto obstaja tudi povezava med vozliščema e_i in e_k. V omrežju prijateljstev (Slika 1) slednje pomeni, da je prijatelj našega prijatelja pogosto tudi naš prijatelj.

Tranzitivnost se kaže v velikem številu trikotnikov v omrežju, in je v realnih omrežjih navadno veliko večja kot v primerljivih naključnih omrežjih. Merimo jo s pomočjo *koeficienta*

gručenja C [18] (ang. *clustering coefficient*), ki je v primeru omrežja prijateljstev dejansko enak verjetnosti, da je prijatelj našega prijatelja prav tako prijatelj.

2.5 Skupnosti

Realna omrežja so pogosto sestavljena iz *skupnosti* [5] (ang. *communities*). Skupnosti so množice vozlišč, ki so med seboj zelo tesno povezana, dočim med samimi skupnostmi skorajda ni povezav. V omrežju na Sliki 1 lahko vidimo štiri jasno definirane skupnosti.

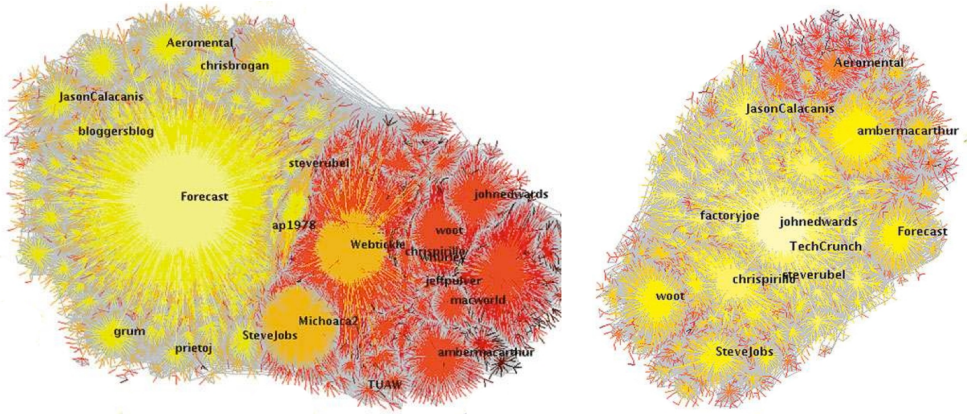
Skupnosti igrajo pomembno vlogo v številnih kompleksnih sistemih – prehranjevale verige, proteinske interakcije, družabna omrežja, in drugo. Raziskave pa kažejo, da so s stališča omrežja skupnosti dobro definirane le pri majhnem številu vozlišč. Natančneje, dobro definirane skupnosti ležijo na obrobju omrežja, in vsebujejo približno 100 vozlišč [9]. Slednje se ujema z Dunbarjevim številom o velikosti obvladljive človeške okolice (med 100 in 200).

3 PRIMERI UPORABE V PRAKSI

V nadaljevanju na kratko predstavimo izbrane primere uporabe analize omrežij v praksi.

3.1 Analiza družabnih omrežij

Najznačilnejši primer uporabe analize omrežij je gotovo preučevanje različnih družabnih omrežij, kot so *Twitter*, *Flickr*, *Facebook* in druga. Na Sliki 3 lahko vidimo socialno omrežje poznanstev med 25 000 uporabniki omrežja *Twitter*. Opazimo, da v omrežju obstajajo uporabniki z značilno centralno vlogo – na primer, napoved vremena (ang. *forecast*), *Steve Jobs*, *John Edwards* in nekateri drugi. Slednji so povezani z večjim številom ostalih uporabnikov, in imajo tako posledično zelo velik vpliv na celotno omrežje.

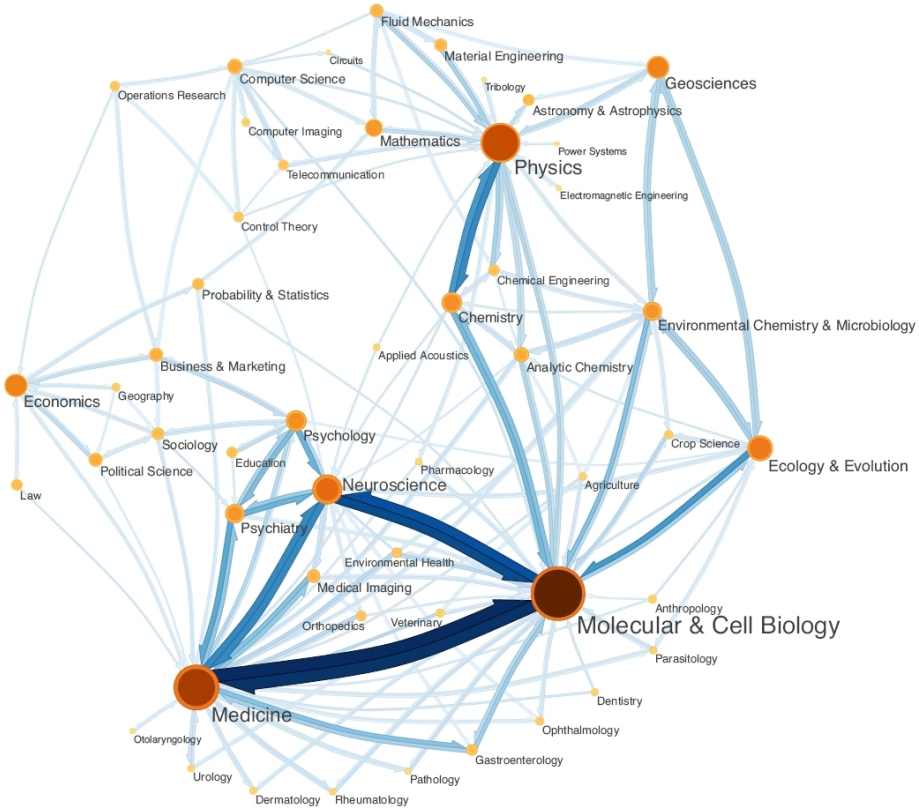


(a) (b)

Slika 3: Omrežje (a) enostranskih in (b) vzajemnih poznanstev med uporabniki omrežja *Twitter*. Opazimo, da imajo nekateri uporabniki značilno centralno vlogo v omrežju. [6]

3.2 Analiza znanstvenih publikacij

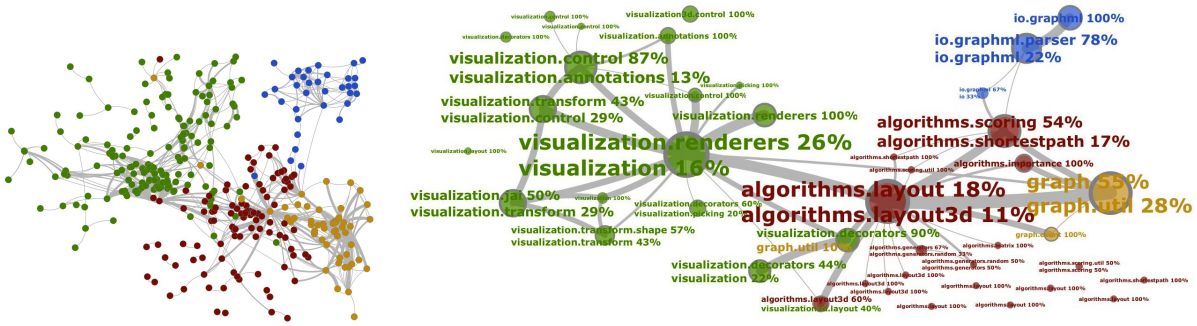
Rosvall in Bergstrom [14] sta zbrala prispevke iz vrste različnih znanstvenih publikacij, ki sta jih povezala glede na medsebojno citiranje. Dobljeno informacijsko omrežje sta analizirala s pomočjo *naključnih sprehodov* (ang. *random walk*). Rezultati kažejo (Slika 4), da ima znanost obliko črke "U" – levo (desno) zgoraj so družboslovne (naravoslovne) znanosti, med seboj pa so povezane prek interdisciplinarnih področij, kot sta medicina in biologija (spodaj).



Slika 4: Omrežje medsebojnega citiranja med znanstvenimi prispevki, kjer vsako vozlišče predstavlja množico prispevkov. Rezultati kažejo, da ima znanost obliko črke "U". [14]

3.3 Analiza kompleksnega programja

V zadnjih letih se analiza omrežij uporablja za preučevanje številnih kompleksnih sistemov. Eden od zanimivejših primerov je analiza programskih sistemov (programja), saj ti sodijo med najkompleksnejše sisteme, ki jih je kdaj ustvaril človek. S pomočjo odkrivanja skupnosti v omrežjih (objektnega) programja lahko razkrijemo razdelitev razredov po paketih, ki minimizira odvisnosti med razredi različnih paketov – načelo *modularnosti* (Slika 5). [15]

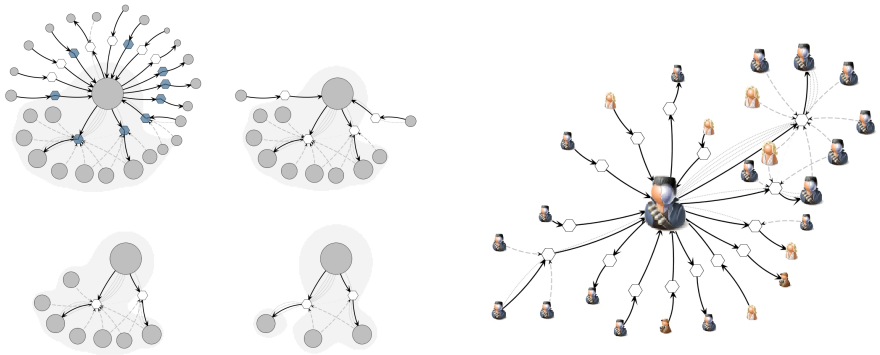


(a) (b)

Slika 5: (a) Omrežje (objektne) knjižnice *JUNG* [13] (vozlišča predstavljajo razrede, povezave pa odvisnosti med njimi). (b) Razdelitev razredov po paketih, ki minimizira odvisnosti. [15]

3.4 Odkrivanje avtomobilskih goljufij

Analiza omrežij se pogosto uporablja tudi za odkrivanje anomalij (odstopanj) v podatkih. V primeru omrežij prometnih nesreč lahko z analizo povezav razkrijemo goljufive skupine posameznikov, ki uprizarjajo prometne nesreče in se tako okoristijo preko svojega zavarovanja (Slika 6). S pomočjo *propagacije* sumljivosti po omrežju lahko razkrijemo goljufive udeležence, nesreče in vozila, ter izpostavimo ključne povezave med njimi. [16]



(a) (b)

Slika 6: Rezultati odkrivanja skupin avtomobilskih goljufov. Okrogla (oglata) vozlišča predstavljajo udeležence (nesreče), velikost vozlišč pa je sorazmerna njihovi sumljivosti. [16]

4 UPORABA V JAVNI UPRAVI

Metode in pristope analize omrežij je moč uporabiti povsod, kjer nas zanimajo relacije med določenimi entitetami. S pomočjo analize (socialnih) omrežij preučujemo globalne lastnosti omrežij, s pomočjo analize povezav odkrivamo ključne ali izstopajoče entitete, množice entitet s skupnimi lastnostmi in značilnostmi pa razkrivamo s pomočjo analize skupnosti.

V javni upravi so zanimiva predvsem omrežja, ki prikazujejo odnose med določenimi posamezniki – socialna omrežja. V nadaljevanju podrobneje predstavimo nekatere možnosti uporabe v javni upravi.

S pomočjo analize omrežij lahko na primer razkrijemo politično usmerjenost ljudi. V tem primeru z metodami analize povezav ocenimo podporo posameznikov določenemu kandidatu, ali pa z identifikacijo skupnosti poiščemo skupine ljudi, ki bodo verjetno glasovali za istega kandidata. Pri tem uporabljamo predvsem omrežja, ki predstavljajo različne odnose med ljudmi, celotna analiza pa temelji na predpostavki, da ljudje s podobnimi lastnostmi pogosto podprejo istega kandidata. Za več glej [1, 8].

V preteklosti so bile razvite številne (predvsem statistične) metode za odkrivanje in preprečevanje različnih vrst goljufij (glej [3]). Omenjeni pristopi pa v večini zanemarijo interakcije med posamezniki, ki so ključne pri odkrivanju določenih vrst goljufij. V primeru pranja denarja lahko sodelujoče skupine goljufov razkrijemo zgolj z ustrezno analizo odnosov med njimi – uporaba pristopov analize omrežij je tako v tem primeru ključna.

Podobno lahko s primerno analizo odnosov med posamezniki razkrijemo različne nepravilnosti v podatkih. Pri določanju lastništva nepremičnin lahko z analizo sorodstvenih vezi in drugih odnosov med ljudmi odkrijemo anomalije, ki ne sovpadajo z omenjenim omrežjem in tako predstavljajo možno nepravilnost v podatkih.

Na koncu omenimo še, da omrežja omogočajo enostaven, a hkrati izjemno pregleden, prikaz podatkov mnogih kompleksnih domen, kjer drugi načini predstavitve pogosto odpovedo.

5 ZAKLJUČEK

V prispevku predstavimo osnovne pojme analize omrežij, ter opišemo bistvene dosežke iz preteklih let. V nadaljevanju nato izpostavimo nekatere primere uporabe v praksi, in na kratko predstavimo možnosti uporabe v javni upravi. Analiza kompleksih omrežij je razmeroma mlado področje, ki pa je v zadnjem desetletju doživelo močan vzpon. Sledenje je omogočilo uporabo pri številnih praktičnih problemih, analiza omrežij pa trenutno velja za eno najobetavnejših področij analize podatkov. Za celovit pregled področja glej [11, 12, 17].

VIRI IN LITERATURA

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election. In *Proceedings of the International Workshop on Link Discovery*, pages 36–43, 2005.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

[3] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–249, 2002.

[4] H. Burch and B. Cheswick. Network of autonomous systems of internet by internet mapping project. URL `http://www.lumeta.com/research/`.

[5] 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*, pages 7821–7826, 2002.

[6] A. Java. *Mining social media communities and content*. PhD thesis, University of Maryland, 2008.

[7] U. Kang, C. E. Tsourakakis, A. P. Appel, C. Faloutsos, and J. Leskovec. Radius plots for mining tera-byte scale graphs: Algorithms, patterns, and observations. In *Proceedings of the SIAM International Conference on Data Mining*, 2010.

[8] V. Krebs. The link between social interaction and political choice. In *Extreme democracy*. 2004. URL `http://www.extremedemocracy.com/`.

[9] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *e-print arXiv:0810.1355*, (1), 2008.

[10] J. Moody. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, 107(3):679–716, 2001.

[11] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45 (2):167–256, 2003.

[12] M. E. J. Newman. *Networks: An introduction*. Oxford University Press, 2010.

[13] J. O'Madadhain, D. Fisher, S. White, P. Smyth, and Y.-B. Boey. Analysis and visualization of network data using JUNG. *Journal of Statistical Software*, 10(2):1–35, 2005.

[14] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of United States of America*, 105:1118–1123, 2008.

[15] L. Subelj and M. Bajec. Community structure of complex software systems: Analysis and applications. *Oddano v Physica A: Statistical Mechanics and its Applications*.

[16] L. Subelj, S. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38(1):1039–1052, 2011.

[17] D. J. Watts. The 'new' science of networks. *Annual Review of Sociology*, 30(1):243–270, 2004.

[18] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.