

A Compilation of Results on Phase Transitions, Scale-Invariance

by

Vamsikrishna Kalapala

B. Tech., Indian Institute of Technology, Chennai, 1998

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2005

©2005, Vamsikrishna Kalapala

Dedication

*To Dr. William Feller, for writing such a wonderful book on Probability Theory
(referring to volume 1 of his Introduction to Probability Theory).*

Acknowledgments

I would like to express my gratitude to Dr. Cristopher Moore, my advisor for his guidance and support throughout my research. I thank him for making me a part his group and for directing me to work on the phase transitions and scale-invariance, and for helping me increase my problem-solving skills. I would like to thank fellow members of the team for creating a friendly, conducive atmosphere for research.

I want to thank Dr. Jared Saia and Dr. Terran Lane for consenting to be the members of my committee and reviewing my thesis. I want thank Dr. David Ackley for teaching me how to program well.

I would like to thank the staff of the Department of Computer Science (esp. Lynne Jacobsen) for its support.

A Compilation of Results on Phase Transitions, Scale-Invariance

by

Vamsikrishna Kalapala

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2005

A Compilation of Results on Phase Transitions, Scale-Invariance

by

Vamsikrishna Kalapala

B. Tech., Indian Institute of Technology, Chennai, 1998

M.S., Computer Science, University of New Mexico, 2005

Abstract

This thesis presents our work on two areas of interest to both computer scientists and physicists.

First, we study the 3-bit Exact Cover problem (EC3), also known as Positive 1-in-3 SAT. Random instances of this problem have been used as a benchmark for simulations of an adiabatic quantum algorithm. Empirical results suggest that EC3 has a transition from satisfiability to unsatisfiability when the number of clauses per variable r exceeds some threshold r^* . This problem is unusual in that numerical experiments can determine its threshold to a greater accuracy than for other variants of 3-SAT; by performing exhaustive search on problems of up to 1700 variables we estimate that $r^* \approx 0.62 \pm 0.01$. Recently it was shown that $r^* \leq 0.644$. Using the method of differential equations, we place a lower bound on r^* 's location by showing when $r < 0.5460$ w.h.p. a random instance of EC3 is satisfiable.

Next, we study road networks as complex networks with an underlying geography.

We find that these networks have topological and geographical scale invariant properties. We employ both primal and dual models of the network. In the primal model, intersections form the nodes and road segments form the edges. In the dual model, roads form the nodes and road intersections form the edges. In the primal model, we find that journeys of widely varying lengths have a scale-invariant structure. In the dual model, the degree distribution of the network is a power law indicating a scale free topology. We give a simple fractal model that not only produces these properties, but also hints at a scaling relation connecting the power-law exponent of the degree distribution to the fractal dimensions of the set of intersections in the plane and along a single road.

Contents

List of Figures	xi
List of Tables	xii
1 Introduction	1
1.1 Phase Transitions: An Overview	1
1.2 Scale Invariance in Networks: An Overview	3
2 Phase Transition in 3-bit Exact Cover	5
2.1 Introduction	5
2.2 The lower bound	7
2.3 Numerical experiments	15
2.4 Conclusion	15
3 Scale Invariance in Road Networks	17
3.1 Introduction	17

Contents

3.2 Primal and Dual Models of the Network	18
3.3 Methodology	20
3.4 Journey Structure	21
3.5 Degree Distribution	23
3.6 A Toy Model	24
3.7 Conclusion	26
References	28

List of Figures

2.1	The largest eigenvalue and clause densities for $r = 0.5460$	14
2.2	The global view and local view of the threshold.	16
3.1	Average distance travelled per step versus total trip distance, up to the fifth largest step for mainland United States, England and Denmark respectively. As we can see each of these steps covers a constant fraction of the journey distance over a wide range of journey distances, from 1.9 miles to approximately 2000 miles for mainland U.S road network.	22
3.2	The cumulative degree distributions of road degree for mainland United States, England, and Denmark. The lines shown suggest the least-squares-fit power law exponents (β) -1.41 for U.S, -1.05 for England and -1.18 for Denmark.	23
3.3	A version of our fractal model used to explain the power-law degree distribution and journey structure. Line-thickness indicates greater road capacity and speed limits. The Sierpinski's Gasket corresponds to recursively subdividing all squares except square 5.	27

List of Tables

2.1	Our Algorithm - SC.	8
3.1	A summary of fraction of path-length travelled by the five largest steps.	22
3.2	Hausdorff dimensions of the population (d_p), intersections along a named road (d_i) and the power-law exponent (α) for different subsets of squares recursively sub-divided.	26

Chapter 1

Introduction

This chapter presents an overview of two interdisciplinary topics on which this thesis is based. The topics are phase transitions in NP-complete problems and scale invariance in networks which are of interest to both physicists and computer-scientists. In the world of physics, phase transitions and scale invariance are closely related, Kenneth Wilson won a Nobel Prize in Physics in 1982 for his work that showed that certain properties of interest near the phase transition become scale invariant [1, 2]. Unfortunately, there is no such relation between the problems we have studied here, and so, in essence, this thesis represents a compilation of work done on two reasonably different problems. We first present an overview of phase transitions in NP-complete problems and then an overview of scale invariance in networks.

1.1 Phase Transitions: An Overview

Phase Transition is a term borrowed from the physics community to indicate a sudden change in a property in random structures (like random boolean formulae or random graphs) as some global parameter (like the clause to variable ratio - r , average degree

Chapter 1. Introduction

- d) is varied. For example, random instances of 3-SAT show a transition from satisfiability to unsatisfiability at a clause to variable ratio $r \approx 4.25$, in random graphs a giant component appears when the average degree $d = 1$. The value of the parameter at which the transition occurs is known as the threshold (r^*, d^*) .

For random instances of NP-complete constraint satisfaction problems, phase transitions provide some insight into both the structure of satisfiable instances and “hardness” of these instances. By hardness, we mean the time-complexity of a complete algorithm like Davis-Putnam-Logeman-Loveland (DPLL) algorithm to determine whether an instance is satisfiable or not. Problem instances located below the threshold, i.e. $r < r^*$, are *under-constrained* and are satisfiable with high probability (w.h.p.), instances located above the threshold, i.e. $r > r^*$, are *over-constrained* and are unsatisfiable w.h.p., and instances located near the threshold, i.e. $r \approx r^*$, are *critically-constrained*, and so the algorithm does a lot of back-tracking before finding either a solution or a contradiction [3]. When $r \ll r^*$, the average-case time-complexity is polynomial, i.e. most instances are “easy” to solve. But for some $r_c < r < r^*$, where r_c is a constant depending on the algorithm used, the average-case time-complexity becomes exponential. As r approaches r^* , the exponent in the time-complexity increases, reaching its peak when $r = r^*$. As r becomes greater than r^* , the time-complexity remains exponential, but the exponent drops off as $1/r^2$. This phenomenon for various algorithms on random instances of 3-SAT is documented in [4].

Existence of phase transition and location of threshold are known rigorously only for relatively “easy” problems. For example, 2-coloring, 2-SAT and 2-XOR-SAT [5]. In all these problems the colorability or satisfiability depends on the presence of a cycle in the constraints which is relatively easy to characterize. For “hard” problems like 3-SAT and 3-Coloring, existence of phase transition is not known rigorously. Approximate locations of the thresholds are computed experimentally. Monasson

and Zecchina [6], using non-rigorous methods of physics, showed 3-SAT has a phase transition at 4.25. In general, most of the work in this area has been in proving rigorous upper and lower bounds on the phase transition assuming its existence. Upper bounds are computed using counting arguments like first moment method [7] or rigorous methods of physics [6, 15]. Lower bounds are computed by analyzing simple linear time algorithms to prove satisfiability with positive probability (w.p.p.) and then using a theorem of Friedgut [22, 26] to obtain a lower bound.

In chapter 2, we study a phase transition in the 3-bit Exact Cover (EC3) problem [16, 17]. Using the method of differential equations, we place a lower bound on the threshold r^* . We also describe experiments conducted to estimate r^* .

1.2 Scale Invariance in Networks: An Overview

Ever since Stanley Milgram's famous "six-degrees of separation" experiment [35], in which he showed that the network of social-contacts has a small diameter i.e. a "small-world" network, it was shown that many social and technological networks of interest like co-authorship and citations in scientific research [8], internet [32], movie actors [9], sexual-contacts [10] etc. were shown to be "small-world" networks. By "small-world", we mean any network whose diameter is proportional to the logarithm of the number of nodes in the network.

Most small-world networks were found to have a power-law degree distribution, $N(d) \sim d^{-\alpha}$, where $N(d)$ is the number of nodes with degree d [31]. Power-laws are interesting for two reasons, first, they are self-similar: if d is rescaled (multiplied by a constant), then $N(d)$ is still proportional to $d^{-\alpha}$, although with a different constant of proportionality [45]. Second, these distributions have "heavy" tails, i.e. these distributions allow for extremely rare nodes with extraordinarily high degree. These two things lead to these networks having a self-similar hierarchical topology with a

Chapter 1. Introduction

few nodes of very high degree (allowed because of the heavy tail), followed by some nodes of slightly less, but still high degree. This structure (due to self-similarity) continues all the way down to nodes with degree one or two. Due to self-similarity, power-law degree distributions lack a characteristic scale, hence networks with power-law degree distributions are called scale-free networks.

Most of the earlier studies of networks focussed exclusively on the topology even when the network of interest was embedded in a geography, e.g. internet [37]. Only recently has the spatial structure of networks been investigated [38, 11, 12]. We study road networks as instances of geo-spatial networks. In particular, in road networks distance can be measured in two ways, topological and physical. Topological distance is the number of roads one has to take to travel from one location to another, while the physical distance is the actual distance travelled along the roads to complete the journey. Also, road networks have a predefined hierarchy of roads ranging from interstate highways to neighborhood streets. Our interest in these networks is to study whether the presence of a hierarchy leads to hierarchical scale-free connectivity of roads and whether the topological and physical distance have any scale-invariant relationship.

In chapter three, we present our work toward answering the above mentioned questions. We not only answer our questions in the affirmative, but also provide a simple model that duplicates the observed behavior.

Chapter 2

Phase Transition in 3-bit Exact Cover

2.1 Introduction

Numerous constraint satisfaction problems are believed to have a “phase transition” in the random case when the ratio r of clauses to variables crosses a critical threshold r^* . For 3-SAT, for instance, this ratio appears to be roughly 4.27; see [13] for a review.

In this paper we study a similar phase transition in 3-bit Exact Cover (EC3) [16, 17]. EC3 is a restriction of the Exact Cover problem (EC). An instance of EC consists of a set $S = \{a_1, a_2, \dots, a_m\}$ and set of subsets of S , $F = \{S_1, S_2, \dots, S_n\}$. The problem is to determine whether there is a cover $C \subseteq F$, such that each element in S is covered exactly once i.e. for each $a_i \in S$ there is exactly one $S_j \in C$ such that $a_i \in S_j$. In EC3, each $a_i \in S$ is restricted to appear in three and only three of the subsets $S_k \in F$.

EC3 is usually formulated as Positive 1-in-3 SAT. Positive 1-in-3 SAT is a variant of 3-SAT, an instance of which consists of a set of boolean variables $V =$

$\{v_1, v_2, \dots, v_n\}$ and a set of clauses $C = \{c_1, c_2, \dots, c_m\}$, where each $c_i \subset V$ and $|c_i| = 3$. Note that the variables appear as positive literals only. A clause is satisfied when one of its variables is set to TRUE and the other two variables are set to FALSE. The problem is to determine whether any of the 2^n assignments to variables in V satisfies all the clauses in C . An instance of EC3 can be transformed into an instance of Positive 1-in-3 SAT by setting $v_i \leftarrow S_i$ and $c_i \leftarrow \{v_i \mid a_i \in S_i\}$. In what follows we use the terminology of Positive 1-in-3 SAT instead of the terminology of EC3.

A related problem is 1-in-3 SAT where variables can appear as negative literals in the clauses. In the case of 1-in-3 SAT the threshold is known exactly [14]. For EC3 an upper bound of $r^* \leq 0.644$ was established in [15]. We provide a lower bound of $r^* > 0.5460$ in this paper.

In addition to the transition phenomenon, our motivation is partly that Farhi et al. recently simulated a quantum adiabatic algorithm [18] on random cases of EC3. They were only able to simulate this algorithm on small numbers of variables (up to 17), but, in this range, the algorithm appeared to work in polynomial time on formulas with a variety of values of r . This is exciting given that EC3 is NP-complete. On the other hand, van Dam and Vazirani [19] showed that such algorithms cannot succeed in polynomial time in the worst case, suggesting that either the experiments in [20] do not capture the asymptotic behavior of the algorithm, or that random formulas are considerably easier than worst-case ones.

An interesting feature of EC3 is that we can carry out exhaustive searches on surprisingly large random formulas. For 3-SAT, while state-of-the-art SAT solvers such as Chaff can solve problems in VLSI verification of up to 1 million variables [21], they can only handle up to few hundred variables for random formulas due to their lack of structure. For EC3, on the other hand, we can solve random formulas of up to about 1700 variables. This allows us to obtain a rather sharp numerical estimate of the threshold r^* .

In the following section we prove a lower bound on the threshold of $r^* > 0.5460$, assuming it exists. More formally, we prove the following theorem:

Theorem 1 *Let ϕ be a Positive 1-in-3 SAT formula consisting of $m = rn$ clauses chosen uniformly with replacement from the $\binom{n}{3}$ possible clauses. If $r < 0.5460$, $\lim_{n \rightarrow \infty} \Pr[\phi \text{ is satisfiable}] = 1$.*

We prove the theorem by analyzing a greedy algorithm using the method of differential equations. For a detailed introduction to the method of differential equations refer to [22]. We conclude by reporting on our numerical experiments, which give an estimate of $r^* \approx 0.62 \pm 0.01$.

2.2 The lower bound

In this section we prove Theorem 1. Before delving into details of the proof, we first describe the mechanics of setting variables in an EC3 formula. We call clauses of length i in the formula “ i -clauses”. 1-clauses are also called unit clauses. Setting a variable to FALSE or TRUE can be viewed as adding a negative or positive unit clause respectively to the formula. Setting a variable v to FALSE replaces each 3-clause $t_i = \{v, x_i, y_i\}$ it appears in with a 2-clause $x_i \oplus y_i$; and replaces each 2-clause $b_i = v \oplus z_i$ it appears in with a positive unit clause z_i . Similarly, setting v to TRUE replaces each 3-clause it appears in with two negative unit clauses $\overline{x_i}, \overline{y_i}$; and replaces each 2-clause it appears in with a negative unit clause $\overline{z_i}$.

We analyze a simple greedy algorithm which is a variant of Unit Clause resolution or UC for short [23]. The algorithm we analyze is known as Short Clause or SC. Algorithms based on UC have so-called “free” and “forced” steps. A free step is one in which the algorithm decides on a variable and the value to which that variable is

```
while there are any unset variables, do {  
    // Free step.  
    if there are any 2-clauses  
        choose a clause  $c$  at random from the 2-clauses  
    else  
        choose a clause  $c$  at random from the 3-clauses  
        choose a variable  $x \in c$  at random  
        set  $x = \text{TRUE}$   
    // Forced steps.  
    while there are unit clauses, satisfy them;  
}
```

Table 2.1: Our Algorithm - SC.

set. Forced steps result from unit propagations due to a variable being set either in a free step or a forced step. It should be clear that variants of UC differ only in free steps. Two of the common ways to pick variables are

1. pick a variable at random,
2. for a fixed i , pick an i -clause at random, then pick a variable at random from one of the i variables in the clause.

We found that setting a variable to TRUE at the free step gives us better lower bounds. The lower bound of $r^* > 0.5460$ was obtained when we picked a variable at random from a 2-clause and set it to TRUE i.e. SC. Our algorithm is shown in table 2.1.

Each iteration of the outer **while** loop (a round) consists of a free step, in which we pick a 2-clause and satisfy it. If there aren't any 2-clauses, we pick a 3-clause at random and satisfy it. A free step is followed by a series of forced steps where we satisfy a cascade of unit clauses. Since resolving a unit clause creates more unit

clauses, the forced steps are described by a branching process. Our first goal will be to show that this branching process will remain subcritical throughout the algorithm. Here, subcriticality means when r is sufficiently small the largest eigenvalue of the transition matrix of the branching process is bounded below 1, and so, the number of variables set in any round will be $O(1)$ w.h.p.

To analyze our algorithm we need to track the change in the number of 2-clauses, the number of 3-clauses and the number of variables set in each round. Note that at the start of the algorithm we have no 2-clauses. It can be shown that after $o(n)$ free steps the number of 2-clauses is w.h.p. greater than zero. The number of 2-clauses returns to zero only after $\Theta(n)$ free steps. A proof of this fact follows from an argument similar to lemma 3 in [24]. The branching process can become supercritical only in the presence of 2-clauses, therefore, it is this phase of the algorithm's execution that we analyze using the differential equations. This means, as far as the analysis is concerned, in the free step we always set a variable in a 2-clause to TRUE. As will be shown later, once the 2-clauses are exhausted the remaining 3-clauses form sparse formula of density less than 0.05. At this density the graph of the clause to variable connectivity of the formula w.h.p. does not contain a giant component (indeed, with positive probability this graph consists of trees only), a simple argument then shows that the remaining formula is satisfiable with positive probability.

The above reasoning shows that the random formula is satisfiable with positive probability. Satisfiability with probability 1 follows from the existence of a non-uniform threshold analogous to those for 3-SAT [26]. In what follows we first introduce some terminology and then describe the branching process. We then note the implications of subcriticality of the branching process on the differential equations that describe the "trajectory" of the algorithm. Finally, we solve the differential equations and obtain a lower bound on r^* .

Let n be the number of variables in the formula. Let $m = rn$ be the number of

Chapter 2. Phase Transition in 3-bit Exact Cover

clauses. Let $T = t \cdot n$ be the number of rounds completed so far. For $i = 2, 3$ let $S_i(T) = s_i(t) \cdot n$ be the number of clauses of length i . Let $X(T) = x(t) \cdot n$ be the number of variables set so far. Let m_T, m_F be the expected number of variables set to TRUE, FALSE respectively in each round (inclusive of the variable set in the free step).

We compute m_T, m_F according to a two-type branching process as analyzed in [27]. The two types here are positive and negative unit clauses. In the free step we set a variable in a 2-clause to TRUE and this forces us to set the other variable in the 2-clause to FALSE. Thus the initial expected population of unit clauses can be represented by a vector

$$p_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (2.1)$$

where the first and second components count the positive and negative unit clauses respectively.

Setting a variable in the free step leads to unit-propagation i.e. forced steps. These forced steps are analyzed using the two-type branching process. We need to determine the transition matrix of the branching process. If X variables have been set so far, the probability of a variable appearing in a given i -clause is $i/(n-X)$. So, setting a variable to TRUE i.e. satisfying a positive unit clause creates, in expectation, $(6S_3 + 2S_2)/(n-X)$ negative unit clauses. Similarly, satisfying a negative unit clause creates, in expectation, $(2S_2)/(n-X)$ positive unit clauses. Thus, we have the following transition matrix M for the branching process

$$M = \frac{1}{n-X} \begin{pmatrix} 0 & 6S_3 + 2S_2 \\ 2S_2 & 0 \end{pmatrix} = \frac{1}{1-x} \begin{pmatrix} 0 & 6s_3 + 2s_2 \\ 2s_2 & 0 \end{pmatrix}. \quad (2.2)$$

Chapter 2. Phase Transition in 3-bit Exact Cover

So, as long as the largest eigenvalue λ_1 of M is less than 1, the expected number of variables set to true or false in each round is given by

$$\begin{pmatrix} m_T \\ m_F \end{pmatrix} = (I + M + M^2 + \dots) \cdot p_0 = (I - M)^{-1} \cdot p_0 \quad (2.3)$$

where I is the identity matrix. Moreover, as long as $\lambda_1 < 1$ throughout the algorithm, i.e. as long as the branching process is subcritical for all x ; m_T and m_F remain $O(1)$ and our algorithm succeeds with positive probability. On the other hand, if λ_1 ever exceeds 1, then the branching process becomes supercritical, the unit clauses proliferate with high probability and the algorithm fails. Note that

$$\lambda_1 = \frac{2}{1-x} \sqrt{s_2(s_2 + 3s_3)} \quad (2.4)$$

Our next step is to write down the expected change in S_2, S_3 and X in a given round as a function of their values at the beginning of the round. We define $\Delta f(T) = f(T+1) - f(T)$. Then:

$$E[\Delta X(T)] = m_T + m_F \quad (2.5)$$

$$E[\Delta S_3(T)] = -(m_T + m_F) \frac{3S_3}{n - X} \quad (2.6)$$

$$\begin{aligned}
 E[\Delta S_2(T)] &= m_F \frac{3S_3}{n-X} - (m_T + m_F) \frac{2(S_2 - 1)}{n-X} - 1 \\
 &= m_F \frac{3S_3}{n-X} - (m_T + m_F) \frac{2S_2}{n-X} - 1 - \frac{2(m_T + m_F)}{n-X} \quad (2.7)
 \end{aligned}$$

During each round, we set $m_T + m_F$ variables inclusive of the variable picked on the free step, and so we get equation 2.5. Any variable set during a round appears, in expectation, in $3S_3/(n-X)$ 3-clauses, and since we set $m_T + m_F$ variables during a round, the expected change in the number of 3-clauses is $-(m_T + m_F) \cdot 3S_3/(n-X)$ (equation 2.6). Among the 3-clauses, those that had a variable in them set to false become 2-clauses, so $m_F \cdot 3S_3/(n-X)$ new 2-clauses are created in each round. At the beginning of a round when a variable is set during the free step a 2-clause is removed leaving $(S_2 - 1)$ 2-clauses. As in the case of 3-clauses, the number of 2-clauses removed by setting $m_T + m_F$ variables during a round is $(m_T + m_F) \cdot 2(S_2 - 1)/(n-X)$. The expected number of 2-clauses removed in each round is $1 + (m_T + m_F) \cdot 2(S_2 - 1)/(n-X)$. Thus the expected change in 2-clauses is given by equation 2.7.

We apply Wormald's Theorem [28] to equations (2.5, 2.6, 2.7) to get a system of differential equations for s_i, x . Wormald's Theorem implies that w.h.p. the random variables $S_i(tn)$ will be within $o(n)$ of $s_i(t) \cdot n$ for all t , where $s_i(t)$ are solutions of:

$$\frac{dx}{dt} = m_T + m_F \quad (2.8)$$

$$\frac{ds_3}{dt} = -(m_T + m_F) \frac{3s_3}{1-x} \quad (2.9)$$

$$\frac{ds_2}{dt} = m_F \frac{3s_3}{1-x} - (m_T + m_F) \frac{2s_2}{1-x} - 1 - \frac{-2(m_T + m_F)}{n(1-x)} \quad (2.10)$$

Note that as long as $\lambda_1 < 1$, the $-2(m_T + m_F)/n(1-x)$ term in equation 2.10 is $o(1)$ and can be ignored. We simplify the above system of differential equations by dividing equations 2.9 and 2.10 by equation 2.8 to eliminate the variable t . We now have:

$$\frac{ds_3}{dx} = -\frac{3s_3}{1-x} \quad (2.11)$$

$$\frac{ds_2}{dx} = \frac{m_F}{(m_T + m_F)} \frac{3s_3}{1-x} - \frac{2s_2}{1-x} - \frac{1}{m_T + m_F} \quad (2.12)$$

The initial conditions are $s_3(0) = r$, $s_2(0) = 0$, even though the differential equations trace the evolution of s_2 and s_3 after a $o(1)$ fraction of the variables have been set. A justification of this is based on an argument similar to one found in section 6 of [22].

When $r = 0.5460$, numerically solving the differential equations gives us, at $x \approx 0.29$, $\max_x(\lambda_1) \approx 0.996 < 1$ (see figure 2.1), and so the branching process remains subcritical. Also, at $x \approx 0.79$, the density of the 2-clauses $s_2(x) = 0$. This means the algorithm succeeds with positive probability in exhausting all the 2-clauses. The density of the remaining 3-clauses is $s_3(x)/(1-x) \approx 0.02$ (see figure 2.1). For EC3 formulas with such low densities, the graph of clause to variable

Chapter 2. Phase Transition in 3-bit Exact Cover

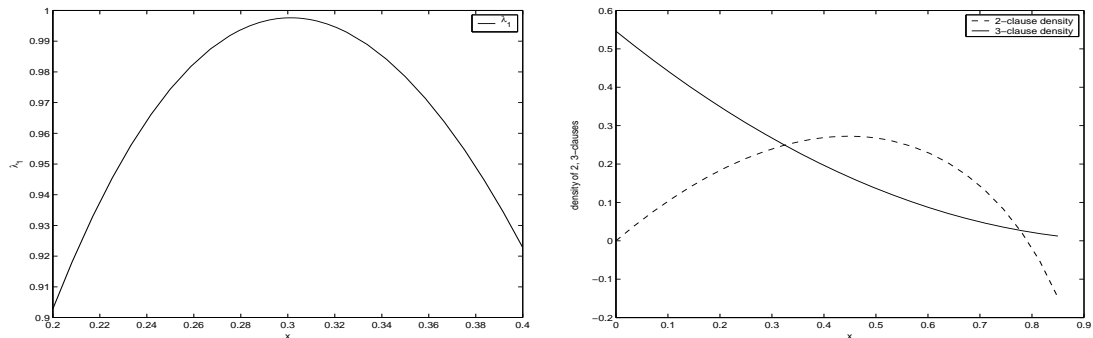


Figure 2.1: The largest eigenvalue and clause densities for $r = 0.5460$.

connectivity (i.e. the graph in which clauses are nodes and clauses that have a variable in common have an edge between them) with positive probability consists of trees only, and so the formula has no “core” in the sense of [15]. The formula can be satisfied by picking a clause in each tree and satisfying it. As a result, the algorithm succeeds with positive probability whenever $r \leq 0.5460$, establishing a lower bound $r^* > 0.5460$.

We also analyzed other algorithms to get lower bounds. Pure UC, i.e. picking a variable at random and setting it to TRUE gave us a lower bound of $r^* > 0.5097$. Picking a 3-clause at random and setting a variable in it to TRUE gave us a better lower bound of $r^* > 0.5386$. The best bound, however, was obtained with SC, as analyzed above. Among the known variants of UC that set only a single variable in free step, SC gives the best lower bound for 3-SAT and 3-Coloring problems [24, 25]. For a detailed analysis of different kinds of free steps available see [22].

2.3 Numerical experiments

In this section we describe the numerical experiments conducted to estimate the threshold of the phase transition in EC3. The clause to variable ratio r^* about which the probability of satisfiability of a random EC3 formula drops from being close to 1 to being close to 0 is called the crossover point or threshold of the phase transition. In what follows, a random EC3 formula has n variables and $m = rn$ clauses.

To estimate the threshold, we measure the probability of satisfiability of a random EC3 formula as a function of r for various system sizes i.e. various values of n , and find a value of r at which these curves appear to intersect. This is a standard approach in finite-size scaling for numerical experiments.

To estimate probability of satisfiability experimentally within an error range of ± 0.01 , we perform 10000 trials. Each trial consists of creating a random EC3 formula and checking whether it is satisfiable or not. To check whether an EC3 formula is satisfiable or not, we convert the EC3 formula into an equivalent 3-SAT formula and use the solver **Satz** [29] to solve it. Probability of satisfiability versus r plots are shown for various values of n in figure 2.2. From the figure we can estimate that $r^* \approx 0.62$.

2.4 Conclusion

We have placed a lower bound of $r^* > 0.5460$ on the threshold of the phase-transition in EC3. Knysh et al. placed an upper bound of $r^* < 0.644$ in [15]. A gap of 0.098 still remains. Our lower bound was obtained by analyzing algorithms that set only one variable either randomly or based on clause length at the free step. Algorithms that pick a variable based on the number of occurrences in the formula give better

Chapter 2. Phase Transition in 3-bit Exact Cover

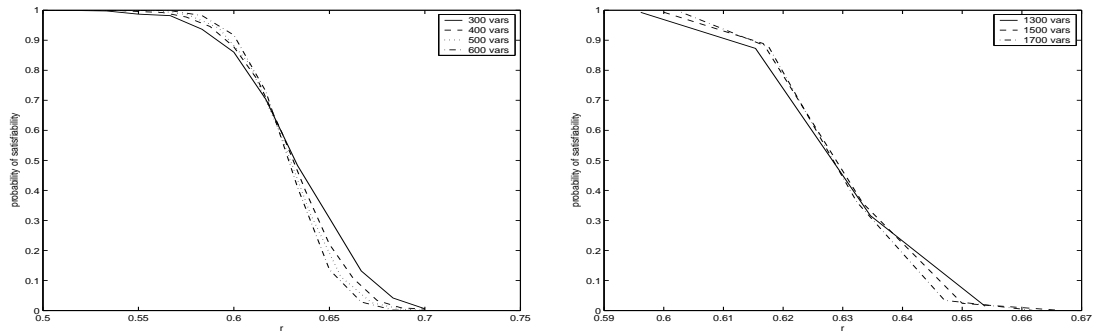


Figure 2.2: The global view and local view of the threshold.

bounds for 3-SAT [30]. Analyzing the performance of a similar algorithm on random instances of EC3 could possibly give us better lower bounds.

Chapter 3

Scale Invariance in Road Networks

3.1 Introduction

The study of complex networks has received much attention from the physics community in the recent past [31, 32]. Most of the earlier networks studied like the network of movie-actors, social-contacts, Internet and world wide web have been studied from a strictly topological point of view paying attention to properties like degree distribution, diameter and clustering coefficient [33, 34]. It was found that many social and technological networks have a power-law degree distribution $N(d) \sim d^{-\alpha}$, where d is the degree of the node and $N(d)$ is the number of nodes with degree d [33, 35, 36, 37].

Geographical networks like road networks introduce metric properties in addition to topological ones. In particular, we can consider distance between nodes in terms of number of edges (topological distance) and also the actual physical distance between them [38]. Also, road networks have a hierarchy of roads like interstate highways, state highways, county roads, city roads and finally neighborhood streets. Given that road networks have these properties not generally found in other networks, our study of road networks focussed on two questions, first, does the presence of a hierarchy of

roads translate into a hierarchical scale-free degree distribution in the dual model. Second, does there exist any scale-invariant relationship between topological and physical (metric) distance.

In this paper we study road networks of mainland United States, England and Denmark. We choose these networks because they serve areas at different scales. Denmark's road network serves a small area (diameter 400 miles), England's road network serves a mid-sized area (diameter 800 miles), while mainland United States road network serves a large area (diameter 3000 miles). We employ both primal and dual models [39, 40]. We find power-law degree distributions in the dual models. In the primal model, we find that journeys of widely varying lengths have a scale-invariant structure. By scale-invariant structure we mean that the fraction of the journey distance covered by each of the five longest roads was constant. Very much in the spirit of models presented in [38, 41] we explain these results using a simple toy model of a self-similar road network.

In what follows we first present primal and dual models, then we present our results on the degree distribution and the journey structure. Finally, we present our toy model.

3.2 Primal and Dual Models of the Network

Geographical networks like road networks are embedded in a Euclidean space, and so they have metric properties in addition to topological properties. The standard models employed in the study of these networks are primal and dual models. In the primal model, road intersections are nodes and road segments between the intersections are the edges [39]. Dual models are created to provide a better view of information not directly accessible in the primal model. Dual models are created based on quantities of interest like named roads [44], lines of sight [42], and angles

of incidence of road segment at the intersection [40]. We use a dual model based on named roads, where named roads become nodes and road intersections are the edges.

The primal model directly represents the geography of the network, and therefore is the model of choice for representing shortest paths and travel times. However, it gives little opportunity to explore scale-free properties since most intersections have a degree of 4 (moreover, the road network is mostly planar and the average degree of any planar graph is 6) [38]. Another drawback of the primal model is that it lacks the concept of a named road. A single road (say Main St.) spans multiple intersections, and it is represented as multiple edges along a path. The fact that these edges represent the same road is missing in the model. Also, a named road becomes a single step in driving directions (such as “stay on Main St. for 10 mi, then make a left turn at Baker St.”) provided by services such as Mapquest, Yahoo and Google [43].

By treating named roads as nodes our dual model captures road connectivity information better. For example, Interstate 5 intersects both Interstate 40 and Interstate 90. As the degree of a node in this model is free from geographical constraints it becomes possible to see a broad-degree distribution. All the networks under study show a broad degree distribution, but only the United States road network is scale-free. The drawback of the dual model is that it is a purely topological model and distance is not well represented. For example, Interstate 40 and Interstate 90 are separated by over a 1000 miles along Interstate 5.

We use the dual model for our studies on degree distribution. Data from both the primal and dual models is used in our study of journey structure, as the quantity of interest is the distance travelled along a named road. The following section describes our data collection methods and discusses the pros and cons of the methods employed.

3.3 Methodology

We obtained data on the primal and dual models of the road networks studied by querying Mapquest.com. For the data on United States, we queried Mapquest for driving directions between 200000 random pairs of zipcodes, while for England and Denmark we queried Mapquest for driving directions between 25000 random pairs of city/towns.

Driving directions are obtained on Mapquest by providing the source and destination addresses. When provided partial information, e.g. zipcode/city/town only, Mapquest picks a default address in that location. This default address is unique for a given location. The driving directions contain a list of roads together with the distances to be travelled along these roads to travel from one location to another. We call this list of roads a “path”. Note that this “path” also happens to be a path in the dual model between the source and the destination. We used the path lists to construct a partial dual-graph of the road networks.

A good thing about our approach is that zipcode/city/towns are distributed according to population ¹, and as a result our data is biased toward travel between more densely populated areas. For example, for mainland United States the average path-length was 1003 miles, reflecting the fact that significant fractions of United States population reside on the east and west coasts.

The downsides of our approach are, first, we treat travel between any two zipcodes/cities/towns as equally likely; this is not true as shorter journeys are much more likely than longer ones, e.g. travel between a city and its suburbs is much more likely than travel between cities. Second, since we obtained paths between zipcodes/cities/towns we miss the connectivity of the roads in the zipcode/city/town.

¹For those interested, this data can be obtained from <http://ftp.census.gov/geo/www/gazetteer/places.html>

So, we have a high-level picture of the road network.

Finally, Mapquest uses a proprietary algorithm to determine paths optimized for travel time. The algorithm employed by Mapquest seems to pick highways over local streets. There is no guarantee that the algorithm used by Mapquest does result in fastest paths. So, our data does not capture full topology of the road networks, instead it captures the connectivity of fastest paths according to Mapquest.

3.4 Journey Structure

In what follows, we use the following terminology. We call travelling along a particular named road in a path a “step,” the distance travelled along the named road is termed “step-length.” We call the sum of all step-lengths the path-length.

Road networks have a hierarchy of roads like interstate highways, state highways, county roads, city roads and finally neighborhood streets. This hierarchy also represents a descending order of both road capacity and speed limits. Since most of the sources and destinations of interest lie on neighborhood streets, and we want to minimize travel time, we expect a journey to originate at neighborhood streets, rise up the hierarchy, take a few large steps on the interstate highways, and then descend back down through the hierarchy as we approach the destination. Furthermore, we expect these large steps to cover a constant fraction of the path-length.

We wanted to confirm whether the largest steps indeed covered constant fractions of the path-length. To that end, we measured the fraction of the path-length covered by the five largest steps. We not only confirmed our hypothesis, but also found an interesting pattern in the fraction of path-length covered by the five largest steps. For United States we found that the largest step covered 40% of the entire distance, while the second largest step covered 20%, third largest step covered 13% and so on.

Chapter 3. Scale Invariance in Road Networks

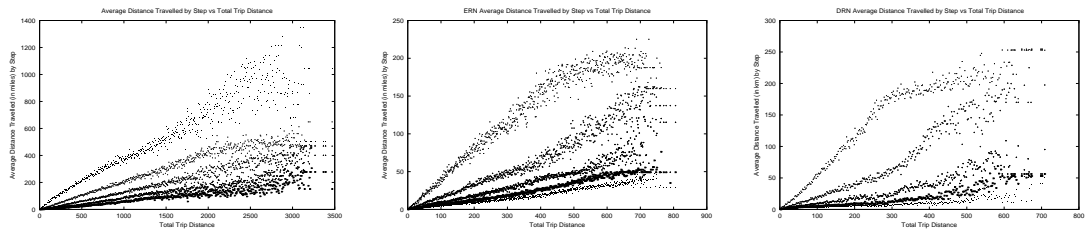


Figure 3.1: Average distance travelled per step versus total trip distance, up to the fifth largest step for mainland United States, England and Denmark respectively. As we can see each of these steps covers a constant fraction of the journey distance over a wide range of journey distances, from 1.9 miles to approximately 2000 miles for mainland U.S road network.

The results for England and Denmark are similar. The data is summarized in table 3.1 and figure 3.1.

Distribution	1 st	2 nd	3 rd	4 th	5 th
U. S.	0.40	0.20	0.13	0.08	0.05
England	0.38	0.15	0.08	0.06	0.03
Denmark	0.60	0.20	0.09	0.04	0.03

Table 3.1: A summary of fraction of path-length travelled by the five largest steps.

We find that the five largest steps together covered 86% of the path-length in the United States road network, the results for England and Denmark are similar. Also, on the average, the largest step covers between 33-66% of the path-length, while the second largest step covers as much distance as between 33-66% of the distance covered by the longest step. This pattern continues through the fifth largest step (figure 3.1). This is interesting as it indicates some self-similarity in the geographic structure of the road network.

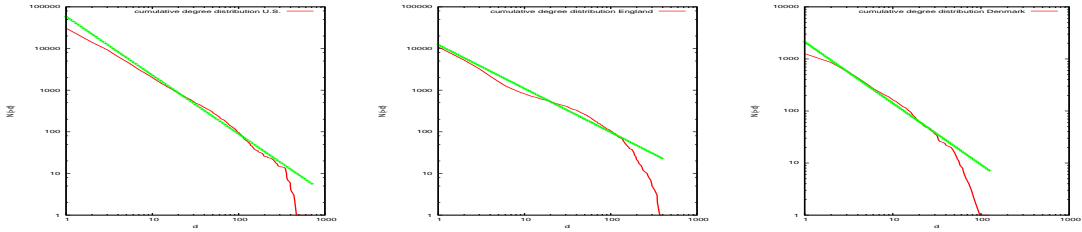


Figure 3.2: The cumulative degree distributions of road degree for mainland United States, England, and Denmark. The lines shown suggest the least-squares-fit power law exponents (β) -1.41 for U.S, -1.05 for England and -1.18 for Denmark.

3.5 Degree Distribution

We test whether the dual models of the road networks have power-law degree distributions of the form $N(d) \sim d^{-\alpha}$, where d is the degree of a node and $N(d)$ is the number of nodes of degree d . Our goal is to estimate the power-law exponent α . A standard method of estimating α is to plot the cumulative degree distribution $N(> d) \sim d^\beta$ on a log-log plot, β is the slope of the least-squares-fit line through the data. Power-laws in the real world suffer from finite-size effects and so we ignore some of the tail data in our estimation of β . The cumulative degree distributions and the suggested power-law exponents are shown in figure 3.2. The estimated exponents α for mainland United States, England and Denmark are 2.41, 2.05 and 2.18 respectively.

The power-law exponents found vary from country to country. A full explanation of the factors affecting the exponents is beyond the scope of this work. We will show below, using our toy model that the power-law exponent is affected by the fractal dimension of the distribution of road intersections in the plane and along a single road.

We also notice that, while all the networks show a broad degree distribution,

only the United States road network seems to be convincingly scale-free over 2.5 decades. The networks of England and Denmark are scale-free over 1.5 decades only. Jiang and Claramunt [44] and Rosvall et al. [43] found broad degree distributions on the named road based dual models at the city level. Both their data and ours (England & Denmark) show power-law behavior, but only over a short range of roughly 1.5 decades, and so these relatively small samples do not give a conclusive power-law; but the United States road network does. The road networks of England and Denmark have fewer levels of hierarchy than the mainland United States road network. Looking at the entire United States road network, from Interstate highways all the way down to local streets allows us to see a power-law over a wider range of degrees. This suggests that dynamics of road planning and building do lead to scale-free structure when carried out on a wide enough range of scales.

We next present a self-similar toy model that attempts to explain the power-law degree distribution found in the dual model and the hierarchical structure of the journeys found in the primal model.

3.6 A Toy Model

The self-similar toy model we present is a generalization of Sierpinski's gasket [45] (shown in figure 3.3). This model can account for both the power-law degree distributions and the hierarchical journey structure. Moreover, it hints at a relationship between the fractal dimension of the road intersections on the plane (d_p) and along a named road (d_i), and the power-law exponent (α). The main drawback of our model is that unlike models presented in [38, 41] this model is structured and does not involve any constraint satisfaction. Also, our model does not take into account the distribution of population, see [46, 47] for a study of fractal dimension of population.

To create a road network according to our model, we start with the outer four

edges of a unit square. These edges represent roads at zeroth level and each has a unit length. We have intersections at the four corners of the unit square. We then proceed to sub-divide the unit square into k^2 squares, where k is an integer. The roads that create these k^2 squares are first level roads each having length of $\frac{1}{k}$. Among the k^2 squares, we pick a subset (see figure 3.3b) and sub-divide it exactly as the original square. This process can be continued to as many levels as desired. Edges created at each level are treated as named roads. The roads created at the zeroth level are the fastest roads (analogous to interstate highways), roads created at the first level are slower than the roads at zeroth level (analogous to state highways) and so on. Observe that in this model the road intersections are fractally distributed both over the unit square and along a named road.

A journey is defined as travel from one intersection to another. As in our variant of the primal model, journey along a named road is treated as a step. We find that the observed journey structure is similar to the ones we observed for the three road networks studied.

As in the dual model when we treat named roads in this model as nodes and intersections between roads as edges, we obtain power-law degree distributions for some choice of subsets. The power-law exponent(α) is determined as:

$$\alpha = 1 + \frac{\log r(k)}{\log d(k)}, \quad (3.1)$$

where $r(k)$ is the number of roads at the k^{th} level and $d(k)$ is the degree accumulated by the zeroth level roads when we have k levels of sub-division. Table 3.2 shows the fractal(Hausdorff) dimension of the road intersections over the unit square (d_p), the fractal dimension of intersections along a named road (d_i) and the power-law exponent (α) for different choices of squares recursively sub-divided.

In the data in table 3.2, we find the following scaling relation between d_p , d_i and α :

Sq. Filled	d_p	d_i	α
all	$\log_3 9$	$\log_3 3$	$1 + \log_3 9$
all - 5	$\log_3 8$	$\log_3 3$	$1 + \log_3 8$
1,3,5,7,9	$\log_3 5$	$\log_3 2$	$1 + \log_2 5$
1,3,7,9	$\log_3 4$	$\log_3 2$	$1 + \log_2 4$

Table 3.2: Hausdorff dimensions of the population (d_p), intersections along a named road (d_i) and the power-law exponent (α) for different subsets of squares recursively sub-divided.

$$\alpha = 1 + \frac{d_p}{d_i}. \quad (3.2)$$

We can derive this relation, at least in our toy model, using a simple counting argument. The number of roads at the k^{th} level, $r(k)$, grows exponentially as the number of squares recursively sub-divided, therefore it is related to d_p as $r(k) \sim \rho^{d_p k}$. ρ is the scaling factor for the population distribution and the intersection distribution, here $\rho = 3$. Similarly, the degree accumulated by the zeroth level roads, $d(k)$, grows exponentially as the number of intersections added at each level, and so it is related to d_i as $d(k) \sim \rho^{d_i k}$. Substituting these relationships in equation 3.1 gives us equation 3.2.

3.7 Conclusion

We have shown that road networks possess both topological and geographical scale-invariance. The road-connectivity graphs have power-law degree distributions and journeys of varying lengths have a similar structure. Also, a highly idealized fractal model that produces these properties was presented.

Few questions still remain, the impact of Mapquest's routing algorithm on the

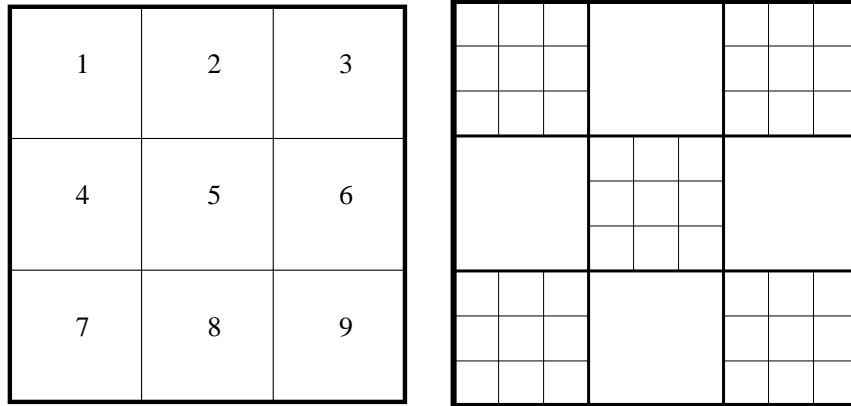


Figure 3.3: A version of our fractal model used to explain the power-law degree distribution and journey structure. Line-thickness indicates greater road capacity and speed limits. The Sierpinski's Gasket corresponds to recursively subdividing all squares except square 5.

connectivity graphs and journey structure remains to be studied. Our toy model suggests a relationship between the power-law exponents of road-connectivity graphs and the fractal dimension of the intersection distribution over the unit square and along a named road. We believe confirming this relationship with real measurements of these fractal dimensions is an interesting direction for future work.

References

- [1] K. Wilson, The Renormalization Group and Critical Phenomena I: Renormalization Group and the Kadanoff Scaling Picture, *Phys. Rev. B* 4 (1971) 3174.
- [2] K. Wilson, The Renormalization Group and Critical Phenomena II: Phase Space Cell Analysis of Critical Behavior, *Phys. Rev. B* 4, (1971) 3184.
- [3] D. Mitchell, B. Selman and H. Levesque, Hard and Easy Distributions of SAT problems, *AAAI* (1992) 459-465.
- [4] C. Corfa, D. Demopoulos, A. S. M. Aguirre, D. Subramanian and M. Vardi, Random 3-SAT: The Plot Thickens, *LNCS* 1894 (2000), 143-159.
- [5] V. Chvatal and B. Reed, Mick gets some (the odds are on his side), *FOCS* (1992), 620-627.
- [6] O. Martin, R. Monasson and R. Zecchina, Statistical mechanics methods and phase transitions in optimization problems, *TCS* 265 (2001), 3-67.
- [7] M. Molloy and B. Reed, *Graph Colouring and the Probabilistic Method*, Springer 2002.
- [8] M. E. J. Newman, Coauthorship networks and patterns of scientific collaboration, *Proc. Natl. Acad. Sci. USA* 101 (2004), 5200-5205.
- [9] The Oracle of Bacon at Virginia, <http://www.cs.virginia.edu/oracle>.
- [10] F. Liljeros, C. R. Edling and L. Amaral, Sexual Networks: implications for the transmission of sexually transmitted infections, *Microbes and Infection* 5 (2003), 183-189.
- [11] M. T. Gastner and M. E. J. Newman, The Spatial Structure of Networks, [cond-mat/0409702](http://arxiv.org/abs/cond-mat/0409702).

References

- [12] Spatial Small Worlds: New geographic patterns for an information economy, [cond-mat/0310426](#).
- [13] *Theoretical Computer Science* 265 (2001), Special Issue on NP-Hardness and Phase Transitions.
- [14] D. Achlioptas, A. Chtcherba, G. Istrate and C. Moore, The phase transition in 1-in- k SAT and NAE 3-SAT, *SODA* (2001) 721-722.
- [15] S. Knysh, V.N. Smelyanskiy and R.D. Morris, Approximating satisfiability transition by suppressing fluctuations, [quant-ph/0403416](#).
- [16] A. M. Childs, E. Farhi and J. Preskill, Robustness of adiabatic quantum computation, *Physical Review A*, 65 (2002), 012322.
- [17] Wenjin Mao, Quantum Algorithm to Solve Satisfiability Problems, [quant-ph/0411194](#).
- [18] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren and D. Preda, A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem, *Science* 292 (2001) 472-474, [quant-ph/0104129](#).
- [19] W. van Dam, M. Mosca and Umesh Vazirani, How Powerful is Adiabatic Quantum Computation? *FOCS* (2001) 279-287, [quant-ph/0206003](#).
- [20] E. Farhi, J. Goldstone and S. Gutmann, A Numerical Study of the Performance of a Quantum Adiabatic Evolution Algorithm for Satisfiability, [quant-ph/0007071](#).
- [21] S. Malik, M. W. Moskewicz, C. F. Madigan, Y. Zhao and L. Zhang, Chaff: Engineering an Efficient SAT Solver, *DAC* (2001).
- [22] D. Achlioptas, A Survey of Lower Bounds for Random 3-SAT via Differential Equations, *Theoretical Computer Science* **265**, Special Issue on NP-Hardness and Phase Transitions (2001).
- [23] M.-T. Chao and J. Franco, Probabilistic Analysis of a generalization of the unit-clause literal selection heuristics for the k -satisfiability problem, *Information Science* 51 (3) (1995) 289-314.
- [24] D. Achlioptas and M. Molloy, "The Analysis of a List Coloring Algorithm on a Random Graph." *FOCS* (1997).
- [25] A. Frieze and S. Suen, Analysis of two simple heuristics on a random instance of k -SAT, *J. Algorithms*, 20 (2) (1996) 312-355.

References

- [26] E. Friedgut, Sharp thresholds of graph properties, and the k-SAT problem, J. Amer. Math. Soc. 12 (1999) 1017-1054.
- [27] D. Achlioptas and C. Moore, Almost all graphs with average degree 4 are 3 colorable, STOC (2002) 199-208.
- [28] N. Wormald, Differential Equations for random processes and random graphs, Annals of Applied Probability 5 (4) (1995) 1217-1235.
- [29] Chu Min Li and Anbulagan, Heuristics based on unit-propagation for satisfiability problems, IJCAI (1997).
- [30] A.C. Kaporis, L. M. Kirousis and E. M. Lalas, The Probabilistic Analysis of a Greedy Satisfiability Algorithm, 10th Annual European Symposium on Analysis of Algorithms (2002).
- [31] R. Albert and A.-L. Barabasi, Statistical mechanics of complex networks, Reviews of Modern Physics, 74, 47-97 (2002).
- [32] A.-L. Barabasi, The Physics of the Web, Physics World, 14(7), 2001.
- [33] Watts, D.J., Small Worlds, Princeton University Press (Princeton), 1999.
- [34] M.E.J. Newman, Models of the Small World, J. Stat. Phys., 101, 819-841, 2000.
- [35] Milgram, S., Psychology Today 1(60), 1967.
- [36] A.-L. Barabasi, Linked, Plume Books (Penguin), 2002.
- [37] M. Faloutsos, P. Faloutsos and C. Faloutsos, Proc. ACM SIGCOMM Comput. Commun. Rev., 29(251), 1999.
- [38] Gastner, M.T., Newman, M.E.J., Spatial Structure of Networks, cond-mat/0407680.
- [39] S. Porta, P. Crucitti, V. Latora, The Network Analysis of Urban Streets : A primal approach. cond-mat/0506009.
- [40] Porta, S., Crucitti, P., Latora, V., The Network Analysis of Urban Streets : A dual approach, cond-mat/0411241.
- [41] A. Fabrikant, E. Koutsoupias and C. H. Papadimitriou, Heuristically optimized trade-offs: a new paradigm for power-laws in the internet, ICALP 2002, LNCS 2380, 110-122.

References

- [42] W. Hillier, The Common Language of Space, <http://www.spacesyntax.org/publications/commonlang.html>.
- [43] M. Rosvall, A. Trusina, P. Minnhagen, and K. Sneppen, Networks and Cities : An Information Perspective, `cond-mat/0407054`.
- [44] B. Jiang, C. Claramunt, Topological Analysis of Urban Street Networks, *Environment and Planning B* 31, 151-162, 2004.
- [45] M. Schroeder, *Fractals, Chaos, Power Laws : Minutes from an infinite paradise*, W.H. Freeman (1992).
- [46] J. Tang, Evaluating the relationship between urban road pattern and population using fractal geometry, *UCGIS* 2003.
- [47] J. Ozik, B. R. Hunt and E. Ott, Formation of Multifractal Population Patterns from Reproductive Growth and Local Resettlement, `nlin/0502008`.