Advances in Intelligent Systems and Computing 246

G. Sai Sundara Krishnan · R. Anitha
R. S. Lekshmi · M. Senthil Kumar
Anthony Bonato · Manuel Graña
Editors

Computational Intelligence, Cyber Security and Computational Models

Proceedings of ICC^3, 2013

Springer

Advances in Intelligent Systems and Computing

Volume 246

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

G. Sai Sundara Krishnan · R. Anitha
R. S. Lekshmi · M. Senthil Kumar
Anthony Bonato · Manuel Graña
Editors

Computational Intelligence, Cyber Security and Computational Models

Proceedings of ICC^3, 2013

Springer

Editors
G. Sai Sundara Krishnan
R. Anitha
R. S. Lekshmi
M. Senthil Kumar
Applied Mathematics and Computational Sciences
PSG College of Technology
Coimbatore, Tamil Nadu
India

Manuel Graña
School of Computing
University of Basque Country
Paseo Manuel De Lardizalbal 1
San Sebastian
Spain

Anthony Bonato
Department of Mathematics
Ryerson University
Toronto, ON
Canada

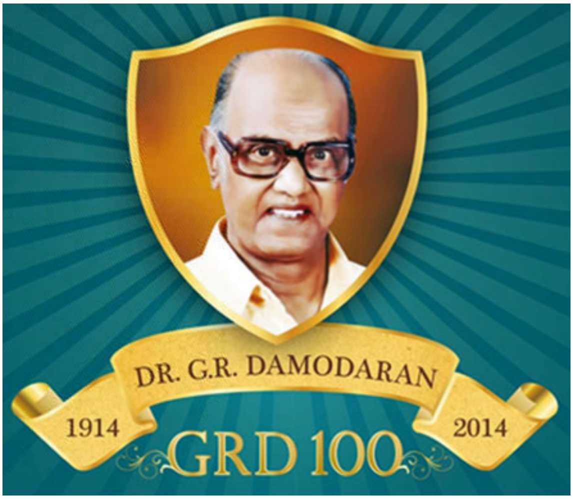
ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-81-322-1679-7 ISBN 978-81-322-1680-3 (eBook)
DOI 10.1007/978-81-322-1680-3
Springer New Delhi Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013954037

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated
To



Dr. G. R. Damodaran
Founder Principal
PSG College of Technology
Coimbatore–641004
India

Preface

The rapid development of network technologies and computing machines has broadened the scope of research and development in computer science and allied areas. To provide a broad interdisciplinary research forum, the International Conference on Computational Intelligence, Cyber Security, and Computational Models (ICC^3-2013) has been organized by the Department of Applied Mathematics and Computational Sciences of PSG College of Technology, during 19–21 December, 2013. We are proud to place on record that this International Conference is a part of the centenary year celebrations of Dr. G. R. Damodaran, Founder Principal, PSG College of Technology, Coimbatore, India.

The primary objective of this conference is to present state-of-the-art scientific results, explore cutting-edge technologies, and promote collaborative research in the areas of revolutionary ideas using computational intelligence, cyber security, and computational models. The conference aims to serve as a platform to establish research relations worldwide.

Computational Intelligence (CI), as a branch of science is applicable in many fields of research, including engineering, data analytics, forecasting, and biomedicine. CI systems are inspired by imitable aspects of living systems. They are used in image and sound processing, signal processing, multidimensional data visualization, steering of objects, expert systems, and many practical implementations. The common feature of CI systems is that it processes information by symbolic representation of knowledge. CI systems have the capability to reconstruct behaviors observed in learning sequences, form rules of inference, and generalize knowledge in situations where they are expected to make predictions or classify objects based on previously observed categories. The CI track comprises research articles which exhibit potential practical applications.

With worldwide escalation in the number of cyber threats, there is a need for comprehensive security analysis, assessment, and action to protect critical infrastructure and sensitive information. Large-scale cyber attacks in various countries threaten information security which, could pose a threat to national security and requires effective crisis management. Such information security risks are becoming increasingly diversified, advanced, and complex and conventional means of security fail to ensure information safety. The cyber security track in this conference aims to bring together researchers, practitioners, developers, and users to arrive at a common understanding of the challenges and build a global framework for security and trust.

Fields such as theory of computation, data analytics, high performance computing, quantum computing, weather forecasting, and flight simulation need computational models like stochastic models, graph models, and neural networks to make predictions about performance of complicated systems. Solutions to many technical problems require extensive mathematical concepts to model the problem and to understand the behavior of associated complex systems by computer simulations. With the advent of efficient computations, solutions to complex problems can be found using computational modeling and research.

ICC^3-2013 received a total of 117 technical submissions out of which only 33 full papers and five short papers were selected for presentation and publication in the proceedings. This selection was done through a stringent blind peer review process. Besides these, research papers by invited speakers have also been included in this proceedings.

The organizers of ICC^3-2013 wholeheartedly appreciate the peer reviewers for their support and valuable comments for ensuring the quality of this proceeding. We also extend our warmest gratitude to Springer for their support in bringing out the proceedings volume in time and with excellent production quality.

We would like to thank all invited speakers, international advisory committee members, and the chairpersons for their excellent contributions. We hope that all the participants of the conference will be benefited academically from this event and wish them success in their research career.

Organization

Patron	Shri L. Gopalakrishnan, Managing Trustee, PSG and Sons Charities Trust, Coimbatore, India
Chairman	Dr. R. Rudramoorthy, Principal, PSG College of Technology, Coimbatore, India
Organizing Chair	Dr. R. Nadarajan, Professor and Head, Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, India
Program Chair	Dr. G. Sai Sundara Krishnan, Associate Professor, Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, India
Computational Intelligence Track Chair	Dr. M. Senthil Kumar, Associate Professor, Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, India
Cyber Security Track Chair	Dr. R. Anitha, Associate Professor, Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, India
Computational Models Track Chair	Dr. R. S. Lekshmi, Associate Professor, Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, India

Advisory Committee Members

Prof. Anthony Bonato	Department of Mathematics, Ryerson University, Canada
Prof. Dan Kannan	Department of Mathematics, University of Georgia, Georgia, USA

Prof. Indrajit Ray	Department of Computer Science, Colorado State University, Fort Collins, USA
Prof. Jeffrey D. Ullman	Department of Computer Science, Stanford University, USA
Prof. Khosrow Sohraby	School of Computing and Engineering, University of Missouri-Kansas City, USA
Prof. Manuel Graña	School of Computing, University of Basque Country, Paseo Manuel de Lardizabal 1, Spain
Prof. Rein Nobel	Vrije University, Amsterdam, The Netherlands
Prof. Srinivas R. Chakravarthy	Department of Industrial and Manufacturing Engineering and Business, Kettering University, Michigan, USA
Prof. Subir Kumar Ghosh	School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India
Prof. Tan Kay Chen	Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Contents

Part I Keynote Address

The Robber Strikes Back 3
Anthony Bonato, Stephen Finbow, Przemysław Gordinowicz,
Ali Haidar, William B. Kinnersley, Dieter Mitsche,
Paweł Prałat and Ladislav Stacho

Some Applications of Collective Learning 13
Balaraman Ravindran

Subconscious Social Computational Intelligence 15
M. Graña

Modeling Heavy Tails in Traffic Sources for Network Performance Evaluation 23
Vaidyanathan Ramaswami, Kaustubh Jain, Rittwik Jana
and Vaneet Aggarwal

The Future of Access Control: Attributes, Automation, and Adaptation 45
Ravi Sandhu

Optimal Control for an $M^X/G/1/N + 1$ Queue with Two Service Modes 47
Rein D. Nobel and Adriaan A. N. Ridder

Part II Computational Intelligence

A Novel Approach to Gene Selection of Leukemia Dataset Using Different Clustering Methods 63
P. Prasath, K. Perumal, K. Thangavel and R. Manavalan

A Study on Enhancing Network Stability in VANET with Energy Conservation of Roadside Infrastructures Using LPMA Agent 71
T. Karthikeyan and N. Sudha Bhuvaneswari

An Intuitionistic Fuzzy Approach to Fuzzy Clustering of Numerical Dataset 79
N. Karthikeyani Visalakshi, S. Parvathavarthini and K. Thangavel

ELM-Based Ensemble Classifier for Gas Sensor Array Drift Dataset 89
D. Arul Pon Daniel, K. Thangavel, R. Manavalan and R. Subash Chandra Boss

Hippocampus Atrophy Detection Using Hybrid Semantic Categorization 97
K. Selva Bhuvaneswari and P. Geetha

Hyper-Quadtree-Based K-Means Algorithm for Software Fault Prediction 107
Rakhi Sasidharan and Padmamala Sriram

Measurement of Volume of Urinary Bladder by Implementing Localizing Region-Based Active Contour Segmentation Method 119
B. Padmapriya, T. Kesavamurthi, B. Abinaya and P. Hemanthini

Improved Bijective-Soft-Set-Based Classification for Gene Expression Data 127
S. Udhaya Kumar, H. Hannah Inbarani and S. Senthil Kumar

Mammogram Image Classification Using Rough Neural Network 133
K. T. Rajakeerthana, C. Velayutham and K. Thangavel

Performance Assessment of Kernel-Based Clustering 139
Meena Tushir and Smriti Srivastava

Registration of Ultrasound Liver Images Using Mutual Information Technique 147
R. Suganya, R. Kirubakaran and S. Rajaram

Sentiment Mining Using SVM-Based Hybrid Classification Model. . . . 155
G. Vinodhini and R. M. Chandrasekaran

Shape Based Image Classification and Retrieval with Multiresolution Enhanced Orthogonal Polynomial Model 163
S. Sathiya Devi and R. Krishnamoorthi

Using Fuzzy Logic for Product Matching 171
K. Amshakala and R. Nedunchezhian

Part III Cyber Security

A Novel Non-repudiate Scheme with Voice FeatureMarking 183
A. R. Remya, M. H. Supriya and A. Sreekumar

A Study of Spam Detection Algorithm on Social Media Networks 195
Jacob Soman Saini

Botnets: A Study and Analysis 203
G. Kirubavathi and R. Anitha

Comparative Study of Two- and Multi-Class-Classification-Based Detection of Malicious Executables Using Soft Computing Techniques on Exhaustive Feature Set 215
Shina Sheen, R. Karthik and R. Anitha

CRHA: An Efficient HA Revitalization to Sustain Binding Information of Home Agent in MIPv6 Network 227
A. Avelin Diana, V. Ragavinodhini, K. Sundarakantham and S. Mercy Shalinie

Interaction Coupling: A Modern Coupling Extractor 237
S. Gomathi and P. Edith Linda

Secure Multicasting Protocols in Wireless Mesh Networks—A Survey 245
Seetha Surlees and Sharmila Anand John Francis

Part IV Computational Models

A New Reversible SMG Gate and Its Application for Designing Two's Complement Adder/Subtractor with Overflow Detection Logic for Quantum Computer-Based Systems 259
S. Manjula Gandhi, J. Devishree and S. Sathish Mohan

An Algorithm for Constructing Graceful Tree from an Arbitrary Tree 267
G. Sethuraman and P. Ragukumar

Characterization of Semi-open Subcomplexes in Abstract Cellular Complex 275
N. Vijaya and G. Sai Sundara Krishnan

Fluid Queue Driven by an *M*/*M*/1 Queue Subject to Catastrophes 285
K. V. Vijayashree and A. Anjuka

Fuzzy VEISV Epidemic Propagation Modeling for Network Worm Attack 293
Muthukrishnan Senthil Kumar and C. Veeramani

Hexagonal Prusa Grammar Model for Context-Free Hexagonal Picture Languages 305
T. Kamaraj and D. G. Thomas

Iso-Triangular Pictures and Recognizability 313
V. Devi Rajaselvi, T. Kalyani and D. G. Thomas

Job Block Scheduling with Dual Criteria and Sequence-Dependent Setup Time Involving Transportation Times 321
Deepak Gupta, Kewal Krishan Nailwal and Sameer Sharma

Modular Chromatic Number of $C_m \square C_n$ 331
N. Paramaguru and R. Sampathkumar

Network Analysis of Biserial Queues Linked with a Flowshop Scheduling System 339
Seema Sharma, Deepak Gupta and Sameer Sharma

Solving Multi-Objective Linear Fractional Programming Problem—First Order Taylor's Series Approximation Approach 347
C. Veeramani and M. Sumathi

Voronoi Diagram-Based Geometric Approach for Social Network Analysis 359
Subu Surendran, D. Chitraprasad and M. R. Kaimal

Part V Short Papers

Comparative Analysis of Discretization Methods for Gene Selection of Breast Cancer Gene Expression Data 373
E. N. Sathishkumar, K. Thangavel and A. Nishama

Modified Soft Rough set for Multiclass Classification 379
S. Senthilkumar, H. Hannah Inbarani and S. Udhayakumar

A Novel Genetic Algorithm for Effective Job Scheduling in Grid Environment. 385
P. Deepan Babu and T. Amudha

A Multifactor Biometric Authentication for the Cloud 395
Shabana Ziyad and A. Kannammal

Accelerating the Performance of DES on GPU and a Visualization Tool in Microsoft Excel Spreadsheet 405
Pinchu Prabha, O. K. Sikha, M. Suchithra and K. P. Soman

About the Editors 413

Author Index 415

Part I
Keynote Address

The Robber Strikes Back

Anthony Bonato, Stephen Finbow, Przemysław Gordinowicz, Ali Haidar, William B. Kinnersley, Dieter Mitsche, Paweł Prałat and Ladislav Stacho

Abstract We consider the new game of Cops and Attacking Robbers, which is identical to the usual Cops and Robbers game except that if the robber moves to a vertex containing a single cop, then that cop is removed from the game. We study the minimum number of cops needed to capture a robber on a graph G, written $\mathrm{cc}(G)$. We give bounds on $\mathrm{cc}(G)$ in terms of the cop number of G in the classes of bipartite graphs and diameter two, $K_{1,m}$-free graphs.

Keywords Cops and robbers · Cop number · Bipartite graphs · Claw-free graphs

1 Introduction

Cops and Robbers is a vertex-pursuit game played on graphs, which has been the focus of much recent attention. Throughout, we only consider finite, connected, and simple undirected graphs. There are two players consisting of a set of *cops* and

A. Bonato (✉) · W. B. Kinnersley · P. Prałat
Ryerson University, Toronto, Canada
e-mail: abonato@ryerson.ca

S. Finbow
St. Francis Xavier University, Antigonish, Canada

P. Gordinowicz
Technical University of Lodz, Lodz, Poland

A. Haidar
Carleton University, Ottawa, Canada

D. Mitsche
University of Nice Sophia-Antipolis, Nice, France

L. Stacho
Simon Fraser University, Burnaby, Canada

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_1, © Springer India 2014

a single *robber*. The game is played over a sequence of discrete time steps or *rounds*, with the cops going first in the first round and then playing on alternate time steps. The cops and robber occupy vertices, and more than one cop may occupy a vertex. When a player is ready to move in a round, they may move to a neighbouring vertex or *pass* by remaining on their own vertex. Observe that any subset of cops may move in a given round. The cops win if after some finite number of rounds, one of them can occupy the same vertex as the robber. This is called a *capture*. The robber wins if he can avoid capture indefinitely. A *winning strategy for the cops* is a set of rules that if followed result in a win for the cops, and a *winning strategy for the robber* is defined analogously.

If we place a cop at each vertex, then the cops are guaranteed to win. Therefore, the minimum number of cops required to win in a graph G is a well-defined positive integer, named the *cop number* of the graph G. We write $c(G)$ for the cop number of a graph G. For example, the Petersen graph has cop number 3. Nowakowski and Winkler [14], and independently Quilliot [19], considered the game with one cop only; the introduction of the cop number came in [1]. Many papers have now been written on cop number since these three early works; see the book [8] for additional references and background on the cop number. See also the surveys [2, 4, 5].

Many variants of Cops and Robbers have been studied. For example, we may allow a cop to capture the robber from a distance k, where k is a non-negative integer [7], play on edges [12], allow one or both players to move with different speeds or teleport, or allow the robber to be invisible. See Chap. 8 of [8] for a non-comprehensive survey of variants of Cops and Robber.

We consider a new variant of the game of Cops and Robbers, where the robber is able to essentially strike back against the cops. We say that the robber *attacks* a cop if he chooses to move to a vertex on which a cop is present and eliminates her from the game. In the game of *Cops and Attacking Robbers*, the robber may attack a cop, but cannot start the game by moving to a vertex occupied by a cop; all other rules of the game are the same as in the classic Cops and Robbers. We note that if two cops are on a vertex u and the robber moves to u, then only one cop on u is eliminated; the remaining cop then captures the robber, and the game ends. We write $\mathrm{cc}(G)$ for the minimum number of cops needed to capture the robber. Note that $\mathrm{cc}(G)$ is the analogue of the cop number in the game of Cops and Attacking Robbers; our choice of notation will be made more transparent once we state Theorem 1. We refer to $\mathrm{cc}(G)$ as the *cc-number of* G. Since placing a cop on each vertex of G results in a win for the cops, the parameter $\mathrm{cc}(G)$ is well defined.

To illustrate that $\mathrm{cc}(G)$ can take different values from the cop number, consider that for the cycle C_n with n vertices, we have the following equalities (which are easily verified):

$$\mathrm{cc}(C_n) = \begin{cases} 1 & \text{if } n = 3, \\ 2 & \text{if } 4 \le n \le 6, \\ 3 & \text{else.} \end{cases}$$

We outline some basic results and bounds for the cc-number in Sect. 2. We consider bounds on cc(G) in terms of $c(G)$ in Sect. 3. In Sect. 4, we give the bound of cc(G) $\leq c(G) + 2$ in the case that G is bipartite; see Theorem 9. In the final section, we supply in Theorem 10 an upper bound for cc(G) for $K_{1,m}$-free, diameter two graphs.

For background on graph theory, see [20]. For a vertex u, we let $N(u)$ denote the neighbour set of u, and let $N[u] \ = N(u) \cup \{u\}$ denote the closed neighbour set of u. The set of vertices of distance 2 to u is denoted by $N_2(u)$. We denote by (G) the minimum degree in G. In a graph G, a set S of vertices is a *dominating set* if every vertex not in S has a neighbour in S. The *domination number* of G, written $\gamma(G)$, is the minimum cardinality of a dominating set. The *girth* of a graph is the length of the shortest cycle contained in that graph and is ∞ if the graph contains no cycles.

2 Basic Results

In this section, we collect together some basic results for the cc-number. As the proofs are either elementary or minor variations of the analogous proofs for the cop number, they are omitted. The first result on the game of Cops and Attacking Robbers is the following theorem; note that the second inequality naturally inspires the notation cc(G). We use the notation $\bar{c}(G)$ for the edge cop number, which is a variant where the cops and robber move on edges; see [12].

Theorem 1 *If G is a graph, then*

$$c(G) \leq \mathrm{cc}(G) \leq \min\{2c(G), 2\bar{c}(G), \gamma(G)\}.$$

The following theorem is foundational in the theory of the cop number.

Theorem 2 [1] *If G has girth at least* 5, *then*

$$c(G) \geq \delta(G).$$

The following theorem extends this result to the cc-number.

Theorem 3 *If G has girth at least* 5, *then*

$$\mathrm{cc}(G) \geq \delta(G) + 1.$$

Isometric paths play an important role in several key theorems in the game of Cops and Robbers, such as the cop number of planar graphs (see Chap. 4 of [8]). We call a path P in a graph G *isometric* if the shortest distance between any two vertices is equal in the graph induced by P and in G. For a fixed integer $k \geq 1$, an induced subgraph H of G is *k-guardable* if, after finitely many moves, k cops can move only in the vertices of H in such a way that if the robber moves into H at round t, then he will be captured at round $t + 1$ by a cop in H. For example, a clique in a graph is 1-guardable.

Aigner and Fromme [1] proved the following result.

Theorem 4 [1] *An isometric path is* 1*-guardable.*

We have an analogue of Theorem 4 for the cc-number.

Theorem 5 *An isometric path is* 2*-guardable in the game of Cops and Attacking Robbers, but need not be* 1*-guardable.*

See Fig. 1 for an example where the robber can freely move onto an isometric path without being captured by a sole cop.

A graph G is called *planar* if it can be embedded in a plane without two of its edges crossing. It was shown first in [1] that planar graphs require at most three cops to catch the robber; see [8] for an alternative proof of this fact. Given the results above, we may conjecture that the cc-number of a planar graph is at most 4 or even 5, but either bound remains unproven.

Outerplanar graphs are those that can be embedded in the plane without crossings in such a way that all of the vertices belong to the unbounded face of the embedding. Clarke proved the following theorem in her doctoral thesis.

Theorem 6 [11] *If G is outerplanar, then* $c(G) \leq 2$.

The counterpart to Theorem 6 is the following.

Theorem 7 *If G is outerplanar, then* $\mathrm{cc}(G) \leq 3$.

Meyniel's conjecture—first communicated by Frankl [13]—is one of the most important open problems surrounding the game of Cops and Robbers. The conjecture states that $c(n) = O(\sqrt{n})$, where $c(n)$ is the maximum of $c(G)$ over all n-vertex, connected graphs. Cops and Robbers has been studied extensively for random graphs (see for example, [3, 9, 15, 16]), partly owing to a search for counterexamples to Meyniel's conjecture. However, it was recently shown that Meyniel's conjecture holds asymptotically almost surely (that is, with probability tending to 1 as the number of vertices tends to infinity) for both binomial random graphs $G(n, p)$ [17] as well as random d-regular graphs [18].

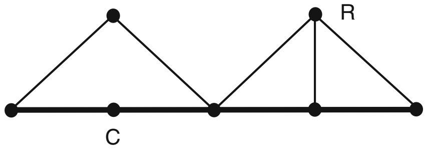


Fig. 1 One cop cannot guard the isometric path (*depicted in bold*). We assume that the robber has just arrived at their vertex, and it is the cop's turn to move

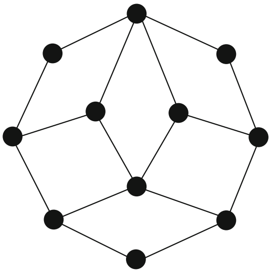


Fig. 2 A graph G with $c(G) = \mathrm{cc}(G) = 2$ and $\gamma(G) = 3$

In [9], it was shown that for dense random graphs, where $p = n^{-o(1)}$ and $p < 1 - \epsilon$ for some $\epsilon > 0$, asymptotically almost surely we have that

$$c(G(n,p)) = (1 + o(1))\gamma(G(n,p)) = (1 + o(1)) \log_{1/(1-p)} n. \qquad (1)$$

Note that (1) implies that $c(G(n, p)) = (1 + o(1))\mathrm{cc}(G(n, p))$ for the stated range of p; in particular, applying (1) to the $p = 1/2$ case (which corresponds to the uniform probability space of all labelled graphs on n vertices), we have that for every $\epsilon > 0$, almost all graphs satisfy $\mathrm{cc}(G)/c(G) \in [1, 1 + \epsilon]$. Unfortunately, the asymptotic value of the cop number is not known for sparser graphs. However, it may be provable that $c(G(n,p)) = (1 + o(1))\mathrm{cc}(G(n,p))$ for sparse graphs, without finding an asymptotic value.

We finish the section by noting that graphs with $\mathrm{cc}(G) = 1$ are precisely those with a universal vertex. However, characterizing those graphs G with $\mathrm{cc}(G) = 2$ is an open problem. Graphs with $\mathrm{cc}(G) = 2$ include cop-win graphs without universal vertices and graphs which are not cop win but have domination number 2. Before the reader conjectures this gives a characterization, note that the graph in Fig. 2 with cc-number equalling 2 is in neither class.

3 How Large Can the cc-Number Be?

One of the main unanswered questions on the game of Cops and Attacking Robbers is how large the cc-number can be relative to the cop number. Many of the results from the last section might lead one to (mistakenly) conjecture that

$$\mathrm{cc}(G) \leq c(G) + 1$$

for all graphs, and this was the thinking of the authors and others for some time. We provide a counterexample below.

By Theorem 1, we know that $\mathrm{cc}(G)$ is bounded above by $2c(G)$. For example, this is a tight bound for a path of length at least 3. However, we do not know an improved bound which applies to general graphs, nor do we possess graphs G with $c(G) > 2$ whose cc-number equals $2c(G)$. In this section, we outline one approach which may ultimately yield such examples. Improved bounds for several graph classes are outlined in the next two sections.

Our construction utilizes line graphs of hypergraphs. For a positive integer k, a *k-uniform hypergraph* has every hyperedge of cardinality k. A hypergraph is *linear* if any two hyperedges intersect in at most one vertex. The *line graph* of a hypergraph H, written as $L(H)$, has one vertex for each hyperedge of H, with two vertices adjacent if the corresponding hyperedges intersect.

Lemma 8 *Let H be a linear k-uniform hypergraph with minimum degree at least* 3 *and girth at least* 5*. If $L(H)$ has domination number at least* $2k$*, then* $\mathrm{cc}(L(H)) > 2k$*.*

Proof Suppose there are at most $2k - 1$ cops. Since the domination number of $L(H)$ is at least $2k$, the robber can choose an initial position that lets him survive the cops' first move. To show that $2k - 1$ cops cannot catch the robber in the game of Cops and Attacking Robbers on $L(H)$, suppose otherwise, and consider the state of the game on the robber's final turn (that is, just before he is to be captured). Let v be the robber's current vertex, E_v the corresponding edge of H, and $w_1, w_2, \ldots, w_k$ the elements of E_v. The neighbours of v in $L(H)$ are precisely those vertices corresponding to the edges of H that intersect E_v; denote by S_{w_i} the set of vertices (other than v) corresponding to edges containing w_{i}. Each S_{w_i} is a clique; moreover, since H has minimum degree at least 3, each contains at least two vertices. By hypotheses for H, it follows that the S_{w_i} are disjoint and that no vertex outside S_{w_i} dominates more than one vertex inside. Finally, since H has girth at least 5, no vertex in G dominates vertices in two different S_{w_i} (that is, the neighbourhoods $N[S_{w_i}]$ only have v in common).

Consider the cops' current positions. The cops must dominate all of $N[v]$, since otherwise the robber would be able to survive for one more round (by moving to an undominated vertex). Since the $N[S_{w_i}]$ only have v in common, for some j, we have at most one cop in $N\,[S_{w_j}]$. If in fact there are no cops in $N[S_{w_j}]$, then no vertices of S_{w_j} are dominated, a contradiction. Thus, S_{w_j} contains exactly one cop. Since each vertex outside S_{w_j} dominates at most one vertex inside and S_{w_j} contains at least two vertices, the cop must actually stand within S_{w_j}. However, since she is the only cop within $N[S_{w_j}]$, the robber may attack the cop without leaving himself open to capture on the next turn. Thus, the robber always has a means to avoid capture on the cops' next turn. Hence, at least $2k$ cops are needed to capture the robber, as claimed. □

We aim to find, for all k, graphs G such that $c(G) = k$ and $\mathrm{cc}(G) = 2k$. This, however, remains open for all $k \geq 3$.

As an application of the lemma, take H to be the Petersen graph. It is easily verified that $c(L(H)) = 2$; see also [12]. Lemma 8 with $k = 2$ shows that $\mathrm{cc}(L(H)) \geq 4$; hence, Theorem 1 then implies that $\mathrm{cc}(L(H)) = 4$. See Fig. 3 for a drawing of the line graph of the Petersen graph.

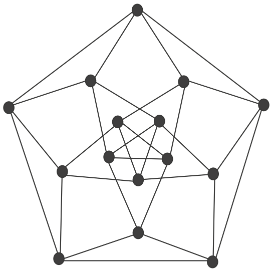


Fig. 3 The line graph of the Petersen graph

4 Bipartite Graphs

For bipartite graphs, we derive the following upper bound.

Theorem 9 *For every connected bipartite graph G, we have that* $\mathrm{cc}(G) \leq c(G) + 2$.

Proof Fix a connected bipartite graph G. Let $k = c(G)$; we give a strategy for $k + 2$ cops to win the game of Cops and Attacking Robbers on G. Label the cops $C_1, C_2, \ldots, C_k, C_1^*, C_2^*$. Intuitively, cops C_1, C_2, …, C_k attempt to follow a winning strategy for the ordinary Cops and Robber game on G; since they must avoid being killed by the robber, they may not be able to follow this strategy exactly, but can follow it "closely enough". Cops C_1^* and C_2^* play a different role: They occupy a common vertex throughout the game, and in each round, they simply move closer to the robber. This has the effect of eventually forcing the robber to move on every turn. (Since the cops move together, the robber cannot safely attack either one.) Further, when the robber passes, the cops $C_1, C_2, \ldots, C_k$ pass. Therefore, we may suppose throughout that the robber moves to a new vertex on each turn.

It remains to formally specify the movements of $C_1, C_2, \ldots, C_k$. To each cop C_i, we associate a *shadow* S_i. Throughout the game, the shadows follow a winning strategy for the ordinary game on G. Let $C_i^{(t)}, S_i^{(t)}$, and $R^{(t)}$ denote the positions of C_i, S_i, and the robber, respectively, at the end of round t. We maintain the following invariants for $1 \leq i \leq k$ and all t:

1. $S_i^{(t)} \in N\left[C_i^{(t)}\right]$ (that is, each cop remains on or adjacent to her shadow).
2. if $C_i^{(t+1)} \neq S_i^{(t+1)}$, then $S_i^{(t+1)}$ and $R^{(t)}$ belong to different partite sets of G.
3. $C_i^{(t+1)}$ is not adjacent to $R^{(t)}$ (that is, the robber never has the opportunity to attack any cop).

On round $t + 1$, each cop C_i moves as follows:

(a) If $C_i^{(t)} \neq S_i^{(t)}$, then C_i moves to $S_i^{(t)}$.
(b) If $C_i^{(t)} = S_i^{(t)}$, and $S_i^{(t+1)}$ is not adjacent to $R^{(t)}$, then C_i moves to $S_i^{(t+1)}$.
(c) Otherwise, C_i remains at her current vertex.

By invariant (1), this is clearly a legal strategy.

We claim that all three invariants are maintained. Invariant (1) is straightforward to verify. For invariant (2), first suppose that $C_i^{(t)} = S_i^{(t)}$, but $C_i^{(t+1)} \neq S_i^{(t+1)}$. By the cops' strategy, this can happen only when $S_i^{(t+1)}$ is adjacent to $R^{(t)}$, in which case, the shadow and robber belong to different partite sets, as desired. Now, suppose that $C_i^{(t)} \neq S_i^{(t)}$ and $C_i^{(t+1)} \neq S_i^{(t+1)}$. By the cops' strategy, we have $C_i^{(t+1)} = S_i^{(t)}$. It follows that $C_i^{(t+1)} \neq C_i^{(t)}$, $S_i^{(t+1)} \neq S_i^{(t)}$, and $R^{(t-1)} \neq R^{(t)}$. Thus, if $S_i^{(t)}$ and $R^{(t-1)}$ belong to different partite sets, then so must $S_i^{(t+1)}$ and $R^{(t)}$; that is, the invariant is maintained. For invariant (3), if $S_i^{(t+1)}$ is adjacent to $R^{(t)}$, then we may suppose that $S_i^{(t+1)} \neq S_i^{(t)}$, since otherwise the shadow would have captured the robber in round $t + 1$. By the cops' strategy, we now have that $C_i^{(t+1)} \neq S_i^{(t+1)}$. But now, the cop and her shadow are in different partite sets by invariant (1), and the shadow and robber are in different partite sets by invariant (2), so the cop and robber are in the same partite set, contradicting adjacency of the cop and the robber.

Since the shadows follow a winning strategy, eventually some shadow S_i captures the robber; that is, for some t, we have that either $S_i^{(t)} = R^{(t)}$ or $S_i^{(t+1)} = R^{(t)}$. In the former case, invariant (3) implies that $C_i^{(t)} \neq S_i^{(t)}$ and invariant (1) implies that C_i captures the robber in round $t + 1$. Now, consider the case when $S_i^{(t+1)} = R^{(t)}$. By invariant (2), since $S_i^{(t+1)}$ is not adjacent to $R^{(t)}$, we in fact have that $C_i^{(t+1)} = S_i^{(t+1)} = R^{(t)}$ so the cops have won. □

5 $K_{1,m}$-Free, Diameter 2 Graphs

We provide one more result giving an upper bound on the cc-number for a set of graph classes.

Theorem 10 *Let G be a $K_{1,m}$-free, diameter 2 graph, where $m \geq 3$. Then,*

$$\mathrm{cc}(G) \leq c(G) + 2m - 2.$$

When $m = 3$, Theorem 10 applies to claw-free graphs; see [10] for a characterization of these graphs. The cop number of diameter 2 graphs was studied in [6].

Proof of Theorem 10 A cop C is *backup* to a cop C' if C is in $N[C']$, note that a cop with a backup cannot be attacked without the robber being captured in the next round.

Now, let $c(G) = r$, and consider $c(G)$ cops labelled $C_1, C_2, \ldots, C_r$. We refer to these r-many cops as *squad 1*. Label an additional $2m - 2$ cops as $\widehat{C_{i,1}}$ and $\widehat{C_{i,2}}$, where $1 \leq i \leq m - 1$; these cops form *squad 2*. The intuition behind the proof is that the cops in squad 2 act as backup for those in squad 1, who play their usual strategy on G. Further, the cops $\widehat{C_{i,j}}$ are positioned in such a way that the cops C_k need only restrict their movements to the second neighbourhood of some fixed vertex.

More explicitly, fix a vertex x of G. Move squad 2 so that they are contained in $N[x]$. Next, position each of the cops $\widehat{C_{i,1}}$ on x. Hence, R must remain in $N_2(x)$ or he will lose in the next round (in particular, no squad 2 cop is ever attacked). Throughout the game, we will always maintain the property that there are $m - 1$ cops on x.

We note that the squad 2 cops in $N(x)$ can move there essentially as if that subgraph were a clique, and in addition, preserve the property that $m - 1$ cops remain on x. To see this, if $\widehat{C_{i,2}}$ were on $y \in N(x)$ and the cops would like to move to $z \in N(x)$, then move $\widehat{C_{i,2}}$ to x, and move some squad 2 cop from x to z. In particular, a cop from squad 2 can arrange things so that she is adjacent to a cop in squad 1 after at most one move. We refer to this movement of the squad two cops as a *hop*, as the cops appear to jump from one vertex of $N(x)$ to another (although what is really happening is that the cops are cycling through x). Note that hops maintain $m - 1$ cops on x.

We now describe a strategy $\mathcal{S}$ for the cops, and then show that it is winning. The cops in squad 1 play exactly as in the usual game of Cops and Robbers; note that the squad 1 cops may leave $N_2(x)$ depending on their strategy, but R will never leave $N_2(x)$. The squad 2 cops play as follows. Squad 2 cops do not move unless the following occurs: a squad 1 cop C_k moves to a neighbour of R, and C_k has no backup from a squad 1 cop. In that case, some squad 2 cop $\widehat{C_{i,j}}$ hops to a vertex of $N(x)$ which is adjacent to C_k. There are a sufficient number of squad 2 cops to ensure this property, since if m (or more) squad 1 cops move to neighbours of R, then some of these cops must be adjacent to each other as G is $K_{1,m}$-free (in particular, the cops in $N(R)$ play the role of backups to each other).

Hence, the squad 1 cops may apply their winning strategy in the usual game and ensure that whenever they move to a neighbour of R, some squad 2 cop serves as backup. In particular, R will never attack a squad 1 cop for the duration of the game. Thus, $\mathcal{S}$ is a winning strategy in the game of Cops and Attacking Robbers. □

Acknowledgments The authors were supported by grants from NSERC and Ryerson University.

References

1. M. Aigner, M. Fromme, A game of cops and robbers, *Discrete Applied Mathematics* **8** (1984) 1–12.
2. W. Baird, A. Bonato, Meyniel's conjecture on the cop number: a survey, *Journal of Combinatorics* **3** (2012) 225–238.
3. B. Bollobas, G. Kun, I. Leader, Cops and robbers in a random graph, *Journal of Combinatorial Theory Series B* **103** (2013) 226–236.

4. A. Bonato, WHAT IS… Cop Number? *Notices of the American Mathematical Society* ***59*** *(2012) 1100–1101.*
5. A. Bonato, Catch me if you can: Cops and Robbers on graphs, In: *Proceedings of the 6th International Conference on Mathematical and Computational Models* (ICMCM'11), *2011.*
6. A. Bonato, A. Burgess, Cops and Robbers on graphs based on designs, *Journal of Combinatorial Designs* **21** (2013) 404–418.
7. A. Bonato, E. Chiniforooshan, P. Pralat, Cops and Robbers from a distance, *Theoretical Computer Science* **411** (2010) 3834–3844.
8. A. Bonato, R.J. Nowakowski, *The Game of Cops and Robbers on Graphs,* American Mathematical Society, Providence, Rhode Island, 2011.
9. A. Bonato, P. Pralat, C. Wang, Network security in models of complex networks, *Internet Mathematics* **4** (2009) 419–436.
10. M. Chudnovsky, P. Seymour, Clawfree Graphs IV - Decomposition theorem, *Journal of Combinatorial Theory. Ser B* **98** (2008) 839–938.
11. N.E. Clarke, *Constrained Cops and Robber,* Ph.D. Thesis, Dalhousie University, 2002.
12. A. Dudek, P. Gordinowicz, P. Pralat, Cops and Robbers playing on edges, preprint, 2013.
13. P. Frankl, Cops and robbers in graphs with large girth and Cayley graphs, *Discrete Applied Mathematics* **17** (1987) 301–305.
14. R.J. Nowakowski, P. Winkler, Vertex-to-vertex pursuit in a graph, *Discrete Mathematics* **43** (1983) 235–239.
15. T. Luczak, P. Pralat, Chasing robbers on random graphs: zigzag theorem, *Random Structures and Algorithms* **37** (2010) 516–524.
16. P. Pralat, When does a random graph have constant cop number?, *Australasian Journal of Combinatorics* **46** (2010) 285–296.
17. P. Pralat, N.C. Wormald, Meyniel's conjecture holds for random graphs, preprint, 2013.
18. P. Pralat, N.C. Wormald, Meyniel's conjecture holds for random *d*-regular graphs, preprint, 2013.
19. A. Quilliot, Jeux et pointes fixes sur les graphes, These de 3eme cycle, Universite de Paris VI, 1978, 131–145.
20. D.B. West, *Introduction to Graph Theory,* 2nd edition, Prentice Hall, 2001.

Some Applications of Collective Learning

Balaraman Ravindran

Abstract Much of the real-world data have complex dependencies between the individual tuples. For example, the chance that a patient has a particular disease depends on the prevalence of the disease in the immediate neighborhood. One approach to handling such linked data is “collective learning.” In collective learning, one deals with a set of data points taken at a time. The dependencies between the data points are modeled as a graph, with the nodes representing the tuples and the edges between them representing the influence of the tuples on one another. A variety of domains lend themselves naturally to such graph-based modeling. There have been a variety of collective learning and inferencing approaches that have been proposed in the literature. In this talk, I will give a brief introduction to collective learning and describe two applications.

keywords Sentiment analysis · Functional site prediction · Interaction network · Label misclassification

The first of these is a sentiment analysis task. Sentiment analysis is the task of identifying the sentiment expressed in the given piece of text about the target entity under discussion. In this work, we look at the problem of analyzing sentiments at different granularities. For example, we want to analyze sentiment about a movie as whole as well as about the acting and directing. Models built for such multigrain sentiment analysis assume fully labeled corpus at fine-grained level or coarse-grained level or both. Huge amount of online reviews are not fully labeled at any of the levels, but are partially labeled at both the levels. We propose a multigrain collective classification framework to not only exploit the information available at all the levels but also use intra dependencies at each level and interdependencies between the levels. We demonstrate empirically that the proposed framework enables better performance at both the levels compared to

B. Ravindran (✉)
Computer Science and Engineering, Indian Institute of Technology Madras,
Chennai 600036, India
e-mail: ravi@cse.iitm.ac.in

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_2, © Springer India 2014

baseline approaches. Part of this work was reported in ECAI 2010, and it is a joint work with S. Shivashankar and Shamshu Dharwez.

The second task is that of functional site prediction in proteins. Functional site prediction is an important problem in the structural genomics era where we have a large number of experimentally determined protein structures with unknown function. The functional sites provide useful insights into protein function. In this paper, we propose a method for prediction of functional residues in a given protein from its three-dimensional (3D) structure. Our method exploits correlation between labels of interacting residues to obtain significant performance improvements over the existing methods on the benchmark dataset. We represent each protein as a weighted undirected residue interaction network, where spatially proximal residues in terms of their van der Waal's radii are connected by an edge. The edge weight captures correlation between the labels of interacting residues. The correlation is estimated based on the features of interacting residues. We then obtain a label assignment by minimizing combined cost of residue-wise label misclassification and violation of label correlation constraints. We solve this problem in two stages, where the first stage minimizes residue-wise label misclassification cost followed by an iterative collective inference scheme that adjusts the labels predicted in the first stage so as to minimize the correlation constraint violations. Our approach significantly outperforms state-of-the-art methods on standard benchmark dataset. This work was reported in ACM BCB 2012, and it is a joint work with Ashish V. Tendulkar, Saradindu Kar, and Deepak Vijayakeerthi.

Subconscious Social Computational Intelligence

M. Graña

Abstract The success of social network Web services mediating social interactions, as well as the increasing observation capabilities of human interactions in real life, has prompted the emergence of new computational paradigms, namely social computing, computational social science, and social intelligence. Subconscious social intelligence appears when the social network service is able to provide solutions, generated by a hidden intelligent layer, to problems posed by the social player. This paper discusses some features of subconscious social intelligence and ensuing challenges for machine learning systems implementing the hidden intelligent layer.

Keywords Social computing · Social intelligence · Subconscious reasoning · Learning systems

1 Introduction

This paper discusses the requirements for machine learning systems contributing to the development of a nascent computational field, which can be identified by the name of subconscious social computing. Reviewing the related fields of social computing and computational social science will help to clarify the subtle distinctive features of this new class of systems. We describe the general scheme of subconscious social computing highlighting its contrast with the previous ones, including the description of one instance, the EU-financed SandS project [1, 3, 4, 6]. Then, we identify requirements posed by this kind of systems on learning subsystems implementing the subconscious intelligent layer, discussing how computational intelligence approaches can cope with them.

M. Graña (✉)
Grupo de Inteligencia Computacional, UPV/EHU, Erandio, Spain
e-mail: manuel.grana@ehu.es
URL: www.ehu.es/ccwintco

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_3, © Springer India 2014

2 Social Computing Paradigms

Computational social science [2] aims to understand the dynamics of social systems form of data that can be extracted from all existing sources of human behavior observation, ranging from surveillance cameras, mobile identification tags to social Web services or electronic commercial transactions. From computational social science point of view, the social players are subjects of observation and experimentation, searching for answers to questions such as:

- which interaction pattern leads to economic success?
- how social interaction influences contagious sickness diffusion?
- what is the best way to promote a product?
- what social hints can be useful to predict the fate of a stock asset?

To this end, computational social science deals with intelligent and efficient hardware/software systems able to process huge amounts of data coming from all kinds of observational sources within some real-time constraints. Big networking data are subject to statistical and data mining analysis, providing answers to the institutional or corporate costumer.

Social computing [5, 8] concerns the development of software for the enhanced interaction between social players and to develop simulation scenarios to forecast the effects of policies and forces, such as technological innovation, on societies. Examples of these systems are entertainment/therapeutic social games involving autonomous intelligent agents, negotiation, recommender and reputation systems, security applications for the detection of criminal social activities, and artificial societies of agents designed to provide adaptation to changing environments (i.e., traffic) through competition, platforms for scientific collaboration offering information about the current state of the research community and research effort planning. Social computing is developing into a productive model where rewarding mechanisms are required to control the desired output of the system [7].

Social intelligence is the emergence of problem-solving behavior out of social interactions from the point of view of the social player. In other words, the social player expects to obtain solutions to his/her problems from the pool of intelligence available from a social network and the computational resources that may be at work behind the social service. We may further distinguish between conscious and subconscious intelligent computing. In the former, social players contribute information and the intelligence to create/discover solutions. In the latter, an underlying intelligent layer is able to provide innovative solutions to old and new problems, following an autonomous process that is not directly controlled by the social players. Figures 1 and 2 illustrate the differences between paradigms. The social interaction layer in both figures includes all means of sharing information between users, but does not contemplate any information transformation.

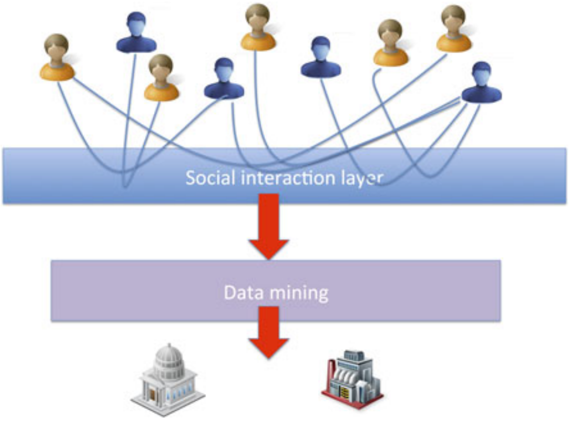


Fig. 1 Social computing and computational social science paradigm

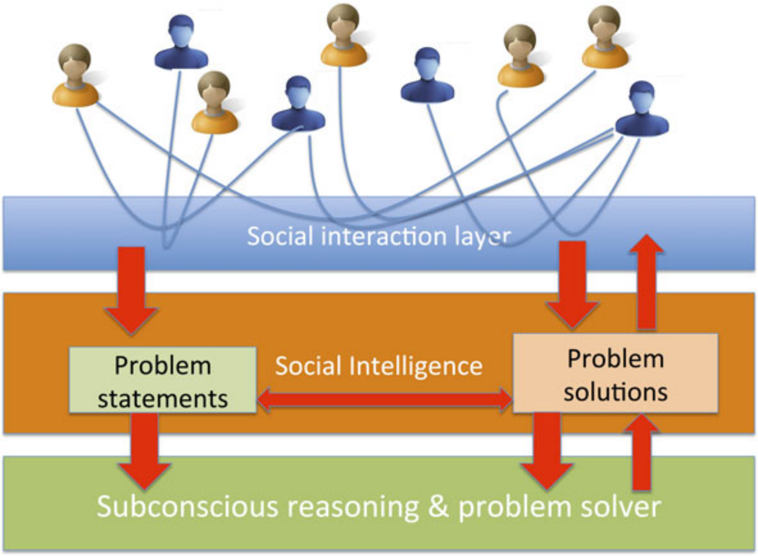


Fig. 2 Subconscious social intelligence paradigm

3 Subconscious Social Intelligence

In the social computing and computational social science paradigms illustrated in Fig. 1, there is an underlying computational layer that performs data mining over observations of the social interactions. The social player is unaware of it and does not directly benefit from it. This lack of benefit motivates research in rewarding mechanisms [7]. The results of these computations are delivered to a third party, either government institutions or industry. The productive model of social computing relies on the technological prowess of this data mining layer that provides the benefit from the investment in the social interaction layer.

On the other hand, in the subconscious social intelligence paradigm illustrated by Fig. 2, the social intelligence layer below the social interaction layer is dedicated to provide solutions to problem statements posed by the social players. To that end, repositories of problem statements and problem solutions are maintained, along with a mapping between them. The upward and downward red arrows model the flow of problem statements and solutions. Problem statements posed by social players flow downward to the social intelligence, matching problem solutions are searched in the repository (horizontal red double-headed arrow), if one is found, it flows upward to the social interaction layer to be retrieved by the interested social player. If there is no solution matching the statement, the statement is percolated further down to the subconscious reasoning and problem solver layer that works to produce a solution that will be pushed upward to the social intelligence layer solutions repository, and the user at the social interaction layer. The subconscious reasoning and problem solver is trained on problem solutions that percolate from the social intelligence layer. The product of the social intelligence goes directly back to the user, and there is no beneficiary institution, either company of government. The aim of the system is empowering the social player to solve his/her real-life problems, maybe against the pressures of some institution, or within its. As a corollary, social players do not need to be rewarded externally to use/contribute the system.

4 Requirements for Learning Systems

The requirements for learning systems meeting the SandS networked intelligence and the general subconscious reasoning and problem solver of Fig. 2 are as follows:

- Quick learning times that allow for quick adaptation to changing environments and supporting the effects of scale that potentially big social communities will introduce. Social network services can experience dramatic rises in user involvement and subsequent computational load. Moreover, changes in problem specification may involve addition/removal of variables with ensuing retraining processes.
- Flexibility to cope with diverse data representations and desired outputs. The desired responses may be categorical and continuous, involving both classification and regression, even in the same problem-solving process.
- Robust performance when dealing with multidimensional heterogenous output. Most machine learning approaches have serious degradation when the desired output is multivariable, and even worse when it is composed of diversely typed variables.

- Minimal uncertainty: In the development of subconscious social intelligence, we want to perform one-shot training with minimal uncertainty about the achieved performance. Machine learning papers often report average or peak results of extensive computational experiments. These results do not provide a performance guarantee for a specific instance of the learning process, nothing prevents it to be catastrophically stupid.
- Robust incremental learning to process incoming batches of user feedback driving the adaptation process. Incremental learning, on the fly adaptation, is not an optional feature in this setting. Social systems and the needs of the social players are continuously evolving. Training systems with a sample of data are meaningless after some period of time.
- Easy implementation/learning of forward and backward mappings, the former to provide solutions and the latter to translate the user feedback into error measures driving learning processes. Social players want to be able to understand why a solution works, which is the chain of reasoning that produces this improved solutions, and to have some control on the responses of the system to required adaptive changes.
- Hybridization of diverse computational paradigms to allow the composition of selection/classification/regression modules to cope with the complex landscape of user problem statement. It is not likely that a single learning paradigm will be able to cope with all kinds of social player requests and needs. Many kinds of intelligence may need to be called upon to provide answers at diverse levels.

5 The SandS Project

The EU-funded SandS project (http://www.sands-project.eu) aims to build an instance of the subconscious social intelligence in the domotic domain. SandS social players are users of household appliances that exchange information about them in the form of “recipes” of use. The software structure under development in the project follows the pattern of Fig. 2. The SandS social network has a repository of household tasks that have been posed by the users and a repository of appliance recipes, which are related by a map between (to and from) tasks and recipes. This map needs not to be one-to-one. User queries interrogate the database of known/solved household tasks. If the queried task is already known, then the corresponding recipe can be send to the user appliance. After recipe execution, the user can express its satisfaction with the results. When the queried task is unknown and unsolved, it is forwarded to the underlying SandS networked intelligence to produce a new recipe by an intelligent system reasoning able to learn and predict new recipes maximizing user satisfaction. The source of recipes filling the repository is, therefore, twofold. On the one hand, engaged user and/or appliance manufacturing companies consciously provide new recipes. On the other hand, the underlying networked intelligence is the subconscious generator of new solutions.

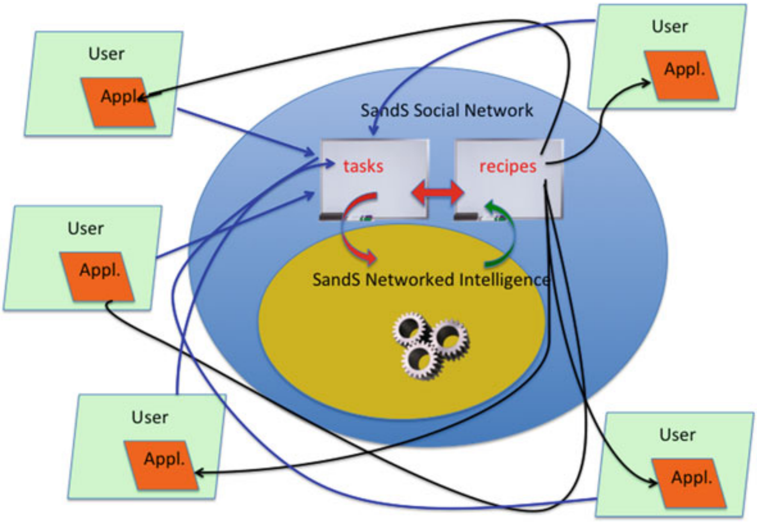


Fig. 3 Social and smart system prototypical architecture

More specifically, Fig. 3 shows an intuitive representation of the architecture and interactions between the system elements. The SandS social network mediates the interaction between populations of users, each owning a set of appliances. The SandS social network has a repository of tasks that have been posed by the eahoukers and a repository of recipes for the use of appliances. These two repositories are related by a map between (to and from) task and recipes. This map needs not to be one-to-one. Blue arrows correspond to the path followed by the eahouker queries, which are used to interrogate the database of known/solved tasks. If the task is already known, then the corresponding recipe can be returned to the eahouker appliance (black arrows). The eahouker can express its satisfaction with the results (blue arrows). When the queried task is unknown and unsolved, the social network will request a solution from the SandS networked intelligence that will consist in a new recipe deduced from the past knowledge stored in the recipe repository. This new solution will be generated by intelligent system reasoning. The eahouker would appreciate some explanation of the sources and how it has been reasoned to be generated; therefore, explicative systems may be of interest for this application.

In the SandS social network, input data should be in the form of household task codifications, while the output may correspond to recipe parameter settings, which may be continuous variables, i.e., water temperature in the washing machine, or categorical, i.e., steps in the washing process. The user feedback may be expressed in simple terms, such a Likert scale of satisfaction, which needs to be translated into an error measure that may drive the recipe learning. Household tasks performed by different appliances need to be solved by specific learned systems, which amounts to perform some partition in the task/recipe space by a selection mechanism driven by the task specification.

6 Conclusions

Subconscious social intelligence is a new way to pose the problem-solving power of social networks, combining conscious social computing built from explicit interactions from social players and subconscious problem solving trained from the experiences percolated from the social interaction down to a subconscious reasoning layer. The consideration of this kind of systems amounts to a radical shift on how social Web services are designed and deployed. It would no longer be the needs and requirements of the large corporations owning huge computational facilities that drive the system computational intelligence. Instead of the social players, the individual users of the system are the ones reaching the benefits of the social interaction for a better personal and social life.

Acknowledgments Work performed under Grant agreement 317947 of the EU, SandS project, the research and funds unit UFI11/07 of the UPV/EHU, and university research group Grant IT874-13 from the Basque Country Government.

References

1. B. Apolloni, M. Fiasche, G. Galliani, C. Zizzo, G. Caridakis, G. Siolas, S. Kollias, M. Grana-Romay, F. Barriento, and S. San-Jose. Social things - the sands instantiation. *In Internet of Things: Smart Objects and Services IoT-SoS 2013*. IEEE PRESS, 2013.
2. D. Lazer et al. Computational social science. *Science,* 323(5915):721–723, 2009.
3. M. Grana, B. Apolloni, M. Fiasche, G. Galliani, C. Zizzo, G. Caridakis, G. Siolas, S. Kollias, F. Barriento, and S. San Jose. Social and smart: towards an instance of subconscious social intelligence. In H. Papadopoulos L. Iliadis and C. Jayne (Eds.), editors, *EANN 2013,* volume part II, pages 302-3011. Springer Berlin Heidelberg, 2013.
4. M Grana and I Rebollo. Instances of subconscious social intelligent computing. In *CASON 2013.* IEEE PRESS, 2013.
5. W. Mao, A. Tuzhilin, and J. Gratch. Social and economic computing. *IEEE Intelligent Systems,* 26(6):19-21, 2011.
6. M. Grana I. Marques, A. Savio, and B. Apolloni. A domestic application of intelligent social computing: the sands project. In *SOCO 2013.* Springer Berlin Heidelberg, 2013.
7. Ognjen Scekic, Hong-Linh Truong, and Schahram Dustdar. Incentives and rewarding in social computing. *Communications of the ACM,* 56(6):72-82, 2013.
8. F.-Y. Wang, K.M. Carley, D. Zeng, and W. Mao. Social computing: From social informatics to social intelligence. *Intelligent Systems, IEEE,* 22(2):79-83, 2007.

Modeling Heavy Tails in Traffic Sources for Network Performance Evaluation

Vaidyanathan Ramaswami, Kaustubh Jain, Rittwik Jana and Vaneet Aggarwal

Abstract Heavy tails in work loads (file sizes, flow lengths, service times, etc.) have significant negative impact on the performance of queues and networks. In the context of the famous Internet file size data of Crovella and some very recent data sets from a wireless mobility network, we examine the new class of LogPH distributions introduced by Ramaswami for modeling heavy-tailed random variables. The fits obtained are validated using separate training and test data sets and also in terms of the ability of the model to predict performance measures accurately as compared with a trace-driven simulation using NS-2 of a bottleneck Internet link running a TCP protocol. The use of the LogPH class is motivated by the fact that these distributions have a power law tail and can approximate any distribution arbitrarily closely not just in the tail but in its entire range. In many practical contexts, although the tail exerts significant effect on performance measures, the bulk of the data is in the head of the distribution. Our results based on a comparison of the LogPH fit with other classical model fits such as Pareto, Weibull, LogNormal, and Log-*t* demonstrate the greater accuracy achievable by the use of LogPH distributions and also confirm the importance of modeling the distribution in its entire range and not just in the tail.

Keywords Network performance · Heavy tailed random variables · LogPH distribution · Markov chain

V. Ramaswami (✉) · R. Jana · V. Aggarwal
Florham Park, New Jersey, USA
e-mail: ram@ramaswami.com

K. Jain
College Park, Maryland, USA

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_4, © Springer India 2014

1 Introduction

The negative impact of heavy tails in work loads on the performance of systems is well known in the queuing literature. Indeed, many new scheduling strategies came to be invented primarily to avoid these bad effects of very large work loads (even if they be infrequent and from a small set of customers) for systems with schedules such as the First-in-First-Out discipline. Concern about heavy tails nevertheless holds even in the context of modern-day systems such as high-speed and wireless networks. Indeed, the increasing presence of bandwidth-intensive video and streaming audio has heightened the concern particularly in wireless networks as evidenced, for example, by the AT&T experience soon after the introduction of the iPhone.

An early work drawing attention to the presence of heavy tails in Internet file sizes is that of Crovella [4]. We use Crovella's data set and model the distribution as a LogPH distribution and also in terms of classical models such as Pareto, Weibull, LogNormal, and Log-*t*. The LogPH distribution was proposed by Ramaswami [6] who identified it to have a power law tail and dense (in the weak convergence metric) in the class of all distributions on $[1, \infty)$. A LogPH random variable Y is a random variable that can be written as $Y = \mathrm{e}^X$ where X is a phase-type random variable as defined by Neuts [9, 10]; see also [7] for a discussion on phase-type (PH) distributions. The first formal reference to LogPH was made in the paper by Ghosh et al. [6] on modeling traffic to a public Wi-Fi network. A detailed mathematical treatment of the LogPH class of distributions has been given in Ahn et al. [1]. That work of Ahn et al. demonstrates the power of the LogPH class to model heavy-tailed distributions in their entire range in the context of several financial examples. This paper demonstrates its power in the context of network performance modeling.

Needless to say, there are many attempts in the literature (see [1, 5, 8]) in queuing, performance analysis, risk theory, and finance to model heavy-tailed distributions; a key idea is to use some distributions (such as Pareto or Weibull) with a known heavy tail to model the tail and then a mixture to obtain a fit across the entire range. Unfortunately, these attempts have not resulted in a single class of models that can be used dependably in a large number of contexts, and furthermore, many aspects of the fitting methodology appear to be adhoc. Also, it would appear prudent to adopt more stringent standards in assessing the statistical quality of the fits in terms of a test data set that is separate from the training data set used for fitting a model. Also, it is desirable not only to consider the fitted random variable, but also to assess the quality of the fit in terms of its ability to predict performance of systems in which the models are used. Judged in this context, this paper may be found interesting and useful by many.

This paper is organized as follows. In Sect. 2, we provide a quick discussion of various classical models used in the context of heavy tails and also of the LogPH distribution. A brief discussion of a method based on the EM algorithm to fit LogPH distributions is given. In Sect. 3, we fit LogPH distributions to two data sets, the Crovella Internet file size data set [4] and a very recent (2012) data set on file sizes downloaded by mobile phone users in a cellular mobility network. We also provide a comparison of the LogPH fit with various other models such as the Pareto, Weibull, LogNormal, and Log-*t*. In addition to visual comparisons of the empirical with fitted distributions, we also provide some quantitative measures that aid such comparison. We wish to note that in the comparisons related to wireless mobility, since several data sets were available, we have taken a more stringent approach of having a "training set" based on which the model fits were made and a separate "test set" for comparing the models. Unfortunately, we had only one data set in the case of the Crovella data; we did comparisons with bootstrapped samples generated from this data set and found the fit to hold good for them as well. This step was undertaken primarily to make sure that we did not run the risk of overfitting. In Sect. 4, we take the Crovella data set and make a trace-driven simulation of a bottleneck Internet link and compare the performance results (queue lengths, throughput) against simulations run with fitted models using LogPH, Pareto, Weibull, LogNormal, and Log-*t*. Our results show that the LogPH gives more accurate results and thereby increase our confidence in the LogPH model class.

Our results give us great confidence in the ability of the LogPH class to model heavy-tailed distributions in a way to yield more accurate performance predictions in the network context. Much further work is needed on this class of models with regard to various issues including metrics for assessing goodness of fit, comparing different fits as well as certain issues related to the use of heavy traffic distributions with an infinite support. We will discuss some of these open issues. It is our hope that this work and the success of the LogPH class reported in Ahn et al. [1] will draw the attention of researchers and help improve our understanding of this class and our ability to model heavy-tailed phenomena more accurately. With this perspective, we will present not only the results obtained by us, but we shall also dwell on some of the gaps that need to be filled through further research.

2 Background

2.1 Phase-Type Distribution

A Phase-Type (PH) distribution is defined as the distribution of the time until absorption of a Markov chain with an absorbing state. This general class was introduced by M.F. Neuts [9, 10]. To be specific, consider a Markov chain with states 0,1,…, n, initial probability vector $(0, \boldsymbol{\tau})$ and infinitesimal generator Q. The row vector $\boldsymbol{\tau}$ is of size n and satisfies $\boldsymbol{\tau}\mathbf{1} = 1$, where $\mathbf{1}$ is a column vector of 1's. Assuming state 0 is an absorbing state, Q can be denoted as

$$Q = \begin{bmatrix} 0 & \mathbf{0} \\ \boldsymbol{t} & T \end{bmatrix}$$

where $\boldsymbol{t}$ is a column vector of size n and T is a $n \times n$ non-singular matrix satisfying, $T(i, i) < 0$, $T(i, j) > 0$ for $i \neq j$, and $T\mathbf{1} + \boldsymbol{t} = 0$. Thus, the Markov chain is completely characterized by parameters τ and T. The random variable X describing the time until absorption of the Markov chain into state 0 is called a PH random variable, denoted PH (τ, T). The number of non-absorbing states (n) is called the order of the PH random variable. The distribution and density functions of the PH random variable defined above are given by

$$F(x) = 1 - \boldsymbol{\tau}\exp(Tx)\mathbf{1}, \quad \text{for } x \geq 0, \tag{1}$$

$$f(x) = \boldsymbol{\tau}\exp(Tx)\boldsymbol{t}, \quad \text{for } x > 0. \tag{2}$$

PH distributions are known to be dense in the class of all distributions on [0, ∞). That is, they can approximate any distribution arbitrarily closely. Furthermore, they have many interesting closure properties and are highly tractable due to the connection with a Markov chain, which makes conditioning arguments easy. For these reasons, they have attracted much attention in applied probability. A property of the PH distribution is that its tail is asymptotically exponential; more specifically, for the distribution above, $P(X > x) \approx K\mathrm{e}^{-\eta x}$ for large x, where $-\eta < 0$ is the eigenvalue of T closest to zero.

2.2 LogPH Distribution

The LogPH distribution, denoted by LogPH($\boldsymbol{\tau}$, T), is defined as the distribution of the random variable Y that can be written as $Y = \exp(X)$ where X has a PH distribution with ($\boldsymbol{\tau}$, T). The LogPH random variable Y has its distribution function and density function as

$$F_Y(y) = 1 - \tau e^{T\log y}\mathbf{1}, \quad y \geq 1$$

and

$$f_Y(y) = \frac{1}{y}\tau e^{T\log y}t, y \geq 1, \quad t = -T\mathbf{1}.$$

From the exponential decay of the tail of the PH distribution, it easily follows that the LogPH random variable has a power law tail. Specifically, for large y, $P(Y > y) \approx K/y^{\eta}$, where η is as defined earlier. Also, from the fact that PH-type distributions are dense on [0, ∞), it follows by standard continuity theorems governing weak convergence (see Whitt [12]) that LogPH distributions are dense in the set of all distributions defined on [1, ∞). These properties make LogPH an attractive candidate class for modeling heavy-tailed random variables. In this context, we wish to note that the restriction of its range to $[1, \infty)$ is not particularly limiting for two reasons: (a) In many cases, one could rescale the data and fit LogPH to the scaled data set or (b) one can use a similar construction based on the bilateral PH random variables, which generalize the PH distribution to the entire real line; see Ahn and Ramaswami [2].

2.3 Some Classical Heavy-Tailed Distributions

Traditionally, modeling of heavy-tailed random variables has been focused mainly on the tail of the distribution. Some of the commonly used distributions are Pareto, Weibull, and LogNormal. Pareto has the following tail distribution:

$$\Pr[Z > z] = \left(\frac{b}{z}\right)^{a}, \quad z \geq b \tag{3}$$

where a and b are the shape and the scale parameters, respectively. Various enhancements of the Pareto distribution have also been used to match the mean, to select the cutoff at which the asymptotic power law takes over. A Pareto random variable Y can be realized as $\exp\,(X)$ where X is an exponential random variable; in this sense, one may consider the LogPH class as a natural generalization of the Pareto distribution since the exponential distribution is the most trivial example of a PH distribution.

The Weibull distribution does not have a power law in the tail distribution, but still the tail decays more slowly than the exponential. Denoted by a and b are the shape and scale of the distribution,

$$\Pr[Z > z] = \exp\{-(z/b)^{a}\}, \quad z > 0, \quad a < 1. \tag{4}$$

LogNormal distribution is modeled as a distribution, which is normal in the log scale. Specifically, Z has a LogNormal distribution if we can write

$$Z = e^{X}, \quad \text{where } X \sim \mathcal{N}(\mu, \sigma^2). \tag{5}$$

The normal distribution has a fast decaying tail $e^{-x^2/2}$, and this has led some researchers to use the t-distribution in place of the normal and define a Log-t distribution as a model for heavy-tailed distributions. Like the normal, the t-distribution also is symmetric about the origin and that could limit some of its applicability.

While the literature abounds in many applied examples where the above distributions and mixtures involving them have been used successfully for specific situations, there are some basic challenges in their use. These distributions do not form a dense class that provides a guarantee that one may effect a fit from any one of the members to a desired accuracy. Also, often one is forced to make a trade-off between matching the tail and matching the head of the distribution and to come up with ad hoc procedures for fitting a mixture that attempts to give a good model in the entire range of the data. We refer to Ahn et al. for a discussion of the issues in light of a famous data set—the Danish fire insurance data—that has been used as an important test data set in the statistical literature.

2.4 Fitting a LogPH Distribution

A LogPH distribution is fitted to the data by fitting a PH distribution to the logarithms of the data values (to the base *e*). If only a very small fraction of data values exist that are less than 1, we may discard them, or alternately, we may rescale the data by dividing all the elements by the minimum value so that all logarithms are positive and then fit a LogPH to these log values.

The standard approach to fitting a phase-type distribution is to use the EM algorithm whose details are provided in a paper by Asmussen et al. [3]. The EM algorithm is based on the following observations. Suppose the true distribution is a phase-type distribution of order n. If one knew the number of visits to each of the n transient states in the Markov chain and the amount of times spent in each of them before the Markov chain gets absorbed, these values together would constitute a set of sufficient statistics for the unknown parameters of the Markov chain. Now, the EM algorithm starts with a trial phase-type distribution of order n and iterates on the following two steps: (1) E-step: Consider the number and duration of visits to the transient states as missing values and replace them with their conditional expectation evaluated with respect to the current estimate of the parameters; (2) M-step: Now considering as though we have a complete sample on the sufficient statistics, maximize the likelihood function to obtain an improved estimate of the Markov chain parameters. The general theory of EM guarantees convergence to (a local) maximum of the likelihood function.

3 Data Sets and Fitted Models

This paper deals with two data sets. The first is the well-known World Wide Web file size traces collected by Crovella in 1995 commonly used by researchers to model heavy-tailed data. We will demonstrate that LogPH provides a much better fit for the entire data range as compared with previously used models, particularly in its ability to predict network performance more accurately as evaluated in the context of a bottleneck Internet link. Secondly, we will also use a very recent data set of file sizes from the *mobile Web.* We show once again that LogPH provides a good fit to the entire range of the mobile data set. Our interest in the second data set is due to our current focus on wireless and mobile networks and the fact that the Crovella data are now quite dated and predate some major new bandwidth-intensive applications such as video and streaming audio that have become much