Stefano Leonardi (Ed.)

Algorithms and Models for the Web-Graph

LNCS 3243

Third International Workshop, WAW 2004
Rome, Italy, October 2004
Proceedings

Springer

Stefano Leonardi (Ed.)

Algorithms and Models for the Web-Graph

Third International Workshop, WAW 2004
Rome, Italy, October 16, 2004
Proceeedings

Springer

Volume Editor

Stefano Leonardi
University of Rome "La Sapienza"
Via Salaria 113, 00198 Roma, Italy
E-mail: leon@dis.uniroma1.it

Library of Congress Control Number: 2004113291

CR Subject Classification (1998): F.2, G.2, H.4, H.3, C.2, H.2.8, E.1

ISSN 0302-9743
ISBN 3-540-23427-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11333302 06/3142 5 4 3 2 1 0

Preface

This volume contains the 14 contributed papers and the contribution of the distinguished invited speaker Béla Bollobás presented at the 3rd Workshop on Algorithms and Models for the Web-Graph (WAW 2004), held in Rome, Italy, October 16, 2004, in conjunction with the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2004).

The World Wide Web has become part of our everyday life and information retrieval and data mining on the Web is now of enormous practical interest. Some of the algorithms supporting these activities are based substantially on viewing the Web as a graph, induced in various ways by links among pages, links among hosts, or other similar networks.

The aim of the 2004 Workshop on Algorithms and Models for the Web-Graph was to further the understanding of these Web-induced graphs, and stimulate the development of high-performance algorithms and applications that use the graph structure of the Web. The workshop was meant both to foster an exchange of ideas among the diverse set of researchers already involved in this topic, and to act as an introduction for the larger community to the state of the art in this area.

This was the third edition of a very successful workshop on this topic, WAW 2002 was held in Vancouver, Canada, in conjunction with the 43rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2002, and WAW 2003 was held in Budapest, Hungary, in conjunction with the 12th International World Wide Web Conference, WWW 2003. This was the first edition of the workshop with formal proceedings.

The organizing committee of the workshop consisted of:

Andrei Broder	IBM Research
Guido Caldarelli	INFM, Italy
Ravi Kumar	IBM Research
Stefano Leonardi	University of Rome "La Sapienza"
Prabhakar Raghavan	Verity Inc.

Papers were solicited in all areas of the study of Web graphs, including but not limited to:

- Mathematical models, topology generators, and dynamic properties;
- Algorithms for analyzing Web graphs and for computing graph properties at the Web scale;
- Application of Web graph algorithms to data mining and information retrieval;
- Clustering and visualization;
- Representation and compression;
- Graph-oriented statistical sampling of the Web;
- Empirical exploration techniques and practical systems issues.

The extended abstracts were read by at least three referees each, and evaluated on their quality, originality, and relevance to the symposium. The program committee selected 14 papers out of 31 submissions. The program committee consisted of:

Dimitris Achlioptas	Microsoft Research
Lada Adamic	HP Labs
Jennifer Chayes	Microsoft Research
Fan Chung Graham	UC San Diego
Taher Haveliwala	Stanford University and Google
Elias Koutsoupias	Univ. of Athens
Ronny Lempel	IBM Research
Stefano Leonardi (Chair)	Univ. of Rome "La Sapienza"
Mark Manasse	Microsoft Research
Kevin McCurley	IBM Research
Dragomir Radev	Univ. of Michigan
Sridhar Rajagopalan	IBM Research
Oliver Riordan	Cambridge University
D. Sivakumar	IBM Research
Panayotis Tsaparas	Univ. of Helsinki
Eli Upfal	Brown University
Alessandro Vespignani	Univ. of Paris Sud

WAW 2004, and in particular the invited lecture of Béla Bollobás, was generously supported by IBM. A special thanks is due to Andrei Broder for his effort in disseminating the Call for Papers, to Ravi Kumar for handling the Web site of the – Workshop, and to Debora Donato for her assistance in assembling these proceedings. We hope that this volume offers the reader a representative selection of some of the best current research in this area.

August 2004

Stefano Leonardi
Program Chair
WAW 2004

Lecture Notes in Computer Science 3243

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Lecture Notes in Computer Science

For information about Vols. 1–3174

please contact your bookseller or Springer

Vol. 3293: C.-H. Chi, M. van Steen, C. Wills (Eds.), Web Content Caching and Distribution. IX, 283 pages. 2004.

Vol. 3274: R. Guerraoui (Ed.), Distributed Computing. XIII, 465 pages. 2004.

Vol. 3273: T. Baar, A. Strohmeier, A. Moreira, S.J. Mellor (Eds.), <<UML>> 2004 - The Unified Modelling Language. XIII, 454 pages. 2004.

Vol. 3271: J. Vicente, D. Hutchison (Eds.), Management of Multimedia Networks and Services. XIII, 335 pages. 2004.

Vol. 3270: M. Jeckle, R. Kowalczyk, P. Braun (Eds.), Grid Services Engineering and Management. X, 165 pages. 2004.

Vol. 3266: J. Solé-Pareta, M. Smirnov, P.V. Mieghem, J. Domingo-Pascual, E. Monteiro, P. Reichl, B. Stiller, R.J. Gibbens (Eds.), Quality of Service in the Emerging Networking Panorama. XVI, 390 pages. 2004.

Vol. 3265: R.E. Frederking, K.B. Taylor (Eds.), Machine Translation: From Real Users to Research. XI, 392 pages. 2004. (Subseries LNAI).

Vol. 3264: G. Paliouras, Y. Sakakibara (Eds.), Grammatical Inference: Algorithms and Applications. XI, 291 pages. 2004. (Subseries LNAI).

Vol. 3263: M. Weske, P. Liggesmeyer (Eds.), Object-Oriented and Internet-Based Technologies. XII, 239 pages. 2004.

Vol. 3261: T. Yakhno (Ed.), Advances in Information Systems. XIV, 617 pages. 2004.

Vol. 3260: I. Niemegeers, S.H. de Groot (Eds.), Personal Wireless Communications. XIV, 478 pages. 2004.

Vol. 3258: M. Wallace (Ed.), Principles and Practice of Constraint Programming – CP 2004. XVII, 822 pages. 2004.

Vol. 3257: E. Motta, N.R. Shadbolt, A. Stutt, N. Gibbins (Eds.), Engineering Knowledge in the Age of the Semantic Web. XVII, 517 pages. 2004. (Subseries LNAI).

Vol. 3256: H. Ehrig, G. Engels, F. Parisi-Presicce, G. Rozenberg (Eds.), Graph Transformations. XII, 451 pages. 2004.

Vol. 3255: A. Benczúr, J. Demetrovics, G. Gottlob (Eds.), Advances in Databases and Information Systems. XI, 423 pages. 2004.

Vol. 3254: E. Macii, V. Paliouras, O. Koufopavlou (Eds.), Integrated Circuit and System Design. XVI, 910 pages. 2004.

Vol. 3253: Y. Lakhnech, S. Yovine (Eds.), Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. X, 397 pages. 2004.

Vol. 3250: L.-J. (LJ) Zhang, M. Jeckle (Eds.), Web Services. X, 301 pages. 2004.

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004. (Subseries LNAI).

Vol. 3246: A. Apostolico, M. Melucci (Eds.), String Processing and Information Retrieval. XIV, 332 pages. 2004.

Vol. 3245: E. Suzuki, S. Arikawa (Eds.), Discovery Science. XIV, 430 pages. 2004. (Subseries LNAI).

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004. (Subseries LNAI).

Vol. 3243: S. Leonardi (Ed.), Algorithms and Models for the Web-Graph. VIII, 189 pages. 2004.

Vol. 3242: X. Yao, E. Burke, J.A. Lozano, J. Smith, J.J. Merelo-Guervós, J.A. Bullinaria, J. Rowe, P. Tiňo, A. Kabán, H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature - PPSN VIII. XX, 1185 pages. 2004.

Vol. 3241: D. Kranzlmüller, P. Kacsuk, J.J. Dongarra (Eds.), Recent Advances in Parallel Virtual Machine and Message Passing Interface. XIII, 452 pages. 2004.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004. (Subseries LNBI).

Vol. 3239: G. Nicosia, V. Cutello, P.J. Bentley, J. Timmis (Eds.), Artificial Immune Systems. XII, 444 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004. (Subseries LNAI).

Vol. 3236: M. Núñez, Z. Maamar, F.L. Pelayo, K. Pousttchi, F. Rubio (Eds.), Applying Formal Methods: Testing, Performance, and M/E-Commerce. XI, 381 pages. 2004.

Vol. 3235: D. de Frutos-Escrig, M. Nunez (Eds.), Formal Techniques for Networked and Distributed Systems – FORTE 2004. X, 377 pages. 2004.

Vol. 3232: R. Heery, L. Lyon (Eds.), Research and Advanced Technology for Digital Libraries. XV, 528 pages. 2004.

Vol. 3231: H.-A. Jacobsen (Ed.), Middleware 2004. XV, 514 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004. (Subseries LNAI).

Vol. 3225: K. Zhang, Y. Zheng (Eds.), Information Security. XII, 442 pages. 2004.

Vol. 3224: E. Jonsson, A. Valdes, M. Almgren (Eds.), Recent Advances in Intrusion Detection. XII, 315 pages. 2004.

Vol. 3223: K. Slind, A. Bunker, G. Gopalakrishnan (Eds.), Theorem Proving in Higher Order Logics. VIII, 337 pages. 2004.

Vol. 3222: H. Jin, G.R. Gao, Z. Xu, H. Chen (Eds.), Network and Parallel Computing. XX, 694 pages. 2004.

Vol. 3221: S. Albers, T. Radzik (Eds.), Algorithms – ESA 2004. XVIII, 836 pages. 2004.

Vol. 3220: J.C. Lester, R.M. Vicari, F. Paraguaçu (Eds.), Intelligent Tutoring Systems. XXI, 920 pages. 2004.

Vol. 3219: M. Heisel, P. Liggesmeyer, S. Wittmann (Eds.), Computer Safety, Reliability, and Security. XI, 339 pages. 2004.

Vol. 3217: C. Barillot, D.R. Haynor, P. Hellier (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004. XXXVIII, 1114 pages. 2004.

Vol. 3216: C. Barillot, D.R. Haynor, P. Hellier (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004. XXXVIII, 930 pages. 2004.

Vol. 3215: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVII, 906 pages. 2004. (Subseries LNAI).

Vol. 3214: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1302 pages. 2004. (Subseries LNAI).

Vol. 3213: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1280 pages. 2004. (Subseries LNAI).

Vol. 3212: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition. XXIX, 862 pages. 2004.

Vol. 3211: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition. XXIX, 880 pages. 2004.

Vol. 3210: J. Marcinkowski, A. Tarlecki (Eds.), Computer Science Logic. XI, 520 pages. 2004.

Vol. 3209: B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, G. Stumme (Eds.), Web Mining: From Web to Semantic Web. IX, 201 pages. 2004. (Subseries LNAI).

Vol. 3208: H.J. Ohlbach, S. Schaffert (Eds.), Principles and Practice of Semantic Web Reasoning. VII, 165 pages. 2004.

Vol. 3207: L.T. Yang, M. Guo, G.R. Gao, N.K. Jha (Eds.), Embedded and Ubiquitous Computing. XX, 1116 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004. (Subseries LNAI).

Vol. 3205: N. Davies, E. Mynatt, I. Siio (Eds.), UbiComp 2004: Ubiquitous Computing. XVI, 452 pages. 2004.

Vol. 3203: J. Becker, M. Platzner, S. Vernalde (Eds.), Field Programmable Logic and Application. XXX, 1198 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004. (Subseries LNAI).

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004. (Subseries LNAI).

Vol. 3199: H. Schepers (Ed.), Software and Compilers for Embedded Systems. X, 259 pages. 2004.

Vol. 3198: G.-J. de Vreede, L.A. Guerrero, G. Marín Raventós (Eds.), Groupware: Design, Implementation and Use. XI, 378 pages. 2004.

Vol. 3196: C. Stary, C. Stephanidis (Eds.), User-Centered Interaction Paradigms for Universal Access in the Information Society. XII, 488 pages. 2004.

Vol. 3195: C.G. Puntonet, A. Prieto (Eds.), Independent Component Analysis and Blind Signal Separation. XXIII, 1266 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004. (Subseries LNAI).

Vol. 3193: P. Samarati, P. Ryan, D. Gollmann, R. Molva (Eds.), Computer Security – ESORICS 2004. X, 457 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004. (Subseries LNAI).

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004. (Subseries LNAI).

Vol. 3190: Y. Luo (Ed.), Cooperative Design, Visualization, and Engineering. IX, 248 pages. 2004.

Vol. 3189: P.-C. Yew, J. Xue (Eds.), Advances in Computer Systems Architecture. XVII, 598 pages. 2004.

Vol. 3188: F.S. de Boer, M.M. Bonsangue, S. Graf, W.-P. de Roever (Eds.), Formal Methods for Components and Objects. VIII, 373 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004. (Subseries LNAI).

Vol. 3186: Z. Bellahsène, T. Milo, M. Rys, D. Suciu, R. Unland (Eds.), Database and XML Technologies. X, 235 pages. 2004.

Vol. 3185: M. Bernardo, F. Corradini (Eds.), Formal Methods for the Design of Real-Time Systems. VII, 295 pages. 2004.

Vol. 3184: S. Katsikas, J. Lopez, G. Pernul (Eds.), Trust and Privacy in Digital Business. XI, 299 pages. 2004.

Vol. 3183: R. Traunmüller (Ed.), Electronic Government. XIX, 583 pages. 2004.

Vol. 3182: K. Bauknecht, M. Bichler, B. Pröll (Eds.), E-Commerce and Web Technologies. XI, 370 pages. 2004.

Vol. 3181: Y. Kambayashi, M. Mohania, W. Wöß (Eds.), Data Warehousing and Knowledge Discovery. XIV, 412 pages. 2004.

Vol. 3180: F. Galindo, M. Takizawa, R. Traunmüller (Eds.), Database and Expert Systems Applications. XXI, 972 pages. 2004.

Vol. 3179: F.J. Perales, B.A. Draper (Eds.), Articulated Motion and Deformable Objects. XI, 270 pages. 2004.

Vol. 3178: W. Jonker, M. Petkovic (Eds.), Secure Data Management. VIII, 219 pages. 2004.

Vol. 3177: Z.R. Yang, H. Yin, R. Everson (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2004. XVIII, 852 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004. (Subseries LNAI).

Vol. 3175: C.E. Rasmussen, H.H. Bülthoff, B. Schölkopf, M.A. Giese (Eds.), Pattern Recognition. XVIII, 581 pages. 2004.

Table of Contents

IBM Invited Lecture

Contributed Papers

The Phase Transition and Connectedness in Uniformly Grown Random Graphs

Béla Bollobás[1,2,*] and Oliver Riordan[2,3]

[1] Department of Mathematical Sciences, University of Memphis, Memphis TN 38152, USA
[2] Trinity College, Cambridge CB2 1TQ, UK
[3] Royal Society Research Fellow, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK

Abstract. We consider several families of random graphs that grow in time by the addition of vertices and edges in some 'uniform' manner. These families are natural starting points for modelling real-world networks that grow in time. Recently, it has been shown (heuristically and rigorously) that such models undergo an 'infinite-order phase transition': as the density parameter increases above a certain critical value, a 'giant component' emerges, but the speed of this emergence is extremely slow. In this paper we shall present some of these results and investigate the connection between the existence of a giant component and the connectedness of the final infinite graph.

1 Introduction

Recently, there has been a lot of interest in modelling networks in the real world by random graphs. Unlike classical random graphs, many (perhaps most) large networks in the real world evolve in time; in fact they tend to grow in time by the addition of new nodes and new connections. Real-world networks differ from classical random graphs in other important ways (for example, they are often 'scale-free', in the sense of having a power-law degree distribution), and, of course, one cannot expect to model any particular network very accurately, as the real mechanisms involved are not amenable to mathematical analysis. Nevertheless, it is important to model these networks as well as one can, and one general approach is to develop mathematical models for important general features. These models should be simple enough that their properties can be analyzed rigorously. Of course, such models will not be accurate for any given network, but they will give insight into the behaviour of many networks.

One important property of real-world networks is their *robustness*, or resilience to random failures. There are many ways in which one might measure robustness; perhaps the most common is to consider deleting edges or vertices

* Research supported by NSF grant ITR 0225610 and DARPA grant F33615-01-C-1900.

S. Leonardi (Ed.): WAW 2004, LNCS 3243, pp. 1–18, 2004.

from the network at random, and ask whether the network fractures into 'small' pieces, or whether a 'giant component' remains, i.e., a component containing a constant fraction of the initial graph. We shall describe a precise form of this question below.

2 Models

The baseline that any new random graph model should initially be compared with is, and will remain, the classical 'uniform' random graph models of Erdős and Rényi, and Gilbert. Erdős and Rényi founded the theory of random graphs in the late 1950s and early 1960s, setting out to investigate the properties of a 'typical' graph with n vertices and M edges. Their random graph model, $G(n,M)$, introduced in [13], is defined as follows: given $n \geq 2$ and $0 \leq M \leq N = \binom{n}{2}$, let $G(n,M)$ be a graph on n labelled vertices (for example, on the set $[n] = \{1,2,\ldots,n\}$) with M edges, chosen uniformly at random from all $\binom{N}{M}$ such graphs.

Around the same time that Erdős and Rényi introduced $G(n,M)$, Gilbert [15] introduced a closely related model, $G(n,p)$. Again, $G(n,p)$ is a random graph on n labelled vertices, for example on the set $[n]$. The parameter p is between 0 and 1, and $G(n,p)$ is defined by joining each pair $\{i,j\} \subset [n]$ with an edge with probability p, independently of every other pair. For a wide range of the parameters, for many questions, there is essentially no difference between $G(n,M)$ and $G(n,p)$, where $p = M/N$. Nowadays, $G(n,p)$ is much more studied, as the independence between edges makes it much easier to work with.

Although the definition of $G(n,p)$ has a more probabilistic flavour than that of $G(n,M)$, it was Erdős and Rényi rather than Gilbert who pioneered the use of probabilistic methods to study random graphs, and it is perhaps not surprising that $G(n,p)$ is often known as 'the Erdős-Rényi random graph'. When studying $G(n,p)$, or $G(n,M)$, one is almost always interested in properties that hold for 'all typical' graphs in the model. We say that an event holds *with high probability* or **whp**, if it holds with probability tending to 1 as n, the number of vertices, tends to infinity.

Perhaps the single most important result of Erdős and Rényi about random graphs concerns the emergence of the giant component. Although they stated this result for $G(n,M)$, we shall state it for $G(n,p)$; this is a context in which the models are essentially interchangeable.

For $x > 0$ a constant let

$$t(x) = \frac{1}{x} \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} \left(xe^{-x}\right)^k . \tag{1}$$

Erdős and Rényi [14] proved the following result.

Theorem 1. *Let $x > 0$ be a constant. If $x < 1$ then* **whp** *every component of $G(n, x/n)$ has order $O(\log n)$. If $x > 1$ then* **whp** *$G(n, x/n)$ has a component with $(1 - t(x) + o(1))n = \Theta(n)$ vertices, and all other components have $O(\log n)$ vertices.*

In other words, there is a 'phase transition' at $x = 1$. This is closely related to the robustness question described vaguely in the introduction, and to the percolation phase transition in random subgraphs of a fixed graph. In this paper, we shall often consider a *random* initial graph, and delete edges (or vertices) independently, retaining each with some probability p, to obtain a random subgraph. The question is, given the (random) initial graph on n vertices, for which values of p does the random subgraph contain a *giant component*, i.e., a component with $\Theta(n)$ vertices? In the context of $G(n,p)$, and in several of the examples we consider, there is no need for this two step construction: if edges of $G(n,p_1)$ are retained independently with probability p_2, the result is exactly $G(n,p_1p_2)$. In these cases, the robustness question can be rephrased as follows: 'for which values of the edge density parameter is there (**whp**) a giant component?' In the case of $G(n,p)$, the natural normalization is to write $p = x/n$ and keep x fixed as n varies. Thus we see that the classical result of Erdős and Rényi stated above, is exactly a (the first) robustness result of this form.

When the giant component exists, one is often interested in its size, especially near the phase transition. In principle, for $G(n,p)$, the formula (1) above answers this question. More usefully, at $x = 1$ the right-derivative of $t(x)$ is -2, so when $x = 1 + \varepsilon$, the limiting fraction (as $n \to \infty$ with $\varepsilon > 0$ fixed) of vertices in the giant component is $2\varepsilon + o(\varepsilon)$.

2.1 The CHKNS Model

Many growing real-world networks have a number of direct connections that grows roughly linearly with the number of nodes. From now on we shall use graph theoretic terminology, so vertices and edges will correspond to nodes and direct connections between pairs of nodes. (Here we consider undirected connections only.) A very natural model for a growing graph which has on average a constant number of edges per vertex (i.e., in the limit, a constant average degree) was introduced by Callaway, Hopcroft, Kleinberg, Newman and Strogatz [9] in 2001. The model is defined as follows: at each time step, a new vertex is added. Then, with probability δ, two vertices are chosen uniformly at random and joined by an undirected edge. In [9] loops and multiple edges are allowed, although these turn out to be essentially irrelevant. We shall write $G_C^{(n)}(\delta)$ for the n-vertex CHKNS graph constructed with parameter δ. Most of the time, we shall suppress the dependence on n, writing $G_C(\delta)$, to avoid cluttering the notation.

The question considered by Callaway et al in [9] is the following: as the parameter δ is varied, when does $G_C(\delta)$ contain a giant component, i.e., a component containing order n vertices? In other words, what is the equivalent of Theorem 1 for $G_C(\delta)$? In [9], a heuristic argument is given that there is a phase transition at $\delta = \delta_c = 1/8$, i.e., that for $\delta \leq 1/8$ there is no giant component, while for $\delta > 1/8$ there is. Callaway et al also suggest that this phase transition has a particularly interesting form: for $\delta = 1/8 + \varepsilon$, they give numerical evidence (based on integrating an equation, not simply simulating the graph) to suggest that the average fraction of the vertices that lie in the giant component is a

function $f_C(\varepsilon)$ which has all derivatives zero at $\varepsilon = 0$. Such a phase transition is called an *infinite-order* transition.

In essence, the question of finding the critical probability for the existence of a giant component in $G_C(\delta)$ had been answered more than a decade before Callaway et al posed the question. As we shall see in section 2.3, a question that turns out to be essentially equivalent was posed by Dubins in 1984, answered partially by Kalikow and Weiss [17] in 1988, and settled by Shepp [20] in 1989.

Dorogovtsev, Mendes and Samukhin [10] analyzed the CHKNS model in a way that, while fairly mathematical, is still far from rigorous. Their methods supported the conjecture of [9], suggesting that indeed the transition is at $\delta_c = 1/8$, and that the phase transition has infinite order.

Before turning to comparison with other models, note that there is a natural slight simplification of the CHKNS model, suggested independently in [11] and [3]. At each stage, instead of adding a single edge between a random pair of vertices with probability δ, for each of the $\binom{n}{2}$ pairs of vertices, add an edge between them with probability $\delta/\binom{n}{2}$, independently of all other pairs. In this way, the number of edges added in one step has essentially a Poisson distribution with mean δ. In the long term, this will make very little difference to the behaviour of the model. The key advantage is that, in the graph generated on n vertices, different edges are now present independently: more precisely, for $\{i,j\} \neq \{i',j'\}$, whether there is an edge (or edges) between i and j is independent of whether there is an edge (or edges) between i' and j'. Note also that the expected number of edges between i and j, $1 \leq i < j \leq n$, is exactly

$$\sum_{k=j}^{n} \delta \binom{k}{2}^{-1} = 2\delta \left(\frac{1}{j-1} - \frac{1}{n} \right). \tag{2}$$

2.2 The Uniformly Grown Random Graph

Although it is not our main focus here, perhaps the most studied growing-graph model is the growth with preferential attachment 'model' of Barabási and Albert, introduced in [1]. The reason for the quotation marks is that the description given by Barabási and Albert is incomplete, and also inconsistent, so their 'model' is not a model in a mathematical sense. Roughly speaking, the Barabási-Albert, or BA, model is defined as follows: an integer parameter m is fixed, and the graph grows by adding one vertex at a time, with each new vertex sending m edges to old vertices, chosen *with probabilities proportional to their degrees*. (This is known as the 'preferential attachment rule'). To prove rigorous results, one must first know exactly what model one is talking about, and this is the reason for the introduction of the *linearized chord diagram* or LCD model in [6]. (See [6] for a description of the problems with the BA model, and [4] for a detailed discussion.) The LCD model is a precisely defined model (one of many) fitting the rough description of Barabási and Albert. It also has an important extra property: although defined as a growing graph process, with rules for how each new vertex attaches to the old graph, it has an equivalent static description, giving the whole n-vertex graph in one go. This makes it much easier to analyze.

The main motivation of Barabási and Albert in [1] was to provide a model explaining the power-law distribution of degrees seen in many real-world networks. They show heuristically that their model does indeed have a power-law degree distribution; this is proved rigorously for the LCD model in [8]. Barabási and Albert note that their new model differs in two fundamental ways from classical uniform models – growth in time, and the preferential attachment rule. They ask whether both these differences are necessary to obtain power-law degree distribution, leading naturally to the study of the following model.

Given an integer m, start with m vertices and no edges. At each time step, add a new vertex to the graph, and join it to a set of m earlier vertices, chosen uniformly at random from among all possible such sets. We shall call this the *growing m-out model*, and write $G_m^{(n)}$, or simply G_m, for the n-vertex graph obtained after $n - m$ steps. Note that this is perhaps the most natural model for a growing graph with (asymptotically) constant average degree.

Barabási and Albert [1] considered G_m briefly, noting that it does not have a power-law degree sequence. Turning to other properties of the BA or LCD models, the question of robustness was considered rigorously in [5] (working with the precisely defined LCD model). It turns out that the LCD model is in some sense 'infinitely robust' in that there is no phase transition: for any $p > 0$, if edges (or vertices) of the $m \geq 2$ LCD graph are retained independently with probability p, there is a giant component, although it may be very small. (Its size is linear in n, but the constant is extremely small when p is small.) Again, it is natural to ask if this striking difference from classical random graphs is due to growth or preferential attachment or both, providing another reason to study the phase transition in growing models, and in particular in G_m. The answer given in [5] is that growth alone is not enough. Much more precise results are given in [7] and [19]; we return to this later.

Just as for the CHKNS model $G_C(\delta)$, there is a natural modification to the growing m-out model G_m that makes it easier to study. Instead of adding exactly m edges from the new vertex, when adding the jth vertex, join it independently to each of the $j - 1$ earlier vertices, joining it to each with probability $m/(j - 1)$. Writing μ instead of m, as there is now no reason for the parameter to be an integer, one obtains a graph on n vertices in which edges are present independently, and for $1 \leq i < j \leq n$ the probability that ij is an edge is

$$\frac{\mu}{j-1}. \tag{3}$$

We call this graph the *uniformly grown random graph* and denote it by $G_U^{(n)}(\mu)$, or simply $G_U(\mu)$. Note that the model makes perfect sense for $\mu > 1$, but we should write $\min\{\mu/(j-1), 1\}$ in place of (3). It will turn out that the interesting values of μ are less than 1; also, the results we shall consider do not depend on the presence or absence of the first few edges. Thus we shall not bother with this detail. As we shall see later, and as one can guess by comparing (2) and (3), $G_C(\delta)$ and $G_U(\mu)$ will turn out to be closely related when $\mu \sim 2\delta$.

The modification described above is rather larger than that suggested for the CHKNS model: it destroys the property that the graph is m-out, i.e., that each vertex sends exactly m edges to earlier vertices. As in the CHKNS model edges are always added between random vertices (rather than the new vertex and a random old vertex), there is no corresponding property in the CHKNS model, and the change to independence is a very small change indeed.

2.3 Dubins' Model

In 1984, Dubins proposed a model for an *infinite* inhomogeneous random graph $G_D(\lambda)$ (see [17, 20]). This is the graph on the vertex set $\mathbf{N} = \{1, 2, 3, \ldots\}$ in which each edge ij is present independently, and the probability of the edge ij, $1 \leq i < j$, is

$$\frac{\lambda}{j}, \tag{4}$$

where $\lambda > 0$ is a real parameter. As before, if $\lambda > 2$ we should write $\min\{\lambda/j, 1\}$ above, but the interest is in smaller values of λ, so we shall not bother. It will come as no surprise that there is a strong connection between $G_D(\lambda)$ and the finite graphs $G_U(\mu)$ and $G_C(\delta)$, when $\lambda \sim \mu \sim 2\delta$. Dubins asked the following question: when $\lambda = 1$, is the graph $G_D(\lambda)$ almost surely *connected*?

At first sight, this question may seem rather strange. For one thing, infinite random graphs are frequently not very interesting. For example, if we take a fixed probability p, $0 < p < 1$, and join each pair $i, j \in \mathbf{N}$ independently with probability p, then with probability 1 we get *the infinite random graph* R. As R is not in fact a random object, and does not depend on p, probabilistic questions about it do not make sense. In fact, R has (in some sense) no structure: for every pair of finite disjoint sets U and W of vertices of R, there are infinitely many other vertices v each of which is joined to every vertex in U and to no vertex in W. Thus, R is trivially connected, having diameter 2. The infinite random graph proposed by Dubins is a very different matter: its structure depends very much on λ, and there are many non-trivial questions one can ask about it, in particular, whether it is connected.

Kalikow and Weiss [17] showed in 1988 that for $\lambda > 1$ the graph $G_D(\lambda)$ is connected (here and from now on, with probability 1 is to be understood), using a very weak form of the classical result of Erdős and Rényi [14] given as Theorem 1 above. They also showed that for $\lambda < 1/4$, $G_D(\lambda)$ is disconnected. They conclude that 'While we are fairly certain that $\frac{1}{4}$ can be replaced by 1, what happens at the critical value $\lambda = 1$ remains for us a mystery.' It is clear that there is some critical value λ_c (a priori, this could be zero or infinity) such that for $\lambda > \lambda_c$, $G_D(\lambda)$ is connected, while for $\lambda < \lambda_c$, $G_D(\lambda)$ is disconnected. Dubins believed that this critical value is equal to 1.

In 1989, Shepp [20] showed that for $\lambda > 1/4$ the graph $G_D(\lambda)$ is connected (with probability 1), establishing that in fact $\lambda_c = 1/4$. He also showed that at $\lambda = 1/4$, $G_D(\lambda)$ is disconnected. A corresponding result for a generalization of the model was proved by Durrett and Kesten [12] in 1990.

3 Connections

It is clear that two of the models considered above are very closely related. In defining the uniformly grown random graph, it makes very little difference if we take j instead of $j-1$ in (3) (this can actually be formalized, see [3]). After making this change, the only difference between the uniformly grown random graph $G_U(\mu)$, $\mu = \lambda$, and Dubins' graph $G_D(\lambda)$ is that the former is finite and the latter is infinite. In fact, $G_D(\lambda)$ can be viewed as the limit of $G_U(\lambda) = G_U^{(n)}(\lambda)$ as n, the number of vertices, tends to infinity.

In many contexts the $-1/n$ correction term in (2) also makes little difference. Comparing (2) and (3), one might expect $G_C(\delta)$ and $G_U(\mu)$ to behave similarly when μ is (approximately) equal to 2δ. Again, this can be formalized; it was shown by Bollobás, Janson and Riordan [3] that the thresholds for emergence of the giant component in G_C and G_U differ by a factor of *exactly* two, and that the phase transition is infinite-order in one if and only if it is in the other.

The models we consider are related, but what about the questions? For $G_C(\delta)$ or $G_U(\mu)$, we are interested in the question of when (i.e., for which values of the edge-density parameter) there is a giant component. Dubins asked when $G_D(\lambda)$ is connected. Translating to a finite context, the question of whether $G_D(\lambda)$ is connected is the same as the following: is it true that any two vertices i, j will eventually become connected (meaning, joined by a path) if we run the process $G_U(\lambda)$ for long enough? In other words, as n increases, does any fixed pair i, j, of vertices eventually become connected? The giant component question can be rephrased as the following: in the n-vertex graph $G_U(\lambda)$, is the number of connected pairs of order n^2? These two questions are related, but the connection is not obviously a very strong one. Indeed, *a priori*, it is possible that they could have different answers either way round: it might be that any two vertices become connected eventually, but not for a very long time, so the number of connected pairs in the n-vertex graph grows, but very slowly. On the other hand, order n^2 pairs could be connected, but there could be a few pairs that never become connected. It is easy to construct models in which either of these possibilities occurs.

It turns out, however, that in the context of uniformly grown graphs the two questions discussed above *are* very closely related. Indeed, the methods of Kalikow and Weiss [17] and Shepp [20] show that for $\lambda \leq 1/4$, **whp** there is no giant component in $G_U(\lambda)$, while for $\lambda > 1/4$, **whp** there is. The latter result is proved explicitly by Shepp, and then used to immediately deduce connectedness, in the same way that Kalikow and Weiss [17] used the classical Erdős-Rényi giant component result. Thus the question of where the phase transition (defined by the emergence of a giant component) happens in the uniformly grown random graph $G_U(\lambda)$ had already been answered in 1989 – the transition is at $\lambda_c = 1/4$. Recently, Durrett [11] pointed out that the methods of Durrett and Kesten [12] give a corresponding answer for the CHKNS model $G_C(\delta)$. Alternatively, as pointed out in [3], one can compare the models $G_C(\delta)$ and $G_U(\lambda)$, $\lambda \sim 2\delta$, directly: it is easy to show that the critical value δ_c at which phase transition in