Dissertation

A Scale-Invariant Spatial Graph Model

ausgeführt zum Zwecke der Erlangung
des akademischen Grades eines Doktors der Naturwissenschaften

unter der Leitung von

O. Univ. Prof. Dipl.-Ing. Dr. techn. Andreas U. Frank
Department für Geodäsie und Geoinformation (E120)

eingereicht an der technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Franz-Benjamin Mocnik
Matr.-Nr. 1229696
Fasangasse 50/11, 1030 Wien

Wien, am 1. Dezember 2015

Kurzfassung der Dissertation

Information wird räumlich genannt, wenn sie Referenzen zum Raum beinhaltet. Die vorliegende Dissertation zielt darauf ab, die Charakterisierung räumlicher Information auf ein strukturelles Level zu heben. Toblers erstes Gesetz der Geographie und die Skaleninvarianz werden weithin zur Charakterisierung räumlicher Information verwendet. Ihre formale Beschreibung basiert jedoch auf expliziten Referenzen zum Raum, was einer Verwendung für die strukturelle Charakterisierung räumlicher Information entgegensteht. Der Autor führt daher ein Graphenmodell ein, welches im Falle einer Einbettung des Graphen in einen Raum typische Eigenschaften räumlicher Information aufweist, d. h. unter anderem Toblers Gesetz befolgt und skaleninvariant ist. Das Graphenmodell weist die Auswirkungen dieser typischen Eigenschaften auf seine Struktur auch dann auf, wenn es als abstrakter Graph interpretiert wird. Daher ist es zur Diskussion dieser typischen Eigenschaften auf einem strukturellen Level geeignet.

Ein Vergleich des Modells mit verschiedenen räumlichen und nicht-räumlichen Datensätzen in der vorliegenden Dissertation legt nahe, dass räumliche Datensätze durch eine gemeinsame Struktur gekennzeichnet sind, weil die betrachteten räumlichen Datensätze im Gegensatz zu den nicht-räumlichen Gemeinsamkeiten mit dem Modell aufweisen. Dies lässt das Konzept einer räumlichen Struktur sinnvoll erscheinen. Das eingeführte Modell ist ein Modell dieser räumlichen Struktur. Die Dimension des Raumes wirkt sich auf räumliche Information und somit auch auf die räumliche Struktur aus. Die Dissertation untersucht, wie die Eigenschaften des Modells, insbesondere im Falle einer Gleichverteilung der Knoten im Raum, von der Dimension des Raumes abhängen und zeigt, wie eine Schätzung der Dimension aus der räumlichen Struktur eines Datensatzes gefolgert werden kann.

Die Ergebnisse der Dissertation, insbesondere das Konzept einer räumlichen Struktur und das Graphenmodell, stellen einen grundlegenden Beitrag für die Diskussion räumlicher Information auf einem strukturellen Level dar: Auf räumlichen Daten operierende Algorithmen können unter Berücksichtigung der räumlichen Struktur verbessert werden; eine statistische Evaluation von Überlegungen zu räumlichen Daten wird möglich, da das Graphenmodell beliebig viele Testdatensätze mit kontrollierbaren Eigenschaften generieren kann; und das Erkennen von räumlichen Strukturen sowie die Schätzung der Dimension und weiterer Parameter kann zum langfristigen Ziel beitragen, Daten mit unvollständiger oder fehlender Semantik zu verwenden.

Abstract of the Thesis

Information is called spatial if it contains references to space. The thesis aims at lifting the characterization of spatial information to a structural level. Tobler's first law of geography and scale invariance are widely used to characterize spatial information, but their formal description is based on explicit references to space, which prevents them from being used in the structural characterization of spatial information. To overcome this problem, the author proposes a graph model that exposes, when embedded in space, typical properties of spatial information, amongst others Tobler's law and scale invariance. The graph model, considered as an abstract graph, still exposes the effect of these typical properties on the structure of the graph and can thus be used for the discussion of these typical properties at a structural level.

A comparison of the proposed model to several spatial and non-spatial data sets in this thesis suggests that spatial data sets can be characterized by a common structure, because the considered spatial data sets expose structural similarities to the proposed model but the non-spatial data sets do not. This proves the concept of a spatial structure to be meaningful, and the proposed model to be a model of spatial structure. The dimension of space has an impact on spatial information, and thus also on the spatial structure. The thesis examines how the properties of the proposed graph model, in particular in case of a uniform distribution of nodes in space, depend on the dimension of space and shows how to estimate the dimension from the structure of a data set.

The results of the thesis, in particular the concept of a spatial structure and the proposed graph model, are a fundamental contribution to the discussion of spatial information at a structural level: algorithms that operate on spatial data can be improved by paying attention to the spatial structure; a statistical evaluation of considerations about spatial data is rendered possible, because the graph model can generate arbitrarily many test data sets with controlled properties; and the detection of spatial structures as well as the estimation of the dimension and other parameters can contribute to the long-term goal of using data with incomplete or missing semantics.

Mathematics Subject Classification (2000): primary 05C82, 91D25; secondary 05C42, 68R10, 91D30

Keywords: spatial structure; spatial data; spatial information; Tobler's law; principle of least effort; graph models; spatial networks; small world networks; complex networks; scale invariance; human activities

to my wonderful family

Preface

> *Les photographes s'occupent des choses qui diſparaissent continuellement et quand ils ont diſparu là n'eſt aucune adaptation sur terre qui peut les faire revenir encore.*
>
> —**Henri Cartier-Bresson**
> french photographer
> (1908–2004)

As a young child, I wanted to become an inventor: I invented rotating rockets that need less fuel, because their angular momentum was preventing them from tilting; invented new advanced arithmetic operations that generalize existing ones; and I invented circuit boards for the independent control of multiple model railway locomotives. Every time I had a new idea, it turned out that the invention was great but already existed. When I went into secondary school, I recognized that inventing existing things is no pleasure and that insights go before an invention. That was when I decided to become a scientist.

I strove for insights, and the philosophical point of view became more and more important to me. I had to decide whether I should proceed with philosophy or natural sciences. One of my greatest teachers, Florian Pop, asked me whether I would like to think about these fundamental philosophical questions, possibly without gaining results, or whether I would like to derive results that turn out to work, even when a philosophical justification is missing. I decided for the latter and became a pragmatic scientist, but never lost the interest in these fundamental philosophical questions completely.

Which topics should I conduct research on? There are many more interesting research topics than I will, in my whole life, have the chance to pay attention to. My decision to become a scientist is strongly linked with the desire to gain insights, and insights itself can be gained by the exploration of principles. The more fundamental the principles are, the more extensive is the scope of the insights. It seems thus logical to explore fundamental laws and principles of a topic that appears to be important to many different fields of science and in many contexts. Spacetime is such a topic.

Space and time have always been incredibly fascinating to me, because all matter is bound to space and time, independent of where and when it exists; there is yet no obvious reason for why matter always has a location. Spacetime is a physical phenomenon, a geographical phenomenon, a psychological phenomenon, and it is a social phenomenon. As a scientist, I am fascinated by the fact that these phenomena nevertheless trace back to a common core – this is why we denote these phenomena by the same name. I hence decided to explore the laws and principles of space and time.

I have a passion for elegant things. Spacetime can be described by a very simple structure, and many theories about spacetime are thus widely regarded as elegant. Spacetime is, moreover, fundamental to many elegant things including art: everything we do, even the most aesthetic and elegant things, require space and time to exist, independent of our culture and language. Space and time do not only render elegant things possible, but space and time are as well limiting art, as becomes apparent in the epigraph: art cannot overcome the influence of space and time, because also art happens in space and time. To pay homage to the role of space and time in art, each chapter opens with an epigraph dedicated to an art.

My research initially focussed on public transport, which happens in space and time. This example suggests an algebraic modelling, because it is very specific and we can understand how many aspects of this example relate. I used category theory, algebraic structures and monoidal homology to model this example but could draw only few general theoretical conclusions from this modelling. I hence extended the focus to spatial and temporal information and started using human activities and public transport only as examples of such information.

During my research on spatial information, I read hundreds of papers, took hundreds of notes, discussed numerous ideas with colleagues and worked thousands of hours. One day, there was the moment every researcher yearns for, the εὕρηκα (eureka) moment: an idea emerged of how to characterize and model spatial information. Ideas often turn out to have a major drawback and are thus, sooner or later, rejected. This idea however turned out to be the right one – no serious drawback appeared, it is captivatingly simple and yet has a wide scope. I am more than happy to present the idea and its context in this thesis, with the aim to impart to the reader some of the model's elegance.

Conducting research successfully and writing a thesis is a long process, which is the result of education, a multitude of discussions, research collaborations and manifold influences. I would like to express my sincere gratitude to all these researchers who have taught me, shared insights and influenced me over the years.

In particular, I would like to express my deepest gratitude to Andrew U. Frank and Werner Kuhn for advising me during my doctoral studies, especially for forming the way I understand research in geographical information science. My thanks also goes to Georg Gottlob that he agreed to review my thesis.

Florian Pop and Jakob Stix are amongst the best teachers I had in my life. They fundamentally influenced my view of science, on abstractions and on pragmatic research. I am grateful for their guidance and support during my first years of my studies, and for their influence that is still present.

I would like to express my gratitude to Simon Scheider for the discussions on cybernetics, grounding and other philosophical topics. These discussions opened my mind to new ideas. I also express my gratitude to Tomi Kauppinen who gave me a pragmatic view on how to do science and successfully publish in geographical information science.

Benedikt Gräler gave me an understanding of how mathematicians can survive in applied science by communicating the results in suitable ways. My thanks goes to him for providing this practical help as well as for discussing mathematical aspects of models and theories in the field of geographical information science.

My thanks also goes to Paolo Fogliaroni, who enhanced my understanding of qualitative aspects of spatial information and related algorithms. I would like to thank Jens Knoop for the discussions about functional languages and the topics of this thesis.

Special thanks are due to my colleagues of the Research Group Geoinformation at the Vienna University of Technology and of the former Münster Semantic Interoperability Lab of the Institute for Geoinformatics at the University of Münster. I participated in many discussions with the researchers of these groups, which broadened my knowledge in the field of geographical information science.

My final thanks are due to my family Gabriele, Hans-Jürgen and Paulus-Johannes as well as to Amelie Kampelmühler for their love, their deep understanding and their continuous support, all the nice words and their endless encouragement. Without you, I would never have completed my thesis in the way I did.

Franz-Benjamin Mocnik
Vienna, December 2015

Contents

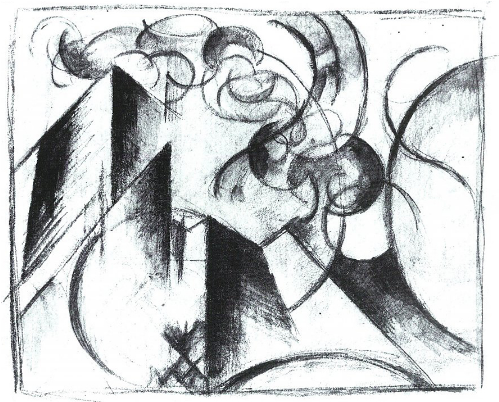
List of Symbols

S	state space	16
A	set of activities	16
$\mathcal{G}(S, A)$	graph representation of human activity systems, consisting of states S and activities A	16
$\mathcal{G}(S, A)/\sim$	graph representation $\mathcal{G}(S, A)$ collapsed by an equivalence relation $\sim$	30
$B_n(p, r)$	n-dimensional ball with radius r centred in p	46
$\mathrm{Vol}_n(r)$	volume of an n-dimensional ball with radius r	46
$d(\cdot, \cdot)$	metric of a vector space	128
$\tau\colon V \to V$	scale transformation	48
$G = (N, E)$	graph consisting of nodes N and edges E	128
$\delta(\cdot, \cdot)$	(undirected) distance in a graph	130
$B_G(p, r)$	ball with radius r centred in p in a graph G	55
$S_G(p, r)$	sphere with radius r centred in p in a graph G	72
$G_0 \subset \ldots \subset G_{s-1}$	series of subgraphs	57
$c_{\text{density}}(G)$	density of a graph G	60
$c_{\text{total density}}(G)$	total density of a graph G	60
$c_{\text{diameter}}(G)$	diameter of a graph G	74
$c_{\text{average shortest path length}}(G)$	average shortest path length of a graph G	75
$c_{\text{max cliques}, \kappa}(G)$	number of maximal cliques with κ nodes in a graph G	76
$c_{\text{subgraphs of min degree } \kappa}(G)$	number of maximal subgraphs of minimal degree κ in a graph G	62
$c_{\text{degree}, \kappa}(G)$	κ-th degree coefficient of a graph G	64
$c_{\text{max degree}}(G)$	maximal degree coefficient of a graph G	64
$\sigma^3(G, p)$	arithmetic mean of the three maximal volumes of spheres centred in p in a graph G	74

$\sigma^3(G)$	arithmetic mean of $\sigma^3(G, p)$ for all nodes p in a graph G	74
$c_{\text{eigen}}(G)$	dominant eigenvalue of the simple undirected adjacency matrix of a graph G	64
$c_{\text{centrality}}(G)$	centrality coefficient of a graph G	67
$c_{\text{clustering}}(G)$	clustering coefficient of a graph G	67
$c_{\text{diversity}}(G)$	diversity coefficient of a graph G	70
ρ	density parameter	44
S	generating set	44
m	minimal dimension	88
$\mathcal{M}_\rho(S, V)$	scale-invariant spatial graph (SISG) model for density parameter ρ and generating set $S \subset V$	44
$\mathcal{M}_\rho(S)$	SISG model for density parameter ρ and generating set $S \subset V$ for some vector space V	45
$\mathcal{M}_\rho^m(S, V)$	SISG model for density parameter ρ, minimal dimension m and generating set $S \subset V$	88
$\mathcal{M}_\rho^m(s)$	SISG model for density parameter ρ and a generating set of s randomly distributed points with uniform distribution in the m-dimensional unit ball	53
$\hat{x}$	estimate of a value x	109

1

Introduction



—**Franz Moritz Wilhelm Marc**
german painter
(1880–1916)

Numerous aspects of spatial information have been examined, but yet, a structural description is missing. Structural realism suggests that a long-lasting view of spatial information could be gained by focusing on its structure. This thesis aims at deciding whether spatial information can be characterized by a common structure, with the aim of contributing to the long term goal of a transdisciplinary concept of space and spatial information. The thesis' main contributions include the concept of spatial structure and the scale-invariant spatial graph model.

A very short overview on this topic has been provided by the author of this thesis at the Vienna Young Scientists Symposium (Mocnik 2015). A more detailed discussion was published by Mocnik et al. (2015) at the 12th Conference On Spatial Information Science. Parts of the thesis are based on these papers.

This chapter begins with a discussion of the concepts necessary to formulate the hypotheses: the concepts of space, time and structure are discussed (section 1.1), and the concept of human activities is introduced (section 1.2). Tobler's first law of geography is reviewed as a typical property of spatial information (section 1.3).

We hypothesize that spatial data sets share, beside Tobler's first law, a common structure, and that this structure reflects the dimension of space (section 1.4). The methodological approach (section 1.5), the relevance of the hypotheses, the contributions of the thesis (section 1.6) and the limitations of the argumentation (section 1.7) are discussed. We finally outline the overall argumentation of this thesis (section 1.8).

1.1 Space, Time and Structure

Our understanding of space and time is characterized by its structure. We argue, in this section, how information with spatial and temporal aspects can be represented, and we introduce the concept of *spatial structure*.

Space and Time. Both, space as well as time, are fundamental to our life: most human activities incorporate space and time, because they are performed somewhere and somewhen. Information thus often describes things that exist or happen in space and time. Existing or happening in space and time means to be bound to space and time, to its existence and its features. It has turned out that spatial aspects of information can be crucial to solve problems, e. g. when investigating and containing the Broad Street cholera outbreak in London (Snow 1854). It is widely assumed that information is of spatial nature in large part (Franklin 1992) but evidence is very rare (Hahmann et al. 2011).

Various concepts of space and time exist, e. g. the concept of a metric space (Lang 2002), the concept of a topological space (Bredon 1993), concepts of space in physics (Basri 1966), concepts of geographical space (Couclelis 2005) and concepts of space in psychology (Uttal 2008). These concepts describe several aspects of what we perceive as space and time. Accordingly, many different features of space and time, also ostensible contradicting ones, exist.

Representations of the World. The examination of information related to space and time requires that the information is represented such that it is accessible to our mind. When we are solving tasks in our environment, we build mental representations of the parts of the world that are important for the solution. Such a representation is task specific and influenced by our mental model, in particular because our mental model is used to judge which parts of the world are useful. Our mental model is influenced by our perception (BonJour 2013), and it is dynamic and develops over time because perception depends on the situational context and we perceive and therewith are able learn continuously (Glaser 1989). Our representations of the world are thus necessarily situational and task-specific.

Representations are usually referring to things, which we will call entities in the following. These entities can be related. We may choose different representations, i. e. different entities and relations[1], of the same reality due to the representations' situational character.

[1] Relations can even be entities in other representations. Entities and relations have, in this case, the same status. A philosophical view on this thematics has been given by Esfeld et al. (2011).

When relations are made explicit by formal representations, we can view these representations as graphs, referring to the world[2]. We will, in the following, refer to these graphs as *graph representations*. If a graph representation is meaningful, the graph should inherit some of the features that the represented reality has: the relations (together with the entities) as a whole should reveal some of the properties of the world's structure. Spatial information should, for example, reveal some properties of the concepts of space. The concept of structure tries to capture how several properties of a thing are related and has turned out to be important in the scientific context.

Structure. In many fields of science, the term *structure* is used to denote how elements of some bigger system are related, e. g. in linguistics, philosophy of science and mathematics. We will use the tangible but still vague definition given by Shapiro (1997, p. 74):

> 'A structure is the abstract form of a system, highlighting the interrelationships among the objects, and ignoring any features of them that do not affect how they relate to other objects in the system.'

Different systems can have the same structure, i. e. the configurations of its elements can be the same, which renders the reuse of formal methods for these systems possible. The elements of the systems however can, at the same time, represent very different things.

Structure constituted by the interpretation of a system is of relative nature because it depends on how we perceive and describe the system. This relative nature has been described by Psillos (2006):

> '[...] the structure of a domain is a relative notion. It depends on, and varies with, the properties and relations that characterise the domain. A domain has no inherent structure unless some properties and relations are imposed on it.'

The concept of structure plays a major role in the evolution of science, especially in logics, mathematics and physics. The rise and use of mathematical structures in the modern formulation of algebra is, for example, discussed by Corry (2004) and Krömer (2007).

Structure Realism and the Evolution of Science. The philosophical position of *structuralism* assumes that we can represent reality best in terms of formal entities and relations, and not in terms of the real entities itself. This approach of formal representations raises the problem of grounding but is very common in information theory.

Worral introduced the position of *structural realism*[3] in contemporary philosophy. He argued that there is no need to accept the form of scientific realism that assumes

[2] We distinguish between the entity, the reference and the symbol used to denote the reference. (Ogden et al. 1923, pp. 6ff) We will denote the entity and the reference by the same symbol if no distinction is necessary.

[3] Several positions of structural realism can be distinguished (Frigg et al. 2011): *epistemic structural realism* (Worral 1989), *ontic structural realism* (Ladyman 1998), *radical ontic structural realism* (van Fraassen 2006), etc.

our theories to correctly describe reality, nor any position of anti-realism. Worral instead claims that there exists a continuity of structural elements of theories, e. g. formal representations, during the evolution of science. This claim suggests that scientific progress is, in large part, based on the use of structures for the formulation of theories, and that it can be convenient to formulate inter- and transdisciplinary concepts in a structural way. The transition from Fresnel's aether theory to Maxwell's theory of the electrodynamic field, in particular to Maxwell's equations (Poincaré 1905, pp. 178ff), is often discussed as an example of such a continuity of structure (Worral 1989).

We will, in this thesis, focus on the structural description of information, with the hope that structure is suitable to characterize information with spatial and temporal aspects. This approach seems to be of special interest[4] due to the fact that the field of geographical information science is interdisciplinary and deals with various competing, sometimes even incommensurable concepts, e. g. concepts of space and time. A clear notation of structure is needed to differentiate between entities and their representations in order to render a discussion of only structural aspects possible.

Spatial Structure. When information, in particular interpreted representations, exposes a number of references to space (or time), it is by definition called *spatial* (or *temporal*). Information which is spatial and temporal at the same time is called *spatiotemporal*[5]. The concept of information implies that we know how entities and their relations are represented and that we can determine whether the representation explicitly or implicitly refers to space. Information describing entities in space refers, for example, implicitly to space.

The structure of spatial information as well as the structure of data which becomes spatial information by interpretation is based on the properties of space and the entities that constitute space: the existence of distance and the effort of travelling leads to a predominance of relations between near things; the similarity of space and physical processes at different scales of tangible reality leads to scale invariance of the spatial structure; and non-uniform distributions of objects in space lead to not necessarily uniform but in many cases bounded distributions of relations (cf. section 3.2).

We call such a structure of data a *spatial structure*, and we say that data *has a spatial structure* (in which case we also speak of *spatial data*) if it exposes some of these properties. It is important to note that data can, by the above definition, have a spatial structure without being interpreted and actually without being related to space; we only require that the data's structure *can* be interpreted such that it is related to space and exposes some of these properties.

This thesis focuses on the examination of the spatial structure. We will discuss an important example of spatial information in the next section, namely information about human activities.

[4] We make *no* claim to whether theories are true as is done by the *miracle argument* (Putnam 1975, pp. 72ff), but rather concentrate on the pragmatic choice to focus on structure.

[5] Information representing things happening in space and time can be non-spatiotemporal. *Going by public transport* can, for example, be understood as an activity of spending money without any spatial or temporal aspects.

1.2 Human Activities

Representations of human activities potentially have a spatial structure. We thus use information about human activities throughout the thesis as an example of spatial information. We review, in this section, several concepts of human activities, and we discuss the prominent example of information about public transport.

Concept of Human Activities. Things that happen potentially change the world: rainfall makes the ground wet, and kicking a ball makes the ball moving. We call the second thing a (human) activity, because rainfall just happens but kicking a ball is something that a human usually does by intention. In remark 621, Wittgenstein (1967) raises the following quest ion: 'when "I raise my arm", my arm goes up. [...] what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?' Even if we do not provide an answer, there seems to be something that distinguishes a human activity from other events that just happen. An overview over human activities in general has been given by Wilson et al. (2012), over spatial human activities by Miller (2004^{a}) and Golledge et al. (1997).

Human activities are present virtually everywhere and anywhere in our daily life. They constitute the way we are interacting with the world and other humans, because interpersonal communication and human interaction are characterized by perception-action cycles, which consist of several human activities (Ortmann 2014, pp. 84ff). These cycles can even be understood in terms of semiotic triangles (Ogden et al. 1923, pp. 6ff): an entity (referent) is observed, yielding an observation (reference). This observation can lead to an action (symbol) which has been motivated by the observed entity and thus refers to it. The action can itself be observed if it results in an observable change in the world and can therefore be seen as another referent, which can evoke another perception-action cycle.

Existing concepts of *activities* and *actions* are very similar. The distinction between both concepts is not clear, because both terms have been used differently by different scientists. The concept of activities seems, in many cases, to refer to the combination of several actions, and the beginning and the end of an activity are not always well defined. Kicking a ball is, for example, in many cases regarded to be an action, but playing soccer is more likely to be referred to as an activity. A category of actions is often referred to as activities. Playing sport, for example, can be referred to as an activity. The distinction between both concepts is, however, not done in the same way by all scientists, and we thus will use the word *activity* whenever a human is doing something that changes the world's state.

Formally, we can understand an activity to be a transition of one state of the world to another one. This transition, especially in space and time, can be described in several ways (Strobach 1998). In the scope of this thesis, such a transition will be understood as some intentional activity that transfers the world from one state to another one without caring about how the transition is actually performed and how it looks like.

A transition happens between two states of the world, mostly different ones. In most cases, not the entire world but only a small environment is changed by a human activity. We thus focus on transitions between different states of a human's environment, not of the entire world. As humans performing the activity can also change themselves (thinking is, for example, the state of one's memory), the human can be considered to be a part of its environment. Human activities influence, in many cases, each other, and activities that belong to the same thematics suggest itself to be examined at once. Such a set of human activities is called a *human activity system* if it also includes affordances of human activities, i. e. human activities that are possible to perform but have not necessarily been performed.

Public Transport. Many examples of human activity systems exist. Public transport is an example of a human activity system that is related to space. We will thus use information about public transport as an example of spatial information in this thesis.

Public transport is simple to represent and information about public transport is widely available, because clear regulations and information is needed to make public transport effective: we are able to choose combinations of vehicles such that the destination can be reached as fast as possible, if we have knowledge of which vehicle will be located at which stop at which time, which direction a vehicle takes and how fast it is travelling. The communication of these regulations to the customers, e. g. by timetables, is thus crucial for a public transport system to work.

We should keep in mind that public transport is a very specific example of human activities because it is planned, e. g. by the creation of integrated periodic timetables (Schöbel et al. 2013, Grujičić et al. 2014). The fact that public transport is planned can lead to a very specific structure, e. g. in the case of periodic timetables to symmetry minutes and other symmetries (Liebchen 2003). These characteristics created by organizing public transport superpose the spatial structure.

The concept of human activities as transitions between states of the world affords a representation of human activity systems by graphs. Such graph representations can be used to discuss the structure of spatial information, e. g. Tobler's first law of geography.

1.3 Tobler's First Law of Geography

Nodes of graph representations are, in many cases, related when they are in the same neighbourhood. This important aspect of spatial structure is known as Tobler's first law of geography. We will, in the following, discuss this law by the example of human activities.

Tobler's Law. Stops of public transport have locations in space, and they are connected by vehicles which are driving between them. Connections between stops of the same neighbourhood occur, in many cases, much more often than between distant stops, e. g. in bus or railway networks. This fact translates to the graph representation, which we gain by a representation of stops by nodes and connections by edges: edges in neighbourhoods of nodes occur more often than edges between non-neighboured nodes.

The correlation between the configuration of edges and the distance between the nodes can also be observed for other types of representations, not only for public transport networks. The generalization of this correlation, as a statistical statement, is known as Tobler's first law of geography (Tobler 1970):

Theorem (Tobler's first law of geography/Tobler's law). *Everything is related to everything else, but near things are more related than distant things.*

When 'costs' of relations depend on the distance between the nodes to connect, e. g. transport and communication costs, Tobler's law can be assumed to be valid to some extent. This suggests Tobler's law to be an important aspect of spatial structure.

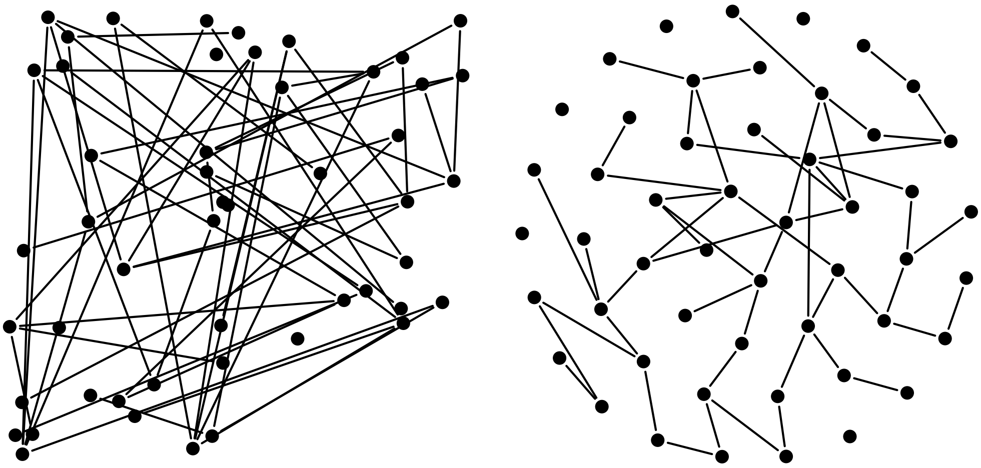
Tobler's Law Without Coordinates. In a graph representation, Tobler's law suggests some constellations of edges to occur more frequently than others. These constellations become visually apparent, when the nodes are naturally embedded in space: the nodes are placed at the locations of the things that they represent, and we can visually test whether Tobler's law is valid.

When the nodes of a graph representation contain no information about their natural location, it is not clear how to embed them in space. We could expect that the constellation of edges becomes, independent of their embedding, visually apparent, but the converse is true: most graphs, even those that do not satisfy Tobler's law, appear to satisfy Tobler's law if a suitable embedding is chosen, as will be motivated by the example of a random graph.

The Gilbert model is a random graph model which assumes an edge between two nodes to exist with a given probability (Gilbert 1959). When such a graph is randomly embedded in space, we can try to move the nodes around until the graph satisfies Tobler's law. Existing force-directed graph drawing algorithms, amongst others the Fruchterman-Rheingold algorithm, assume linear attractive and inverse-quadratic repulsive forces[6] and minimize, by the nodes' movements, the energy of the system (Fruchterman et al. 1991, Kobourov 2013). The attractive force aims at minimizing the distance between nodes, which is expected for a graph satisfying Tobler's law, and the repulsive force ensures that the graph is not collapsed to one point in space. The resulting graph satisfies Tobler's law, as becomes visually apparent (cf. figure 1.1).

Tobler's law implies specific constellations of edges to be more probable. We discussed that it is, in contrast to our intuition, hard to visually distinguish between random Graphs and abstract graph representations which satisfy Tobler's law. We will hypothesize, in the next section, that such a distinction is however possible.

[6] This situation can be compared to the physical system of electrically charged particles which are coupled by springs. The attractive force can be characterized by Hooke's law and the repulsive force, by Coulomb's law.



(a) Gilbert model

(b) Gilbert model, nodes rearranged

Figure 1.1
Random graphs visually satisfy Tobler's law for a suitable embedding; (a) Gilbert model, and (b) Gilbert model after application of two force-directed graph drawing algorithms: the ForceAltas2 algorithm (Jacomy et al. 2014) and the Fruchterman-Rheingold algorithm (Fruchterman et al. 1991)

1.4 Hypotheses

The definition of a spatial structure only makes sense if the structural properties exposed by spatial data sets are similar. In this case, the structure can be used to characterize data in a meaningful way. We demonstrated, in the last section, that it is hard to visually detect such a spatial structure (cf. figure 1.1). Methods to computationally detect the spatial structure may though exist, and we thus hypothesize that spatial data sets share structural properties:

(1) The concept of spatial structure is meaningful, because most spatial data sets share structural properties.

If the hypothesis turns out to be valid, we can lift the discussion of spatial information from a semantic to a structural level. This lift would open new possibilities of comparing and analysing spatial data, because the hypothesis implies that we can categorize data sets by their structure in a meaningful way.

Since the dimension of space influences many of the properties of space, we expect it to have an impact on the spatial structure of data. We thus raise the following hypothesis:

(2) Spatial structure implicitly reflects the dimension of space.

It is not clear to which extent the dimension of space is decisive for the spatial structure, and whether the dimension of space can be concluded. If the hypothesis is valid, some semantics can, in principle[7], be inferred from the structure.

In higher dimensions, there exist more interrelations because the volume per surface ratio is larger. If the hypothesis is valid, i. e. if a spatial structure reflects the dimension of space, we nevertheless expect the dimension to have a different effect on the spatial structure than just a higher ratio of edges to nodes.

[7] We can only infer semantics if the representation is chosen in a meaningful way.

1.5 Methodological Approach

We summarize, in this section, the approach that will be used to decide whether the hypotheses are valid. A detailed view of the thesis' outline, which explains how the approach is implemented, is provided in section 1.8.

Hypothesis (1) claims that most data sets share structural properties, and a comparison of the structure of spatial data sets is needed for a validation of the hypothesis. The comparison of structures can be carried out in many different ways, and it is only efficient if we know which aspects to focus on. We will discuss typical properties of spatial information as examples of such aspects.

These typical properties are formulated in terms of things, which are embedded in space, and relations between them. It is, in consequence, not possible to verify that most spatial data sets have these properties, when the locations of the things in space are not known. Instead, we can try to trace the effect of these properties on the structure, in particular on the constellation of the relations. We will, for this purpose, introduce a graph model that constructs edges for a set of nodes embedded in space. The model has the typical properties of spatial information, as can be proven by an analysis of its construction. It can, by the comparison of the constellation of edges in the model to the constellation of relations of a data set, implicitly be checked whether the data set exposes these typical properties. Such a comparison will be performed for numerous spatial and non-spatial data sets, amongst others by testing whether certain properties, which approximately coincide for the proposed model, also coincide for the data sets.

The dimension of space has an influence on the structure of spatial data. Hypothesis (2) claims that the dimension of the space is reflected by the spatial structure, i. e. that the structure uniquely determines, with some uncertainty, the dimension. When a method to conclude the dimension of the proposed model just by the constellation of edges exists, the hypothesis is corroborated. We will discuss numerous properties of the proposed model and argue which of them is best suited to conclude the dimension. It will turn out that only combinations of the dimension and the density parameter, which is used for the generation of the proposed model, can directly be concluded. The dimension can finally be concluded by the comparison of different combinations of the dimension and the density parameter.

The hypotheses, which are corroborated by the argumentation of the thesis, as well as some novel definitions, concepts and algorithms are contributions of this thesis. They will, together with their relevance, be discussed in the next section.

1.6 Relevance and Contributions

The structure of spatial information has been argued in section 1.1 to be of high relevance to the advance of the formal foundations in geographical information science. The proposed model of spatial structure is, as far as I know, the first model of spatial structure in general, that is, of the similarities of most spatial data sets.

Its simplicity should not hide the fact that it can play a major role in applied and theoretical research, much like other network models, e. g. the Barabási-Albert model, do in other fields of science. The simplicity is rather an important factor for the applicability and the viability of the model.

The thesis makes the following contributions which are, as far as I know, new in the field of geographical information science:

(1) the concept of spatial structure,

(2) a model of spatial structure in general, namely the (uniform) SISG model, including some analytical properties of the model,

(3) the concepts of scale invariance of a spatial graph and, of transformations of relative scale[8],

(4) algorithms to measure to which degree the structure of a data set is spatial, including an evaluation,

(5) the concept of total density,

(6) the concept of series of subgraphs, and

(7) a new view on graph representations of human activity systems, including the collapsing of the state space, the definition of relevant interchange facilities, the concept of packing a graph representation and the more advanced concept of packing a graph representation of public transport.

These contributions are accompanied by a literature review of several aspects of data sets about human activities, a discussion of typical properties of spatial information, and a discourse on existing graph models and methods, and why they do (not) work.

1.7 Limitations

The hypotheses and the methodological approach of how to corroborate the hypotheses have limitations. We discuss these limitations, in this section, to provide the context in which the statements of this thesis can and should be understood.

The claims of the hypotheses are formulated on a structural level. Statements can, without grounding, not be transferred from a structural level to the semantic level of things and their properties: information about relations between things explains how the things and their properties relate, but neither the objects nor the properties can be identified with their real counterparts. The hypotheses can thus only be used to draw structural conclusions, but these conclusions can, with a grounding, be interpreted in terms of things and their properties.

We assume, in large part of the thesis, that information is represented by graphs. This methodological limitation is crucial when a concrete data set shall be tested

[8] Both concepts have been introduced in geographical and mathematical contexts, e. g. by Aldous et al. (2013), but they are, in this thesis, introduced in the context of spatial information for the first time.

for a spatial structure. As the hypotheses are formulated on a structural level which assumes information to consist of relations, the assumption of information to be represented by graphs is, however, no real limitation.

The hypotheses cannot analytically be corroborated for all existing spatial data sets; they can only be proven valid for single data sets. A validation of the hypotheses on a large number of data sets is, by the absence of counter examples, able to suggest that the hypotheses statistically are valid. Numerous data sets from different domains are examined in this thesis. It yet remains to test the hypotheses on a higher number of data sets to put the hypotheses on a firmer ground. This can however only weaken the limitation but not invalidate it.

None of the limitations is crucial, and suggestions on how to improve and generalize the approach will be discussed in the conclusion (cf. section 6.2).

1.8 Outline of the Thesis

Graph representations of human activity systems are, throughout this thesis, examined as examples of data sets which expose relations between things, and they are used for the evaluation of the hypotheses at a later point. We discuss, for this purpose, concepts of human activities and introduce the concept of graph representations of human activity systems (section 2.1). The creation and use of data sets is discussed (section 2.2), and existing data sets are reviewed (section 2.3). Graph representations of public transport can be gained by timetables, but modifications are necessary in order to use them as an almost prototypical example of spatial information (section 2.4).

We discuss spatial structure by having graph representations of public transport in mind: we motivate Tobler's first law of geography by the principle of least effort (section 3.1) and review additional typical properties of spatial information (section 3.2). As existing graph models cannot be used to model spatial structure (section 3.3), we propose a graph model of spatial structure (section 3.4) and prove that it has the required properties (section 3.5). In addition, analytical results are proven for the model (section 3.6).

The proposed model of spatial structure assumes a set of nodes which are placed in space. We detailedly examine the model that is gained for a set of randomly distributed nodes with uniform distribution. Statistical methods for the examination of the model's properties (section 4.1) and the effect of the finiteness and the non-connectedness on these properties (section 4.2) are discussed. We introduce the concept of series of subgraphs to define additional statistical properties that are less affected by the finiteness of the model (section 4.3). The examination of various properties (sections 4.4 to 4.6) can be used to classify the model (section 4.7).

As an evaluation of the proposed model, we compare the model to spatial and non-spatial data sets. We tackle the question of how to test data for spatial structures and discuss possible methods to compare data sets to the model (section 5.1).

Algorithms for this comparison are provided (section 5.2). The algorithms are evaluated by the proposed model (section 5.3) and used for the evaluation of the provided algorithms on real data sets and the validation of the hypotheses (section 5.4). This evaluation of the proposed model does not depend on how we derived the model, but it is only based on the model as an abstract structure.

Mathematical definitions (appendix A) and computational aspects (appendix B) are provided at the end of the thesis. The reader who is unfamiliar with mathematical notations or computational aspects is referred to these appendices, in order to understand the notations and algorithms used in the thesis.

2

Graph Representations of Human Activity Systems

> *Der Spielmann richtet sich, da nimmt Löchlein sich eine Jungfrau an die Hand, juheia! wie er springt! Herz, Milz, Lung' und Leber schwingt sich in ihm um, er fällt in den Anger, dass ihm Ohren, Nase und Maul von Blut überwallen, zu beiden Seiten sieht man sein Herz heftig klopfen, ihm hat gedünkt, als wären sieben Sonnen am Himmel und lief er um wie ein gedrehter Topf, ihm schwindelte es um den Kopf, und er meinte zu versinken.*
>
> —**Franz Magnus Böhme**
> german composer and scientist
> (1827–1898)

Representations of human activities are, in many cases, examples of spatial information. We will use them, in subsequent chapters, for the discussion and evaluation of the proposed graph model. The evaluation is crucial for the validation of the thesis' hypotheses, and the discussion of the data sets and their semantics an important part of the argumentation.

This chapter begins with a discussion of representations of human activities (section 2.1). In particular, we introduce the concept of *graph representations*, which is necessary for the structural discussion in the following chapters, (section 2.1.2) and show how it applies to public transport (section 2.1.3). A review of methods to create data sets, of existing data sets and of their use is provided (section 2.2). Graph representations are introduced and will, in later chapters, be used as examples of data sets (section 2.3). As graph representations are in many cases very large, we discuss methods to reduce them in size with the aim of preserving the relevant structural information (section 2.4); some of the provided algorithms have high complexity and can only be executed on the reduced data sets in a reasonable time.

2.1 Representations of Human Activities

We give, in this section, an overview of how human activities can be conceptualized and represented. This overview provides the basis to introduce data sets in a later section.

Representations of human activities are widely used, and various concepts of human activities can be found in literature (section 2.1). Representations of human activities by graphs are discussed (section 2.1.2), and a more detailed view of representations of public transport is provided (section 2.1.3).

2.1.1 Concepts of Human Activities

Extensive research has been conducted on various aspects of the description and conceptualization of human activities. We review, in this section, some of these aspects and concepts.

A widely used concept is to understand human activities as transitions between two states of the world. Such transitions can be concatenated and combined to more complex activities, e. g. for describing the use of public transport (Frank 2007, 2008). Activities can also be described at the level of gestures, poses, movements of segments of the body, and interactions (Ryoo 2008, pp. 52ff) as well as by statistics and patterns (Chen et al. 2011). It is not trivial to perform a query on data sets of human activities, because several concepts of human activities are, at least in parts, incompatible. A method of searching human activities without visual examples has, for example, been presented by İkizler Nazlı et al. (2008).

The activities which a person can perform are restricted by the environment the person is placed in, and the environment may offer certain activities to be performed. These opportunities of activities that the environment offers are called *affordances* (Turvey 1992, Sanders 1997). Affordances can lead to activities when they are performed, and affordances can be perceived by simulating activities, as was discussed by Raubal (2001, pp. 39ff) and Ortmann (2014, pp. 84ff). The interaction possibilities described by affordances have successfully been applied to mobile robots (Raubal et al. 2008). The concept of affordances is broader than the concept of activities, because it includes the environment. We will, in the scope of this thesis, understand an activity as an operation that *can be* performed. This concept of human activities includes activities that *were* performed, but also those that are afforded by the environment.

Performed activities can have an impact on the environment. As the environment determines the affordances and potentially also influences the activities' results, activities can indirectly influence other activities via the environment. This influence of an activity on other activities or even the activity itself is a feedback loop and can be interpreted as context of activities. Since this context can, in many cases, only be grasped when the results of other activities can be observed or are known, it can be hard or even impossible to completely understand how human activities influence and constrain each other. It is beneficial for most applications if conceptualizations of human activities take the effect of mutual dependencies into account, e. g. when describing daily activity patterns (Hemmens 1970).

Conceptualizations of dependencies between human activities have been studied by Abdalla et al. (2014) in order to support prospective memory for planning our daily routines. Spatiotemporal, hierarchical and conceptual relations between tasks and activities that are important for trip planning have been discussed by Abdalla et al. (2013). Physical, social and mental affordances constraining decision-making in the context of spatial tasks have been discussed by Raubal et al. (2004). We call a set of human activities with mutual dependencies a human activity system:

Definition 2.1. A *human activity system* is set of affordances of human activities which have many interdependencies and are influencing and constraining each other.

Human activity systems can be relatively small and restricted to only one domain, e. g. the system of traditional diving, fishing and hunting activities in parts of Japan (Watanabe 1977), but also very large systems as, for example, the set of human activities which have have an effect on our natural environment (Goudie 2013). Climate change is, at least in large part, caused by human activities, and it is hard to understand, because the number of involved activities and their interdependencies is very large (Allen et al. 2014).

We discussed the concept of activities as transitions between states of the world and how activities depend on each other. The notation of human activity system was introduced as a set of human activities that have many interdependencies. We will introduce a formal representation of human activity systems in the next section.

2.1.2 Graph Representations

Representations of human activity systems that emphasize relations between activities are called graph representations. We introduce, in this section, a formal definition of graph representations, and discuss their shortcomings.

Representations of human activities are very common in our daily lives, e. g. assembly instructions for furnitures, cooking recipes, legislative texts, etc. When we try to solve a task, the task itself determines which formal representation is suitable, because the use of the formal representation has observable consequences to the real world: whenever furnitures are assembled, a meal is cooked or one acts in a social context, it has observable consequences to the real world: the furniture may be assembled well, the meal may be delicious and one may impinge upon someone else's rights. We will, in the following, introduce a graph representation of human activity systems that is suitable for the discussion of the structure of a human activity system.

Graph Representations of Human Activity Systems. We will, in this thesis, understand an activity as a transition between states of the world, as was conceptualized in the last section. Regarding the states[1] of the environment as nodes and the activities[2] (i. e. the transitions) between them as directed edges, we obtain a directed abstract graph.

Definition 2.2. A *graph representation of a human activity system* $\mathcal{G}(S, A)$ is the hypergraph consisting of the set of *states* S as nodes and the set of *activities* A as edges. The set of states S is called the *state space* of the graph representation.

A graph representation of a human activity system is a hypergraph, because there potentially exists more than one activity that leads from one to another state, e. g. due to different modes of transportation. Edges that all start in a node p and end in a node q cannot be distinguished any longer when the graph representation is considered as an abstract graph, i. e. a graph whose nodes and edges have no additional semantics.

We implicitly assume that we can concatenate two activities of a graph representation, as long as the end state of the first and the starting state of the second one coincide. Concatenation defines an algebraic structure[3] on the set.

Non-Representable Phenomena. Graph representations cannot cover every aspect of human activities, which prevents some phenomena from being modelled. The following discussion aims at providing an intuition of which phenomena can be represented by graph representations, and which common phenomena of human activities can only be represented by a more complex algebraic representation.

We implicitly assumed that the concatenation of two activities of a graph representation always exists, as long as the end state of the first and the starting state of the second activity coincide. This implicit assumption is, in general, wrong because concatenations are not always possible: the activities 'walking from A to B' and 'cycling from B to C' can only be concatenated as long as there is a bicycle available in B (Abdalla et al. 2012). There would not occur any problem with the concatenation, if the first activity would not be 'walking' but 'cycling'. At second glance, it however becomes clear that 'having (not) a bicycle with you' is a state that can be taken into account for the construction of the state space. When the state is taken into account, we do not any longer expect that the activities 'walking from A to B' and 'cycling form B to C' can be concatenated, because the end state 'being at A without bicycle' and the starting state 'being at A with a bicycle' do not coincide.

Human activities can be performed in parallel, but graph representations cannot describe this circumstance properly: activities that are performed in parallel can be represented as one compound activity, but it cannot be represented how the

[1] In the following, we consider the class of states and the class of activities to meet the axioms of a set.

[2] An activity refers, in this context, not only to performed but also to possible activities, as was discussed in section 2.1.1.

[3] A graph representation of a human activity system embodied with the concatenation meets the axioms of a partial semigroup.

activity is composed of other activities. Activities that start at the same point in time but end at different ones cannot be represented at all.

Graph representations depend on two choices: the set of states, and the set of activities. Similar states and similar activities are usually grouped together, when a graph representation is constructed and these choices are performed. We can, for example, walk down a road in many different ways (skippingly, lumbering, etc.), at different speeds, at the left or the right pavement, or even forwards or backwards. All these modes may, in a graph representation, be combined to a single activity 'walking down a road'. It is however not clear which states and activities shall be combined, and which shall considered to be different. This problem of categorization arises as it does with every other representation.

Some activities may be performed more regularly or with a higher probability than others. Graph presentations as introduced above cannot represent this fact. We could however use weighted graphs to denote the probability of an activity to be performed but would still not be able to represent that some concatenation of activities happen more often than other ones. 'Inserting a CD into the player' and 'pressing the player's start button' are activities that usually are performed with a much higher probability in combination than the single activities.

The concept of graph representations captures some basic properties of how activities are related. We introduced a formal definition and discussed some of its shortcomings. When graph representations of public transport are constructed, specific problems occur. We will discuss these problems in the next section to gain a better understanding of graph representations.

2.1.3 Graph Representations of Public Transport

Information about public transport is widely used. Several types of timetables exist and are used for different purposes (e. g. travelling, waiting for someone at a stop) and travelling habits (e. g. commuters, irregular travellers): stop-specific timetables, route-specific timetables, commuter timetables with typical route combinations, or combination of them. Route planning systems are, in addition, used to search for the best solutions for transport tasks. We discuss, in this section, how graph representations of public transport can be constructed and which problems occur due to missing interchange facilities in the data. The discussion of this section facilitates the transformation of existing data sets about public transport into graph representations, which will be used for the evaluation of the proposed graph model in section 5.4.

Graph Representations. The majority of the representations of public transport names transport modes, stops, trips (i. e. sequences of stops), routes (i. e. sets of trips usually involving the same stops, e. g. a bus line), and the times when a vehicle of a certain trip will come to a stop. Such information can be represented by a graph representation: stops are interpreted as states, and pairs of successive stops,

i. e. stops p and q such that at least one vehicle travels from p to q without stopping in between, as activities. The resulting graphs are directed, and the edges can be named by a trip identification.

Alternatively, tuples consisting of stops and corresponding stop times can be interpreted as states, which leads to more extensive sets of nodes. The activities are again pairs of successive stops. Such a graph representation is called *time-considering* because the states include temporal information, whereas the graph representation with only stops as states is called *time-ignoring* because the states do not include temporal information.

Definition 2.3. A graph representation is called *time-considering* if the states incorporate time, otherwise *time-ignoring*.

Time-considering graph representations can be transformed into time-ignoring ones by grouping all states corresponding to the same stop. We will, in this thesis, only consider time-ignoring graph representations of public transport, but usually, add interchange facilities to the time-considering representation before transforming it into a time-ignoring one.

Adding Interchange Facilities to Graph Representations. Timetables contain, in many cases, no information about interchange facilities between different vehicles and modes, even if they are part of timetable planning and important for public transport to work. We will define interchange facilities properly in order to include them into the graph representation of public transport.

An interchange facility is the activity of staying at a stop for a certain amount of time such that the span of time between the arrival with a vehicle at the stop and the departure with another vehicle at the same stop at a later point in time is seamlessly bridged. We formally define:

Definition 2.4. An edge $e = ((p, t), (p, t'))$ with p a stop and t and t' stop times is called an *interchange facility between vehicles d and d'* if and only if there exists an arriving vehicles d (coming to p at t) and a departing vehicles d' (coming to p in t') such that the following requirements are met:

(a) the edge e is *temporal*, i. e. $t < t'$, and
(b) the vehicles d and d' are not associated to the same trip.

Requirement (a) ensures that first vehicle is arriving at the stop before the second one is leaving, which provides the chance to change the vehicle. The 'change' from a vehicle to itself is not regarded as a change, because it has the same result as staying in the vehicle without changing; hence requirement (b).

Relevant Interchange Facilities. We can add all possible interchange facilities to a time-considering graph representation, also the ones that are never used in reality. A bus line with hourly service, for example, affords, at each stop and for each hour, interchange facilities of one hour, two hours, etc. (cf. figure 2.1a).

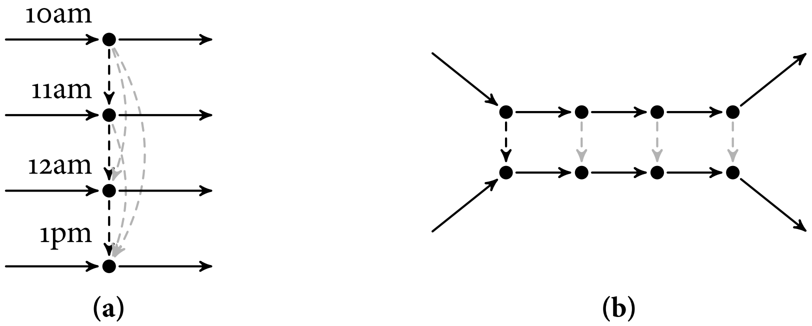


Figure 2.1
Examples of interchange facilities; the dashed edges are interchange facilities; non-relevant interchange facilities are marked in grey

The issue lies with how to judge which interchange facilities are relevant in order to gain a meaningful graph representation. This judgement can be made having different use cases of the graph representation in mind.

We will approach the question of whether an interchange facility is relevant by considering its relevance for travelling as fast as possible between two arbitrarily chosen stops. A set of interchange facilities may, at the same time, lead to the same travelling time, e. g. when to routes have a sequence of stops in common and one can change the vehicle at all of these stops (cf. figure 2.1b). In such cases, we only consider one of these interchange facilities as relevant, because this minimizes the size of the group representation but does not change travelling time. We formally define:

Definition 2.5. An interchange facility $((p, t), (p, t'))$ between vehicles d and d' is called *relevant* if and only if the following requirements are met:

(a) there is no interchange facility $((q, s), (q, s'))$ starting at the previous stop q of the vehicle d such that an edge $((q, s'), (p, t'))$ exists.
(b) there exist no relevant interchange facilities $((p, t), (p, t''))$ and $((p, t''), (p, t'))$, and
(c) one of the following criteria is met:
   (i) d and d' are associated to the same trip, and d is not antiparallel to d' (i. e. the previous stop for d is the next stop for d'), or
   (ii) the vehicle d is ending in p (i. e. there is no vehicle associated to the same trip proceeding from p).

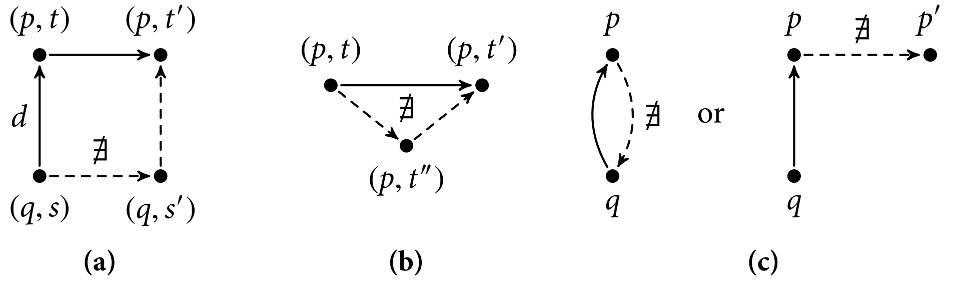


Figure 2.2
Requirements to an interchange facility to be regarded as *relevant*;
in (c), all depicted edges are associated to the same trip; cf. definition 2.5

Requirement (a) ensures that, in case of vehicles driving in parallel, interchange facilities are not considered relevant at every stop but only at the first one (cf. figure 2.2a). The choice of interchange facilities to be relevant at the first stop is not necessarily the choice that we would choose in reality: we may prefer one stop to another one for some reasons, e. g. because of a longer time span of the interchange facility; because of the number of alternative connections if the change was missed; because of the way we have to walk to change the vehicle; or because of the environment where the change takes place in (it can be nice, unpleasant, etc.). Independent of this choice, the effect on how to travel from one to another stop in the transport network is roughly the same.

The number of relevant interchange facilities is minimized by requirement (b), because changes are only considered relevant if they are not a concatenation of two other relevant interchange facilities (cf. figure 2.2b).

Finally, requirement (c) ensures that interchange facilities between vehicles of the same trip are not considered relevant if the vehicles are driving in the opposite direction. This includes, in particular, the case that a vehicle is travelling a route, turns around, and travels in the opposite direction: the effect of getting off, waiting, and getting on the same vehicle is in this case the same as just staying at the vehicle (cf. figure 2.2c).

When we add relevant interchange facilities to a graph representation of public transport, we usually add only those interchange facilities that are not already part of the representation.

Various concepts of human activities exist. We discussed the representation of human activities by states of the world and transitions between these states, which lead to the definition of graph representations. It was discussed how graph representations of public transport can be build by timetables. We will discuss in the next section, how data sets about human activities can be created in general, which data sets exist and how they can be used.

2.2 Creation and Use of Data Sets

Data sets about human activities have been created in many contexts. This section contains a review of existing methods to create and use data sets, with the aim to promote the understanding of which properties data sets about human activities have. This knowledge is necessary for the understanding of the conceptual modelling of spatial information and the validation of the hypotheses.

Environmental and body-worn sensors as well as sensors which are integrated in smartphones and other mobile devices can be used to build data sets about human activities (section 2.2.1). Methodological shortcomings influence the quality of the resulting data sets (section 2.2.2). Numerous data sets about human activity systems can be found in literature (section 2.2.3), and such data sets can, as formal representations of human activities, be beneficial, as can be argued by existing applications (section 2.2.4).

2.2.1 Data Collection

Data about human activities, which are performed by single persons or groups of persons, can be collected by the use of various types of sensors: environmental sensors, body-worn sensors, smartphones, and many more. We discuss, in this section, how these types of sensors have been used in different contexts and studies.

Environmental Sensors. Video cameras and other environmental sensors can be placed in the environment in order to observe a person performing activities (Moeslund et al. 2006, Poppe 2007). Cooking activities, for example, have been recorded in a realistic cooking environment (Rohrbach et al. 2012[a]).

Body-Worn Sensors. Body-worn sensors have the advantage that they always are following the user and can thus be used to aid in more complex situations and in situations when the user is moving between many places. Body-worn sensors have however high requirements: they should be small, light, insusceptible and easy to use. An overview over human activity recognition using body-worn sensors has been given by Huynh (2008). It is possible to monitor and recognize the activities in real-time (Karantonis et al. 2006). An example of activity recognition has been given by Lukowicz et al. (2004), where progress in workshop activities was monitored by acceleration sensors.

Smartphones. Smartphones usually contain sensors like accelerometers, gyroscopes, magnetometers, photometers, microphones, GPS sensors, etc., These sensors have a high availability, because smartphones are, in many cases, carried around with the user. This high availability makes smartphones interesting for applications of daily life, because no additional device needs to be carried around. Many algorithms for activity recognition take acceleration and gyroscope data into account (Reiss et al. 2013, Anguita et al. 2013, Kwapisz et al. 2010, Brezmes et al. 2009). Even other sensors can be incorporated into the recognition of human activities to improve performance (Shoaib et al. 2014). Algorithms for human activity recognition, especially for smartphones and other mobile sensors, have to be efficient, because energy and computing power are limited on smartphones (Anguita et al. 2012).

After the collection of raw data, a graph representation can be constructed. This construction needs the data to be processed and analysed, e. g. in respect to semantic aspects. We will discuss the quality of the resulting graph representation as well as how the analysis can influence the completeness of the representation in the next section.

2.2.2 Completeness and Quality of Collected Data

Graph representations cannot be construct from the collected raw data, when it is unclear how to represent the data as a graph, or only performed activities are collected. We discuss, in this section, factors that can prevent collected data from being represented as graphs.

Incomplete Semantics. Collected data contains, in many cases, no semantic data about human activities, e. g. a video showing a person performing several activities. The video contains information about the activities that are performed but the information is not easily accessible. A formal description of the occurring activities can be gained by processing the data, recognizing human activities and describing them semantically.

The collection of sensor data does not explain *why* activities have been performed. Knowledge of the user's intentions can improve the effectiveness when making use of the data, e. g. when a system assists a user by providing reminders or helpful information for future activities (Abdalla et al. 2014). Activities are, in many cases, motivated by a need and are hence goal-oriented. We can reason the goal of a sub-activity by checking to which degree the sub-activity of a compound activity is necessary to achieve the goal of the compound activity: walking to a bus stop, for example, is necessary for travelling by bus, and the walk is much shorter than the bus ride. The walk to the bus stop is, in conclusion, performed for the reason of travelling by bus. Similar considerations, based on the degree of fulfilment of a compound task, can provide insights in why a sub-activity is performed (Nieves et al. 2013).

Missing Grounding. Activity recognition is well tested in controlled environments, but knowledge about the ground truth is necessary to apply activity recognition in real environments. Otherwise, it is not possible to understand the complete meaning of the data because no grounding can be made (Hossmann et al. 2012).

Missing Information About Affordances. Graph representations state that certain activities *can* be performed, i. e. that there is the possibility to perform these activities, as was discussed in sections 2.1.1 and 2.1.2. When data contains no information about affordances, the data does not afford a graph representations.

Affordances that have not been performed can by definition not be directly observed. When performed activities are observed, only a subset of all possible activities is collected. The more observations are made, the more probable it becomes that a large fraction of possible activities is covered. A year-long observation of all vehicles driving through a town, for example, will reveal most roads that a vehicle is allowed to drive on and even those that a vehicle is not allowed but able to drive on.

The planning of activities presumes knowledge about affordances, e. g. when a road is constructed or a timetable created. If we know how to interpret the resulting data of the planning process (e. g. road signs, timetables), we can simulate the activities by imagining what happens while performing them[4]. This simulation provides

[4] Timetables for public transport were useless without having this capability.

the possibility to estimate how activities are changing the state of the world and to, in principle, construct a graph representation without having observation data.

We discussed how the quality of data about human activities is affected by the used creation methods. Existing data sets about human activities and a great number of studies have thus been created with much effort, as will be discussed in the next section.

2.2.3 Existing Data Sets

Data sets about human activity systems have been created in various contexts. We review, in this section, these data sets. Most of them have no graph representation but are examples of representations of extensive human activity systems.

Daily routines are of interest for many research projects as well as for many applications. The *Multinational Time Use Study (MTUS)* provides human activity data from numerous countries (Fisher et al. 2000). The *National Human Activity Pattern Survey (NHAPS)* contains representative U. S. human activity patterns that can be used to assess the exposure to environmental pollutants (Klepeis et al. 2001). Similar data has been collected in California to detect indoor pollutant sources. Minute-by-minute activity collections have been published as the *California Activity Pattern Survey (CAPS)* (Jenkins et al. 1992). A corresponding data set has been collected for Canada, published as the *Canadian Human Activity Pattern Survey (CHAPS)* (Leech et al. 1996). Data from 19 studies about daily routines have been combined into one data set, the *Consolidated Human Activity Database (CHAD)* (McCurdy et al. 2000).

Cooking activities have been investigated, because they are prototypical for simple and clearly-restricted indoor activities. Cooking activities are restricted to the kitchen space and can be controlled by handing a recipe to the participants of the study. Amongst others, video, infrared video, audio, accelerometer, gyroscope and trajectory data was collected in several studies, which resulted, for example, in the *MPII Cooking Activities Data Set* (Rohrbach et al. 2012^a). Scalability issues arise because many activities are compound. The *MPII Cooking Composite Activities* was built for studying these phenomenon (Rohrbach et al. 2012^b). A collection of videos and natural language descriptions of cooking activities was created as the *The Saarbrücken Corpus of Textually Annotated Cooking Scenes*. This data was used for approaching the problem of grounding textural descriptions by visual information (Regneri et al. 2013). The *Carnegie Mellon University Multimodal Activity Database (CMU-MMAC)* contains several types of data, which was collected by asking participants of a study to cook different recipes (De la Torre et al. 2008).

Human morning activities have been examined by diary notes as well as by recording motion tracking data of persons and objects (Karg et al. 2014). The *Opportunity Data Set* about morning activities has been conducted with the use of 72 sensors of

10 modalities. The large number of sensors was used to establish a benchmarking for such a scenario (Chavarriaga et al. 2013).

We reviewed data sets about human activities that have been build for several purposes, amongst others for research and for practical ones. We will discuss in the next section how data sets about human activity systems can be used.

2.2.4 Use of Data Sets

Human activities are an important factor for the solution of many everyday problems as well as for more complex ones. We argue, in this section, that the representation of human activities is beneficial, because it enables us to solve problems in various fields, as numerous applications demonstrate.

The monitoring of humans health and activity-levels can help to understand and forecast how diseases spread, and individual monitoring can improve patients health, especially when the monitoring system is context-aware (Tentori et al. 2008, Choudhury et al. 2006). In addition to a monitoring, knowledge of communication and movements helps us to understand epidemics much better (Salathé et al. 2012, Machens et al. 2013).

Medical healthcare systems measuring vital signs and recognizing human activities can help in patient monitoring (Van Laerhoven et al. 2004, Paradiso et al. 2005). Assisted living can be improved by the extraction and modelling of human activity patterns (Lymberopoulos et al. 2008, Tunca et al. 2014).

An increased context-awareness of mobile computing can lead to improvements also in other fields of life. Besides healthcare applications, context-aware mobile computing can assist in industrial applications like aircraft maintenance, car production or emergency response operations by monitoring trainees' learning progress, supplementing natural senses and providing navigation aid (Lukowicz et al. 2007).

Humans are interacting, amongst others by social networks. It is relatively easy to trace whether humans are in contact in an (online) social network. Predicting whether humans will get into contact face-to-face, based on knowledge of their behaviour in online social networks, is much more challenging. Such a prediction has been performed by Scholz et al. (2013) by the use of activities that have been performed by a group of persons, e. g. publishing of scientific papers.

Forecasting traffic is important for transportation planning and for estimating the impact on the environment. Amongst others, traffic forecasts are created by predicting travel activities based on travel demand. Phenomena like traffic jams or time-dependent average speeds occur because travel activities strongly influence each other. These phenomena make the simulation of a large number of agents complex. The simulation thus requires a convenient conceptualization of human activities as well as suitable algorithms. Such simulation systems are called *transport forecasting systems*.

The *Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions (STARCHILD)* is one of the first transport forecasting systems (Recker et al. 1986$^{a,\,b}$). Other well-known transport forecasting systems are the *Toronto Area Scheduling Model for Household Agents (TASHA)* (Miller et al. 2003) and the *Transport Analysis and Simulation System (TRANSIMS)* (Rilett et al. 2001, Los Alamos National Laboratory et al. 1998).

Warehouse logistics, as an example of logistic processes, is complex. High quality information is needed to optimize the logistics, but the information is, in large part, yet manually collected. Hildebrandt et al. (2010) have argued how such information can, by autonomous robotic observers, be collected. Performed activities can, by methods from spatial cognition, be recognized in order to optimize the logistic processes.

The understanding of human activities can help us to understand the phenomenon of urbanization and to provide solutions for resulting problems. This better understanding can lead to improvements in urban design (Blanchard et al. 2009).

These examples demonstrate that formal representations, which can be interpreted by computers and mobile phones, can be advantageous for solving problems in many different fields. As information technology and algorithms become increasingly powerful, data is more and more used in the personal environment, which renders data about human activities even more important.

We discussed how raw data can be collected to create data sets about human activities and which problems can occur in this process. Existing data sets and their use demonstrate how much work is needed to build extensive data sets and how fruitful their use can be. We will, in the next section, introduce graph representations of human activities that can be created by available formalizations, which avoids the problem of the raw data collection.

2.3 Examples of Graph Representations

When affordances and human activities are planned, we have knowledge about the affordances. It is hence much less demanding to build graph representations of planned activities.

We discuss, in this section, graph representations of several examples of planned human activities: graph representations of public transport can be created from existing timetable data (section 2.3.1), recipes are formal documents of how a meal can be prepared (section 2.3.2), and graph representations of games can be created by the rules of the game (section 2.3.3). These data sets will be used for the validation of the thesis' hypotheses in a later chapter.

2.3.1 Data Sets About Public Transport

We use timetable information, in this thesis, as an example of a formal representation of human activities due to the following reasons: timetable information is widely available (in various formats), often even as open data; it is a good formalization of affordances of human activities because timetable information is planned and not observed; and only a well-defined type of activities are represented in timetable information. Timetable information from public transport in Sweden (Trafiklab 2013), for example, is provided in the General Transit Feed Specification (GTFS) format[5]. We use, in this thesis, time-ignoring graph representations (with interchange facilities) of public transport in Sweden (cf. section 2.4.1 for details on interchange facilities).

2.3.2 Data Sets About Recipes

Cooking or baking is supported by recipes, which are instruction sets of activities leading to the desired meal as a result. Recipes mostly include only one way of how to achieve the desired result, but in general there exist additional ways. A recipe is, in the context of this thesis, expected to include various ways of how to achieve the desired result, because these ways can easily be inferred from a traditional recipe.

Ontologies have been introduced for the description of cooking, e. g. for planning food preparation in industrial scale (Houba et al. 2000) and annotation processes (Ribeiro et al. 2006, Dufour-Lussier et al. 2012). These descriptions use, in many cases, graphs to depict activities as well as the constraints and needs to gain certain states, i. e. configurations of ingredients.

The recipe for *Pizza Napoletana* has, for example, been written down by the European Union in order to enter the name in the register of traditional specialities (European Comission 2010). The recipe is very concise because it describes in detail which activities have to be performed and in which order. We use, in this thesis, a recipe describing many possibilities of how to bake Pizza Napoletana: the recipe written down by the European Union is enhanced by permutations of activities that do not constrain each other.

2.3.3 Data Sets About Games

Games have explicit rules and can be formalized without much efforts. The *Game Description Language* serves this purpose (Love et al. 2006), because it provides a language to formalize rules which have to be met during the game. The Game Description Language can be used for deterministic and for non-deterministic games (Thielscher 2011). In the latter case, there exists a large number of affordances for each state, and the rules do not determine that a certain activity has to, but only that it can be performed. Reasoning based on formal rules can provide knowledge of the game without doing statistics for randomly played games (Haufe et al. 2012).

[5] http://developers.google.com/transit/gtfs, accessed at 2015-10-08

Starting with a valid state (e. g. the initial state of the game), we can check for each activity whether the rules are met, and whether the activity is therewith valid. An explicit description of all states and transitions can however be hard to achieve, because the number of states can be very high and because knowledge of the rules does not necessarily imply that all states are known (Thielscher 2010).

We will consider three games in this thesis: Rubik's Cube, Tic-tac-toe and a dice that is thrown a number of times. These examples will turn out to be very different, because they are influenced by space, time, and other aspects and constraints in very different ways.

Rubik's Cube. The Rubik's Cube was invented by Ernő Rubik in 1974. It is a cube consisting of smaller cubes such that each face of the Rubik's Cube consists of nine faces of the smaller coloured cubes. Each layer of smaller cubes can be rotated independently. The aim of the game is to perform rotations such that the faces of the Rubik's Cube are unicoloured, i. e. such that all smaller cubes which are visible at a face have the same colour.

As a possible formalization of the Rubik's Cube game, we can consider the configuration of colours on the cube's faces as state space, and the rotations as transitions between different states.

Tic-Tac-Toe. A well known game is Tic-tac-toe, which is similar to the Roman game *terni lapilli*. The game is played by two players who are alternately marking the spaces in a 3×3-grid. The player who has marked either three spaces in a row, column or on the diagonal wins.

As a possible formalization of the Tic-tac-toe game, the locations of the marks can be considered as a state space, and the activity of placing a mark corresponds to transitions between different state.

Throwing the Dice. When a dice is thrown, it shows a number between one and six with equal probability. The randomness is used in many games to make the game less deterministic and predictable. Consider that a person is throwing a dice repeatedly. In every step, the dice shows a certain number, is thrown and shows another number. The person who throws the dice may be able to memorize only a certain number of results. We only discuss the case of a person that is able to memorize one result at a time.

The number shown by the dice can be interpreted as a state. The state space consists of six states which can be represented by the numbers one to six. The activity of throwing the dice leads from one state to another one. As the result is random and every state can be gained, the graph representation is the complete graph with six nodes.

We discussed some examples of graph representations of human activity systems. We will discuss in the next section, how graph representations can be modified in order to reduce their size.

2.4 Modification of Graph Representations

The creation of graph representations includes choices, and these choices are usually made such that the graph representations are suitable for a certain purpose, e. g. for the examination of a spatial or temporal structure. Graph representations can thus contain information that is necessary for some purposes, but not for other ones.

Two methods of how to modify graph representations, with the aim of reducing the number of nodes and edges while preserving the relevant structural information, are discussed in this section: one method to emphasize the structure by replacing chains of edges by a new edge (section 2.4.1), and a second method which reduces the size of the state space by making it coarser (section 2.4.2).

Some of the algorithms used for the evaluation in chapter 5 have high complexity, and computations can thus only be made for smaller data sets. The discussed methods to reduce the size of a graph representation are necessary to evaluate the proposed model of spatial information on a number of data sets. Particular attention is paid to the example of public transport networks as they expose a (almost) prototypical spatial structure.

2.4.1 Packing of Edges

An essential aspect of the structure is, in case of a human activity system, which affordances are offered, i. e. which edges join a node in a graph representation. We propose, in the following, a method to reduce the number of edges while preserving many of the edges that join a node.

Packing of Graph Representations of Human Activities. The structure of graph representations can be examined for different purposes. Some nodes and edges may be more distinctive for the structure than others, depending on the purpose. We use the following principles as a general approach for emphasizing the structure of graph representations of human activity systems:

(a) nodes that have only few edges are eliminated, i. e. states of a graph representation are eliminated if it is predictable which activity is performed next, and
(b) nodes with many edges are conserved.

There exist several possibilities of how to pack a graph representation of human activity systems. We will consider the following one:

Definition 2.6. A graph representation of human activity systems is said to be *packed* if the following steps have been performed:

(a) for two edges (p, q) and (q, r) whose nodes are pairwise non-equal and the property that no other edge is joining the node q, the node q is removed from the graph and the two edges are replaced by an edge (p, r), and

(b) for two edges (p, q) and (q, r) as well as their opposite oriented edges whose nodes are pairwise non-equal and the property that no other edge is joining the node q, the node q is removed from the graph and the edges are replaced by the two edges (p, r) and (r, p).

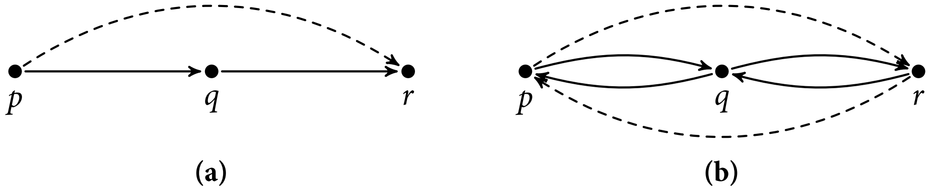


Figure 2.3
Steps executed for creating a packed graph representation; cf. definition 2.6

Step (a) merges sequences of consecutive edges whenever there exist no edges joining inner nodes. The same is performed in step (b) for sequences of consecutive edges if sequences in the opposite direction exist. Trips in public transport are, for example, usually operated in both directions, leading to opposite directed sequences in a time-ignoring graph representation.

Packing of Graph Representations of Public Transport. Additional information about activities is, in case of public transport, available, e. g. the trip that a transport activity belongs to. Graph representations of public transport activities can, due to this additional information, be packed more efficiently by paying attention to its specific structure:

Definition 2.7. A graph representation of public transport activities is said to be *timetable-packed* if the following steps have been performed:

(a) for edges $e_0, \dots, e_k$ from p to q and edges $e'_0, \dots, e'_k$ from q to r whose nodes are pairwise non-equal and who have the property that no other edge is joining the node q and that e_j and e'_j belong to the same trip for $j \in \{0, \dots, k\}$, the node q is removed from the graph and every pair (e_j, e'_j) of edges is replaced by an edge (p, r) of the same trip, and

(b) for edges $e_0, \dots, e_k$ from p to q, edges $e'_0, \dots, e'_k$ from q to r, edges $f_0, \dots, f_l$ from r to q and edges $f'_0, \dots, f'_l$ from q to p whose nodes are pairwise non-equal and who have the property that no other edge is joining the node q, that e_j and e'_j belong to the same trip for $j \in \{0, \dots, k\}$ and that f_j and f'_j belong to the same trip for $j \in \{0, \dots, l\}$, the node q is removed from the graph, every pair (e_j, e'_j) of edges is replaced by an edge (p, r) of the same trip and every pair (f_j, f'_j) of edges is replaced by an edge (r, p) of the same trip.

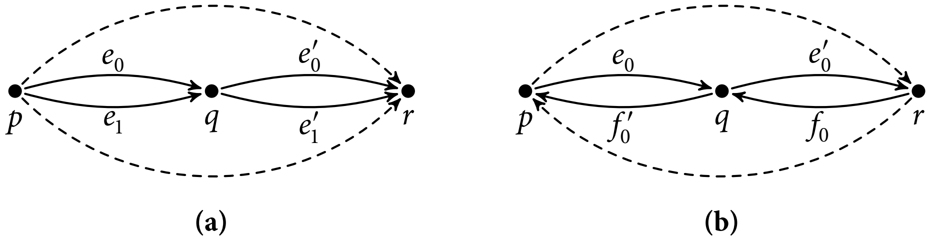


Figure 2.4
Steps executed for creating a timetable-packed graph representation; cf. definition 2.7

This packing is very similar to the packing of general graph representations, but even works for a much denser graph where many trips exist between a sequence of nodes. As each activity is associated with a certain trip, we are able to filter for parallel transport activities where a change between different transport modes is not very likely; two activities associated to the same trip are, most likely, of the same transport mode. Time-ignoring graph representations do not distinguish between vehicles stopping at different points in time at the same stop. The general algorithm for packing hence produces worse packing results.

We discussed how graph representations can be packed by replacing certain configurations of edges. This approach was justified in case of human activities and public transport. We will discuss a more general approach in the next section.

2.4.2 Collapsing the State Space

The state space of a graph representation specifies which states of the world can be distinguished by the representation. The state space of all possible states of the world would be considerably too large for real applications[6]. We discuss, in this section, how the state space of a graph representation can be collapsed in order to emphasize the relevant aspects of the graph representation for certain applications.

A graph representation consists of states and activities, i. e. transitions between the states. Both, the states and the activities, are choices that can be made when a representation is build. The considered activities of a graph representations determine which states are relevant. The state of 'being located at a stop' is, for example, relevant for public transport activities. When less activities are represented in the graph, potentially less states have to be represented.

Different sets of states may be used for the same set of activities, e. g. the location, or the location in combination with time as states for public transport activities. We discussed in section 2.1.3 a method to convert a time-considering graph representation (e. g. a graph representation with location and time as states) into a time-ignoring one (e. g. a graph representation with location as states), resulting in a smaller representation.

The same principle also applies in general: the number of states can be reduced by identifying certain states and thus making them indistinguishable. This approach can formally be realized by introducing an equivalence relation on the state space:

Definition 2.8. Let $\sim$ be an equivalence relation on the state space of a graph representation $\mathcal{G}(S, A)$. We define the graph representation $\mathcal{G}(S, A)/{\sim}$ to be the graph consisting of the nodes $S/{\sim}$ and the edges

$$A/{\sim} = \{([p], [q]) \mid (p, q) \in A, p \not\sim q\}$$

where $[p] \in A/{\sim}$ denotes the equivalence class corresponding to the node p. We call $\mathcal{G}(S, A)/{\sim}$ the *graph representation collapsed by* $\sim$.

[6] The state space would be infinit (or at least virtually infinit) and thus, potentially a proper class. It would, in this case, not be a set and the graph representation, not a proper graph.

By the very definition, the number of activities that are joining a state cannot decrease[7], and the resulting graph is a hypergraph. Activities in the collapsed state space are transitions from a state to the same state if the starting and end state are equivalent.

A time-ignoring graph representation can, as was motivated in section 2.1.3, be gained by collapsing the state space of a time-considering graph representation. The corresponding equivalence relation is, in case of public transport activities, the relation that identifies all states corresponding to the same stop, ignoring the time information.

The transformation of a time-considering to a time-ignoring graph (cf. section 2.1.3) as well as the packing of graphs (cf. section 2.4.1) are examples where a graph is collapsed by some equivalence relation. The collapsing of the state space does, in contrast to the packing of edges, not focus on how the number of edges shall be reduced but on how states are identified.

We have discussed two methods of modifying graph representations: one more specific method, which only applies to graph representations of human activities but requires no further knowledge about which activities are represented; and a more general method, which can be more effective but requires specific choices of which information shall be kept and which shall be lost.

Conclusion

Graph representations can describe structural aspects of human activities. Graphs are general enough to even represent other types of data, but they are specific enough to characterize the structure. This is why we introduced graph representations as an example of data that has, in many cases, a spatial structure. More specifically, the main contributions of this chapter are as follows:

(1) We introduced the conceptual foundations necessary to understand how representations are created and used.

(2) The creation and use of data sets was reviewed, including a discussion of the completeness and the quality of created data sets.

(3) Examples of graph representations were introduced by reusing existing formalizations of human activity systems.

(4) Possibilities of emphasizing the structure, by reducing the size of the graph representation while preserving the core structure of the representation, were explored.

We will use graph representations, in particular the ones discussed in section 2.3, as examples of spatial as well as non-spatial information, in subsequent chapters.

[7] The number of edges potentially decreases if all nodes with the same start and the same end node are identified.

3

A Scale-Invariant Spatial Graph Model

Att vara så rädd för en ringklocka! Ja, men det är inte bara en klocka – det sitter någon bakom den – en hand sätter den i rörelse – och något annat sätter handen i rörelse – men håll för örona bara – håll för örona! Ja, så ringer han ändå värre! ringer bara ända tills man svarar – och då är det för sent! […] Det är rysligt! Men det finns intet annat slut! – Gå!

—**August Strindberg, *Fröken Julie***
swedish playwright, poet and painter
(1849–1912)

Spatial structure is characterized by characteristic properties that are exposed by many spatial data sets. We will, based on these properties, develop a model of the spatial structure.

Tobler's first law of geography, as a typical property of spatial information, describes the relations between things dependent on their distance. The law can be deduced by the principle of least effort in case of human activities (section 3.1). Additional properties of spatial information include scale invariance and the distribution of the density of nodes in the graph (section 3.2). Existing graph models cannot serve as models of spatial structure in general, because they do not expose these properties (section 3.3). We introduce a model that has these typical properties of spatial information (section 3.4) and formally prove that it has these properties (section 3.5). When the number of nodes approaches infinity, we can prove further analytical properties of the spatial graph model (section 3.6). Parts of this chapter are based on a paper by Mocnik et al. (2015).

3.1 Principle of Least Effort and Tobler's First Law of Geography

Spatial structure is, amongst others, characterized by a very simple principle: near things are more related than distant ones. This principle is however not universally but only statistically true.

We discuss, in this section, the principle of least effort and show how Tobler's first law of geography can be deduced from this principle. In the second part of this section, we discuss the role of metrics on Tobler's first law of geography.

Principle of Least Effort. Zipf claimed that there is, in many cases, no reason why humans should take more effort than needed to reach an aim. Zipf introduced the following hypothesis to reflect this observation (Zipf 1947):

Theorem 3.1 (Principle of least effort). *Effort is always minimized in human behaviour.*

When movements in space and time are costly, the principle of least effort claims that movements are minimized, because effort is. In case of compound activities, the total effort is minimized. This does not mean that the effort of every included activity is minimized but that it is very likely. Movements in space and time are, in particular, only performed if they are necessary or minimize the effort in compound activities. This effect on movements, sometimes called *cohesion*, leads to accumulations of people, and it was conceptualized by Stewart (1948) as *demographic force*.

Analogously to gravitational force leading to extended theories which incorporate distance, force, work and energy, *gravity models* for human activities have been build (Fotheringham et al. 1980). These models incorporate distance decay functions, which predict interactions between two places to be more likely the shorter the distance is between both (Halás et al. 2014). This behaviour can, for example, be observed for cities and their surroundings (Kopczewska 2013).

A model of gravitational force was used by Tobler et al. (1971) to predict the relative locations of ancient Hittite villages in central Turkey which were named on the Cappadocian tablets: when village names are assumed to more often cooccur on the tablets in case they are related, gravitational forces can be assigned to the towns. These forces can be used to predict the relative locations of the villages. Constantine et al. (1993) applied these considerations to the example of departments in modern France; instead of the cooccurrence, the relation of having a common border was examined.

Tobler's First Law of Geography. Gravity models predict, as discussed before, human interactions between two places to be more likely the shorter the distance is between both. Tobler observed that this prediction is also true for many other relations (Tobler 1970):

Theorem 3.2 (Tobler's first law of geography/Tobler's law). *Everything is related to everything else, but near things are more related than distant things.*

The law cannot be understood as a strict law but rather as a universal statement that is true in numerous cases, as was e. g. discussed by Hayashi (2006) and Hecht et al. (2009). The law is however not necessarily true for all types of information whose entities are related to space. There however exist networks that can naturally[1] be embedded in space but do not follow Tobler's law, e. g. qualitative constraint networks of spatial entities (Fogliaroni 2013).

Tobler's law relates to the concept of neighbourhood: things in the same small neighbourhood are near and, in many cases, more related than others, while distant things are not in the same small neighbourhood and thus presumably less related. The concept of neighbourhood is a central topological concept of space (Kuhn 2012), demonstrating that Tobler's law describes parts of the nature of space.

Scope of Tobler's First Law of Geography. Tobler's law claims that a correlation exists between relations in space, by the concept of 'near', and relations of some other domain. There is no reason why the claim should not be true for two arbitrary domains, but just in case that one of the domains is space. As long as things of two domains are related, it seems likely that also the relations of these two domains *can* correlate. Such a generalization of Tobler's law is not true in general but applies to many examples.

One example of a generalization of Tobler's law is the correlation between intra-industry trade and GDP. The exchange of similar products, which belong to the same type of industry, between two countries in both directions is called intra-industry trade. It has been shown that intra-industry trade is correlated to numerous factors, amongst others to per capita income, country size, (spatial) distance of the involved countries, common borders, trade barriers, common languages and product standardizations (Balassa et al. 1987). Analogously to Tobler's law, which claims that things are more related if they are near *in space*, Taegi et al. (2001) proved that countries have a high share of intra-industry trade if their economics are of similar size, i. e. if their GDPs are 'near' in the one-dimensional vector space representing the size of the GDP.

Tobler's law describes the structure of information. As the generalization of Tobler's law applies to at least some examples, Tobler's law is however not a unique property of spatial information.

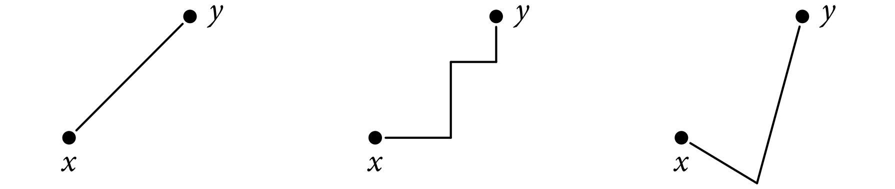
Meaning of 'Near' in Tobler's First Law of Geography. Tobler's law leaves vague what is meant by 'near'. Tobler (2004) and Miller (2004^b) point out that several concepts of 'near' are appropriate in the context of Tobler's law. These concepts relate, more or less, to some distance function in space. Such a distance function is not necessarily metric (cf. definition A.1), but it is expected to share at least some properties with a metric. Examples of such distance functions that are not metrics are travelling time and fuel consumption. The choice of the distance function has an influence on how spatial structure can be modelled by Tobler's law. This influence can be estimated and is not very decisive in many cases, as will be discussed in the following.

[1] The term *naturally* refers to the fact that a distinguished embedding exists, and that no choice has to be made to distinguish this embedding. There exist many embeddings, but only one of them is distinguished because the embedding maps things to their real location in space.

Different metrics assign to a pair of points different distances. It has been proven that the metric distance between points in a real vector space of finite dimension differs only to a constant factor for different metrics induced by p-norms[2] (Dieudonné 1969, p. 106). This fact shows that the difference between such metrics is limited.

A metric gives a meaning to the word 'near', because it defines neighbourhoods[3] by the induced topology. All metrics induce the same topology on real vector spaces of finite dimension, because a homeomorphism can be constructed between any pair of real vector spaces of the same dimension, even when endowed with different topologies (Rudin 1991, pp. 16f). In particular, the Euclidean metric, the Manhattan metric, the French railway metric and others all induce the same topology.

Figure 3.1
Visualization of the construction of different metrics; (a) Euclidean metric, (b) Manhattan metric, and (c) French railway metric



(a) Euclidean metric **(b)** Manhattan metric **(c)** French railway metric

Mathematical results have, in the context of Tobler's law, only limited validity, because the concepts of 'near', neighbourhood and distance function in Tobler's law do not completely coincide with the corresponding mathematical concepts. The similarity between the concepts however suggests that models of spatial structure which incorporate Tobler's law depend only little on the choice of a distance function.

We discussed the principle of least effort and Tobler's law as well as the scope of Tobler's law and its role in spatial information. We argued that Tobler's law is a typical property of spatial information and that the choice of a metric is not crucial when Tobler's law is used for modelling spatial structure, because the metric influences the meaning of 'near' only very limitedly. We will, in the next section, discuss additional typical properties of spatial information that the proposed model of spatial structure is expected to have.

3.2 Typical Properties of Spatial Information

Spatial information is statistically characterized by some core properties, and we expect a model of spatial structure to expose these properties. Tobler's law has already been discussed in the last section, because it is the most prominent one. Graph representations of data were motivated in section 2.1.2 to be beneficial for the purpose of discussing the data's structure. Spatial information will thus in the following be assumed to be represented as graphs.

[2] The class of metrics induced by p-norms is the most important one and includes the Euclidean and the Manhattan metric (cf. chapter A.1).

[3] The term *neighbourhood* is used with slightly different meanings in mathematics and spatial sciences, but both concepts share the idea of representing regions that are partially ordered by a containment relation.

We review, in this section, Tobler's law as a core property of spatial structure (section 3.2.1). Scale invariance is another property of spatial structure. It claims that the structure of data is invariant, when space is scaled and the distance between nodes in space is thus scaled by a constant factor (section 3.2.2). The outdegree of nodes in a graph representation of spatial information is typically bounded, in regions of low node density as well as in regions of high node density (section 3.2.3).

The properties and the structure of spatial information are based on the properties of space and the entities that constitute space: the existence of distance and the effort of travelling leads to a predominance of relations between near things; the similarity of space and physical processes at different scales of tangible reality leads to scale invariance of spatial information; and non-uniform distributions of objects in space lead to not necessarily uniform but in many times bounded distributions of relations. We call such properties of spatial information based on space a *spatial structure*[4] in this paper, and we say that a data set *has a spatial structure* if it exposes some of these properties.

3.2.1 Tobler's First Law

Tobler's law is a correlation of information and space, which is typical for geographical information, but also for many examples of spatial information in general. This correlation is an autocorrelation, because the configuration of things, i. e. their relations, is influenced by their location in space, and the location of things in space is influenced by their configuration. The law was discussed in section 3.1.

3.2.2 Scale Invariance

Space, conceptualized as a Euclidean vector space, has no preferred unit. After rescaling space, it cannot be distinguished from the unscaled one, and many physical processes of our tangible world remain (nearly) the same when rescaled. Classical mechanics holds, for example, for everyday items as well as for solar systems. As soon as objects are placed in space, they define a unit and a scale. If interrelations between objects only depend on relative distances and the Euclidean structure, the objects and their interrelations do not change with rescaling. This effect of *scale invariance*[5] can be observed in several data sets, e. g. for metro and railway networks (Louf et al. 2014) and road networks (Kalapala et al. 2006).

3.2.3 Bound Outdegree

The average edge degree in a planar graph can be proven to be strictly less than six, which can be seen by Euler's formula and the fact that a face has at least three edges and each edge has at most two faces. A result by Chrobak et al. (1991) shows that the edges of a planar graph can be oriented such that the outdegree is bounded by three. We expect that the outdegree of a graph embedded in space behaves similar,

[4] Time has a similar effect on information as space, because it can also be modelled by (one-dimensional) Euclidean vector spaces.

[5] The concept of *scale invariance* of a collection, possibly denoted as a graph, of objects and their interrelations embedded in space, should not be confused with the concept of *scale-freeness* of a graph; scale-freeness is characterized by a power law distribution of the nodes' edge degrees and hence, by the invariance of the distribution's shape, invariant under rescaling of the total number of edges.

even if it is not completely planar, and we expect the outdegree, in consequence, to have an upper bound which is considerably lower than the one of a complete graph.

When nodes are non-uniformly distributed in space, we could expect the outdegree to be non-uniformly distributed for different nodes as well. Following the above argument, we however expect the outdegree to be bounded, as is true in the example of public transport: nodes representing stops of public transport are usually more dense in city centres than in the countryside, but there exist edges in the countryside, and the outdegree is not arbitrarily high in city centres.

3.2.4 Summary

The discussed properties are not valid for every spatial data set but for many ones, and they can be motivated by the structure of space. A graph model of spatial information should thus have these typical properties of spatial information:

Typical Properties of Spatial Information 3.3. Graph representations of spatial information have, in many cases, the following properties:

(a) nodes in the same neighbourhood are more likely to be adjacent than others,
(b) edges exist in regions of low node density,
(c) only a limited number of edges exists in regions of high node density, and
(d) the distribution of entities in space is independent of scale.

We discussed typical properties of spatial information. These properties will be used in section 3.4 to motivate the construction of the proposed model of spatial structure. We will review existing graph models in the next section and motivate, why they are not suitable to model spatial structure in general, based on the typical properties of spatial information, which were discussed.

3.3 Existing Graph Models

Graph models with several structures exist. It has turned out that they can be successfully used for many applications. There exists, yet, no model of spatial structure.

We review, in this section, existing graph models, in particular random graph models (section 3.3.1) and structural graph models (section 3.3.2). These graph models can be classified as complex networks, scale-free networks and small-world networks (section 3.3.3). Further graph models are modelling spatial aspects (section 3.3.4) and temporal aspects (section 3.3.5). These models are not suitable to model spatial structure in general, as we will argue by the typical properties of spatial information, which were discussed in the last section.

3.3.1 Random Graph Models

Euler's discussion of the *Königsberg bridge problem* (Euler 1741) is considered to be the first publication on graphs. Since then, the construction of graphs has been studied in connection with its properties, e. g. by Barabási (2002). Graphs have been constructed as formalization of reality, e. g. by Kirchhoff (1845), who studied the current and the potential difference in electrical circuits, resulting in Kirchhoff's circuit laws.

In 1959, Erdős and Rényi as well as Gilbert independently introduced randomly generated graphs (Erdős et al. 1959, Gilbert 1959). Both models were the starting-point for a new generation of simple graphs that can be described by probability distributions. Graphs having this property are called *random graphs*. The following models are examples of random graphs:

Erdős-Rényi Model. A graph of the model $\mathcal{G}_{\text{Erdős-Rényi}}(n, e)$ is a randomly chosen graph of all simple graphs consisting of n nodes and e edges (Erdős et al. 1959).

Gilbert Model. A graph of the model $\mathcal{G}_{\text{Gilbert}}(n, p)$ is a simple graph consisting of n nodes and an edge between two nodes with probability p (Gilbert 1959).

Random graph models are not suitable for modelling spatial structure in general, because spatial structure is not completely random: the structure of spatial information is influenced by space, and some configurations of edges are expected to occur more often than others.

3.3.2 Structural Graph Models

Graph models of specific aspects of information, modelling its specific structure, have been introduced. They have been successfully applied to various use cases.

Barabási-Albert Model. A simple graph of the model $\mathcal{G}_{\text{Barabási-Albert}}(n)$ is incrementally constructed by adding nodes to an existing graph, beginning with a graph consisting of one node and no edge. In each of the $n - 1$ subsequent steps, a node p is added as well as an edge (p, q) for each node q of the graph with the probability $\deg q \big/ \sum_r \deg r$ where r runs over the set of all nodes. The majority of nodes is, in consequence, joined by a very low number of edges, whereas only a very low number of nodes is joined by a large number of edges, resulting in a power-law degree distribution (Barabási et al. 1999). This model and similar ones have been used to model internet links (Barabási et al. 2000, Yook et al. 2002), citation networks (Barabási et al. 2002) and social networks (Sala et al. 2010). The construction of Barabasi-Albert models does not reflect Tobler's law and is not able to model spatial structure in general.

Watts-Strogatz Model. For the construction of a simple graph of the model $\mathcal{G}_{\text{Watts-Strogatz}}(n, k, p)$ with $k \in 2\mathbb{Z}$ and $0 \leq p \leq 1$, two steps are executed. In the first step, a regular ring lattice with nodes p_i, $i \in \{0, \dots, n-1\}$, is constructed. In such a lattice, two nodes p_i and p_j are adjacent if and only if $0 < |i - j| \bmod (n - k/2) \leq k/2$. Each node has, thus, k edges. In the second step, each edge is rewired with probability p. We can construct completely regular graphs ($p = 0$), completely random graphs ($p = 1$) and graphs which are in between ($0 < p < 1$) by choosing a suitable value for the parameter p (Watts et al. 1998). Spatial graphs usually have longer path length than this model, because edges tend to exist only in neighbourhoods.

Exponential Random Graph Models. The exponential family of distributions is an important class of probability distributions including normal, exponential, Poisson and many more distributions. Any graph model whose edges follow a distribution of the exponential family is called an *exponential random graph model* (Holland et al. 1981). Exponential random graph models have been used to model social networks (Hunter et al. 2008). These graphs are tailored to fit statistical properties, but they do not refer to spatial properties and are incapable to model spatial structure.

Hierarchical Network Model. The class of *hierarchical network models* includes all graph models that are based on an iterative way of replicating an initial graph and adding edges between the replicates and the initial graph (Barabási et al. 2001). These models are suitable to model hierarchical aspects, which spatial data, in principle, can have. Spatial data however is, at the core, not solely characterized by hierarchies but primarily by Tobler's law and other properties.

Additional graph models have been introduced for specific properties, e. g. the pairing model (Wormald 1999) as a model of random regular graphs, i. e. graphs where each node has the same degree. These models are less well-known and do not expose typical properties of spatial information.

3.3.3 Classification of Graph Models

The discussed graph models can loosely be classified by a number of graph properties. We review, in this section, widely used classes of graph models and provide an overview of how existing graph models can be classified.

Complex Networks. A graph is called a *complex network*[6] if it has non-trivial topologic features, i. e. features only depending on the abstract graph that can neither be found in completely regular nor in completely random graphs. Features that have been of interest are a degree distribution with high values for high degrees, a high clustering coefficient and a hierarchical structure. Examples are scale-free and small-world graphs (Strogatz 2001, Albert et al. 2002, Newman 2003^{b}).

Scale-Free Networks. A graph is called *scale-free* if the number $P(k)$ of nodes with k edges follows a power law for large values of k, i. e. if there exists a real

[6] The terms *graph* and *network* will be used synonymously in this thesis. For a detailed view on the difference, see section A.2.

Model	complex	scale-free	small-world	ultra-small-world
Erdős-Rényi	no	no	yes	no
Gilbert	no	no	yes	no
Barabási-Albert	yes	yes	no	yes
Watts-Strogatz	yes	no	yes ($p = 1$)	no
Hierarchical	yes	yes	no	yes

Table 3.1
Classification of graph models according to widely used classes of networks. The Exponential random graph model is not included, because it heavily depends on the choice of the distribution.

number γ such that

$$P(k) \sim k^{-\gamma}.$$

The Barabási-Albert model and the hierarchical network model are examples of scale-free graph models.

In a scale-free network, there exist relatively many nodes, called *hubs*, with a degree that exceeds the average degree. These hubs are adjacent to nodes with a smaller degree, which itself usually are adjacent to nodes with even smaller degree, etc.

The term 'scale-free' refers to the fact that the power law distribution does not depend on the number of edges in the graph, because

$$P(ck) \sim c^{-\gamma}k^{-\gamma} \sim k^{-\gamma} \sim P(k).$$

It has been shown that this behaviour comes in many cases, but not necessarily, along with an inductive creation of a graph where new edges are introduced more likely between nodes of higher degree. Such a process of inductive creation can be found in the links of the web (Barabási et al. 1999).

Small-World and Ultra-Small-World Networks. A graph is called a *small-world graph*, if the average shortest path length $L(n)$ between randomly chosen pairs of nodes grows proportionally to the logarithm of the number of nodes n in the graph, i. e.

$$L(n) \sim \log n.$$

If the average shortest path length grows even slower with $L(n) \sim \log\log n$, the graph is called an *ultra-small-world graph*. Scale-free graphs are examples of ultra-small-world graphs (Cohen et al. 2003). Some neural networks, power grid networks and collaboration networks of film actors have been shown to be small-world graphs (Watts et al. 1998).

An overview of the classification of existing graph models is provided in table 3.1. We will use this classification as a basis to discuss in section 4.7 how the scale-invariant spatial graph model, which is introduced in section 3.4, can be classified.

3.3.4 Graphs Related to Space

An overview over graphs which are related to space and their properties has been given by Barthélemy (2011). The following classes of graphs relate to space:

Planar Graphs. Mathematical research has been conducted on *planar graphs*, i. e. graphs which can be embedded in $\mathbb{R}^2$ such that their edges do, apart from their endpoints, not intersect.

Many characterizations of planar graphs have been discussed (Kuratowski 1930, Whitney 1931, MacLane 1937, Wagner 1937, de Fraysseix et al. 1982, Schnyder 1989, de Verdière 1990, Archdeacon et al. 1995). Graphs have been proven to be planar if and only if they neither contain, after the contraction of edges, the complete graph K_5 with 5 nodes nor the complete bipartite graph $K_{3,3}$ with 6 nodes as a subgraph (Wagner 1937). Spatial data sets usually cannot be represented by planar graphs, because spatial structures characterize data sets globally and spatial data sets thus remain spatial after local modifications, in particular after the introduction of K_5 or $K_{3,3}$ as subgraphs. The study of planar graphs however can help us to gain a deeper understanding of graphs in the plane.

(a) K_5

(b) $K_{3,3}$

Figure 3.2
Graphs that do not occur as subgraphs of planar graphs; (a) K_5, and (b) $K_{3,3}$

Spatial Graphs. Another important class contains graphs whose nodes can be embedded in space (Haggett et al. 1969), either by their natural location or by an explicitly chosen embedding, e. g. in the case of conceptual spaces (Gärdenfors 2000, pp. 15ff). Graphs of this class are called *spatial graphs*, because they are, by its features, related to space, usually Euclidean space. These features are either explicit by semantic annotation, or implicit by the structure of the graph (Barthélemy 2011), and they are, in many cases, empirically collected. These graphs usually model specific applications, and empirical data about the application is needed.

Spatial Graphs with Given Statistical Properties. Kosmidis et al. (2008) construct graphs which are embedded in space and have given statistical properties. The constructed graphs have, in particular, a given degree distribution and a given distribution of distances between adjacent nodes. The properties of such graphs strongly depend on the given distributions, and they do, in general, not coincide with the ones of spatial structures.

Spatial Generalizations of Complex Graph Models. The Erdős-Rényi and the Gilbert model can be modified in two ways in order to generate only planar graphs: edges can, after the construction, be removed such that a planar graph is achieved (Barthélemy 2011), or only planar graphs can be considered during the construction (Denise et al. 1996, McDiarmid et al. 2005). The model has even been generalized to graphs that are embeddable on fixed surfaces (McDiarmid 2008).

Spatial generalizations of the Barabási-Albert model require each node to have a location in a Euclidean space, and new nodes are usually randomly distributed with uniform distribution. The probability to add an edge (p, q) from the new node to an old one is $\deg q \big/ \sum_r \deg r \cdot f(d(p, q))$ where r runs over the set of all nodes and f is a function, e. g. an exponential function (Barthélemy 2003) or a power function (Xulvi-Brunet et al. 2002, Yook et al. 2002).

The Watts-Strogatz model has been generalized by rewiring the graphs' edges not only by a fixed probability p but also by the distance between the nodes (Jespersen et al. 2000).

These generalizations share aspects of spatial information, but as most of their characteristics originate from the non-generalized models, they are not suitable as models of spatial structure in general.

Geometric Graph Models. This class of models assumes nodes to have explicit locations in a Euclidean space, and edges are modelled by the nodes' location in space. These models usually start with a random set of points (uniformly distributed) and introduce an edge between two nodes p and q with distance $d(p, q)$ with probability $f(d(p, q))$, where $f\colon \mathbb{R} \to [0, 1]$ is a probability function. Huson et al. (1995) uses the probability function

$$f(l) = \begin{cases} 1 & \text{if } l < r \\ 0 & \text{otherwise} \end{cases}$$

to model a network of radio transmitters and receivers[7]. Waxman (1988) discusses a similar model with a smoothened, continuous probability function

$$f(l) = \beta \exp\left(-\frac{l}{r}\right).$$

Both models depend on the absolute distance between points and are, in consequence, not scale-invariant. They are not suitable to model spatial structure.

Aldous et al. (2013) discussed the class of scale-invariant graphs. A scale-invariant variant of this model was discussed by Aldous et al. (2010): for a given $k > 0$, edges to the k nodes with minimal distance are introduced for each node. This model does not reflect the fact that the distribution of the edges in spatial information, in particular the number of edges per node, usually depends on the location of the nodes in space. The number of bus routes leading through a certain stop, for example, is not constant.

Graph models do not necessarily expose a spatial structure, even if they are related to space, because they do not necessarily have the structure that spatial information commonly has. The discussed models are, thus, no models of spatial structure.

[7] In case that devices with a higher range are chosen for areas of low device density, the model does not apply: it fits well to handheld transceivers (all of the same power) but not to radio-relay systems.

3.3.5 Graphs Related to Time

Time shares some features with (one-dimensional) space, because both have a Euclidean structure. Graphs describing temporal data may, hence, suggest itself to be familiar to spatial graphs.

Evolving Networks. Change in networks can be modelled over time: nodes and edges can be added, modified or removed. Such an evolving network is not suitable to model spatial structure, because it is only able to model dynamic structures, e. g. the nervous system or modes of communication. A more detailed view has been given by Holme et al. (2012).

Existing graph models are, as was argued, tailored to model specific structures but not spatial structure in general. They thus cannot serve as models of spatial structure.

3.4 The SISG Model

A graph model of spatial structure can be expected to have the typical properties of spatial information, which were discussed in section 3.2. Such a model is *not* a model of public transport nor a model of human activities. It is a model that shares some properties with many types of spatial information, namely these properties that seem to be central in the concept of spatial information.

We propose, in this section, a model of spatial structure. The model is motivated by properties 3.3, and a formal proof will be provided in section 3.5.

For a given set of nodes embedded in space, we ask which edges have to be introduced such that the resulting graph meets properties 3.3. If the graph has property 3.3(a), the configuration of the edges depends on the distance between nodes. On the other hand, the configuration must not depend on the absolute distance between nodes in order to have property 3.3(d), and the graph model is only allowed to depend on the *relative* distances between nodes. In order to have properties 3.3(b) and (c), the number of edges joining a node may vary for regions of different node density, but the number of edges has to be bounded. The following graph model has these properties, as we will prove in section 3.5.

Definition 3.4. (1) Let V be an n-dimensional Euclidean vector space with metric d. To a finite set of points $S \subset V$ and a real number $\rho > 1$, we associate the abstract (directed and simple) graph $\mathcal{M}_\rho(S, V)$ consisting of

(i) a node for every point $p \in S$, and
(ii) a directed edge (p, q) if and only if

$$d(p, q) \leq \rho \cdot \min_{p_0 \in S \setminus \{p\}} d(p, p_0)$$

where the minimum is assumed for p_0 being a nearest neighbour of p.

The graph $\mathcal{M}_\rho(S, V)$ is called the *scale-invariant spatial graph model (SISG model) of the generating set $S \subset V$ of dimension* $\dim V$ *and density parameter ρ*. We call $\mathcal{M}_\rho(S, V)$ to be *generated* by the set S. If the vector space is apparent from the context, we will even write $\mathcal{M}_\rho(S)$.

(2) The simple undirected graph associated to $\mathcal{M}_\rho(S)$ is called the *undirected scale-invariant spatial graph model.*

A model $\mathcal{M}_\rho(S, V)$ is, by definition, an abstract graph, i. e. a node $\tilde{p}$ is an abstract object that is not placed in the vector space V; each node $\tilde{p}$ can however canonically be identified with a point $p \in V$ in space. We will denote both, the node $\tilde{p}$ and the corresponding point p, by the same symbol as long as no confusion arises.

We did only motivate that SISG models have the typical properties of spatial information, which were discussed in section 3.2. A formal proof will be provided in the next section.

3.5 The SISG Model Has Typ. Properties of Spatial Information

The SISG model is a model of spatial structure, and it is thus expected to have properties 3.3. We provide, in this section, formal proofs that the model has these properties.

3.5.1 Property (a)

Tobler's law claims that near things are more related than distant ones. The construction of the SISG model reflects this law by introducing edges only in neighbourhoods. We can formally prove:

Proposition 3.5. *(1) If a node p of $\mathcal{M}_\rho(S)$ is adjacent to a node q, then it also is to every node q' with $d(p, q') \leq d(p, q)$.*

(2) If a node p of $\mathcal{M}_\rho(S)$ is not adjacent to a node q, then it is neither to any node q' with $d(p, q') \geq d(p, q)$.

Proof. (1) Let p, q and q' be nodes with $d(p, q') \leq d(p, q)$ and (p, q) an edge. Then

$$d(p, q') \leq d(p, q) \leq \rho \cdot \min_{p_0 \in S \setminus \{p\}} d(p, p_0).$$

Thus, also (p, q') is an edge.

(2) Let p, q and q' be nodes with $d(p, q') \geq d(p, q)$ such that no edge (p, q) exists. Then

$$d(p, q') \geq d(p, q) > \rho \cdot \min_{p_0 \in S \setminus \{p\}} d(p, p_0).$$

Thus, there exists no edge (p, q') either. □

The proposition proves SISG models to have property 3.3(a), because for each node of the graph there exist edges to all other nodes of a certain neighbourhood, i. e. to all nodes that are nearer than a certain distance, and no edges to all other nodes of the graph.

3.5.2 Property (b)

When the density of nodes in space is lower, there are, statistically, less nodes in a neighbourhood of a fixed size. The low number of neighboured nodes can lead to single nodes that are not connected to other nodes. By the construction of the SISG model, there however exists for each node at least one outgoing edge, namely the one starting in that node and ending in a node at minimal distance. We can prove:

Proposition 3.6. *If there exist at least two nodes in $\mathcal{M}_\rho(S)$, every node has outdegree of at least* 1.

Proof. Assume that there exist at least two nodes. For an arbitrary node p, there exists a node $p' \neq p$ such that $d(p, p') \leq d(p, q)$ for all nodes $q \in S \setminus \{p\}$, because S is finite. As $\rho > 1$, we have $d(p, p') < \rho \cdot \min_{p_0} d(p, p_0)$. This proves (p, p') to be an edge. □

The proposition proves SISG models with at least two nodes to have property 3.3(b), because every node has outdegree of at least 1. If the density parameter is significantly larger than 1, the average outdegree can be expected to be, too.

3.5.3 Property (c)

In regions of higher node density, there are, statistically, more nodes in a neighbourhood of a fixed size. The number of edges between such neighboured nodes is bounded by the number of edges in the complete graph. The expectation value of the number of edges is however much lower for uniformly distributed nodes, as is formally proven by the following theorem:

Theorem 3.7. *In a SISG model $\mathcal{M}_\rho(S, V)$ with S uniformly distributed, the expectation value of the outdegree of a node converges to $\rho^{\dim V}$ for $|S| \to \infty$.*

Proof. Consider points to be uniformly distributed in a vector space of dimension $n = \dim V$. For an arbitrarily chosen point p and a real number $L > 0$, let S be the set of all points in the n-dimensional ball $B_n(p, L)$ of radius L centred in p. We denote the minimal distance between p and the remaining points by $r = \min_{p_0 \in S \setminus \{p\}} d(p, p_0)$.

If for an $R < L$ the n-dimensional open ball $B_n(p, R)$ does not contain any point of S apart from p, the points of $S' = S \setminus \{p\}$ are in $B(L, R) = B_n(p, L) \setminus B_n(p, R)$. Denoting the volume of the n-dimensional ball of radius L by $\mathrm{Vol}_n(L)$, the density of points in the ball $B_m(p, L)$ equals $s/\mathrm{Vol}_m(L)$ with $s = |S|$ for $L \gg R$. We thus expect

$$\mu = \frac{s}{\mathrm{Vol}_n(L)} \cdot [\mathrm{Vol}_n(\rho R) - \mathrm{Vol}_n(R)] + 1 = s(\rho^n - 1)\frac{\mathrm{Vol}_n(R)}{\mathrm{Vol}_n(L)} + 1 \qquad (*)$$

points in $B(\rho R, R)$, namely the one[8] at minimal distance r and the ones in the inner of $B(\rho R, R)$. (The second equality is due to the fact that $\mathrm{Vol}_n(\rho R)$ equals $\rho^n\,\mathrm{Vol}_n(R)$.) If $R \leq r$, we expect at least μ points in $B(\rho r, r)$, i. e. at least μ edges starting in p.

For a given R, the probability of $R \leq r$, i. e. the probability that all $s - 1$ points S' have distance greater than R to the point p, is

$$\left(1 - \frac{\mathrm{Vol}_n(R)}{\mathrm{Vol}_n(L)}\right)^{s-1}.$$

Inserting equation $*$ proves that the probability of at least μ edges starting in p is

$$\nu(\mu) = \left(1 - \frac{\mu - 1}{s(\rho^n - 1)}\right)^{s-1}.$$

The probability that at most μ edges are starting in p equals $1 - \nu(\mu)$, and the corresponding probability density function is given by $-\frac{d}{d\mu}\nu(\mu)$. To compute the expectation value for the number of edges starting in p, we first compute

$$\begin{aligned}\pi(\mu) &= -\int \mu \frac{d}{d\mu}\nu(\mu)\,d\mu \\ &= -\mu \cdot \nu(\mu) + \int \nu(\mu)\,d\mu \\ &= \left[(\mu - 1)\left(\frac{1}{s} - 1\right) - \rho^n\right] \cdot \nu(\mu).\end{aligned}$$

The expectation value of the number of edges starting in p can be computed as

$$\pi(\mu)\Big|_1^{s-1} = \rho^n + \left[(s-2)\left(\frac{1}{s} - 1\right) - \rho^n\right] \cdot \left(1 - \frac{s-2}{s} \cdot \frac{1}{\rho^n - 1}\right)^{s-1}.$$

The second summand vanishes for $s \to \infty$. □

This proves SISG models to have property 3.3(c). We can conclude:

Corollary 3.8. *The graph $\mathcal{M}_\rho(S, V)$ with S uniformly distributed is expected to have $|S| \cdot \rho^{\dim V}$ edges for $|S| \to \infty$.*

3.5.4 Property (d)

The SISG model introduces edges for a given set of points in space, and the configuration of the edges only depends on the location of the points in space. The

[8] The cases where more than one point is at minimal distance is a null set.

change of scale in the vector space, which is a transformation of the vector space, leaves the configuration of edges invariant, as has been discussed in section 3.2.2. We formally define a scale transformation as follows:

Definition 3.9. Let V be a Euclidean vector space. A map $\tau\colon V \to V$ is called a *scale transformation* or a *transformation of relative scale* σ if and only if for every pair p and q, their distance is increased by the factor $\sigma > 0$ under the application of τ:

$$d(\tau(p), \tau(q)) = \sigma \cdot d(p, q).$$

The transformations of relative scale 1 are exactly the isometries, i. e. distance-preserving transformations $\tau\colon V \to V$. Other examples of scale transformations are changes of the bases of a vector space that scale each base vector by the same factor, i. e. transformations of the form $\sigma \cdot \mathrm{id}$. As the factor is not vanishing, scale transformations are bijective:

Proposition 3.10. *Every scale transformation is bijective.*

Proof. Let τ be a scale transformation of relative scale σ. Assume $\tau(p) = \tau(q)$. Then, the distance between $\tau(p)$ and $\tau(q)$ is vanishing. As $\sigma > 0$, the distance between p and q is also vanishing. The points p and q are hence equal. This proves that τ is injective.

Define a scale transformation $\iota\colon V \to V$ as the linear map $1/\sigma \cdot \mathrm{id}$. As ι is a scale transformation of relative scale $1/\sigma$, the concatenation $\tau \circ \iota$ is a scale transformation of relative scale 1, i. e. an isometry. A result by Wobst (1975) shows that $\tau \circ \iota$ is surjective, and thus, also τ. □

The Mazur-Ulam theorem proves that scale transformations are combinations of translations and linear transformations:

Proposition 3.11. *Every scale transformation is affine.*

Proof. Let τ be a scale transformation of relative scale σ. Define $\iota\colon V \to V$ as the linear map $1/\sigma \cdot \mathrm{id}$. Then $\tau \circ \iota$ is an isometry. The Mazur-Ulam theorem proves that $\tau \circ \iota$ is affine (Mazur et al. 1932, Wobst 1975). As ι is linear and bijective, the map τ is affine. □

Scale-transformations leave SISG models invariant, because the construction of SISG models does not refer to the absolute locations of the points of the generating set in space, but only to the distances between these points. We are thus able to prove that SISG models have property 3.3(d):

Theorem 3.12. *SISG models are invariant under scale transformations, i. e.*

$$\mathcal{M}_\rho(S) = \mathcal{M}_\rho(\tau(S))$$

for every scale transformation $\tau\colon V \to V$.

Proof. A scale transformation τ maps the nodes of $\mathcal{M}_\rho(S)$ to the one of $\mathcal{M}_\rho(\tau(S))$, and as τ is injective by proposition 3.10 and by the definition of $\mathcal{M}_\rho(\tau(S))$ surjective, it is an isomorphism between the sets of nodes.

There exists, by definition of the SISG model, a directed edge (p, q) in $\mathcal{M}_\rho(S)$ if and only if

$$d(p, q) \leq \rho \cdot \min_{p_0 \in S \setminus \{p\}} d(p, p_0).$$

As τ is a scale transformation of relative scale $\sigma > 0$, this is equivalent to

$$d(\tau(p), \tau(q)) \leq \rho \cdot \min_{\tau(p_0) \in \tau(S) \setminus \{\tau(p)\}} d(\tau(p), \tau(p_0)).$$

This equation is the condition for the existence of an edge $(\tau(p), \tau(q))$, which proves that an edge (p, q) in $\mathcal{M}_\rho(S)$ exists if and only if an edge $(\tau(p), \tau(q))$ exists in $\mathcal{M}_\rho(\tau(S))$. This proves τ to induce an isomorphism of graphs. □

The theorem proves SISG models to have property 3.3(d). It applies, in particular, to translations because translations are scale-transformations:

Corollary 3.13. *SISG models are invariant under translations, i. e.*

$$\mathcal{M}_\rho(S, V) = \mathcal{M}_\rho(v + S, V)$$

for every vector $s \in V$.

We formally proved SISG models to have properties 3.3. Further analytical results are proven in the next section.

3.6 Propositions on SISG Models

Analytical properties of SISG models are hard to prove, because the structure of a model depends on the locations of its nodes, and these locations are irregular. When no information about the locations is known, only very little conclusions can be drawn.

We provide, in this section, simple analytical results for SISG models. Many properties are, due to their complexity, not analytically discussed. A statistical discussion of these properties is postponed to the next chapter.

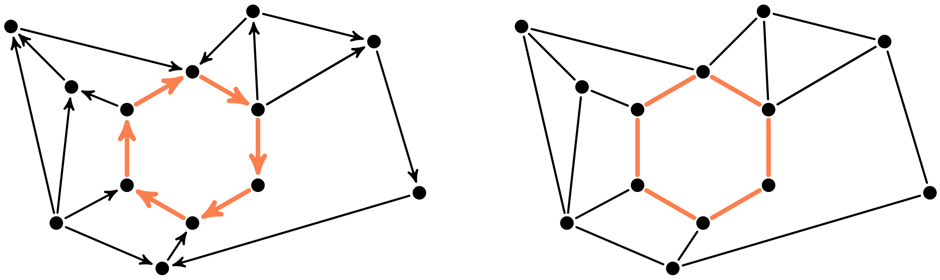
The following proposition relates the local structure, i. e. the structure of subgraphs, to the global structure of SISG models:

Proposition 3.14. *Every subgraph of a SISG model is a subgraph of the SISG model with the same density parameter and the subgraph's nodes as generating set.*

Proof. Let G be a SISG model and $H \subset G$ a subgraph. Denote the distance from a node n to its closest neighbour in the graph G by $d_{\min}(G, n)$. As every node of H is a node of G, the minimal distance $d_{\min}(H, n)$ is larger or equal than $d_{\min}(G, n)$ for every node n in H. Every edge in H is, thus, also an edge in the SISG model generated by the nodes of H. □

Proposition 3.15. *Every connected component of a SISG model is a SISG model with the same density parameter and its nodes as generating set.*

Proof. Let G be a SISG model and $H \subset G$ a connected component. The minimal distance $d_{\min}(H, n)$ of a node n in H is larger or equal than $d_{\min}(G, n)$, as shown in the proof of proposition 3.14. As there exist no edges from H to another connected component H', the closest neighbours of the nodes of H are in again H: if a node n in H would have n' in H' as a closest neighbour, an edge from n to n' would exist proving H and H' to be connected. The minimal distances $d_{\min}(H, n)$ and $d_{\min}(G, n)$ are thus equal for all nodes n in H. This proves the SISG model generated by the nodes of H to be exactly the connected component H. □



(a) Directed hole **(b)** Undirected hole

Figure 3.3
Example of holes of size 6;
(a) directed hole, and
(b) undirected hole

Proposition 3.16. *In a SISG model, there exist no directed holes of size greater than 2.*

Proof. Assume that a hole exists, which consists of a directed cycle along the nodes $p_0, \ldots, p_m$. The distance between consecutive nodes of the cycle decreases, because otherwise there would exist edges in the opposite direction:

$$d(p_{k-1}, p_k) > d(p_k, p_{k+1}) \quad \text{for all } 0 < k < m,$$
$$d(p_{m-1}, p_m) > d(p_m, p_0) \quad \text{and} \quad d(p_m, p_0) > d(p_0, p_1).$$

This yields, by induction,

$$d(p_0, p_1) > d(p_1, p_2) > \ldots > d(p_m, p_0),$$

contradicting $d(p_m, p_0) > d(p_0, p_1)$. □

Corollary 3.17. *In a SISG model, a node with indegree of at least 2 exists in each connected component with at least three nodes.*

Proof. By proposition 3.15, a connected component in the model is a SISG model itself, generated by its nodes. We thus assume that the model is connected, without loss of generality. If each node would have indegree less than 2, each node would have outdegree 1 (each node has, by proposition 3.6(3), outdegree of at least 1) and hence, also indegree 1, because the sum of outdegrees equals the sum of indegrees. As the graph is connected, the model is a hole, contradicting the proposition. □

Proposition 3.18. *In an undirected SISG model of dimension* 1*, there exist no holes of size greater than* 3*.*

Proof. Assume that a hole of size greater than 3 exists. As the nodes of the hole are associated to points on the line, an order relation is naturally defined by their location, i. e. $p_0 < p_1 < \dots < p_m$ with edges between p_k and p_{k+1} for every $0 < k < m - 1$. As a hole is closed, there exists an edge between the minimal and the maximal element, i. e. either an edge (p_0, p_m) or an edge (p_m, p_0) in the directed graph. In the first case, there exist edges (p_0, p_k), in the latter case edges (p_k, p_m) for every $0 < k < m$. In both cases, nodes of the hole in the undirected model would be connected by edges that are not part of the hole, because either p_0 or p_m would be adjacent to more than 2 nodes, contradicting the definition of a hole. □

We were able to analytically discuss subgraphs of SISG models and to prove that connected components of SISG models are again SISG models. Some configurations of edges cannot occur in SISG models. We provided such configurations that cannot occur, and when such a configuration occurs in a graph, we can conclude that the graph is no SISG model.

Conclusion

Spatial structure captures the core concepts of spatial information. It shares many properties with various examples of spatial information, and can thus be used to lift the discussion of spatial information from a semantic to a structural level. We introduced, for this purpose, a graph model of spatial structure. More specifically, the main contributions of this chapter are as follows:

(1) Tobler's law was motivated by the principle of least effort. We discussed the scope of the model and addressed the issue of the meaning of 'near' in the context of the law. This discussion provides the context in which Tobler's law is valid.

(2) Typical properties of spatial information were discussed, including Tobler's law, scale invariance and a bound outdegree of the nodes of a graph representation.

(3) We defined the SISG model as a graph model of spatial structure, and we provided proofs that it has typical properties of spatial information.

(4) The role of subgraphs of SISG models has been discussed. We argued, in particular, in how far subgraphs are SISG models again and showed that some graphs cannot occur as induced subgraphs of SISG models.

One of the main contributions of the thesis, the SISG model, is simple to define and use. It provides, together with the other contributions of this chapter, a mathematical foundation for the discussion of spatial structure, as we will see in subsequent chapters.

4

The Uniform Scale-Invariant Spatial Graph Model

$$\mathrm{H}^{2p}(X, \mathbb{Q})_{\mathrm{alg}} = \mathrm{H}^{2p}(X, \mathbb{Q}) \cap \mathrm{H}^{p,p}(X)$$

—**William Vallance Douglas Hodge**
scottish mathematician
(1903–1975)

The generation of a SISG model presumes a generating set, i. e. a set of points which are located in a vector space (cf. section 3.4), but no presumptions about the distribution of the points are made. An important case is the one of randomly distributed points, with a uniform distribution.

Definition 4.1. A SISG model with a generating set of s randomly distributed points with uniform distribution in the m-dimensional unit ball is called a *uniform scale-invariant spatial graph model* $\mathcal{M}_\rho^m(s)$.

We discuss statistical properties of uniform SISG models in this chapter. Non-statistical methods are, in many cases, not suitable to analyse spatial data and SISG models, because they can, in contrast to the spatial structure, react sensitive to local modifications of the data (section 4.1). The examination is affected by the finiteness and non-connectedness, and hence the concepts of inner and outer regions are introduced (section 4.2). The dependency of a graph's properties on the number of nodes can be analysed by examining induced subgraphs of different sizes. The concept of series of subgraphs addresses this approach (section 4.3). The properties of uniform SISG models can be classified into ones that only depend on the combination ρ^m of the density parameter ρ and the dimension m (section 4.4), and the ones that depend on the density parameter and the dimension separately (section 4.5). Additional properties are reviewed (section 4.6), and the uniform SISG model is classified by its properties (section 4.7).

4.1 Statistical Methods

The characterization of SISG models requires stable properties, i. e. properties that are insensitive to local modifications. Statistical methods can, in many cases, be used to compute stable properties.

We discuss, in this section, the role of statistics for the examination of the uniform SISG model's properties, and we introduce a method that turns global properties into local ones by making them statistical.

The Role of Statistical Properties. The examination of the properties of SISG models aims at characterizing these models and at understanding how these models behave under operations, e. g. which nodes can be reached when moving from one to another node. These properties are influenced by the generating set, but in many applications, only very general assumptions on the generating set are made. The distribution of the points in the generating set, for example, is in many cases known. Properties are, in this context, of special interest, if they are insensitive to small modifications of the generating set and to local modifications of the graph in general.

The properties discussed in section 3.6 are not suitable for the characterization of SISG models, because they are, by and large, sensitive to local modifications. Statistical properties[1] aim at making properties insensitive to local modifications by summarizing their values. The properties discussed in this chapter are of statistical nature due to two reasons: (1) they are based on values that are insensitive to small modifications, e. g. the number of nodes and edges for the definition of density in section 4.4.2, or (2) they are defined for every node, or very small subgraphs, and the resulting values are summarized, e. g. centrality, clustering and diversity coefficients in sections 4.4.6 to 4.4.8. In the latter case, a mechanism for the computation of properties on subgraphs is needed. Such a mechanism is discussed in the following.

Localization of Global Properties. Some properties, called *global* properties, describe a graph as a whole. Examples of global properties are topological properties, such as connectivity. A small modification of the graph can result in a significant change of global properties, as can be seen in the example of connectivity: a graph with two connected components becomes connected if an edge between these components is introduced. Global properties are, in many cases, not suitable to describe graphs whose structure is inhomogeneous.

In contrast to global properties, *local* properties describe only the neighbourhood of single nodes. Centrality, clustering and diversity coefficients (cf. sections 4.4.6 to 4.4.8) are, for example, by construction only considering neighbourhoods of a certain diameter. Small modifications of the graph usually lead to significant changes of some neighbourhoods' properties, but the statistical distribution of the properties for all neighbourhoods does not change significantly.

[1] Statistical properties can become even more insensitive to local modifications, if they are not computed for the directed graph but for the associated undirected one.

When a global property is of interest but insensitivity to small modifications is required, we can examine the distribution of the property on small subgraphs. Balls are such small subgraphs that can easily be constructed and that are suitable to compute global properties for:

Definition 4.2. In a graph G, the *ball*[2] *(or subgraph)* $B_G(p, r)$ *with centre node* p *and radius* $r \in \mathbb{N}$ is the set of nodes q with $\delta(p, q) \leq r$ where δ denotes the undirected distance.

Instead of considering a global property of the whole graph, we can examine the property for each ball of a certain radius, which turns the property into a local one. The property computed for each ball can be examined statistically, e. g. by examining its distribution.

We argued why statistical properties are suitable for characterizing SISG models, and we discussed how global properties can be turned into statistical and local properties. We will discuss in the next section, how these properties are influenced by the finiteness and the non-connectedness of the graph.

4.2 Effect of Finiteness and Non-Connectedness

Properties are, in many cases, analytically easier to compute and to understand for infinite uniform SISG models than for finite ones, because the density of nodes in space is only uniform for infinite models. As data sets are finite, it is however important to examine finite SISG models.

We examine, in this section, how the finiteness and the non-connectedness influence the properties of SISG models. Concepts of inner and outer regions are introduced for this purpose.

Representations of spatial information which are used by computers are necessarily finite. In finite graphs, the density of points in space cannot be uniform. The points are placed in a region U of space, and the density of points decreases near the boundary of this region. We will call these regions $V \subset U$ near the boundary *outer regions*, and all other regions $V' \subset U$ *inner regions*. The inner regions are characterized by the fact that their density does not depend on the size of the whole graph.

The average distance to the next node differs for inner and outer regions of the model, because nodes are not uniformly distributed in finite models. The more nodes the model contains, the relatively less of them are located in the outer regions.[3]

Understanding the outer regions is more complicated than understanding the inner regions, especially for analytical reasoning, because uniform distributions are less complicated to describe than non-uniform ones. (We have, for this reason, proved analytical results only for infinite SISG models in sections 3.5 and 3.6.) This

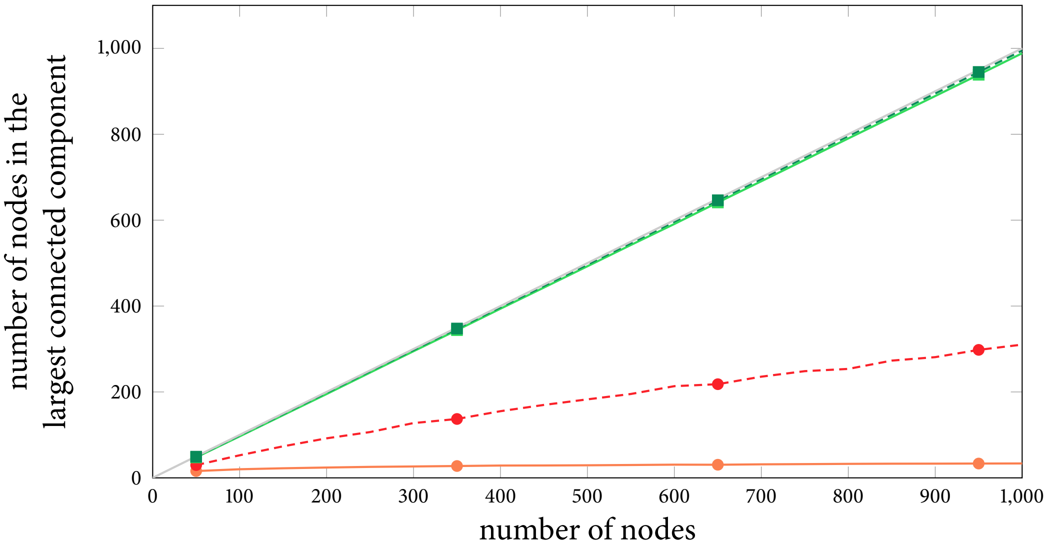
[2] The notation of a ball is common in graph theory, but the author could not trace who introduced it.

[3] This is due to the fact that the ratio of the surface area to the volume is decreasing with increasing diameter.

fact makes it complicated to understand small SISG models. The properties of small models are, in addition, expected to have higher variance than the ones of greater models because the statistical population is greater.

Many real world representations are connected, e. g. many graph representations of public transport; SISG models are however not necessarily connected. For extensive generating sets and low density parameters and dimensions, SISG models are usually not connected: the number of nodes in the largest component is statistically growing much slower than the number of nodes in the graph (cf. figure 4.1), because the number of edges is low for low density parameters and dimensions according to corollary 3.8. It is, due to this behaviour, hard to generate large connected SISG models for low density parameters and dimensions. We consider, in many cases, only the largest connected component of a model, because many properties are only of interest for connected models.

Figure 4.1
Number of nodes in the largest connected component in SISG models; mean value for 1000 models



We argued that finite uniform SISG models are harder to understand, and that it is hard to generate large connected SISG models with low density parameters and dimensions. These problems can, in parts, be overcome by choosing suitable methods to compute properties, as will be argued in the next section.

4.3 Series of Subgraphs

Many properties depend on the number of nodes of the graph. Such dependencies can be examined by computing the property for subgraphs of different sizes. The computations of a property for subgraphs is only very little influenced by the graph's finiteness, because small subgraphs of finite uniform SISG models cannot be distinguished from the ones of infinite models.

We discuss, in this section, how to choose suitable subgraphs for the purpose of examining the described dependency and introduce the notation of a series of subgraphs. The described approach applies, in particular, to SISG models, but also to other graphs.

Inner regions of uniform SISG models have, in the previous section, been argued to be less complicated to understand than outer regions, because the configuration of the nodes and edges is uniform in inner and non-uniform in outer regions. When we examine a property on subgraphs, it seems advantageous to choose subgraphs that are part of the inner region of the graph and to choose subgraphs that contain nodes that are related. The importance of this argument becomes apparent, when an induced subgraph consisting of arbitrarily chosen nodes is considered for a sparse graph: the more nodes the graph has, the less likely it is that the subgraph is connected. It can even happen that the induced subgraph contains no edges at all, even if the original graph is connected.

A node in the inner regions can be found by the 2-dSweep algorithm (cf. section B.3): the algorithm computes a centre node, which is most likely located in the inner region. We can, in a second step, construct a subgraph with a desired number of nodes, placed around the computed centre node, by consecutively adding for each node an adjacent node. Such a subgraph can be expected to have a large number of edges, because only adjacent nodes are added, and in case of a SISG model, nodes with a short distance in space, which are more likely related than arbitrary ones, are added with a high probability.

The method of adding only adjacent nodes works as long as the number of nodes in the subgraph is smaller than the number of nodes in the connected component. The constructed subgraph is, in this case, connected. When the number of nodes in the subgraph equals the connected component, we can proceed by adding an arbitrary node of another connected component. This algorithm defines a series of subgraphs. We formally define:

Definition 4.3. For a graph G, a *series of subgraphs* $G_0 \subset \ldots \subset G_s$ is constructed as follows: the subgraph G_0 consists of a starting node, computed using the 2-dSweep algorithm. Having already defined a subgraph G_k, we consecutively find for each node $p \in G_k$ a node $p' \notin G_k$, if existent, such that an edge (p, p') exists, or if this is not possible, such that an edge (p', p) exists, and add the node p' to the subgraph. This defines a series of subgraphs $G_k \subset G_{k+1} \subset G_{k+2} \subset \ldots$ If for a single node p such a node p' can be found, we just proceed. If no node p' can be found at all, G_k is a connected component. In this case, an arbitrary node $p' \notin G_k$ is added. Applying the algorithm inductively yields a series of subgraphs with $G_s = G$.

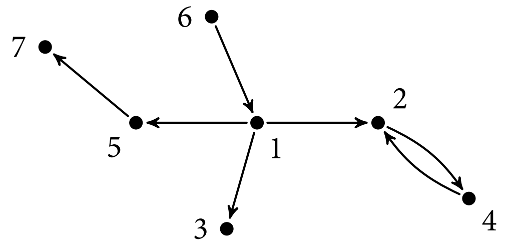


Figure 4.2
Series of subgraphs (G_i); the subgraph G_i contains the nodes $1, \ldots, i$; note that only one possible series is depicted, because choices have to be made

In a series of subgraphs, each step increases the number of nodes by 1, and the connected components are added to the graph G_0 one after another. There exist, in general, many different series of subgraphs, because the definition contains some choices. To compensate for these choices, we can evaluate a property on several series of subgraphs and take the mean value for all subgraphs of the same size. This mean value is more meaningful for statistical considerations than the single values, because the variance is decreased and the properties are more similar to the expectation values.

We introduced the notation of a series of subgraphs, which can be used to examine the dependency of the graph's properties on the number of nodes. Series of subgraphs will be considered for several properties in subsequent sections.

4.4 Properties Depending Only on ρ^{m}

Many properties of uniform SISG models $\mathcal{M}_\rho^m(s)$ only depend on ρ^m, i. e. they statistically coincide for two SISG models $\mathcal{M}_\rho^m(s)$ and $\mathcal{M}_{\rho'}^{m'}(s)$, if $\rho^m = \rho'^{m'}$. We can, in consequence, not distinguish between two SISG models, if the value of ρ^m is equal for both models, and we can only guess ρ^m, but not ρ or m independently, for a given abstract uniform SISG model $\mathcal{M}_\rho^m(s)$.

The correlation between the metric distance of one node to another and the distance in the graph, by the means of the length of the shortest path, is the reason for why many properties only depend on ρ^m: the number of nodes within a distance r is expected to scale by the volume $\gamma_m r^m$ of the m-dimensional ball of radius r, where γ_m is a coefficient depending on the dimension m. The radius of the ball r scales with the density parameter ρ, because the maximal distance, at which an edge between a considered node and another one exists, linearly depends on the density parameter ρ. Whenever a property of the uniform SISG model depends on the number of nodes within a fixed distance, it thus scales by ρ^m.

We discuss, in this section, such properties of the uniform SISG model that only depend on ρ^m. The number of nodes and edges (section 4.4.1), the density and the total density (section 4.4.2) are of special interest, because they characterize uniform SISG models well. They will be used in chapter 5 to test data for spatial structures.

4.4.1 Number of Nodes and Edges

We expect the uniform SISG model $\mathcal{M}_\rho(s)$ to have $s \cdot \rho^m$ edges for $s \to \infty$, according to corollary 3.8. We thus expect:

Proposition 4.4. *For the SISG model $\mathcal{M}_\rho(s)$ with e edges and s nodes, we expect $e/s = \rho^m$ for $s \to \infty$.* □

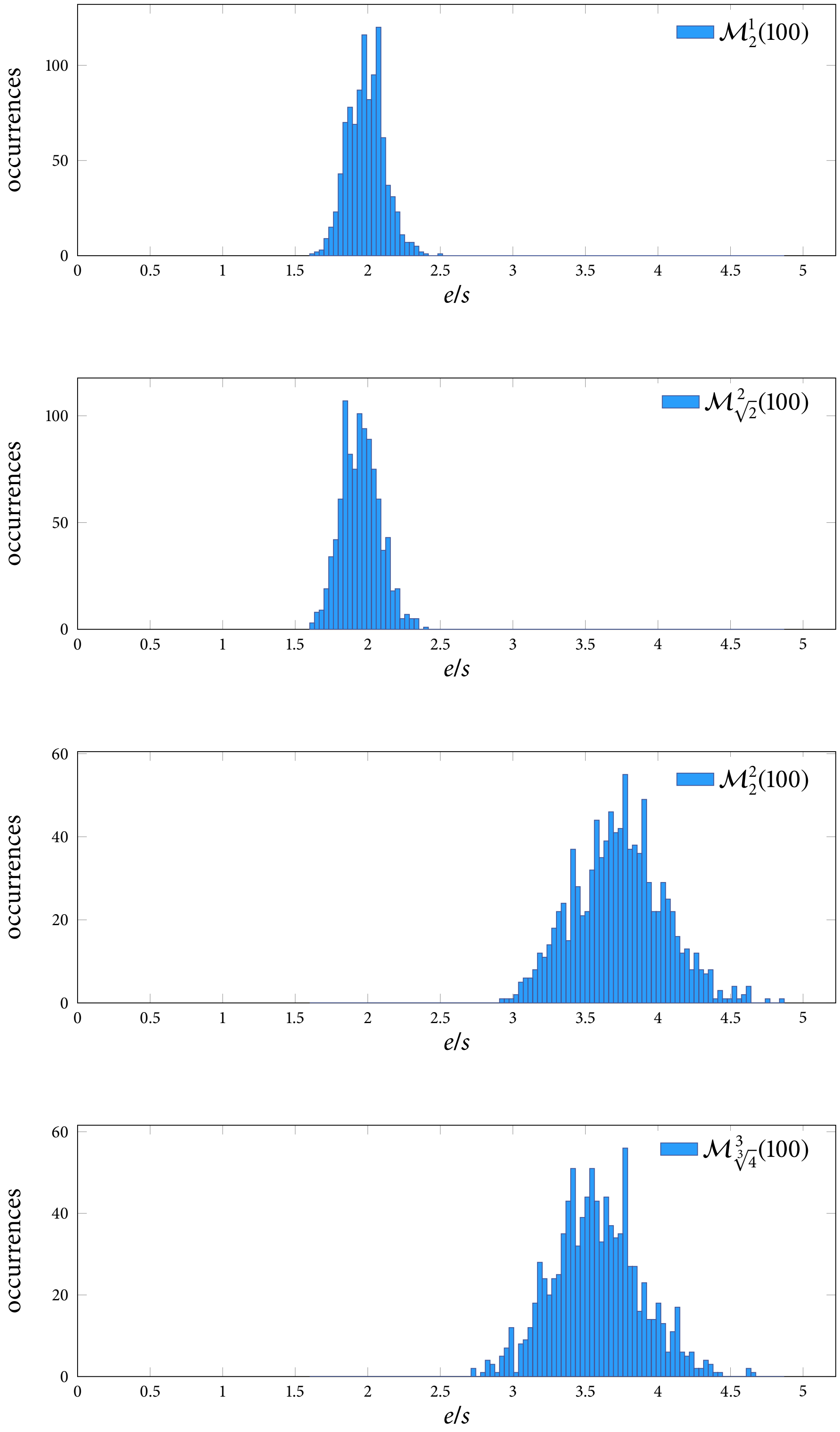


Figure 4.3
Distribution of e/s where s the number of nodes and e the number of edges; aggregated for 1000 graphs

The proposition claims that the ratio between the number of edges and the number of nodes is constant for fixed density parameter and fixed dimension. In particular, the number of edges grows linearly with the number of nodes, unlike it does for many other graph models[4], because edges only connect nodes in the same neighbourhood: when a node is added to the model, only a small neighbourhood is affected. The ratio between the number of edges and the number of nodes has low variance, and the variance is lower for smaller values of ρ^m (cf. figure 4.3).

4.4.2 Density and Total Density

Density characterizes the ratio between edges and nodes in a graph. It is defined such that complete simple undirected graphs have density 1, and graphs without edges have density 0. The following definition of the density has been given by Coleman et al. (1983):

Definition 4.5. The *density* of a graph G consisting of $n > 1$ nodes and e edges is defined as

$$c_{\text{density}}(G) = \frac{e}{n \cdot (n-1)}.$$

By proposition 4.4, the expected density for a SISG model is as follows:

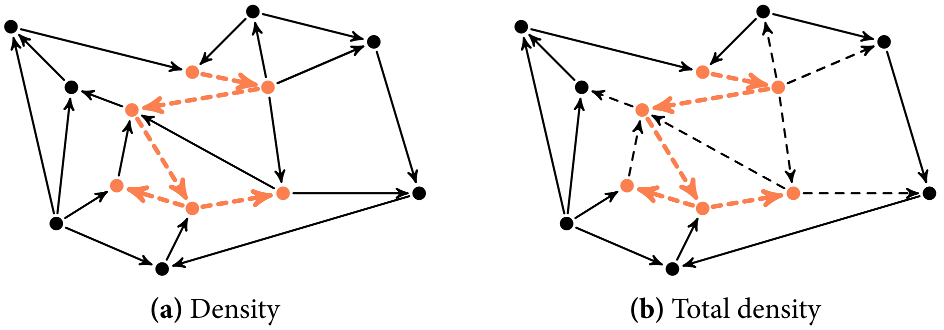
Proposition 4.6. *The uniform SISG model $\mathcal{M}_\rho(s)$ is expected to have density $\rho^m/(s-1)$ for $s \to \infty$.* □

When the density of a finite graph is examined, the effect of finiteness has to be taken into account, as was discussed in section 4.2. This effect occurs, in particular, when the density is evaluated for a series of subgraphs (cf. section 4.3). We can modify the definition of the density, in case of a subgraph, to overcome this problem: we take not only edges inside the subgraph into account, but also these edges that start at a node of the subgraph:

Definition 4.7. The *total density* of a subgraph $H \subset G$ consisting of $n > 1$ nodes is defined as

$$c_{\text{total density}}(H, G) = \frac{e}{n \cdot (n-1)},$$

where G has e edges starting at a node in H.



(a) Density

(b) Total density

Figure 4.4
Computation of the (total) density for the subgraph which is depicted in red; the dashed edges are taken into account for the computation of the density; (a) density, and (b) total density

[4] The number of edges in the complete graph, for example, grows quadratically with the number of nodes.

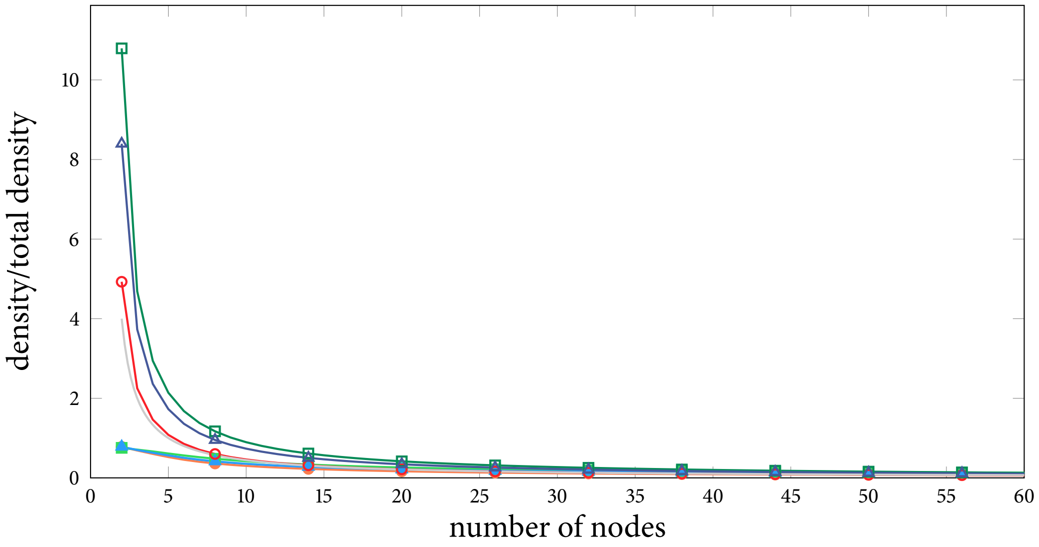


Figure 4.5
Density and total density for a series of subgraphs; mean value for 1000 models with 10 series each

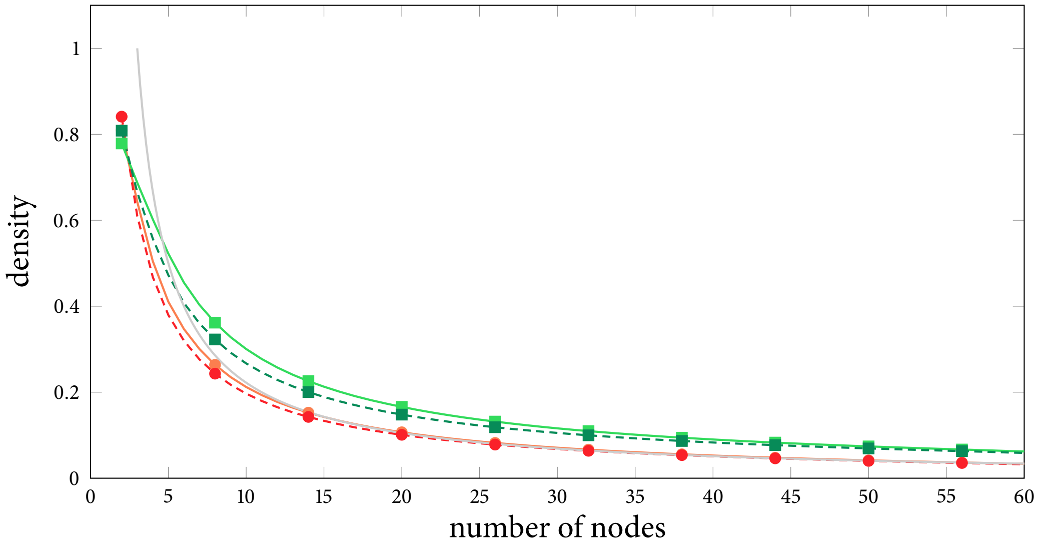


Figure 4.6
Density for a series of subgraphs; mean value for 1000 models with 10 series each

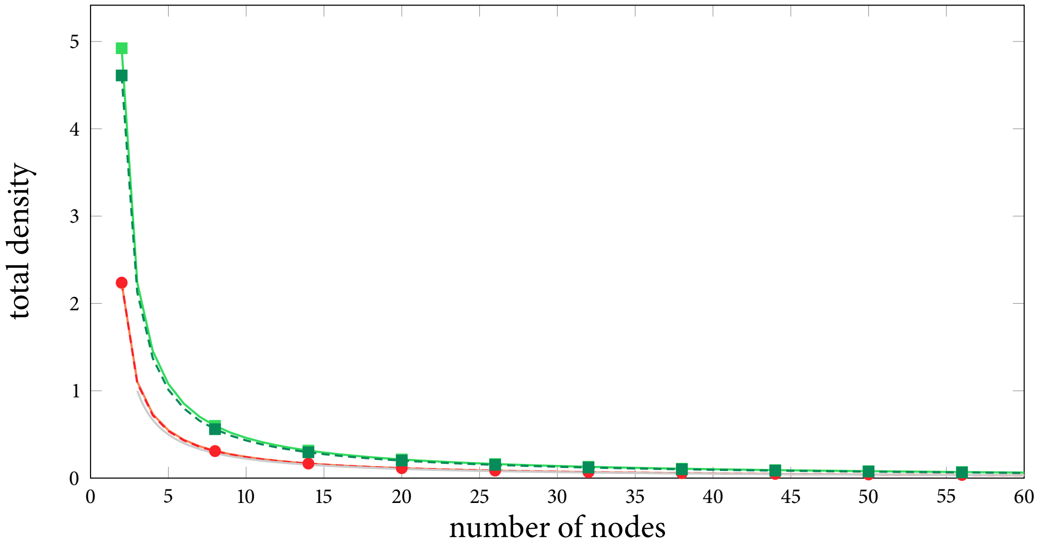


Figure 4.7
Total density for a series of subgraphs; mean value for 1000 models with 10 series each

A comparison of the density and the total density of subgraphs shows why the total density avoids, by and large, the problem of finiteness. Let $N = \bigcup_i N_i$ be a partition of the nodes of a finite graph G, and $G_i \subset G$ the subgraphs induced by the N_i. The definition of the density of a subgraph G_i takes only edges between nodes of N_i into account, but all edges that start at a node $n \in N_i$ and and at another node $n \in N_j$ with $i \neq j$ are left out (cf. figure 4.4). This fact leads to smaller density values than would be expected if the finiteness would be taken into account, because only a subset of all edges are considered.

Every edge of G is, in the computation of the total density of all subgraphs G_i, considered exactly once, namely in the computation for the subgraph G_i that contains the starting node of the edge. Assume each node of G to statistically have the same in- and outdegree. If an N_i is completely contained in the inner region of the graph, the total density of the subgraph induced by N_i is, in contrast to the density, not affected by the effect of finiteness. Proposition 4.6 refers to the estimation value of the density in case that the number of nodes approaches infinity. The density is, in the limit, not affected by the effect of finiteness. The proposition is thus even valid for the total density, and the total density can be expected to even faster converge. The density and the total density thus converge for the number of nodes approaching infinity (cf. figure 4.5).

The choice of the sets N_i has a strong influence on the properties of the induced subgraphs G_i, as was discussed in section 4.3. The density of a subgraph whose nodes are *independent*, i. e. not related, vanishes for example, even if the graph contains numerous edges. This influence can however be expected to be limited for randomly chosen induced subgraphs, because such special cases rarely occur if the nodes' indegrees are approximately equal, and the outdegrees as well.

The densities and total densities of $\mathcal{M}^1_2(s)$ and $\mathcal{M}^2_{\sqrt{2}}(s)$ converge for $s \to \infty$, and they also do for $\mathcal{M}^2_2(s)$ and $\mathcal{M}^3_{\sqrt[3]{4}}(s)$ (cf. figure 4.6 and 4.7). This convergence was expected by proposition 4.6, because the values of ρ^m are equal for these models. The density of SISG models has low variance, and the variance is lower for smaller values of ρ^m (cf. figure 4.8).

4.4.3 Subgraphs of a Minimal Degree

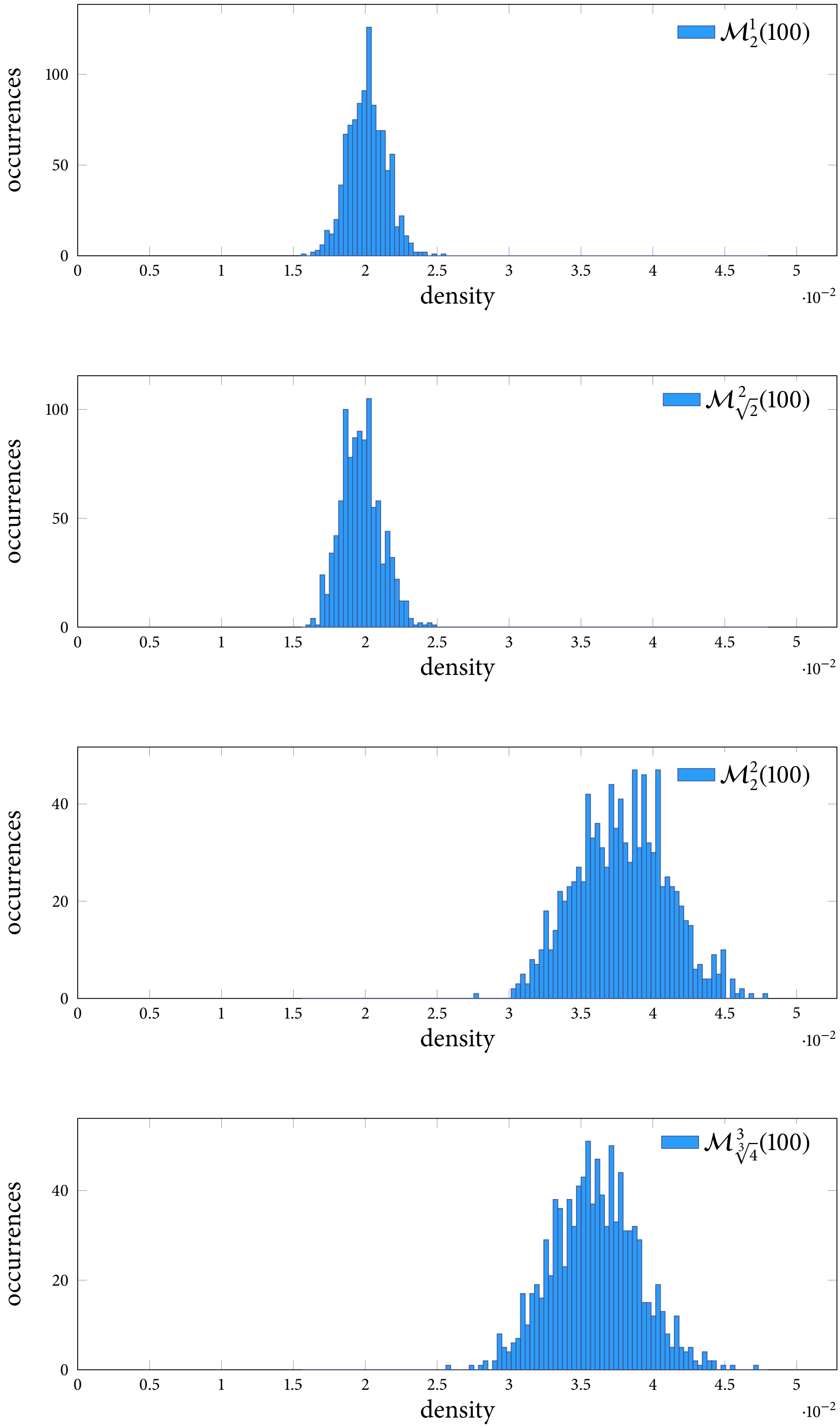
The edge degree is not necessarily equally distributed in a graph: some nodes may be joined by a high number of edges, whereas others only by a low one. Subgraphs of higher or lower edge degree can, in particular, exist. The minimal degree of a graph has been mentioned (as minimum degree) by Albert et al. (2002).

Definition 4.8. A graph is called to have *minimal degree* κ if every node in the graph has at least degree κ and a node of degree κ exists. We denote the *number of maximal subgraphs of minimal degree* κ of a graph G by $c_{\text{subgraphs of min degree }\kappa}(G)$.

The number of subgraphs of minimal degree κ seems[5] to only depend on κ and ρ^m for $s \to \infty$ (cf. figure 4.9).

[5] The dependency has only been evaluated for four different combinations of density parameters and minimal dimensions. We can thus not conclude that the property only depends on ρ^m, but the considerations suggest this dependency.

Figure 4.8
Distribution of the density; aggregated for 1000 graphs



4.4.4 Degree Coefficient

The distribution of the degrees of nodes is characteristic for a graph. It has, for example, been described by Albert et al. (2002).

Definition 4.9. The *degree coefficient* $c_{\text{degree}, \kappa}(G)$ of a graph is the number of nodes that have degree κ. The *maximal degree coefficient* $c_{\text{max degree}}(G)$ of a graph is the maximal degree of a node in the graph.

The degree coefficients seem to only depend on ρ^m for $s \to \infty$ (cf. figure 4.12), and the maximal degree coefficient as well (cf. figure 4.10) The distribution of the nodes' degrees seems to only depend on ρ^m (cf. figure 4.11).

A graph is called *scale-free*, if $c_{\text{degree}, \kappa}(G)$ is, for $\kappa > \kappa_0$, proportional to $\kappa^{-\gamma}$ for some $\gamma > 0$ and some $\kappa_0 > 0$ (Barabási et al. 1999). It can be seen in figure 4.11 that SISG models are approximately scale-free. For many graphs, the exponent γ is in the interval $[2, 3]$ (Hayashi 2006, Albert et al. 2002). The exponent is however larger for some SISG models.

4.4.5 Spectral Graph Properties

The behaviour of random walks on a lattice depends on the dimension of space: in a one- or two-dimensional lattice, the set of random walks which return to its starting point only finitely many times is a null set, i. e. we can expect a random walk to return infinitely many times; in a lattice of dimension greater than three, the set of random walks which return infinitely many times is a null set, and we expect a random walk to only return finitely many times (Polya 1921). Whether random walks are stationary[6], as well as other properties of random walks, is determined by the largest and the second largest eigenvalue of the graphs adjacency matrix (Lovasz 1993, Spielman 2012):

Definition 4.10. We define the *(hypergraph) adjacency matrix A* of a graph to be the matrix with one row and one column for each node, and the number of edges between two nodes n and m as entry A_{nm}. The *simple undirected adjacency matrix* of a graph G is the adjacency matrix of the simple undirected graph associated to the graph G.

The general theory of eigenvalues of the adjacency matrices of graphs (called *spectral graph theory*) reveals, in addition to the stationarity of walks, even more properties (Spielman 2012, Brouwer et al. 2012). We discuss, in this thesis, only the behaviour of the dominant eigenvalue, because it is related to random walks and because it can be computed efficiently.

Definition 4.11. The *dominant eigenvalue* $c_{\text{eigen}}(G)$ of a graph G is the dominant eigenvalue of its simple undirected adjacency matrix.

The power iteration algorithm that computes the dominant eigenvalue is discussed in section B.6. The dominant eigenvalues for the simple as well as for the hy-

[6] We call a random walk *stationary*, if the probability distribution P_i of the walk to be at a certain node in the ith step equals the probability distribution P_{i+1}.

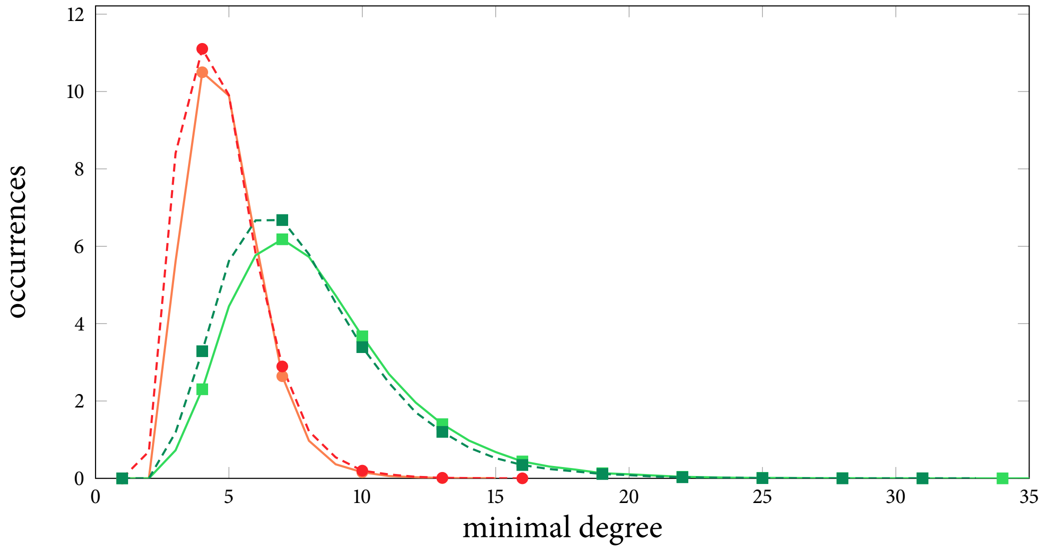


Figure 4.9
Number of subgraphs of a minimal degree; mean value for 1000 models

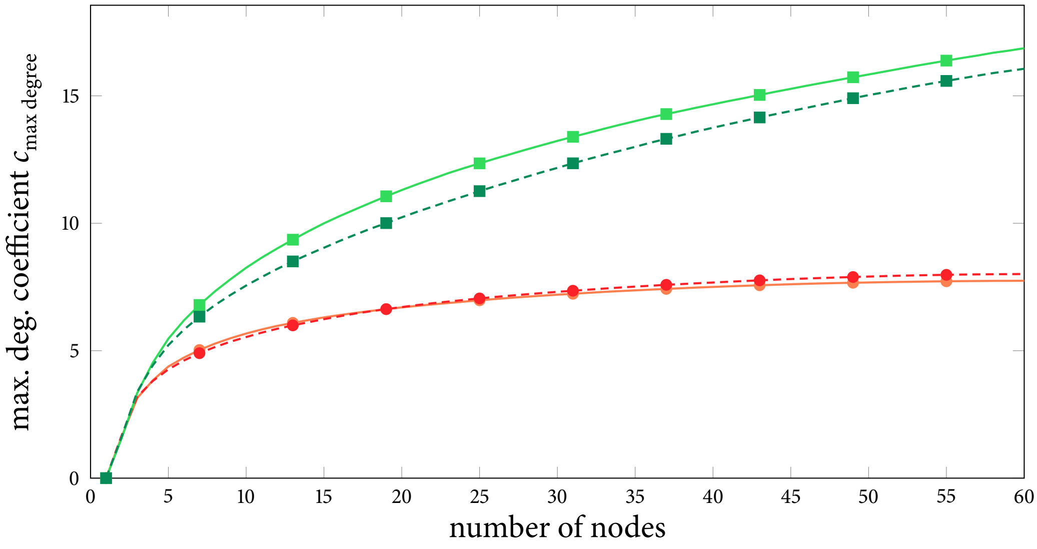


Figure 4.10
Maximal degree coefficient $c_{\text{max degree}}$ for a series of subgraphs; mean value for 1000 models with 10 series each

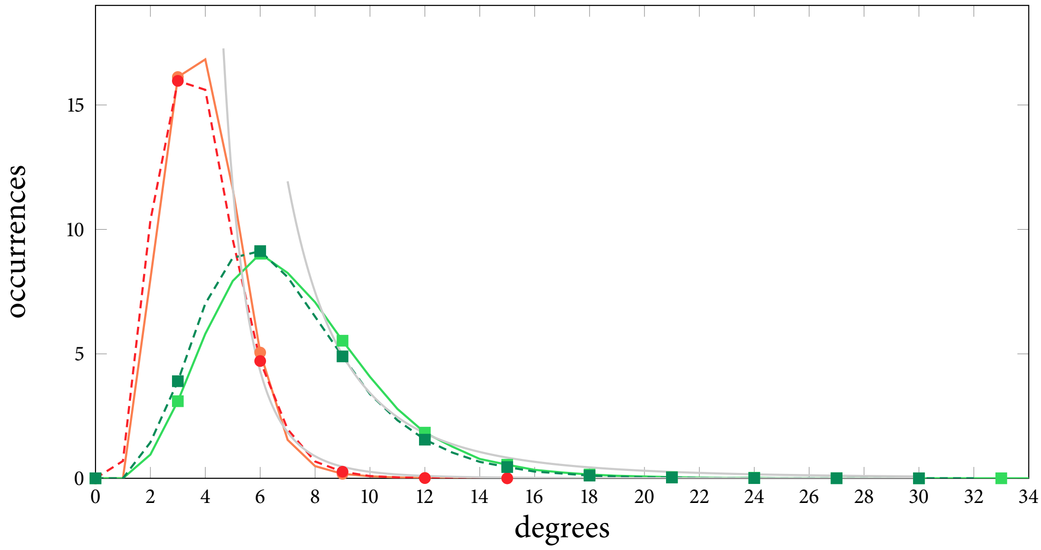
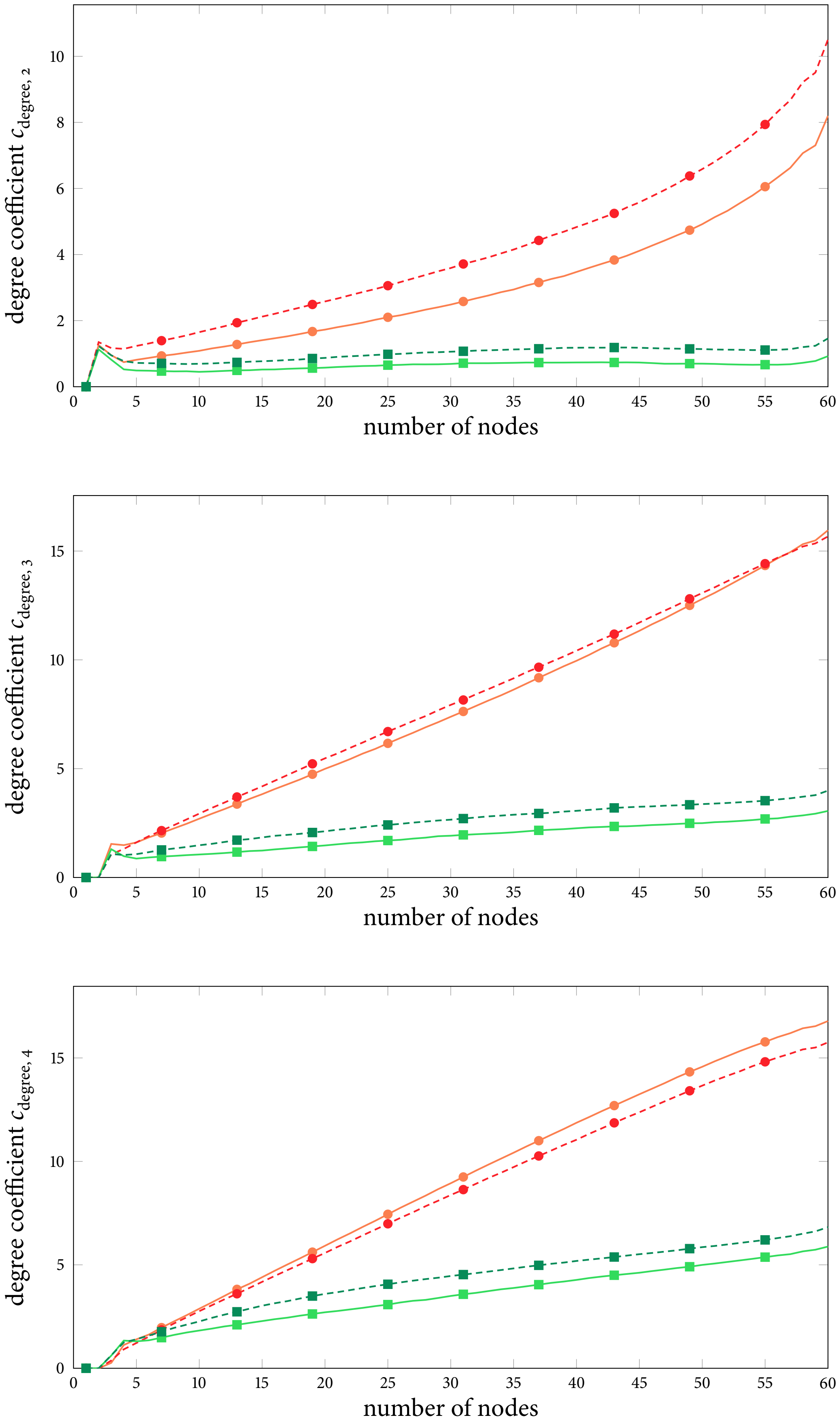


Figure 4.11
Distribution of the degree of nodes; mean value for 1000 models

Figure 4.12
Degree coefficients for a series of subgraphs; mean value for 1000 models with 10 series each



pergraph adjacency matrix seem to only depend on ρ^m (cf. figure 4.13 and 4.14). Barthélemy (2011) makes a short comment about spectral theory of spatial graphs and Albert et al. (2002), for complex networks.

4.4.6 Centrality

Nodes can be very central for a graph when many shortest paths contain them. There exist differing notations of centrality for graphs. A common one has been introduced by Freeman (1979, 1977):

Definition 4.12. The *centrality coefficient* of a graph $G = (N, E)$ is defined as

$$c_{\text{centrality}}(G) = \sum_p \frac{\max_q \deg q - \deg p}{(|N| - 2) \cdot (|N| - 1)}.$$

The centrality coefficient seems to only depend on ρ^m (cf. figure 4.15). The distribution of the centrality coefficient of subgraphs seems, at large, to only depend on ρ^m, and the coefficient varies notably for different subgraphs (cf. figure 4.16).

4.4.7 Clustering

In a graph, the existence of regions with a large number of edges is called clustering. There exist differing definitions of clustering, e. g. the one given by Watts et al. (1998) and Barrat et al. (2000). We use the following definition, which has been introduced by Newman (2003^b):

Definition 4.13. The *clustering coefficient* of a graph $G = (N, E)$ is defined as

$$c_{\text{clustering}}(G) = 3 \cdot \frac{n_{\text{triangles}}}{n_{\text{tuples}}},$$

where $n_{\text{triangles}}$ denotes the number of directed cycles of length 3 and n_{tuples} denotes the number of open walks of length 2.

Closed walks are, according to the definition, not counted for n_{tuples} to ensure that a complete directed graph has maximal clustering coefficient, i. e. a clustering coefficient of 1.

The clustering coefficient does not solely depend on ρ^m (cf. figure 4.17). It has high variance, as can be seen in figure 4.18, and can hence not be used to reliably reconstruct the parameters.

We would expect the variance to be low because clustering is, in general, a robust[7] measure: local modifications do not change the number of clusters and does not considerably change statistical properties. The fact that the variance is, against our expectations, high shows that SISG models with the same number of points in

[7] A measure is called *robust* if it is insensitive to local modifications of the graph.

Figure 4.13
Dominant eigenvalue for the simple adjacency matrix for a series of subgraphs; mean value for 1000 models with 10 series each

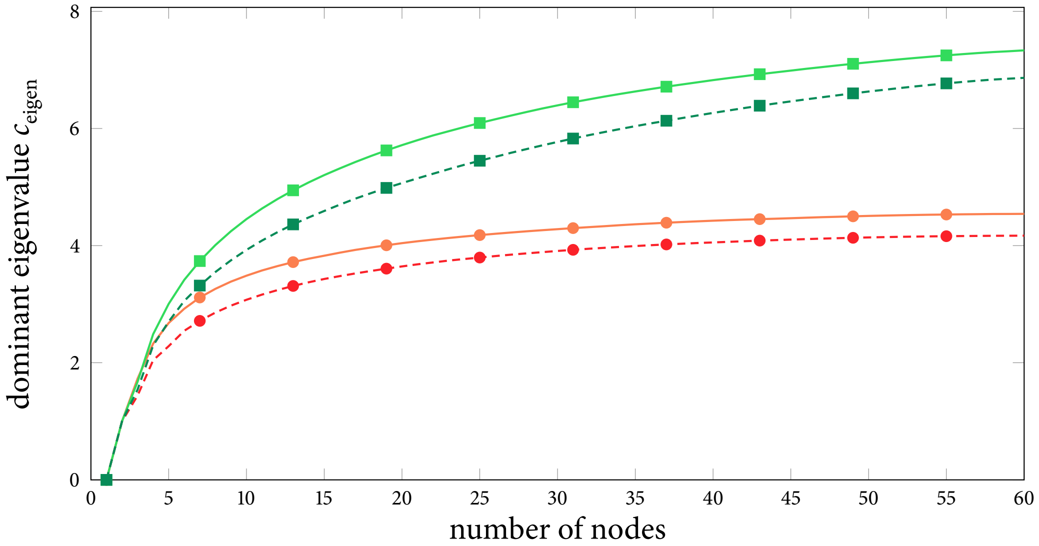


Figure 4.14
Dominant eigenvalue for the hypergraph adjacency matrix for a series of subgraphs; mean value for 1000 models with 10 series each

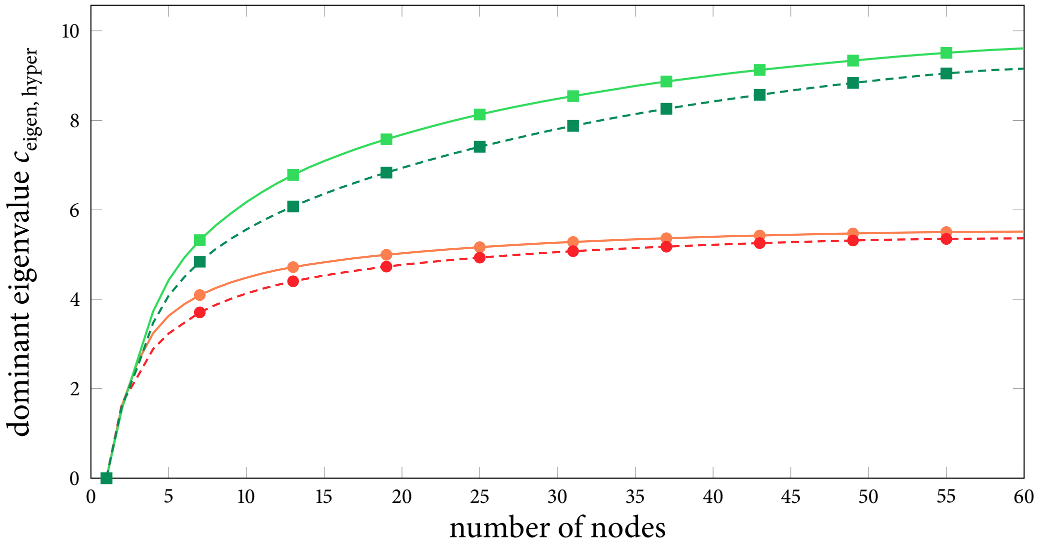
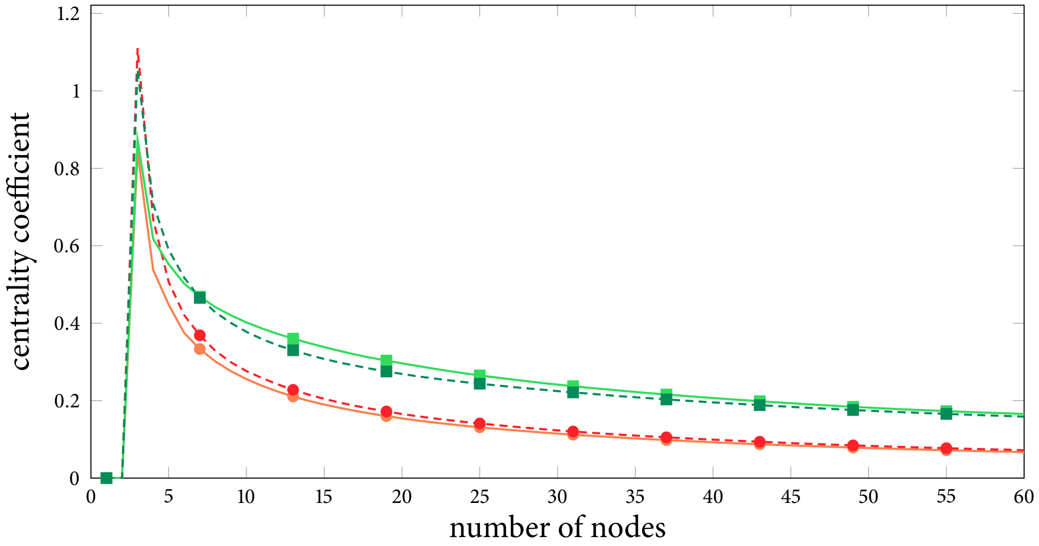


Figure 4.15
Centrality coefficient for a series of subgraphs; mean value for 1000 models with 10 series each



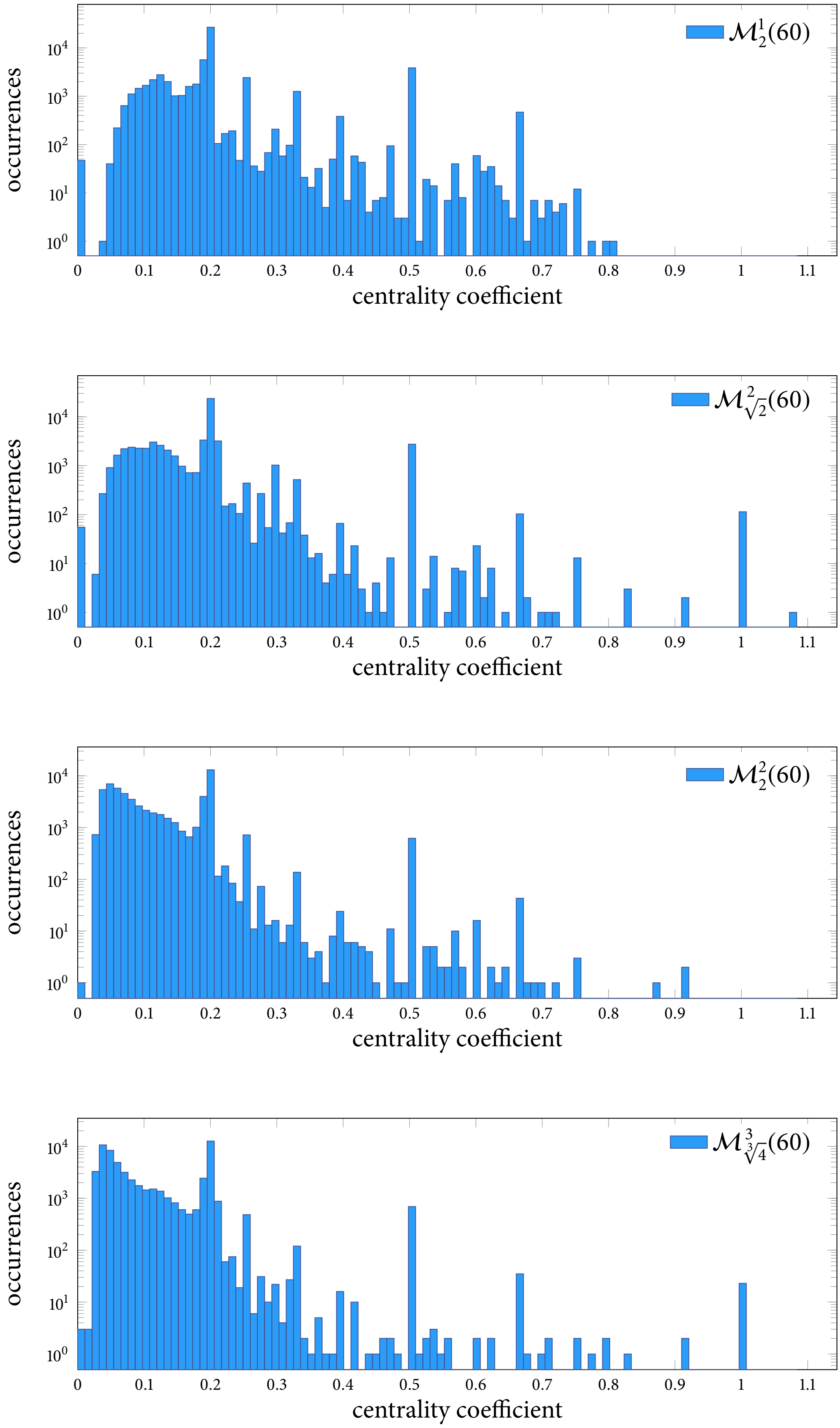


Figure 4.16
Distribution of the centrality coefficient for balls of radius 10; aggregated for 1000 graphs

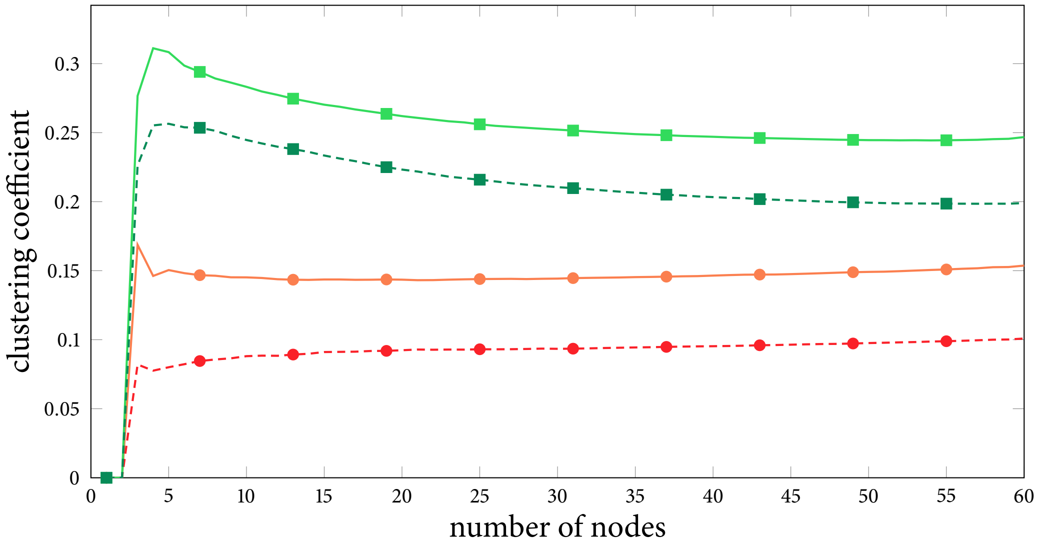


Figure 4.17
Clustering coefficient for a series of subgraphs; mean value for 1000 models with 10 series each

the generating set, the same minimal dimension and the same density coefficient can differ considerably in this aspect. The clustering coefficient does thus not characterize SISG models well.

The distribution of the clustering coefficient of subgraphs seems, at large, to only depend on ρ^m, and the coefficient varies notably for different subgraphs (cf. figure 4.18).

4.4.8 Diversity

Diversity measures how independent the edges of a graph are, by determining how many edges share no common end points. The diversity coefficient vanishes when all edges of a simple graph share a common end point and equals 1 if no edge does. We use the following definition that has been given by Macindoe et al. (2010) and Richards et al. (2009):

Definition 4.14. The *diversity coefficient* of a graph $G = (N, E)$ is defined as

$$c_{\text{diversity}}(G) = 8 \cdot \frac{\sqrt{n_{\text{independent dipoles}}}}{|N| \cdot (|N| - 2)},$$

where $n_{\text{independent dipoles}}$ denotes the number of pairs (e, e') of edges where no edge (p, q) exists with e joining p and e' joining q.

The diversity coefficient seems to only depend on ρ^m (cf. figure 4.19). The distribution of the diversity coefficient of subgraphs seems, at large, to only depend on ρ^m, and the coefficient varies notably for different subgraphs (cf. figure 4.20).

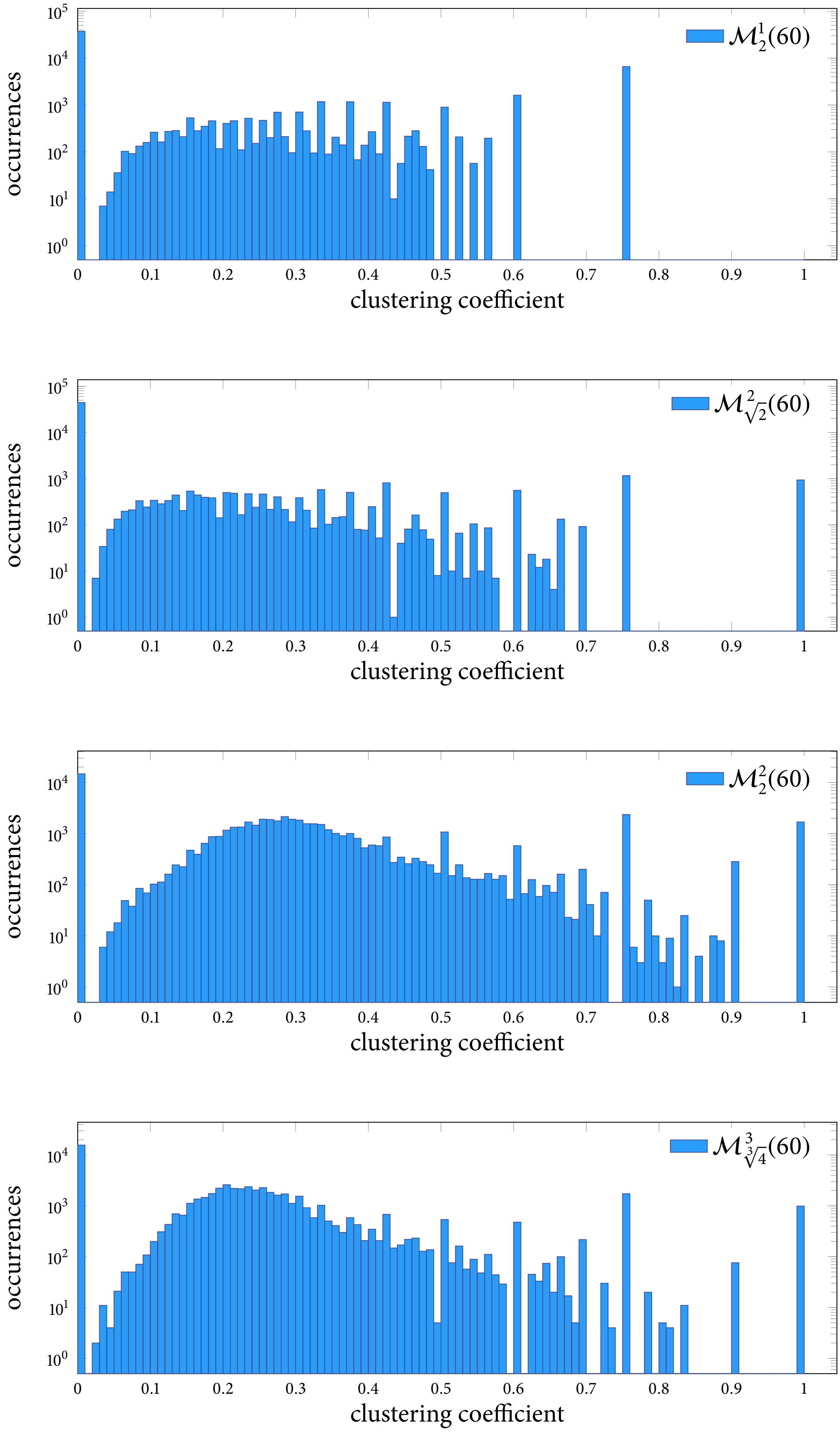


Figure 4.18 Distribution of the clustering coefficient for balls of radius 10; aggregated for 1000 graphs

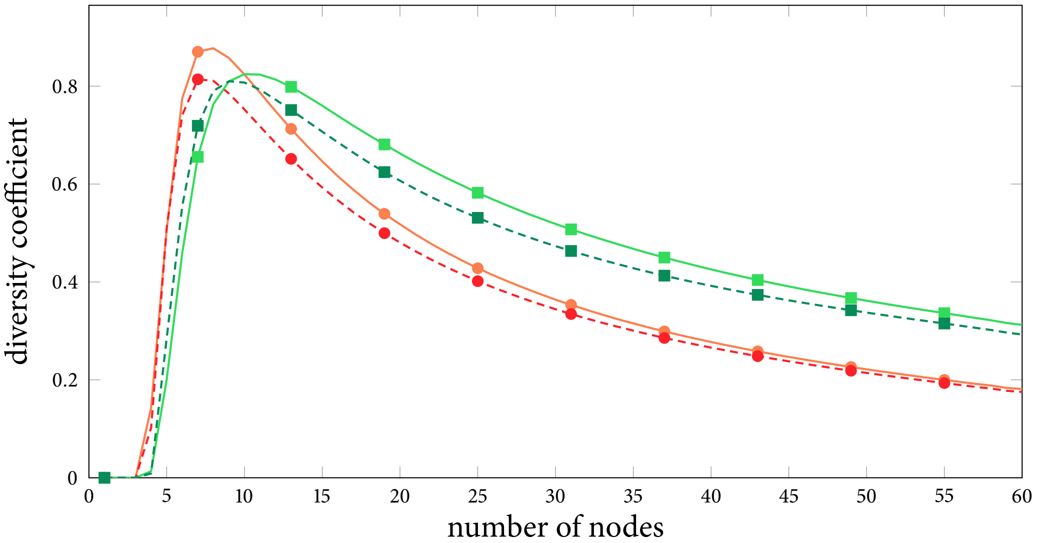


Figure 4.19
Diversity coefficient for a series of subgraphs; mean value for 1000 models with 10 series each

4.5 Properties Depending on ρ and m

We have motivated, in the last section, why properties depend, in many cases, only on ρ^m, where ρ is the density parameter and m the dimension of the SISG model. Properties of the uniform SISG model depend, in general, independently on the density parameter ρ and the dimension m: properties differ, in general, for two parameter sets (ρ, m) and (ρ', m'), even if ρ^m and $\rho'^{m'}$ are equal.

We discuss, in this section, various properties that depend on ρ and m independently. The volume of the sphere (section 4.5.1) is of special interest because it characterizes uniform SISG models well. It will be used in chapter 5 to test data for spatial structures.

4.5.1 Volume of the Sphere

The volume of the ball and the volume of the sphere depend on the dimension of space. Similar to the definitions in metric spaces, we define for graphs:

Definition 4.15. In a graph G, the *sphere* $S_G(p, r)$ *with centre node* p *and radius* $r \in \mathbb{N}$ is the set of nodes q with $\delta(p, q) = r$, where δ denotes the undirected distance. The *volume of the sphere* denotes the number of nodes of the sphere.

By the very definition, a ball (cf. definition 4.2) is the union of spheres contained in the ball:

Proposition 4.16. *For a graph G and an integer $r \geq 0$, the following holds:*

$$B_G(p, r) = \bigcup_{i=0}^{r} S_G(p, i).$$

It is hence sufficient to only consider the volume of the sphere in the discussion which aims at the reconstruction of the parameters.

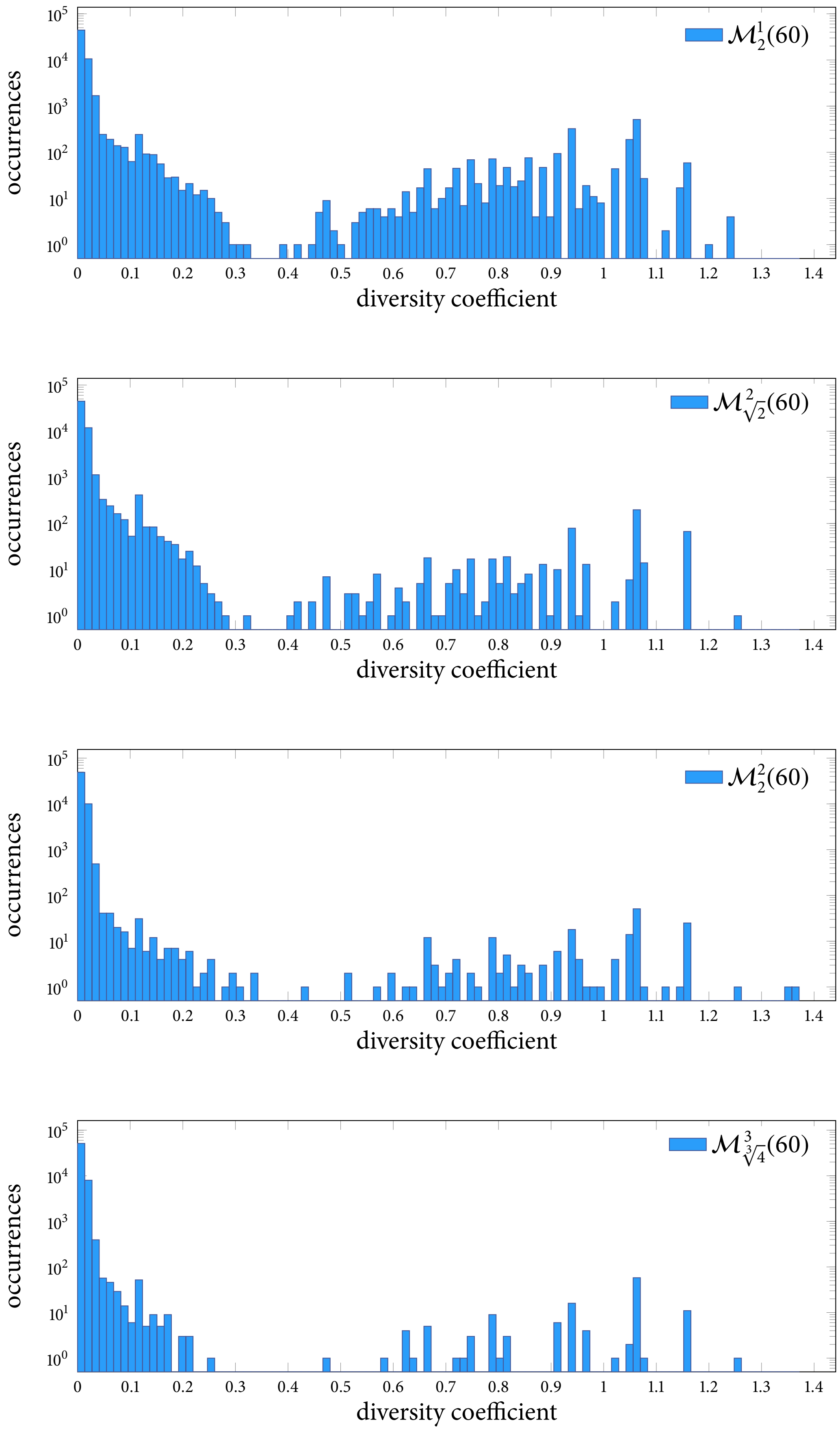
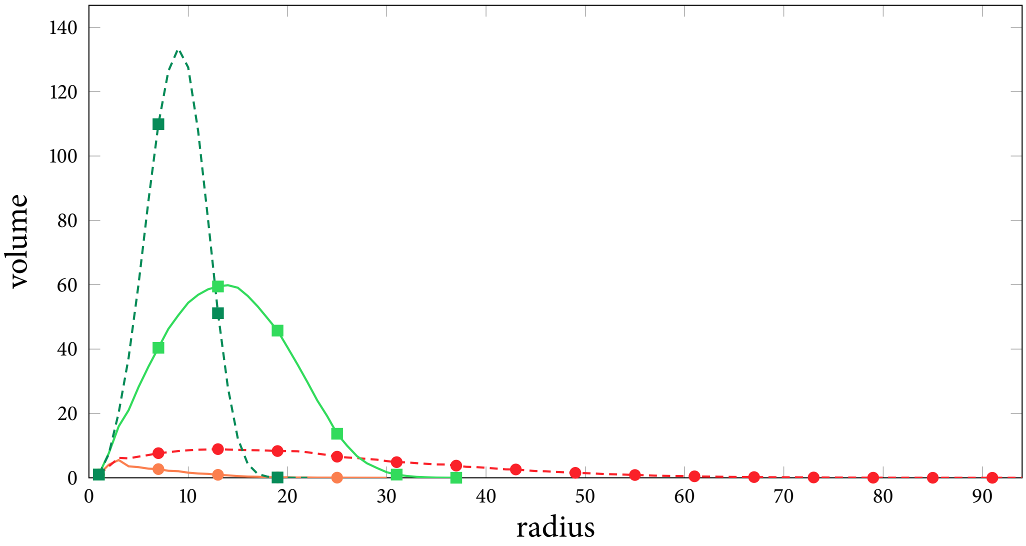


Figure 4.20
Distribution of the diversity coefficient for balls of radius 10; aggregated for 1000 graphs

The volume of the sphere does not solely depend on ρ^m (cf. figure 4.21). It is a robust measure, because the average shortest path length is a robust measure and the modification of a single edge does not considerably change the number of nodes located at a certain distance from the centre node.

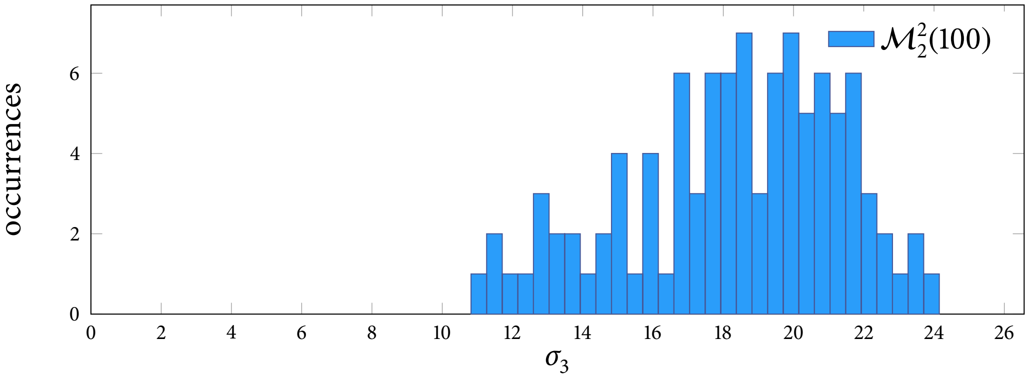
Figure 4.21
Volume of the sphere for several radii for the largest component of the associated undirected graph; mean value for 100 models and 10 centre nodes



During the reconstruction of the parameters, it can be advantageous to represent properties by a single number instead of a distribution. We define:

Definition 4.17. For a graph G, we define $\sigma^3(G,p)$ as the arithmetic mean of the three maximal volumes of spheres with centre node p, and $\sigma^3(G)$ as the arithmetic mean of $\sigma^3(G,p)$ for all nodes p in the graph.

Figure 4.22
Distribution of the mean value of σ_3 for 100 centre nodes; aggregated for 100 graphs



The arithmetic mean was chosen in the definition of σ_3, because it is slightly more stable than the maximal volume alone. The variance of σ_3 is high, but the values of σ_3 for different parameters differ even more (cf. figure 4.22). The property σ_3 is hence a suitable candidate for the reconstruction of the parameters.

4.5.2 Diameter

The diameter of a graph specifies the maximal length of the shortest paths in a graph. It has, for example, been defined by Diestel (2005, p. 8):

Definition 4.18. The *diameter* of a graph G is defined as

$$c_{\text{diameter}}(G) = \max_{p,\,q} \delta(p, q),$$

where δ denotes the undirected distance in G.

Algorithms to compute the diameter are discussed in section B.4, e. g. the Floyd-Warshall algorithm, the iFub and DiFub algorithms as well as the 2-Sweep and 2-dSweep algorithms.

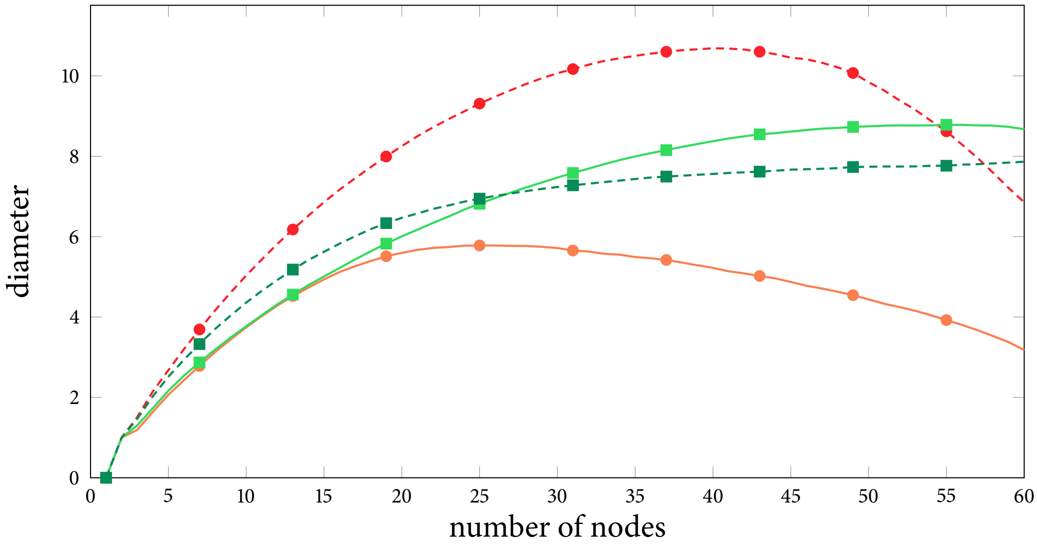


Figure 4.23
Diameter for a series of subgraphs; mean value for 1000 models with 10 series each and 10 values for each diameter computation using the 2-Sweep algorithm

The diameter does not solely depend on ρ^m (cf. figure 4.23), but the behaviour of the diameter for $s \to \infty$ is not in an obvious way related to the density parameter nor to the minimal dimension. The diameter is increasing for larger subgraphs but becomes again smaller when the subgraph approaches the whole graph. It can be shown that the decrease of the diameter, when the subgraph approaches the whole graph, is an effect of outer regions due to the finiteness of the models. This effect renders the data useless for the reconstruction of the parameters. In addition, it cannot be used to reliably reconstruct the parameters, because the diameter of a graph can fundamentally change under edge operations.

4.5.3 Average Shortest Path Length

The average shortest path length is a measure for how many edges are, in average, needed to relate two nodes in a graph. It has, for example, been defined by Albert et al. (2002).

Definition 4.19. The *average shortest path length* of a connected graph $G = (N, E)$ is defined as

$$c_{\text{average shortest path length}}(G) = \frac{1}{|\Delta|} \cdot \sum_{(p,\,q) \in \Delta} \delta(p, q),$$

where δ denotes the undirected distance in G and $\Delta = \{(p, q) \mid 0 < \delta(p, q) < \infty\}$.

The average shortest path length can, as the average of all shortest path lengths, e. g. be computed by the use of partial shortest path trees, which are discussed in section B.2.

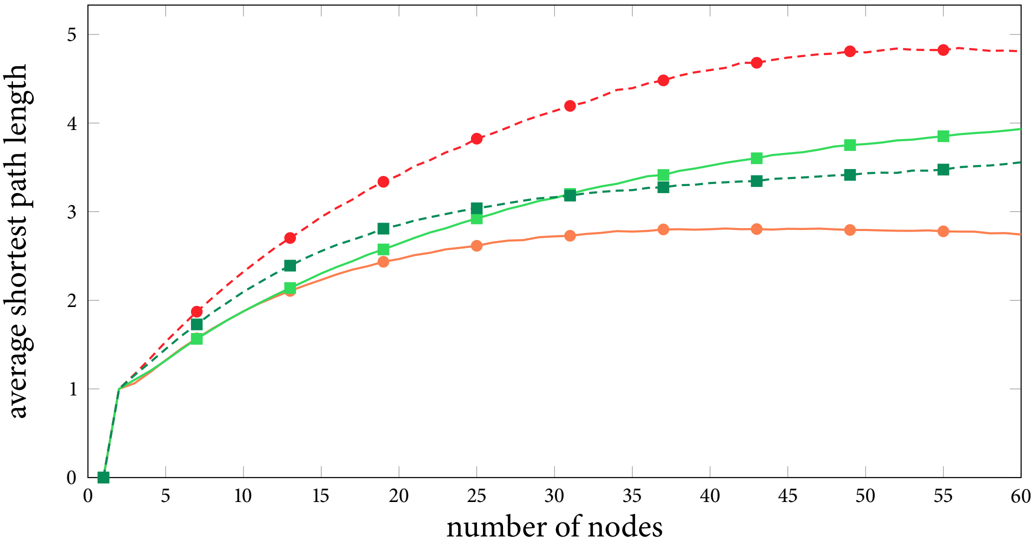


Figure 4.24
Average shortest path length for a series of subgraphs; mean value for 1000 models with 10 series each and computation of the shortest path for 10 randomly chosen pairs

The average shortest path length behaves, for different sizes of models, similarly to the diameter. The average shortest path length is more suitable than the diameter for the purpose of recovering the density parameter and the minimal dimension, because it is more stable than the diameter. There are however two reasons for why the average shortest path length cannot be used: (1) the behaviour of models with the same ρ^m can be similar, as in the example of $\mathcal{M}_2^2(60)$ and $\mathcal{M}_{\sqrt[3]{4}}^3(60)$ in figure 4.24; and (2), the average shortest path length has a high variance, as can be seen in figure 4.25. This fact shows that SISG models with the same number of points in the generating set, the same minimal dimension and the same density coefficient can differ considerably in this aspect. Average shortest path lengths are hence not characteristic for the combination of the density parameter and the dimension that is used to generate SISG model.

4.5.4 Cliques

Subgraphs where all nodes are directly related by edges are called cliques, and their number and sizes are characterizing the graph. Cliques have, for example, been defined by Luce et al. (1949):

Definition 4.20. A *clique* of an undirected graph $G = (N, E)$ is a set of nodes $N' \subset N$ such that the induced subgraph is complete. A *maximal clique* is a clique N' such that no clique $N'' \supsetneq N'$ exists. The *number of maximal cliques with κ nodes* $c_{\text{max cliques}, \kappa}(G)$ of a graph G is defined as the number of maximal cliques with κ nodes in the associated undirected graph.

The Bron-Kerbosch algorithm, which computes maximal cliques, is discussed in section B.5. The number of nodes in maximal cliques is low, because the density of

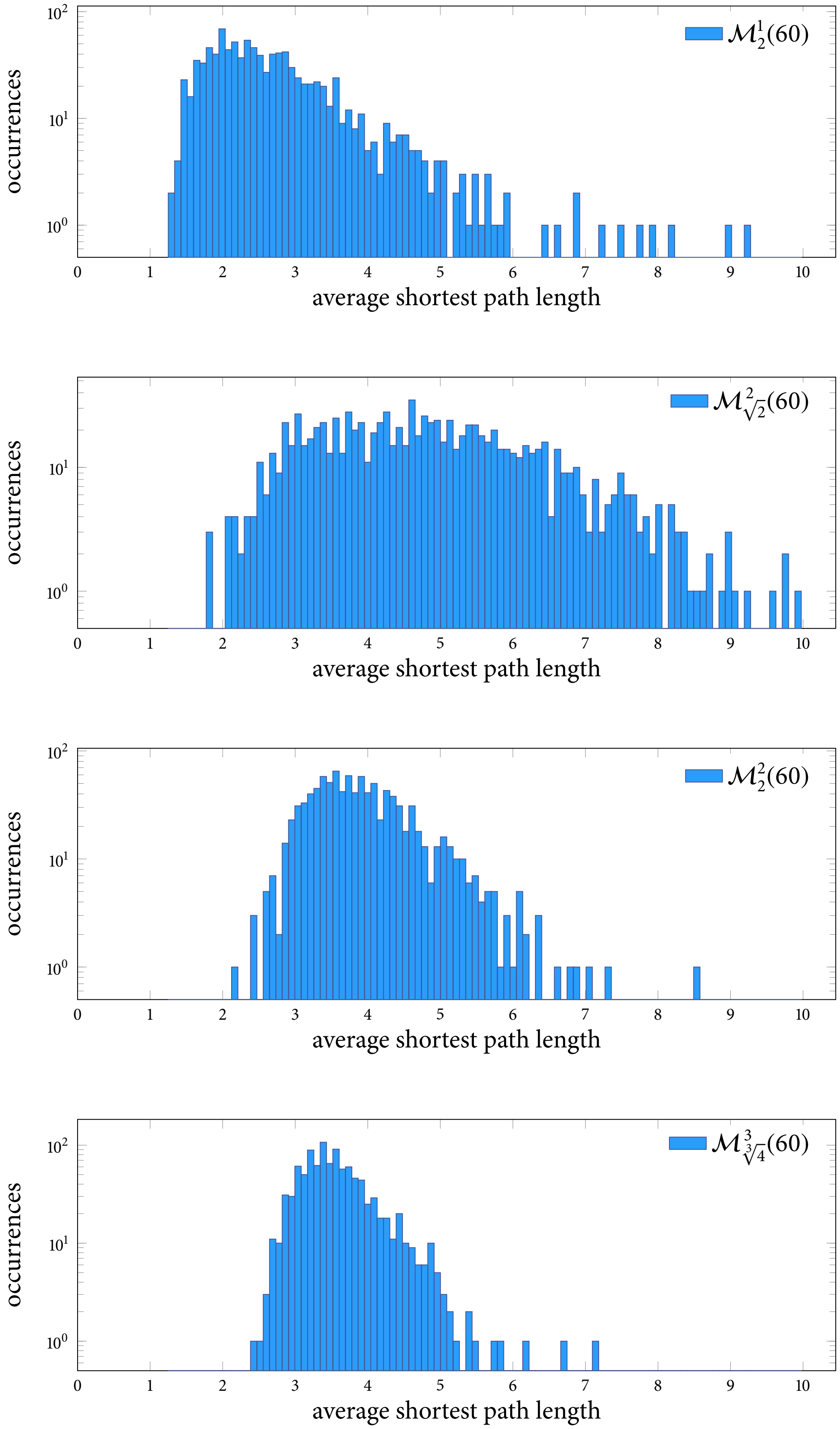


Figure 4.25
Distribution of the average shortest path length in the largest component of the associated undirected graph; aggregated for 1000 graphs

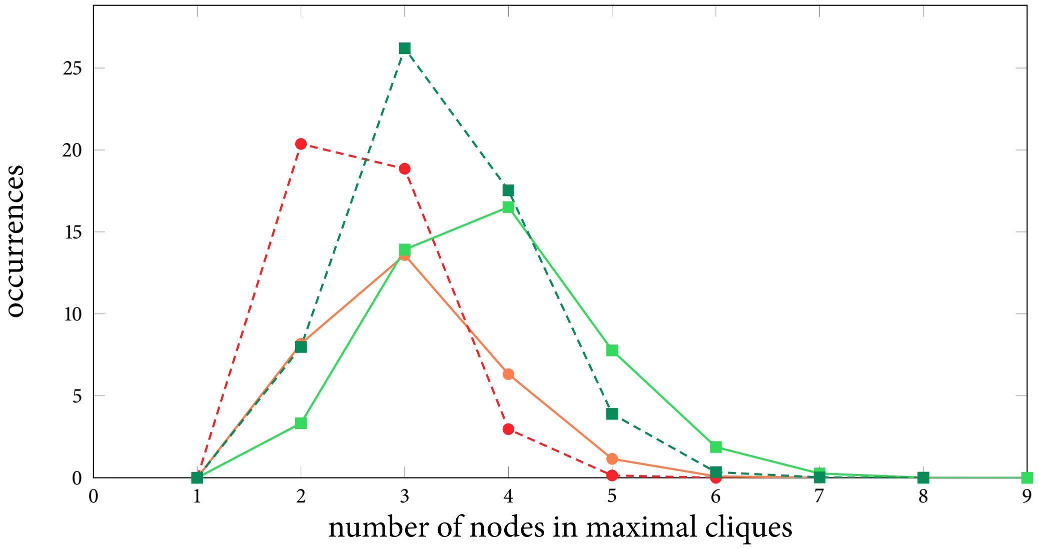


Figure 4.26
Number of maximal cliques; mean value for 1000 series

uniform SISG models is low, and it does not solely depend on ρ^m (cf. figure 4.26). The number of nodes in maximal cliques for $s \to \infty$ is however not in an obvious way related to the density parameter nor to the minimal dimension. The distribution of maximal cliques can only be fitted with high uncertainty, because most cliques contain only very few nodes. It hence cannot be used to reliably reconstruct the parameters.

4.6 Other Properties

A number of graph properties was discussed in the last two sections, but others were not. The selection of the discussed properties was made for two reasons: (1) the chosen property is either characteristic of uniform SISG models, or (2) the property is correlated to the density parameter and the dimension in a simple and obvious way. Properties of the SISG model that usually differ from the one for graph representations of human activities, e. g. the number of connected components, were not discussed. An overview of numerous graph properties has been given by Wasserman et al. (1994), Barthélemy (2011), Albert et al. (2002) and Newman (2003^b). We provide, in this section, an overview on some of the graph properties that were not discussed.

The tree-width of a graph was introduced by Robertson et al. (1986) as the size of the largest vertex set minus one in an optimal tree-decomposition of the graph. While the tree-width is commonly used to characterize graphs in respect to their tree-decompositions, the computation of the tree-width is, in general, an NP-hard problem (Arnborg et al. 1987). Fast algorithms for similar problems exist for specific classes of graphs, e. g. an algorithm by Seymour et al. (1994) to compute the branch-width of planar graphs in $O(n^2)$ time, and an algorithm by Kammer et al. (2015) to compute a tree-decomposition of width $O(k)$ for a planar graph of tree-width k with n nodes in $O(nk^2 \log k)$ time. Though SISG models are not planar, they share many properties with planar graphs. We thus expect the treewidth to be unbound and its computation with existing algorithms to be very time-consuming.

Graphs can be compared by transforming them into each other using node or edge replacement operations according to a grammar (Rozenberg 1997). The induced isomorphism classes characterize the graph. Algebraic and category theoretic approaches, like pull backs and push outs, can, in a similar way, be used to relate graphs (Rozenberg 1997, Kahl 2002). Alternatively, graphs (called motifs) that regularly appear as subgraphs characterize the graph (Barthélemy 2011). These properties can react sensitive to the local structure of graph representations and are time-consuming to compute.

The numbers of colours needed to colour the nodes of a graph such that any edge only joins nodes of different colours is called the chromatic number (Brooks 1941). It can react very sensitive to local modifications of the graph.

A node cover is a set of nodes such that each edge joins at least one node of the cover. The set of minimal node covers is indirectly discussed, because it is a complement to a maximal independent set[8] which itself is a clique in the graph's complement (Tarjan et al. 1977).

Graphs can be characterized by communities, which are subgraphs whose nodes are linked by a large number of edges but have only few edges to nodes outside the subgraph (Fortunato 2010). Similarly, we can characterize graphs by identifying large subgraphs with high densities (Gibson et al. 2005, Lee et al. 2010). In addition, a graph is characterized by its degree correlation, i. e. by the number of edges between nodes of high and low degrees (Maslov et al. 2004, Pastor-Satorras et al. 2001, Vázquez et al. 2002, Newman 2003[a], 2002). All these properties are, however, loosely related to the idea of the clustering coefficient and thus not discussed in this thesis.

Resilience is the property of a graph to become less and less connected when edges are removed. Graphs can behave different when edges are removed, and the diameter can be more or less stable under these modifications (Albert et al. 2000). Resilience is not explicitly discussed in this thesis, because it is related to other properties such as the diameter, the average shortest path length, the degree coefficient and many more.

In geographical literature, numerous indices for graphs have been defined (Haggett et al. 1969, Kansky et al. 1989, Rodrigue et al. 2013, Xie et al. 2007). These indices are not discussed in this thesis, because they usually are combinations of already known properties or related to them.

We argued that only few properties of graphs are suitable for characterizing SISG models or graph representations, and we discussed why only few are suitable. The examination of the properties in sections 4.4 and 4.5 as well as the overview of the non-examined properties make no claim to be complete.

[8] A set of nodes is called independent if they are pairwise non-adjacent.

4.7 Classification

Classifications of existing graph models are widely used. The most prominent classes were discussed in section 3.3.3: the class of complex, of scale-free and of small-world networks. We classify uniform SISG models in this section, according to these prominent classes.

As the topological properties of SISG models and uniform SISG models are non-trivial, they are *complex networks*: they have a distinct local structure, i. e. each node is only adjacent to a small number of other nodes, and the expectation value for this number is independent of the size of the graph; the number of edges is linear to the number of nodes; and the clustering coefficient is almost independent of the number of nodes, if the number of nodes is larger than 20, as can be seen in figure 4.17. The model is, in spite of these non-trivial topological properties, far from being regular if the nodes are not regularly arranged.

Uniform SISG models are, essentially, *scale-free*: the number of nodes with a degree k is approximately proportional to $k^{-\gamma}$ for k larger than a threshold which depends on ρ^m, as can be seen in figure 4.11.

The average shortest path length of uniform SISG models grows much faster than logarithmic if the number of nodes in the model is small, as can be seen in figure 4.24. Uniform SISG models are hence *neither small-world nor ultra-small-world networks*.

We argued that uniform SISG models are complex networks and, essentially, scale-free. They are, however, neither small-world nor ultra-small-world networks.

Conclusion

Uniform SISG models are models of spatial structure that assume the points of the generating set to be randomly distributed with uniformly distribution. When no further assumptions are made on the distribution of the nodes, we may use uniform SISG models as prototypes of spatial structures. We examined properties of uniform SISG models in this chapter. More specifically, the main contributions of this chapter are as follows:

(1) Statistical methods are suitable to characterize uniform SISG models but non-statistical ones are, in most cases, not because they are sensitive to local modifications of the model.

(2) The concept of localization of global properties was introduced for graphs. It can render global properties less sensitive in respect to local modifications of the model, because the localized property can statistically be examined.

(3) The effect of finiteness and non-connectedness impedes the analysis of uniform SISG models. We introduced the concept of inner and outer regions to

understand this effect. The examination of series of subgraphs avoids this problem by examining only inner regions of the model.

(4) Most properties of uniform SISG models are either not characteristic of uniform SISG models, i. e. they cannot be used to distinguish uniform SISG models from other graphs, or they are not correlated to the density parameter and the dimension in a simple and obvious way. The number of nodes and edges, the density, the total density and the volume of the sphere are exceptions.

(5) Uniform SISG models are complex and scale-free networks, but they are neither small-world nor ultra-small-world networks.

We will, in the next chapter, compare data sets to uniform SISG models by the number of nodes and edges, the density, the total density and the volume of the sphere. The comparison can shed light on the question whether a data set has a spatial structure. The quality of the results of this comparison is predicated on the discussion of this chapter, showing the importance of this chapter's results.

5

Testing Data for Spatial Structures

—**Johann Sebastian Bach,** ***BWV847***
german composer and musician
(1685–1750)

Data sets inherit a characteristic structure when they have references to space. We can only indirectly decide whether a data set has such a spatial structure if semantics, and in consequence also explicit references to space, is missing: a comparison of the data set to uniform SISG models reveals whether both share properties and whether their structure is similar.

Spatial information exposes references to space, but the underlaying data set does, by definition, not. The comparison of the data set to uniform SISG models can, in consequence, not assume the data set to expose references to space. The issue lies with the comparison of the structure of the abstract SISG model to the ones of data sets: which structural properties shares the (abstract) graph model with spatial data sets, and shares it the same properties with non-spatial data sets? Are we, in consequence, able to conclude whether a data set has a spatial structure due to the fact that it is similar to a SISG model?

We discuss the problem of how to test whether data has a spatial structure and argue how such a test can be performed (section 5.1). Algorithms are provided to estimate the density parameter and the minimal dimension of an abstract uniform SISG model (section 5.2). These algorithms can be evaluated by applying them to uniform SISG models: the estimated parameters approximately coincide with the ones that were used to generate the SISG models (section 5.3). When the algorithms are applied to real data sets, we can estimate how similar they are to uniform SISG models and to which extent they expose a spatial structure. This approach provides the possibility to characterize real data sets by their (spatial or non-spatial) structure (section 5.4).

5.1 The Problem

The question of whether a data set has a spatial structure is, as simple as it may seem, not trivial: it is not clear, how to detect a spatial structure and which properties of a spatial structure, e. g. the dimension of the space, can be detected at all. When a SISG model is considered as an abstract graph, it is not self-evident which dimension, which embedding of the generating set in space and which density parameter were used to generate the model.

We argue, in this section, that the comparison of a data set with SISG models can, at least in parts, answer the question of whether the data set has a spatial structure (section 5.1.1). All possible values of the parameters that were used to generate a uniform SISG model can, in principle, be determined analytically (section 5.1.2). This approach is, however, not applicable for testing graphs for their spatial structure. Statistical approaches are better suited for this purpose and can be shown to successfully estimate the density parameter and the (minimal) dimension (section 5.1.3).

5.1.1 Testing Whether Data Has a Spatial Structure

Things, in particular objects, processes, events, etc., are in many cases related to the structure of space and to locations in space. A representation of these things can explicitly contain descriptions of these relations, e. g. by describing an object's location by the use of coordinates. The representation is, in this case, by definition to at least some extent spatial. When a representation does not contain such explicit references to space, it is much harder to decide whether it is spatial.

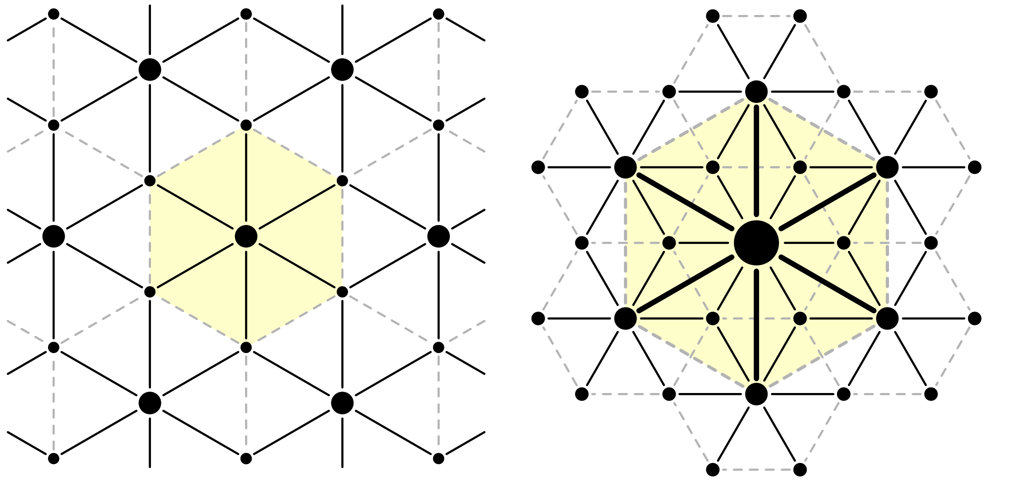
We discuss, in this section, how to decide whether a data set is spatial and which restrictions for a possible answer exist. Only graph representations are considered as examples of data sets, because the following discussion is based on things and relations.

Structures of Data Sets. When the nodes of a graph representation are related to locations in space, e. g. by a natural embedding, the graph typically[1] has to some extent the properties that were discussed in section 3.2. Data sets are usually influenced by several aspects, and space is only one of them. Spatial structure occurs only in very rare cases isolated, and real data sets are practically never identical to some SISG model.

[1] There exist graphs whose nodes are related to locations in space but which do not expose these typical properties, e. g. qualitative spatial networks. These graphs are, in consequence, very different from SISG models.

Examples of structures are manifold: the structure of a town, in particular the configuration of rivers and bridges, has an impact on timetable information; the importance of controlling a number of persons with clear responsibilities leads in many organizations to hierarchical structures, even if the organizations are spatially organized, e. g. by affiliates; and the preference of nodes with a large number of edges during the growth of a network can lead to a power law distribution of the nodes' edge degrees, e. g. in case of social networks.

Central place theory provides an example of a graph representation that has more than one structure (Christaller 1933). Relations for marketing, transport and administration are described by dividing space into hexagons and placing towns in a regular pattern in and around these hexagons. The town in the centre of each hexagon is larger and fulfills marketing, transportation and administrative functions in respect to its neighboured towns. In case of marketing, the neighboured towns are placed at the corners of the hexagon (cf. figure 5.1a). Christaller describes this arrangement at several scales by subdividing space in smaller hexagons and applying the same arrangement for these smaller hexagons (cf. figure 5.1b). The procedure of subdividing space and introducing towns and relations can be repeated, and the towns that are introduced in each step are smaller than the ones before. The resulting graph is called the *graph of the marketing principle*.



(a) One hierarchical level

(b) Two hierarchical levels, with only one hexagon in the higher level

Figure 5.1
Graph of the marketing principle ($K = 3$) according to the central place theory; (a) one hierarchical level, and (b) two hierarchical levels, with only one hexagon in the higher level

The graph that is produced by a finite number of repetitions of this process has a spatial structure. Tobler's law, for example, is met for most relations: there exist seven times as many shorter than longer relations for two hierarchical levels, as can be seen in figure 5.1b. In addition, the graph is scale-invariant by its construction, and the outdegree is bound for each node.

The spatial structure is however not the only one that determines the graph of the marketing principle. The dependencies that occur at several levels, due to the different sizes of towns, lead to a hierarchical structure. A large number of edges of the graph of the marketing principle, namely the ones that incorporate towns of the smallest level, also occur in a SISG model that uses all towns, independent of their size, as generating set. Many edges however occur only in the SISG model or in the graph of central place theory, and the difference can easily be explained by the hierarchy that is described by central place theory.

The example of central place theory illustrates that the structure of a data set can be determined by many aspects, e. g. by space and hierarchy. It is thus meaningful to ask *how important* the aspect of space is for the structure of a data set. The question of whether a data set has a spatial structure is hence a gradual one.

Spatial Structure is a Question of Representation. Representations are build to examine reality: they represent things and their interrelations by a system of formal symbols, with the aim that the system of formal symbols inherits some properties of the things and their interrelations. We can therefore conclude some of the properties of the things and their interrelations by examining a representation. Which properties can be concluded depends on the chosen representation, in particular on which things and which interrelations are and which are not represented, and on which things and relations are identified with each other because the representation maps them to the same symbol.

When it shall be concluded whether a system of things and its interrelations are related to space, usually representations of the system, e. g. timetables or mental representations in case of public transport, are examined. As representations of the same system can be very different, it is important to be aware of that we, in the first instance, examine the representation and not the represented system of things and interrelations itself, and that we can only indirectly conclude whether the system is spatial. When a representation is spatial to some extent, we can conclude that the system is, too. The reverse can however not be concluded by implication.

An analysis of the Spanish social network *Tuenti* showed that social relationships only weakly correlate to geographical distance (Kaltenbrunner et al. 2012). A representation of only the social relationships will thus not reveal a spatial structure, but the representation of the members of the network and their geographical locations would, by definition, have a spatial structure.

The Identity of Indiscernibles[2] is Wrong for Representations. A representation consists of symbols, and a mapping of objects and interrelations to these symbols. When a representation only implicitly contains the mapping, we gain an *abstract representation* which does not explicitly contain information of how it is related to the objects and interrelations. Data sets include, in many cases, only abstract representations, even if we have additional knowledge of the mapping, which is not included in the data set. Timetables contain, for example, no information on which vehicle is represented by which symbol, but as the vehicles are labelled with identifiers, e. g. line numbers, we can use this additional knowledge to establish a relation[3] between the vehicles and the timetable.

[2] The identity of indiscernibles is a principle of analytic ontology. It claims that two things are identical if they have exactly the same properties. This principle is true in some contexts but wrong in others.

[3] In case of vehicles and line numbers, there exists no bijection between the vehicles and the identifiers, but each vehicle is assigned to a line number which makes it possible to conclude, by the use of a timetable, to conclude the route of the vehicle.

We can test whether an abstract representation has a spatial structure and could, in consequence, be related to space, because we know which structure relations to space typically induce. Space and time as physical concepts both have the structure of a real vector space, a three-dimensional and a one-dimensional one. Other concepts have the same structure, and we hence cannot distinguish between the concepts just by their structure. Even if we consider space and time as more complex concepts, as we did in section 3.2, it is not clear whether an abstract graph is related to space, if it has a structure identical to a spatial structure.

Abstract representations can be identical but yet represent different things and interrelations. The positions of a light switch, *up* and *down*, and the transitions between them are, for example, spatial because they describe locations in space. The representation of the same situation by the states of either *emitting light* or *not* is not spatial but yet, as an abstract representation, identical to the previous representation.

The more extensive an abstract representation is, the more reasonable it is to conclude by the existence of a spatial structure, in particular properties 3.3, that the representation is related to space, because the probability that this happens by chance is smaller. The comparison of an abstract representation with a spatial structure cannot determine with absolute certainty whether an abstract representation or a data set is related to space, but it can determine whether this is probable.

Comparing Graph Representations to the Uniform SISG Model. The question of whether a graph representation is spatial can be approached by comparing its structure to a typical structure of spatial information, i. e. to the uniform SISG model. Such a comparison can be conducted in several ways. A nearby possibility is to compare the properties of the graph representation to the ones of the uniform SISG model, i. e. the properties shared by all uniform SISG models. A more adjusted possibility is to find a uniform SISG model that is most similar to the graph representation by comparing the representation to different models. The issue lies with how to determine this uniform SISG model, i. e. the density parameter and the dimension of the model and how to compare it to the graph representation.

We necessarily have to restrict the examination of the SISG model and the graph representation to a finite number of properties. We discussed in sections 4.4 and 4.5 how the model's properties are correlated to the density parameter and the dimension. We can, due to the correlation, estimate the density parameter and the dimension that were used to generate the model, by the computation of suitable properties of a uniform SISG model. The same strategy can also be applied for graph representations: we can, by computing the same properties, in the same way as before try to estimate which density parameter and which dimension can be used to generate a uniform SISG model that is similar to the graph representation.

We discussed principle considerations of how to test data sets for spatial structures and which answers can be expected. Methods and algorithms for this purpose are discussed in subsequent sections.

5.1.2 Reconstruction of the Parameters of a SISG Model

Abstract unlabelled graphs consist of nodes and edges but neither the nodes nor the edges have labels or locations. It is not self-evident whether an abstract graph is a SISG model, i. e. whether there exist a dimension, a generating set of points embedded in space and a density parameter such that the SISG model generated with these parameters equals the abstract graph, and it is not self-evident how to conclude these parameters.

We introduce, in this section, the notion of a minimal dimension in order to state the question of these parameters more precisely. An approach of how to conclude these parameters is introduced.

The Problem of Reconstruction. Nodes of an abstract graph have no location in space, and it is not straightforward to check whether the graph is a SISG model:

Question 5.1. Can we conclude whether a given abstract graph G is a SISG model?

When an abstract graph equals a SISG model, we may be able to find a dimension, a generating set of points embedded in space and a density parameter that generate the abstract graph as a SISG model. The dimension, the generating set and the density parameter are in many cases not unique, because different combinations generate the same SISG model.

The embedding of the generating set and the density parameter are never unique: a set of nodes N embedded in space and a density parameter ρ generate the same SISG model as the set $\tau(N)$ and the density parameter $\sigma \cdot \rho$ for a transformation τ of relative scale $\sigma > 0$. The density parameter is, for a fixed embedding of the nodes in space, almost never unique if the number of nodes is finite, because the set of distances between points is finite: the distances of a node to the other nodes define an interval of possible values of the density parameter. Such an interval of possible values is defined for every node, and as the number of nodes is finite, the intersection of all these values has almost always positive length, i. e. there exist infinitely many possible density parameters, which all generate the same model.

The dimension is never unique, because a vector space can always be embedded in a vector space of higher dimension such that the distances between points do not change. In particular, there always exists, for a given SISG model of dimension m and for every number $n > m$, a SISG model of dimension n that is (as an abstract graph) identical to the original SISG model. As the dimension has 0 as a lower bound, there exists for every SISG model a minimal dimension m such that the model can be generated with the generating set embedded in the m-dimensional space. We formally define:

Definition 5.2. The *minimal dimension* of a SISG model $\mathcal{M}_\rho(S, V)$ is the minimum of all $\dim W$ such that $\mathcal{M}_\rho(S, V) = \mathcal{M}_\rho(S', W)$. We even denote $\mathcal{M}_\rho^m(S, V)$, where $m = \dim \mathcal{M}_\rho(S, V)$ is the minimal dimension.

The minimal dimension is unique, because it exists and is per definition the lowest integer that meets the criteria. We can now formally raise the question of all possible combinations of parameters that were used to generate a SISG model:

Question 5.3. A set of points $S \subset V$ and a density parameter $\rho > 1$ defines a model $\mathcal{M}_\rho(S)$. Can we discover the minimal dimension $\dim \mathcal{M}_\rho(S)$, the set of points S and the density parameter ρ knowing only the abstract graph $\mathcal{M}_\rho(S)$?

We discuss, in the following, whether and how an exact answer to this question could be found. This discussion is however of theoretical and not of practical nature, because the considerations include infinite sets.

Exact Reconstruction. We can compute the minimal dimension, all possible choices of generating sets and all possible values of the density parameter by inductively reconstructing the parameters for subgraphs. After having enumerated the nodes by $p_0, \dots, p_{s-1}$, we can consecutively reconstruct the parameters for the subgraphs induced by the nodes $\{p_0, \dots, p_k\}$ for $k = 0, \dots, s-1$, as is discussed in the following.

For $k = 0$, we can place the node p_0 in a zero-dimensional space, or equivalently at an arbitrary location in one-dimensional space because the SISG model is translation invariant by corollary 3.13. The minimal dimension is 0 and the density parameter can be any number $\rho > 1$.

For $k = 1$, we can place the node p_1 at an arbitrary location, non-equal to the one of p_0, in the one-dimensional space, because the SISG model is invariant under scale transformations by theorem 3.12. The minimal dimension is 1, because the zero-dimensional space only consists of one point but two nodes have to placed in space, and the density parameter can still be any number $\rho > 1$.

In each induction step, we compute possible locations for the new node p_k, resulting in one or more regions that p_k can be placed in. The location of p_k itself restricts the locations of the nodes $p_0, \dots, p_{k-1}$, and valid combinations of locations for the nodes $p_0, \dots, p_k$ can iteratively be computed. Each of these combinations of locations defines combinations of a minimal dimension and intervals of possible density parameters.

When $k = s - 1$, all combinations of locations of the points in the generating set, the minimal dimension and all possible values of the density parameter for the whole graph are found. The number of combinations can, in practice, be very high, in most cases even infinite. An exact reconstruction is thus, in many cases, not practical.

We discussed how to analytically conclude all possible combinations of the minimal dimension, the density parameter and the embedding of the generating set in space that generate a given uniform SISG model. We will discuss in the next section, why the exact reconstruction of these parameters is only of theoretical interest, and why statistical methods should be favoured.

5.1.3 Finding a SISG Model Similar to a Given Graph

We have discussed, in the last section, how the parameters of a SISG model can, in principle, be reconstructed. For the examination of real data sets, we however need to answer the following, more general question:

Question 5.4. Can we conclude whether a given abstract graph G is similar to a SISG model, and can we conclude which minimal dimension, which generating set and which density parameter generate a SISG model that is similar to G?

We discuss, in this section, why statistical methods are, in contrast to analytical ones, most suitable for answering this question and propose a statistical approach to answer the question.

Exact Reconstruction is not Suitable for Real Data. Graph representations are not necessarily equal to a SISG model, even if both may be very similar and may share many statistical properties. An exact reconstruction of the parameters is thus incapable of answering question 5.4, because the graph representation is, in general, not a SISG model, and analytical discussions can lead to (very) wrong conclusions if they contains contradictions.

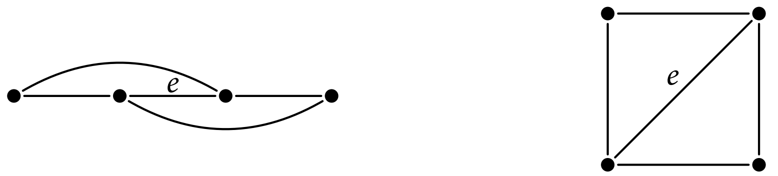
Exact reconstruction cannot provide meaningful answers to question 5.4 for real data sets, even if it would be generalized to meet the generalized context: the parameters depend very sensitive to local modifications, and SISG models only differing in some nodes or edges can have very different density parameter and minimal dimension. When the characteristics of the data sets and its structure as a whole is of interest, the answer to question 5.4 is expected to statistically depend on the structure, and thus to only insensitively react to local modifications.

A connected component of minimal dimension m, for example, proves by proposition 3.15 the whole SISG model to be of minimal dimension of at least m. This shows that the properties of one connected component can have an effect on the properties of the hole model. Proposition 3.14 suggests similar effects when edges inside a connected component are modified, because the proposition relates the local structure (the structure of a subgraph) to the global structure of a SISG model. Local modification of edges can hence result in a change of the properties of the whole graph.

This behaviour can be illustrated for an undirected SISG model G of dimension 1 which contains the induced subgraph of figure 5.2. If we remove the edge e, the minimal dimension becomes greater than 1 by proposition 3.18, because the resulting graph has a hole of size greater than 3. The minimal dimension of the modified graph G' would thus be greater than 1, or the modified graph would not be a SISG model at all.

A similar consideration can be made for a directed SISG model G of dimension 2 that contains the induced subgraph of figure 5.3. If we reverse the orientation of the edge e, the resulting graph G' would not any longer be a SISG model by proposition 3.16.

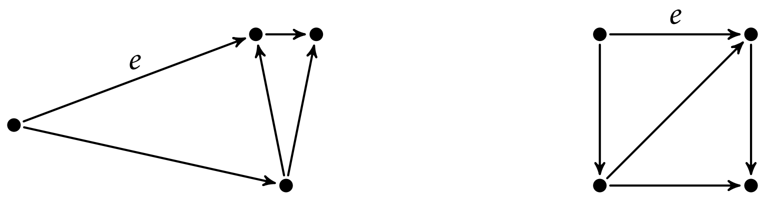
An exact reconstruction of the parameters is neither suitable nor possible for real data sets, as was discussed. Since only an approximate answer to question 5.4 that is insensitive to local modifications of the graph is needed, we will, in the following, propose a statistical approach to answer the question.



(a) Embedded in the line

(b) Abstract graph

Figure 5.2
Example of an undirected SISG model; (a) embedded in the line to illustrate the construction, and (b) as an abstract graph



(a) Embedded in the plane

(b) Abstract graph

Figure 5.3
Example of a directed SISG model; (a) embedded in the plane to illustrate the construction, and (b) as an abstract graph

Statistical Approach. For a reconstruction of the density parameter and the minimal dimension, we can compare statistical properties of the considered graph G to statistical properties of a number of SISG models. It is not clear, without further assumptions, how the generating set S of the model should be embedded in space and which distribution of the generating set S in space is convenient. There is no preferred distribution because physical space is uniform. We hence assume the points of the generating set to be randomly distributed with a uniform distribution. The mean value of the properties for models $\mathcal{M}_\rho(S)$ with several randomly generated sets S can be used to compensate for statistical variance. We can decide whether the parameters generate a SISG model similar to the graph G by comparing this mean value to the properties of the graph G.

A graph is not completely determined by its statistical properties and may thus share statistical properties with a SISG model but still be different from any SISG model, as we can see in the example of the density: the density of a SISG model depends on the number of nodes, the density parameter and the minimal dimension (cf. section 4.4.2). Due to this dependency, we can, in principle, conclude (a combination of) the density parameter and the minimal dimension of a SISG model by the computation of the density. Graphs with a given density may be similar to the SISG model, but they may also be very different from any SISG model. The more properties of a graph coincide with the ones of a SISG model, the higher is the probability that the graph is similar to a SISG model.

We have examined properties of uniform SISG models in sections 4.4 and 4.5. Some properties are less sensitive to local modifications of the graph than others, and are therefore called *robust*. The average shortest path length, the clustering coefficient and the degree distribution, for example, are regarded as robust. Properties of SISG models can however differ considerably, i. e. they can have a high variance (cf. section 4.5.3), even for the same number of points in the generating set, the same density parameter and the same minimal dimension. It is necessary to find properties that are stable and exhibit a low variance for SISG models to enable a meaningful reconstruction of the density parameter and the minimal dimension.

The number of nodes and edges, the density and the volume of spheres have turned out to be robust and to have relatively low variance, as was discussed in sections 4.4 and 4.5. These properties expose a dependency on the density parameter and the minimal dimension, and the number of nodes and edges as well as the density is simple to understand and to describe. Furthermore, the dependencies of these properties are different, which is why we can, by these dependencies, conclude the density parameter and the minimal dimension independently, which answers question 5.4 at least in parts; algorithms to draw these conclusions are yet needed.

We discussed the question of how to test data sets for spatial structure and argued why statistical approaches are most suited. A statistical approach of comparing a graph to uniform SISG models and reconstructing the parameters of the SISG model was proposed. Algorithms for this approach will be provided in the next section.

5.2 Algorithms

A general approach for the estimation of a density parameter ρ and a minimal dimension m, such that a uniform SISG model $\mathcal{M}_\rho^m(|G|)$ is similar to a given graph G, was discussed in the last section. Possible values for the density parameter and the minimal dimension can only be estimated because of the statistical nature of the discussion.

A possible value of ρ^m can be estimated by the number of nodes and edges (section 5.2.1). The algorithm can be improved by taking outer regions into account (section 5.2.2). A value of ρ^m can alternatively be estimated by the density (section 5.2.3). This algorithm can, similarly to the considerations for the first algorithm, be improved by taking outer regions into account (section 5.2.4). We can decide whether a graph is similar to a uniform SISG model by comparing the estimate by the number of nodes and edges with the estimate by the density (section 5.2.5). The parameters ρ and m can independently be estimated by comparing the estimate of ρ^m to the volume of spheres of the graph (section 5.2.6).

5.2.1 Estimation of ρ^m by the Number of Nodes and Edges

We introduce, in this section, an algorithm to estimate, for a given graph G, the value of ρ^m with ρ the density parameter and m the dimension, such that the graph

G is similar to a uniform SISG model with density parameter ρ and dimension m. This algorithm provides, at least in parts, answers to question 5.4.

For a graph $\mathcal{M}_\rho^m(s)$ with e edges, we expect $\rho^m = e/s$ for $s \to \infty$ according to proposition 4.4. Algorithm 5.1 describes this estimation. The algorithm is expected to not only work for SISG models but also for graphs that are similar to uniform SISG models, because the number of edges is robust and has low variance for uniform SISG models (cf. section 4.4.1).

Algorithm 5.1
Estimate ρ^m by the number of nodes and edges

```
Algorithm: EstimateKM_{nodes,edges}($G$)

   Input:  Graph $G$ similar to a uniform SISG model $\mathcal{M}_\rho^m(s)$
   Output: Estimate of $\rho^m$ (only correct for $s \to \infty$)

   $s \leftarrow$ NumberOfNodes($G$)
   $e \leftarrow$ NumberOfEdges($G$)
   return $e/s$
```

The discussed algorithm is based on an analytical result that is only valid if the number of nodes approaches infinity. The effect of the graph's finiteness is not considered. We will address this problem by a heuristic approach in the next section.

5.2.2 Improved Estimation of ρ^m by the Number of Nodes and Edges

We improve algorithm 5.1 in this section by heuristically addressing the problem of the graph's finiteness. The heuristics uses the secant method (cf. section B.1): the deviation of ρ^m from the theoretical value of ρ^m, which describes the limit when the number of nodes approaches infinity, is inductively used to improve the estimate of ρ^m.

The expectation value of the density is only given for $|G| \to \infty$ by proposition 4.4, and algorithm 5.1 does hence not consider the effect of the graph's finiteness. The estimates of ρ^m are, in most cases, lower than the real value, as can be seen in figure 5.4.

In order to compensate for this problem, we are looking for a SISG model $\mathcal{M}_\rho^m(|G|)$ such that

$$\texttt{EstimateKM}_{\texttt{nodes,edges}}(G) \approx \texttt{EstimateKM}_{\texttt{nodes,edges}}(\mathcal{M}_\rho^m(|G|)). \tag{5.5}$$

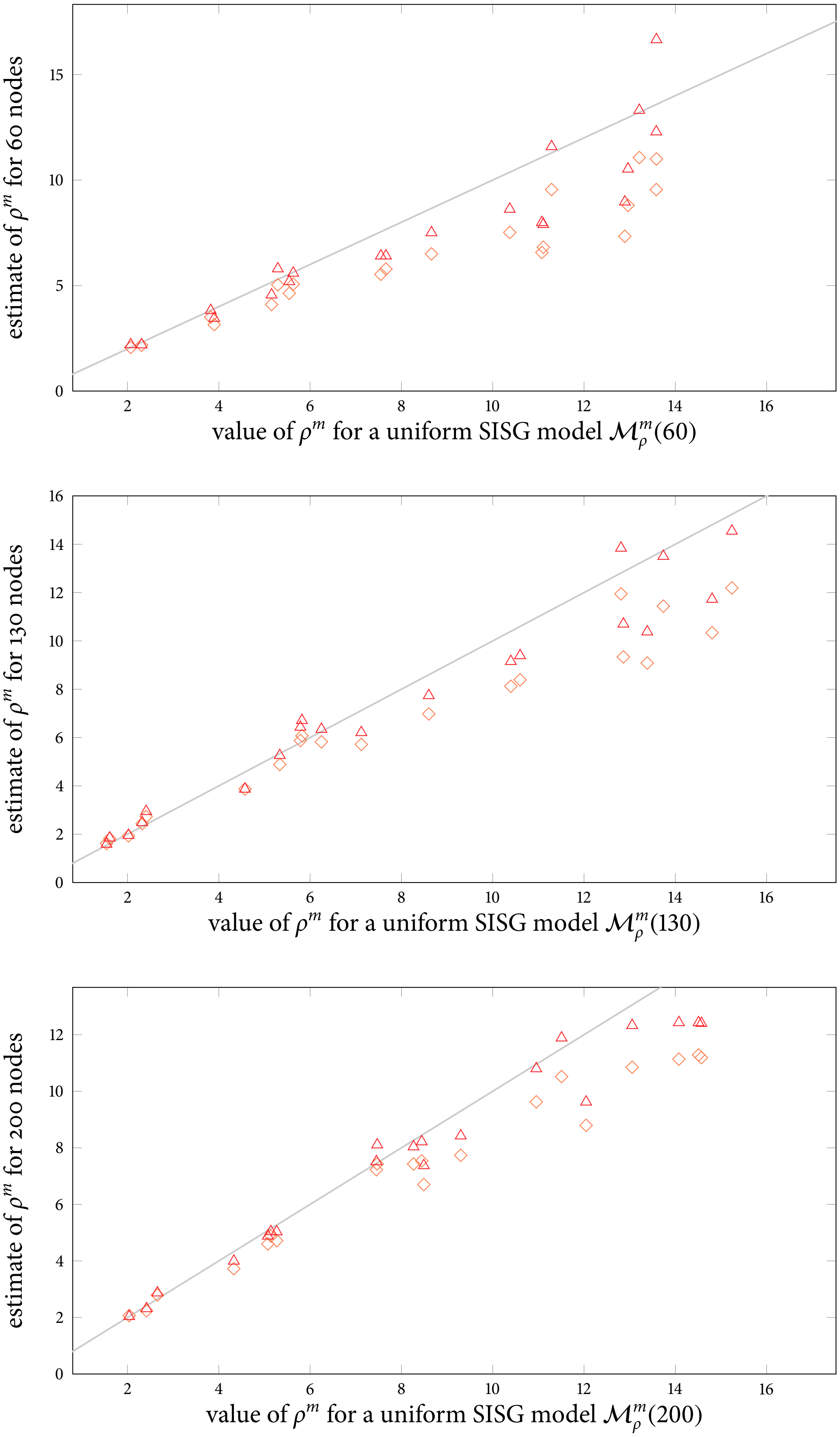
The value ρ^m is an improved estimate which is, for uniform SISG models $G = \mathcal{M}_{\rho'}^{m'}(s)$, by definition correct, i. e. $\rho = \rho'$ and $m = m'$.

The computation of $\texttt{EstimateKM}_{\texttt{nodes,edges}}(\mathcal{M}_\rho^m(|G|))$ in equation 5.5 depends on the generating set which is randomly distributed with uniform distribution. The

Figure 5.4
Estimation of the density parameter and the minimal dimension for uniform SISG models

◇ non-improved (algorithm 5.1)

△ improved (algorithm 5.2)



mean value of the number of edges for many SISG models with the same density parameter and minimal dimension can be used to compensate for the effect of the choice of the generating set.

Algorithm 5.2 aims at providing a solution for equation 5.5, i. e. for finding ρ and m such that they satisfy the equation. At the beginning of the algorithm, an estimate of ρ^m for the graph G is computed by algorithm 5.1, i. e. the left side of equation 5.5. In the following, the algorithm inductively tries to guess estimates of ρ^m. In each step, it tries to improve the estimation and test whether the improved estimate satisfies equation 5.5. If it does sufficiently well, a final estimate is found.

The new estimate is computed in each step by first computing the right side of equation 5.5 for the current estimate of ρ^m, and by secondly guessing a possibly improved estimate using the comparison of both sides of equation 5.5. For the computation of the right side of equation 5.5, we need to compute values of ρ and m, but only an estimate of their combination ρ^m is known. If more than one choice of (ρ, m) is possible, we choose the one which satisfies $\rho \approx m$ best. This choice has, however, only very little influence on the value of the right side of the equation, as can be expected by proposition 4.4. The algorithm of choosing[4] a ρ and m is abbreviated by `BalanceKM` in algorithm 5.2. The possibly improved estimate is finally guessed by the secant method (Forsythe et al. 1977, pp. 159f).

It may happen that algorithm 5.2 does not converge, because the secant method does neither. The value ρ^m is, for a uniform SISG model $G = \mathcal{M}^{m'}_{\rho'}(s)$, correct, because it is checked at the end of the algorithm that the estimate ρ^m satisfies equation 5.5.

A comparison of algorithm 5.2 to algorithm 5.1 is provided by figure 5.4. The results of the improved algorithms are better, as can be seen in the figure, and the improvements are greater for smaller graphs.

We will, in the next section, discuss an alternative method of how to estimate ρ^m that uses the density of the graph.

5.2.3 Estimation of ρ^m by the Density

We introduce, in this section, an alternative to algorithm 5.1. Both, the algorithm introduced in this section and algorithm 5.1, estimate the same value and return very similar results, if the considered graph is similar to a uniform SISG model. We will use this circumstance later in section 5.2.5 to decide whether a graph is similar to a uniform SISG model.

The expected density of a SISG model $\mathcal{M}^m_\rho(s)$ is $\rho^m/(s-1)$ for $s \to \infty$, according to proposition 4.6. We hence expect the density of a subgraph with t nodes to be slightly less than $\rho^m/(t-1)$ due to proposition 3.14, and in case of $t \to s$, the expectation value of the density converges to $\rho^m/(t-1)$. When a uniform SISG model G is given as an abstract graph, we can compute the density for a series of

[4] This choice can cause a slow convergence of algorithm 5.2 when ρ^m is in the neighbourhood of a discontinuity of `BalanceKM`.

Algorithm 5.2
Estimate ρ^m by the number of nodes and edges; improved version of algorithm 5.1

```
Algorithm: EstimateKM'_{nodes,edges}(G, η, τ)

Input: Graph $G$ similar to a uniform SISG model $\mathcal{M}_\rho^m(s)$, the number of
       models $\mu$, a maximum number of iterations $\eta$ and a threshold $\tau$
Output: Estimate of $\rho^m$

$s \leftarrow$ NumberOfNodes($G$)
$\kappa_{\text{aim}} \leftarrow$ EstimateKM_{nodes,edges}($G$)            // algorithm 5.1
$\kappa \leftarrow \kappa_{\text{aim}}$
$\kappa' \leftarrow 1.4 \cdot \kappa$
$(\rho, m) \leftarrow$ BalanceKM($\kappa'$)    // find $(\rho, m)$ with $\rho^m = \kappa'$ and $\rho \approx m$
$L \leftarrow \{\}$
for $i = 0$ to $\mu$ do
    $\tilde{G} \leftarrow \mathcal{M}_\rho^m(|G|)$                      // generate a model
    Append EstimateKM_{nodes,edges}($\tilde{G}$) to $L$         // algorithm 5.1
$\kappa'_{\text{test}} \leftarrow$ ArithmeticMean($L$)                // real number
$c \leftarrow 0$
while $c \leq \eta$ do
    $c \leftarrow c + 1$
    if $\kappa' \leq 1$ then
        error 'parameters out of bound'
    $(\rho, m) \leftarrow$ BalanceKM($\kappa$)   // find $(\rho, m)$ with $\rho^m = \kappa$ and $\rho \approx m$
    $L \leftarrow \{\}$
    for $i = 0$ to $\mu$ do
        $\tilde{G} \leftarrow \mathcal{M}_\rho^m(|G|)$                  // generate a model
        Append EstimateKM_{nodes,edges}($\tilde{G}$) to $L$     // algorithm 5.1
    $\kappa_{\text{test}} \leftarrow$ ArithmeticMean($L$)             // real number
    if $|\kappa_{\text{aim}} - \kappa_{\text{test}}| < \tau \cdot \kappa_{\text{aim}}$ then
        return $\kappa$
    $\tilde{\kappa} \leftarrow \kappa + (\kappa_{\text{test}} - \kappa_{\text{aim}}) \cdot \frac{\kappa - \kappa'}{\kappa_{\text{test}} - \kappa'_{\text{test}}}$
    $\kappa' \leftarrow \kappa$
    $\kappa'_{\text{test}} \leftarrow \kappa_{\text{test}}$
    $\kappa \leftarrow \tilde{\kappa}$
error 'no convergence'
```

subgraphs and estimate ρ^m by fitting the density values to the function $\rho^m/(t-1)$. This computation yields good estimates for $s \to \infty$ but not necessarily for finite graphs.

We can compensate the effect of the graph's finiteness on the density by using the total density instead of the density, as was discussed in section 4.4.2. We hence fit the function $\rho^m/(t-1)$ to the total density for a series of subgraphs. The estimation can be statistically improved when more than one series of subgraphs is considered in order to minimize the variance. Algorithm 5.3 is a direct result of these considerations. The algorithm is expected to not only work for uniform SISG models but also for graphs that are similar to uniform SISG models, because the total density is robust and has low variance for SISG models (cf. section 4.4.2).

Algorithm 5.3
Estimate ρ^m by fitting the total density for a series of subgraphs

```
Algorithm: EstimateKM_density(G, μ)

  Input:  Graph $G$ similar to a uniform SISG model $\mathcal{M}_{\rho}^{m}(s)$, and the number
          of series $\mu$
  Output: Estimate of $\rho^m$ (only correct for $s \to \infty$)

  $L \leftarrow \{\}$
  for $i = 0$ to $\mu$ do
      $S \leftarrow$ SeriesOfSubgraphs(G)                  // list of subgraphs
      $\tilde{D} \leftarrow$ Map $c_{\text{total density}}$ to $S$          // list of real numbers
      Append $\tilde{D}$ to $L$
  $D \leftarrow$ ElementwiseArithmeticMean(L)    // list of real numbers
  $\kappa \leftarrow$ FitInverse1(D)          // fit by $\kappa/(1-d)$ (least square)
  return $\kappa$
```

We discussed in section 4.4.2 that the total density of a subgraph compensates for the subgraph's finiteness as long as the subgraph is completely contained in the inner region of the graph. In this section, we considered however all subgraphs, because no good measure of whether a subgraph is completely contained in the inner region is known. We will address this problem by a heuristic approach in the next section.

5.2.4 Improved Estimation of ρ^m by the Density

We improve, in this section, algorithm 5.3 by heuristically addressing the problem of the subgraph's finiteness. The heuristics uses the secant method and is very similar to the approach that was used in section 5.2.2.

The expectation value for the density is, in proposition 4.6, only given for $|G| \to \infty$, and the use of the total density compensates only in parts for the effect of the graph's finiteness when the subgraph is not completely contained in the inner region, as can be seen in figure 5.5. In order to compensate for this effect, we are looking for a SISG model $\mathcal{M}_{\rho}^{m}(|G|)$ such that

$$\texttt{EstimateKM}_{\texttt{density}}(G,\mu) \approx \texttt{EstimateKM}_{\texttt{density}}(\mathcal{M}_{\rho}^{m}(|G|),\mu). \tag{5.6}$$

The value ρ^m is an improved estimate which is, for uniform SISG models $G = \mathcal{M}_{\rho'}^{m'}(s)$, by definition correct, i. e. $\rho = \rho'$ and $m = m'$.

The computation of $\texttt{EstimateKM}_{\texttt{density}}(\mathcal{M}_{\rho}^{m}(|G|),\mu)$ in equation 5.6 depends on the generating set, which is randomly distributed with uniform distribution. The

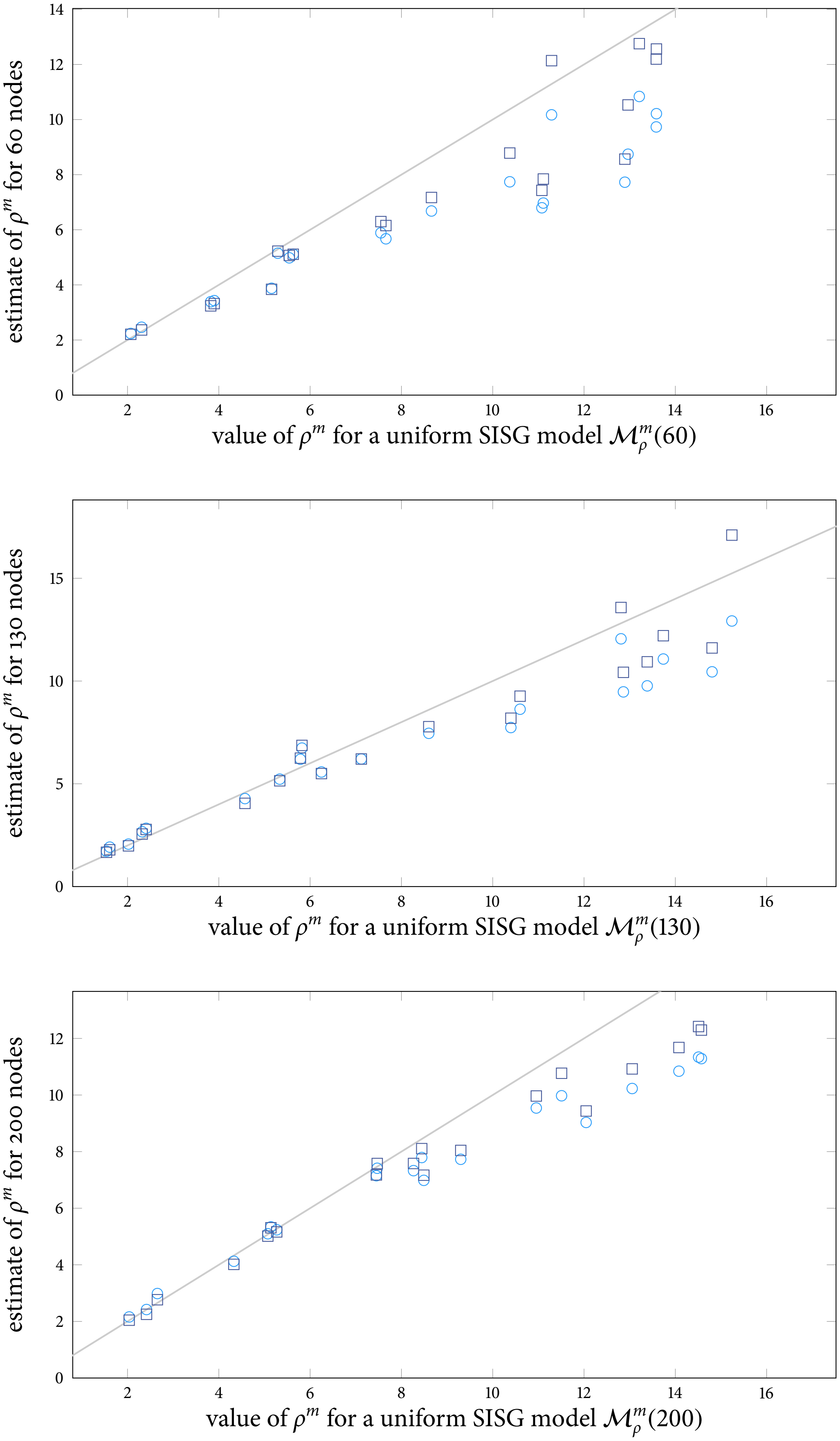


Figure 5.5
Estimation of the density parameter and the minimal dimension for uniform SISG models

○ non-improved (algorithm 5.3)

□ improved (algorithm 5.5)

mean value of the total density of subgraphs for many SISG models[5] with the same density parameter and minimal dimension can be used to compensate for the effect of the choice of the generating set. The modification of algorithm 5.3, which computes the mean value for several SISG models, is described in algorithm 5.4. We are hence looking for ρ and m such that

$$\mathtt{EstimateKM}_{\mathtt{density}}(G,\mu) \approx \mathtt{EstimateKM}^{\mathtt{SISG}}_{\mathtt{density}}(\rho, m, |G|, \mu). \qquad (5.7)$$

Algorithm 5.4
Average of the result of algorithm 5.3 for multiple uniform SISG models

```
Algorithm: EstimateKM$^{\mathtt{SISG}}_{\mathtt{density}}$($\rho$, $m$, $s$, $\mu$)

Input:  Density parameter $\rho$, dimension $m$, a number of nodes $s$ and the
        number of models $\mu$
Output: Average value of EstimateKM$_{\mathtt{density}}$($\mathcal{M}^m_\rho(s)$, $\mu$) for $\mu$ uniform
        SISG models $\mathcal{M}^m_\rho(s)$

$L \leftarrow \{\}$
for $i = 0$ to $\mu$ do
    $G \leftarrow \mathcal{M}^m_\rho(s)$                          // generate a model
    $S \leftarrow$ SeriesOfSubgraphs ($G$)              // list of subgraphs
    $\tilde{D} \leftarrow$ Map $c_{\text{total density}}$ to $S$        // list of real numbers
    Append $\tilde{D}$ to $L$
$D \leftarrow$ ElementwiseArithmeticMean($L$)   // list of real numbers
$\kappa \leftarrow$ FitInverse1($D$)        // fit by $\kappa/(1-d)$ (least square)
return $\kappa$
```

Algorithm 5.5 is an improved version of algorithm 5.3, because it aims at providing a solution for equation 5.7, i. e. for finding ρ and m such that they satisfy the equation.

The improvement is gained in a very similar way to algorithm 5.2: an estimate of ρ^m for a given graph G is computed by algorithm 5.3, and it is compared to the estimate of ρ^m for uniform SISG models by algorithm 5.4. The parameters used to generate the SISG models are adjusted inductively until the estimate of ρ^m is equal for G and the SISG models. If both estimates approximately coincide, equation 5.7 is satisfied. The algorithm then returns the parameters that were used to generate the SISG models.

It may happen that algorithm 5.5 does not converge, because the secant method does neither. The value ρ^m is, for a uniform SISG model $G = \mathcal{M}^{m'}_{\rho'}(s)$, correct, because it is checked at the end of the algorithm that ρ^m satisfies equation 5.7.

A comparison of algorithm 5.5 to algorithm 5.3 is provided by figure 5.5. The results of the improved algorithms are better, as can be seen in the figure, and the improvements are greater for smaller graphs.

Algorithm 5.5
Estimate ρ^m by fitting the total density for a series of subgraphs;
improved version of algorithm 5.3

```
Algorithm: $\texttt{EstimateKM}'_{\texttt{density}}(G, \eta, \tau)$

Input:  Graph $G$ similar to a uniform SISG model $\mathcal{M}^m_\rho(s)$, the number of
        models/series of subgraphs $\mu$, a maximum number of iterations $\eta$
        and a threshold $\tau$
Output: Estimate of $\rho^m$

$s \leftarrow \texttt{NumberOfNodes}(G)$
$\kappa_{\text{aim}} \leftarrow \texttt{EstimateKM}_{\texttt{density}}(G, \mu)$        // algorithm 5.3
$\kappa \leftarrow \kappa_{\text{aim}}$
$\kappa' \leftarrow 1.4 \cdot \kappa$
$(\rho, m) \leftarrow \texttt{BalanceKM}(\kappa')$     // find $(\rho, m)$ with $\rho^m = \kappa'$ and $\rho \approx m$
$\kappa'_{\text{test}} \leftarrow \texttt{EstimateKM}^{\texttt{SISG}}_{\texttt{density}}(\rho, m, |G|, \mu)$        // algorithm 5.4
$c \leftarrow 0$
while $c \leq \eta$ do
    $c \leftarrow c + 1$
    if $\kappa' \leq 1$ then
        error ‘parameters out of bound’
    $(\rho, m) \leftarrow \texttt{BalanceKM}(\kappa)$   // find $(\rho, m)$ with $\rho^m = \kappa$ and $\rho \approx m$
    $\kappa_{\text{test}} \leftarrow \texttt{EstimateKM}^{\texttt{SISG}}_{\texttt{density}}(\rho, m, |G|, \mu)$        // algorithm 5.4
    if $|\kappa_{\text{aim}} - \kappa_{\text{test}}| < \tau \cdot \kappa_{\text{aim}}$ then
        return $\kappa$
    $\tilde{\kappa} \leftarrow \kappa + (\kappa_{\text{test}} - \kappa_{\text{aim}}) \cdot \frac{\kappa - \kappa'}{\kappa_{\text{test}} - \kappa'_{\text{test}}}$
    $\kappa' \leftarrow \kappa$
    $\kappa'_{\text{test}} \leftarrow \kappa_{\text{test}}$
    $\kappa \leftarrow \tilde{\kappa}$
error ‘no convergence’
```

Algorithm 5.5 provides, at least in parts, answers to question 5.4. We will, in the next section, use this algorithm to approach the question of whether a graph has properties similar to a uniform SISG model.

5.2.5 Deciding Whether a Graph Is Similar to a SISG Model

Algorithms 5.1 and 5.4 as well as their improved versions, algorithms 5.2 and 5.5, assume that the examined graph is similar to a uniform SISG model. We use this fact to discuss the question of how to decide whether a given graph is similar to a uniform SISG model (cf. question 5.4).

It is not clear how similarity of graphs shall be measured, and the question of whether a graph is similar to a uniform SISG model is hence vague. We regard two (abstract) graphs as equal, if the number of nodes is the same and the edges relate the nodes in the same way. Two graphs have the same properties if they are equal, and we expect them to share many properties if they are similar. We will approach the question of how similar two graphs are by comparing their properties and deciding how similar the properties are. It is important to choose, for this purpose, robust properties that are characteristic for SISG models and spatial graphs. We will exemplify the approach using the already discussed algorithms.

Algorithms 5.2 and 5.5 both provide estimates of ρ^m for a given graph G that is similar to some uniform SISG model $\mathcal{M}_\rho^m(s)$. This coincidence is of particular interest, if the graph G is an abstract graph and if the density parameter and the dimension of the SISG model are not known: we expect both algorithms to approximately result the same estimate for ρ^m in case that the graph G is similar to a uniform SISG model. When the returned values differ, the graph G cannot be similar to a uniform SISG model.

Equality of the computed values does not necessarily imply the graph to be similar to a uniform SISG model: we expect, for example, algorithms 5.1 and 5.3 both to approximately return the same value even for arbitrary graphs, when the number of edges linearly depends on the number of nodes for subgraphs, because a subgraph with s nodes and e edges would have density $\kappa/(s-1)$ for $\kappa = e/s$. The improved algorithms 5.2 and 5.5 may also approximately return the same value, but as the improvements in algorithms 5.2 and 5.5 incorporate further properties of SISG models that determine the effect of the boundary regions, they may yield different values in some cases.

We will see in section 5.4 that the answers obtained by the comparison of the estimates by the algorithms of this section are reasonable for the considered data sets. Further properties could be compared in order to more precisely answer the question of whether a graph is similar to a SISG model.

We discussed estimates for ρ^m in sections 5.2.1 to 5.2.4. Independent estimates of ρ and m could be used to even more precisely approach the question of whether a data set is spatial: when algorithms 5.1 and 5.3 (and algorithms 5.2 and 5.5) return approximately the same estimates for a data set but the estimated values of ρ and m are very different from the one's that we expect for spatial data (e. g. when m is much greater than 3), the data set may have some properties of the SISG model but has, most probably, no spatial structure.

5.2.6 Estimation of ρ and m

We discussed, in the previous sections, how we can estimate ρ^m for a given graph G that is similar to a uniform SISG model $\mathcal{M}_\rho^m(s)$. We use, in this section, the volume of the sphere (cf. section 4.5.1) to independently estimate ρ and m.

The number of nodes and edges as well as the density depend only on the combination ρ^m of the density parameter ρ and the minimal dimension m. We thus cannot independently estimate ρ and m by these properties. When we examine a property that depends on ρ and m in a different way, i. e. not solely on ρ^m, we

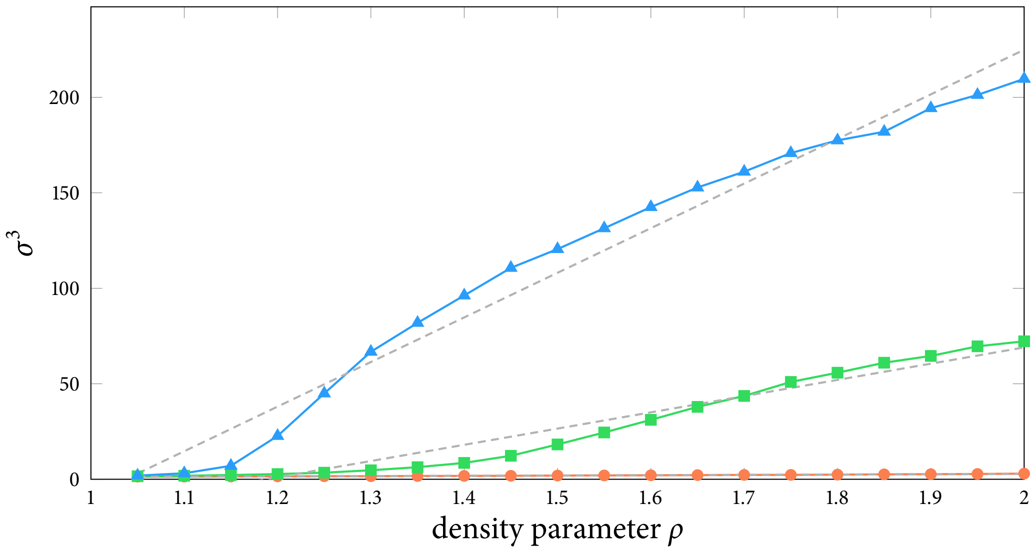


Figure 5.6
Parameter σ^3; mean value for 100 models and 10 centre nodes

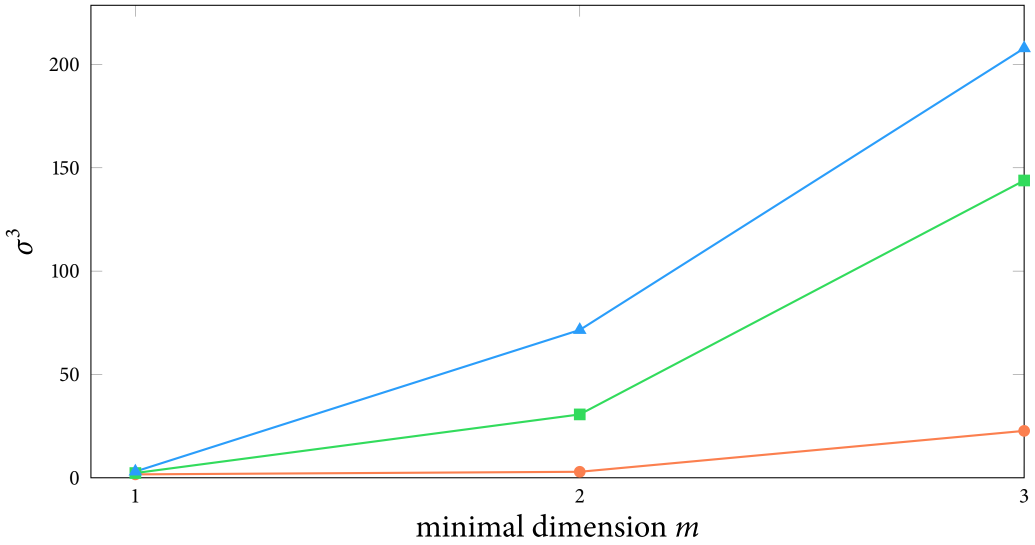


Figure 5.7
Parameter σ^3; mean value for 100 models and 10 centre nodes

can independently estimate ρ and m. We require the property, in addition, to be robust and to have little variance for SISG models to reliable estimate ρ and m. A property that meets these requirements is, for example, the volume of the sphere and σ_3 (cf. section 4.5.1).

An analytical computation of how σ_3 depends on ρ and m is complicated, but the dependency can be fitted and heuristically be used. The dependency of σ_3 on ρ is, with some deviations, linear in the range that is relevant for our considerations (cf. figure 5.6). The dependency of σ_3 on m is almost polynomially (cf. figure 5.7). It is hence reasonable to fit the dependency of σ_3 on ρ and m by

$$f(\rho, m) = \sum_{i\in\{0,1\},\, j\in\{0,1,2\}} \alpha_{ij}\rho^i m^j.$$

Algorithm 5.6 uses σ_3 to estimate ρ and m for a given graph G that is similar to a uniform SISG model. In a first step, σ_3 is computed for uniform SISG models

Algorithm 5.6
Estimate ρ and m by fitting the total density and σ_3 for a series of subgraphs

```
Algorithm: EstimateKAndM($G$, $\mu$, $\eta$, $\tau$, $R$, $M$)

  Input:  Graph $G$ similar to a uniform SISG model $\mathcal{M}_\rho^m(s)$, the number of
          models/series of subgraphs $\mu$, the maximum number of iterations $\eta$,
          a threshold $\tau$, a list of density parameters $R$ in a relevant range and
          a list of minimal dimensions $M$ in a relevant range
  Output: Estimates of $\rho$ and $m$

$F \leftarrow \{\}$
foreach $\tilde{\rho} \in R$, $\tilde{m} \in M$ do
    $L \leftarrow \{\}$
    for $i = 0$ to $\mu$ do
        $\tilde{G} \leftarrow \mathcal{M}_{\tilde{\rho}}^{\tilde{m}}(|G|)$                     // generate a model
        $\tilde{p} \leftarrow$ 2dSweep($\tilde{G}$)      // centre node, cf. algorithm B.4
        Append $\sigma_3(\tilde{G}, \tilde{p})$ to $L$
    $V \leftarrow$ ArithmeticMean($L$)
    Append $(\tilde{\rho}, \tilde{m}, V)$ to $F$
$f \leftarrow$ Fit($F$)      // fit by $f(\rho, m) = \sum_{i \in \{0,1\},\, j \in \{0,1,2\}} \alpha_{ij} \rho^i m^j$ (least square)
$p \leftarrow$ 2dSweep($G$)             // centre node, cf. algorithm B.4
$\kappa \leftarrow$ EstimateKM$'_{\text{density}}$($G$, $\eta$, $\tau$)                   // algorithm 5.5
$V_G \leftarrow \sigma^3(G, p)$
$m \leftarrow$ NewtonMethod($x \mapsto f(\kappa^{1/x}, x)$, $V_G$)        // solve $f(\kappa^{1/m}, m) = V_G$
return $(\kappa^{1/m}, m)$
```

$\mathcal{M}_{\tilde{\rho}}^{\tilde{m}}(|G|)$ with several values of $\tilde{\rho}$ and $\tilde{m}$. The computed values are fitted, in a second step, by $f(\rho, m)$. The fit describes the approximate value of σ_3 of uniform SISG models with density parameters and dimensions in the relevant range. The uniform SISG model $\mathcal{M}_\rho^m(|G|)$ similar to the graph G is hence expected to satisfy

$$f(\rho, m) \approx \sigma_3(G). \tag{5.8}$$

In a third step, an estimate of ρ^m, denoted by κ, is computed by algorithm 5.5. When an estimate of the dimension m is known, the estimate κ also defines an estimate of the density parameter ρ, namely $\kappa^{1/m}$. The estimate of the dimension is, due to equation 5.8, expected to satisfy

$$f(\kappa^{1/m}, m) = \sigma_3(G). \tag{5.9}$$

An estimate of m can be determined by solving the equation, e. g. by Newton's method (Forsythe et al. 1977, pp. 157ff). When an estimate for m is determined, the value $\kappa^{1/m}$ can serve as an estimate of ρ.

We introduced algorithms for estimating ρ^m, ρ and m for a given graph G which is similar to a uniform SISG model. It remains to evaluate these algorithms by checking whether they result reasonable estimates for uniform SISG models and for other graphs.

5.3 Evaluation for Uniform SISG Models

Several algorithms to estimate ρ^m, as well as ρ and m independently, for a given graph G which is similar to a uniform SISG model $\mathcal{M}^m_\rho(|G|)$, were introduced. We argued that the algorithms return the correct expectation value for uniform SISG models $G = \mathcal{M}^m_\rho(|G|)$ for $|G| \to \infty$, but we cannot prove the algorithms to be correct for finite graphs.

We evaluate, in this section, the algorithms by testing whether they return reasonable estimates for finite graphs, and we examine the estimates' variance.

Algorithms 5.1 and 5.3 estimate ρ^m for a given graph model G that is similar to a uniform SISG model. Improved versions of these algorithms have been introduced, namely algorithms 5.2 and 5.5. The improved versions yield better estimates compared to the non-improved versions, as can be seen in figures 5.4 and 5.5.

A comparison of algorithms 5.2 and 5.5 can be found in figure 5.9. The figure shows that the results of both algorithms are approximately equal for uniform SISG models, and that the variance is higher for smaller generating sets.

The results of the reconstruction of the parameters by algorithm 5.6 is depicted in figure 5.8. The figure shows that the estimates with the same $\rho^m = \kappa$ are located on the graph of the function $m(\rho) = \log \kappa / \log \rho$, which is expected because the reconstruction of ρ^m has been demonstrated to be reasonable well (cf. figure 5.9). The independent estimates of ρ and m are worse: a higher[6] variance of σ^3 for different SISG models causes a higher variance of the estimates of the density parameter and the minimal dimension. The estimates for SISG models with the same ρ^m but different $m \in \mathbb{N}$ are however located in mostly disjunct regions. This fact proves that the reconstruction of the parameters is, in principle, possible. This answers question 5.3.

[6] The variance of σ^3 is higher than the one of the density but still small compared to the variance of other properties.

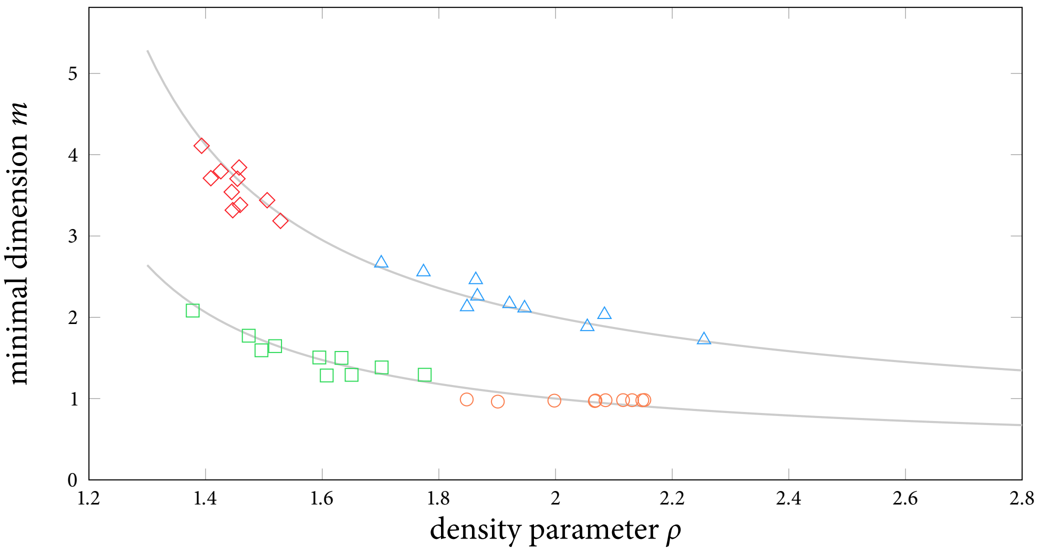


Figure 5.8
Estimation of the density parameter and the minimal dimension for uniform SISG models by algorithm 5.6

○ $\mathcal{M}^1_2(200)$
□ $\mathcal{M}^2_{\sqrt{2}}(200)$
△ $\mathcal{M}^2_2(200)$
◇ $\mathcal{M}^3_{\sqrt[3]{4}}(200)$

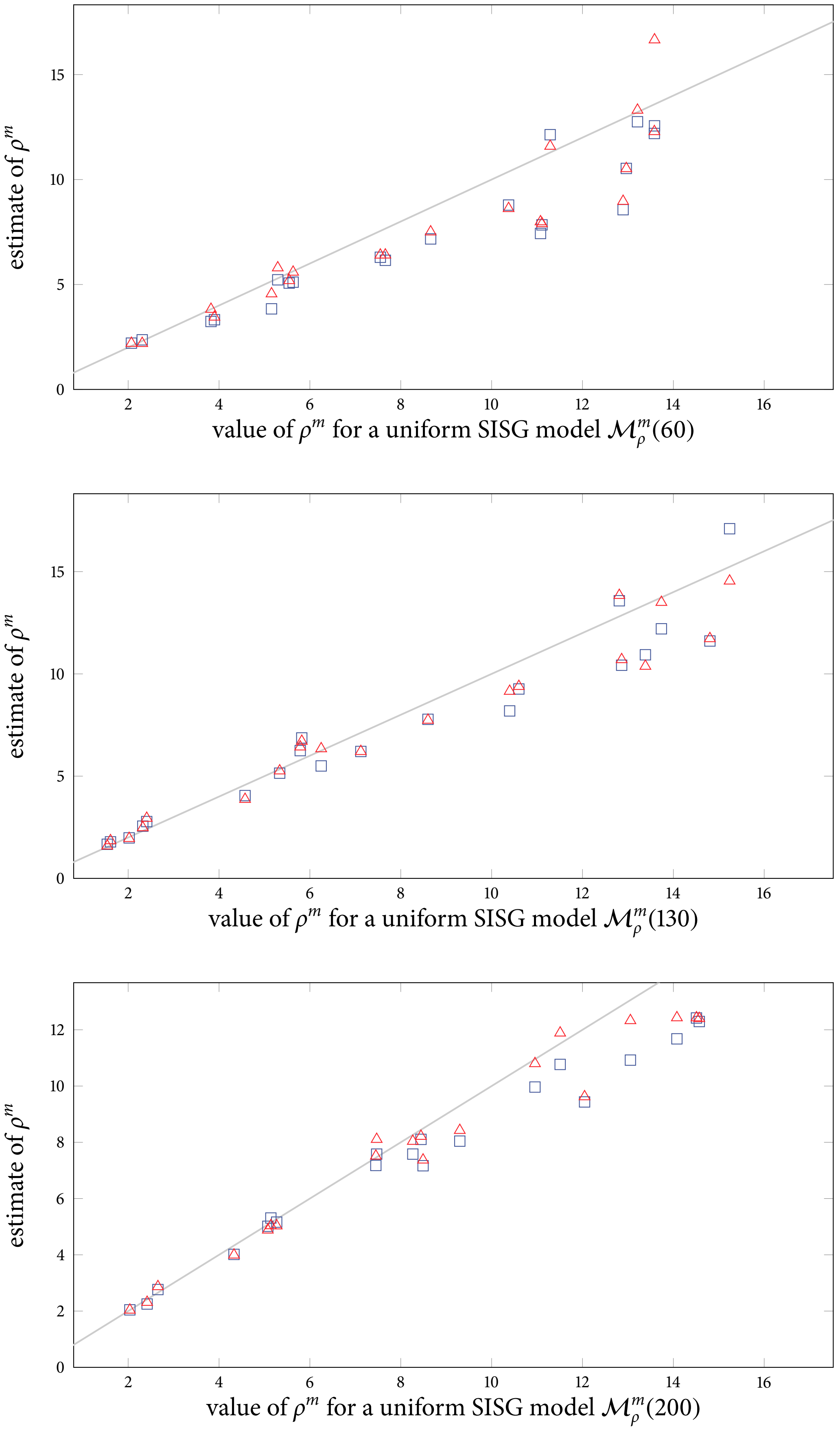


Figure 5.9
Estimation of the density parameter and the minimal dimension for uniform SISG models

□ improved (algorithm 5.5)

△ improved (algorithm 5.2)

We evaluated algorithms 5.1 to 5.6 on uniform SISG models. The algorithms have been shown to return meaningful estimates for ρ^m, ρ and m. The estimates of ρ^m have been shown to have much smaller variance than the one of ρ and m. It remains to evaluate the algorithms on real data sets.

5.4 Evaluation on Real Data Sets

We discussed in section 5.1 the question of how to check whether a data set is spatial. The algorithms that were developed in section 5.2 can be used as a basis of decision making to approach this question: based on analytical results, for the number of nodes approaching infinity, estimates of ρ^m can be computed. When a data set has a spatial structure, i. e. when it is similar to a uniform SISG model, we expect both estimates to coincide. Since data sets are finite, improved versions for finite graphs, which are expected to serve for the same purpose, were introduced. The estimates of the density parameter ρ and the minimal dimension m are expected to be additional indicators of whether the data set has a spatial structure. The estimates of ρ, m, and ρ^m can slightly differ for each computation step, because they depend on the random choice of the generating sets and on the random choice of subgraphs.

We approach, in this section, the question of whether the proposed algorithms yield reasonable results for real data sets. We evaluate, in particular, whether the improved algorithms yield reasonable results for small data sets[7]; whether the results are reasonable in spite of the influence of the random choices during the computation of the estimates; and whether spatial information has a spatial structure according to the conclusions that can be drawn by the algorithms. We can, yet, only evaluate these questions for a restricted number of data sets; a statistical evaluation on a much larger number of data sets would be required for conclusions that apply to other data sets than the examined ones.

Data Sets. The evaluation is conducted for SISG models as well as for real data sets with spatial and non-spatial structures:

SISG Models. We evaluate the hypotheses on SISG models with several parameters.

Graph Representations of Public Transport. We use several public transport networks in Sweden as examples of public transport (cf. sections 2.1.3 and 2.3.1). The data includes different modes of transport for different regions, e. g. nation-wide transport by the Swedish national railway provider (SJ) as well as by bus (Swebus); region-wide transport by bus, train and boat (Länstrafiken Sörmland, Östgötatrafiken, Blekingetrafiken, Hallandstrafiken, Värmlandstrafiken, Västmanlands Lokaltrafik, Dalatrafik); and local transport by bus in a town (Karlstadsbuss, Luleå Lokaltrafik, Stadsbussarna Östersund).

Power Grid in the USA. The graph representation of the high-voltage power grid in the Western States of the USA represents transmission lines by undirec-

[7] The deviation from the analytical results is expected to be larger for smaller data sets than for more extensive ones (cf. section 5.3).

ted edges and the places where transmission lines start, end or meet, e. g. transformers, substations and generators, by nodes (Watts et al. 1998).

Network of Airports in the USA. The graph representation of a network of airports in the USA represents airports by nodes, and an edge exists between two nodes if a flight was scheduled between the corresponding airports in 2002 (Colizza et al. 2007).

Water Distribution Networks. These networks consist of one or more sources and a number of sinks. The network has a flow direction, because the network aims to distribute water. Pipes are thus represented by directed edges. In addition to the directed graphs, the associated 'undirected' graphs are examined: algorithms 5.3 and 5.5 are applied to the directed graphs that are gained from the associated undirected graphs by adding directed edges (p, q) and (q, p) for each undirected edge (p, q). Walski et al. (1987) introduced the hypothetical water distribution network of Anytown, which has been used as a prototypical example in many studies. Another example is the water distribution network of the Wolf-Cordera Ranch, which distributes water to about 370,000 persons (Lippai 2005).

Graph Representation of the Recipe of Pizza Napoletana. Graph representations of recipes have been discussed in section 2.3.2.

Graph Representations of Games. For a given number of moves or rotations and a given size of the board or cube, we can represent a game by a graph (cf. section 2.3.3).

A Peer-To-Peer Gnutella Network. This network is a computer network whose nodes are located in space (Ripeanu et al. 2002)[8].

Metabolic Networks. Chemical transformations inside the cells of a living organism are, amongst others, important for the growth and the reproduction of cells. These transformations can be represented by edges and the corresponding states of the cells, by nodes. Such metabolic networks will be examined for Archaeoglobus fulgidus (single celled microorganism; in the domain of archaea), Caenorhabditis elegans (roundworm; in the domain of eukaryotes) and Escherichia coli (occurs in the human large intestine; in the domain of bacteria) (Jeong et al. 2000)[9].

Graph of Wikipedia Votes. The Wikipedia community votes on the promotion to administratorships. The corresponding graph represents users by nodes and votes, by edges (Leskovec et al. 2010$^{a,\,b}$)[10].

Directed Gilbert Model. The directed Gilbert model $\mathcal{G}_{\text{Gilbert, directed}}(n, p)$ is a simple directed graph consisting of n nodes and an edge between two nodes with probability p (cf. section 3.3).

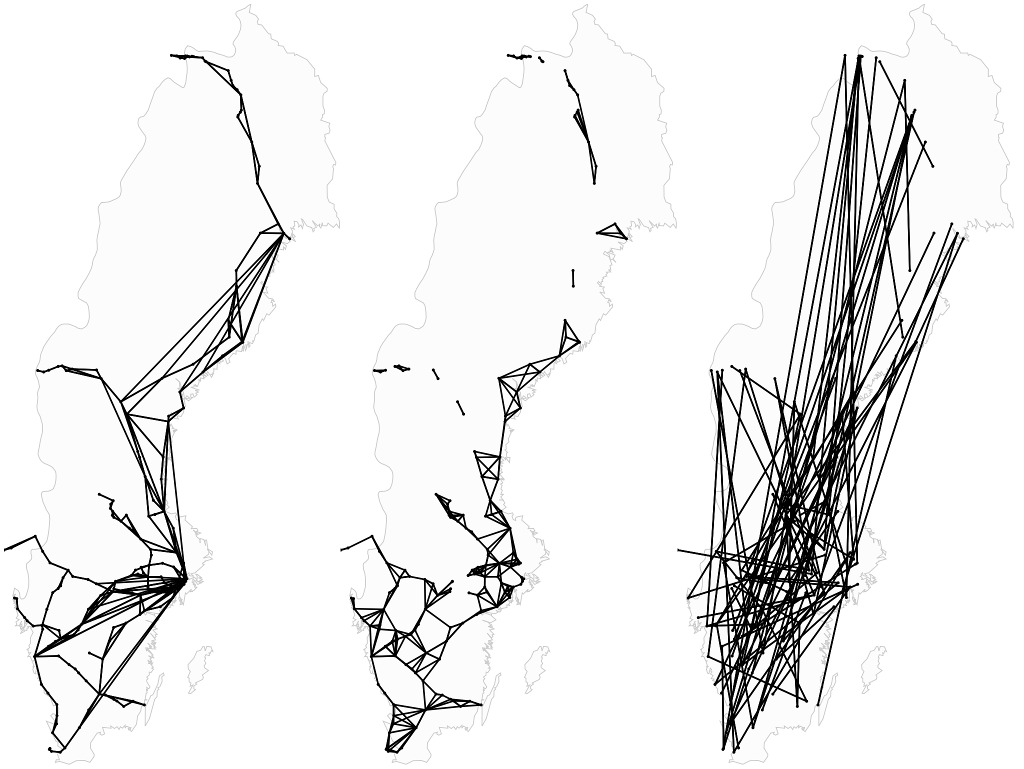
Complete Graph. A directed complete graph of size n consists of n nodes and directed edges between any pair of nodes.

[8] http://snap.stanford.edu, accessed at 2015-02-09

[9] http://www3.nd.edu /~networks/resources /cellular, accessed at 2015-02-11

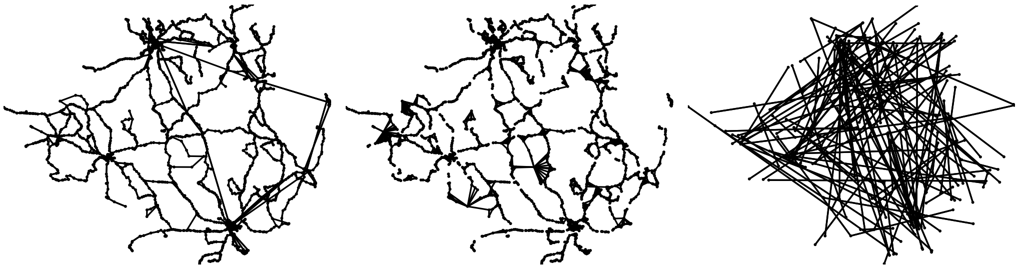
[10] http://snap.stanford.edu, accessed at 2015-02-09

A Visual Comparison. The graph representations of public transport are expected to have a spatial structure which plays a decisive role. A SISG model and a Gilbert model can be build with the same nodes as the ones in a graph representation. The visual comparison of the graph representation of the data set, the SISG model and the Gilbert model shows that the SISG model is much more similar to the graph representation than the Gilbert model is (cf. figures 5.10 and 5.11). The visual similarity suggests that SISG models are, as expected, not very different from these examples of spatial data.



(a) graph representation **(b)** $\mathcal{M}_{1.9}(S)$ **(c)** $\mathcal{G}_{\text{Gilbert}}(S, 6 \cdot 10^{-3})$

Figure 5.10
Graphs whose nodes S are the stops of the data set *SJ* (cf. table 5.1 on page 110); (a) graph representation of the data set, (b) SISG model, and (c) a Gilbert model; the parameters are chosen such that the similarities and dissimilarities visually stand out



(a) graph representation **(b)** $\mathcal{M}_{1.9}(S)$ **(c)** $\mathcal{G}_{\text{Gilbert}}(S, 6 \cdot 10^{-5})$

Figure 5.11
Graphs whose nodes S are the stops of the data set *Länstrafiken Sörmland* (cf. table 5.1 on page 110); (a) graph representation of the data set, (b) SISG model, and (c) a Gilbert model; the parameters are chosen such that the similarities and dissimilarities visually stand out

A Computational Comparison. Some algorithms that were discussed in section 5.2 are nondeterministic. They include uniform SISG models whose generating set is randomly generated according to a given distribution. These algorithms return different results each time they are executed, but the results have only relatively low variance. This effect has been been evaluated for uniform SISG models in section 5.3, and the variance is hence not evaluated in this section. The results of the algorithms are yet affected by the effect of the variance.

Table 5.1 contains the estimates for the data sets. Estimates of ρ^m are included whenever the algorithms converged. Estimates of ρ and m are only included if the comparison of the estimates of ρ^m approximately coincide[11], i. e. if the estimations suggest that the data set may be similar to a SISG model.

A comparison of the estimates of ρ^m by algorithms 5.1 and 5.3 is depicted in figure 5.12 and a comparison of the estimates by the improved algorithms 5.2 and 5.5, in figure 5.13. The positions of the data sets in the diagrams are very similar, apart from the data set about the peer-to-peer network. The similarity demonstrates that the influence of the variance is relatively low for the estimation of ρ^m. The estimates of the density parameter ρ and the minimal dimension m are depicted in figure 5.14.

The density for the series of subgraphs can be fitted with relatively low residual for all considered graphs. The low residual demonstrates that the use of series of subgraphs can yield meaningful results.

The estimates of ρ^m are approximately equal for the considered *SISG models* due to the spatial structure of the models. The density parameters and minimal dimensions of the SISG models could approximatively be reconstructed, which was expected due to the results of section 5.3.

For the *graph representation of public transport* and the *power grid in the USA*, the estimates of ρ^m approximately coincide, as was expected due to their spatial structure. We expect the minimal dimension to be between 2 and 3, because the spatial data sets are embedded in a two- or three-dimensional space, and as time is relevant for public transport, the minimal dimension may even be between 3 and 4. The estimates of the minimal dimensions for the graph representations of public transport are between 2.33 and 4.40 and thus within a reasonable range[12]. The estimate of the minimal dimension for the power grid in the USA is 4.24 and thus higher than expected but still within a reasonable range.

The estimates of ρ^m for the *water distribution networks* using the non-improved algorithms approximately coincide, which was expected because the network has a temporal structure. The improved algorithm to estimate ρ^m by the number of nodes and edges does not converge. The reason for the divergence in case of the water distribution network of Anytown is the very low number of nodes and edges that leads to a high variance of the number of edges of the random SISG models generated during the computation, according to algorithm 5.2. In case of the water distribution network of the Wolf-Cordera Ranch, the divergence is due to the very low ratio of the edges to the nodes. The algorithm converges for the associated simple 'undirected' graphs that contains, for each edge (p, q), also the directed edge (q, p), because the number of edges is about twice as high as in the directed graph. The estimates of the minimal dimension are within a reasonable range.

[11] Formally, we include estimates of ρ and m if the estimates of ρ^m differ less than a factor of 1/2. The choice of the factor is arbitrary and has no relevance to the fact that the question of how similar the data sets are to the diagonal is a gradual one. In the corresponding figures, we even display the area corresponding to this choice to visually illustrate how near data sets are depicted to the diagonal.

[12] We say that an estimate is *within a reasonable range* or *reasonable*, if the difference between the estimate and the expected value can, in principle, be reasoned. In this case, the variance of the results due to random effects and the presence of other structures in the data set are possible reasons.

Table 5.1
Estimations of the density parameter and the minimal dimension for several data sets

$|N|$ number of nodes

$|E|$ number of edges

$\widehat{\rho m}^{N}$ estimate by the number of nodes and edges (algorithm 5.1)

$\widehat{\rho m}^{N'}$ estimate by the number of nodes and edges (algorithm 5.2 with $\mu = 10$, $\eta = 1000$ and $\tau = 0.003$, using only the first 50 densities in the series of subgraphs for the fitting)

$\widehat{\rho m}^{D}$ estimate by density (algorithm 5.3 with $\mu = 10$, using only the first 50 densities in the series of subgraphs for fitting)

$\widehat{\rho m}^{D'}$ estimate by density (algorithm 5.5 with $\mu = 10$, $\eta = 1000$ and $\tau = 0.003$, using only the first 50 densities in the series of subgraphs for fitting)

χ^2 residuals for $\widehat{\rho m}^{D}$

$\hat{m}$ and $\hat{\rho}$ estimates (algorithm 5.6 with $\mu = 10$, $\eta = 100$, $\tau = 0.003$, $R = [1.1, 1.2, 1.4, 1.6, 1.8, 2]$ and $M = [1, 2, 3]$, using only the first 50 densities in the series of subgraphs for fitting)

graphs marked by * are timetable-packed

graphs marked by † are the associated undirected graphs

the following types of data sets are examined: ○ SISG models, □ transport networks, ◇ other spatial graphs, ☆ recipes, △ games, ▽ other data sets, ⬠ existing graphs models

	Graph	$\|N\|$	$\|E\|$	$\widehat{\rho m}^{N}$	$\widehat{\rho m}^{N'}$	$\widehat{\rho m}^{D}$	$\widehat{\rho m}^{D'}$	$\widehat{\rho m}^{D'}/\widehat{\rho m}^{N'}$	χ^2	$\hat{m}$	$\hat{\rho}$	References
⊗	$\mathcal{M}_2^2(1000)$	1 000	3 914	3.91	4.39	4.04	3.85	0.88	$6.51{\cdot}10^{-4}$	1.96	2.12	section 3.4
⊕	$\mathcal{M}_{2,3}^2(1000)$	1 000	5 158	5.16	5.38	5.89	5.43	1.01	$3.72{\cdot}10^{-3}$	2.25	2.11	section 3.4
⊙	$\mathcal{M}_{2,5}^2(1000)$	1 000	6 278	6.28	6.77	7.41	6.83	1.01	$7.46{\cdot}10^{-3}$	2.01	2.59	section 3.4
□	SJ* (Swedish national railway provider)	128	420	3.28	3.46	3.84	3.58	1.04	$1.07{\cdot}10^{-4}$	3.68	1.4	section 2.3.1, Trafiklab 2013
	Länstrafiken Sörmland*	640	1 593	2.49	2.58	2.79	2.60	1.01	$1.15{\cdot}10^{-3}$	3.51	1.31	section 2.3.1, Trafiklab 2013
	Östgötatrafiken*	1 039	2 773	2.67	2.73	2.91	2.58	0.94	$5.19{\cdot}10^{-4}$	3.76	1.31	section 2.3.1, Trafiklab 2013
	Blekingetrafiken*	420	1 096	2.61	2.63	2.73	2.36	0.90	$3.81{\cdot}10^{-3}$	3.88	1.28	section 2.3.1, Trafiklab 2013
⊠	Hallandstrafiken*	684	1 740	2.54	2.57	2.70	2.67	1.04	$1.84{\cdot}10^{-3}$	2.33	1.5	section 2.3.1, Trafiklab 2013
	Värmlandstrafiken*	590	1 569	2.66	2.76	2.74	2.42	0.88	$3.28{\cdot}10^{-4}$	3.42	1.35	section 2.3.1, Trafiklab 2013
	Västmanlands Lokaltrafik*	599	1 487	2.48	2.54	3.00	2.70	1.06	$2.70{\cdot}10^{-3}$	3.23	1.33	section 2.3.1, Trafiklab 2013
	Dalatrafik*	1 239	3 262	2.63	2.75	3.29	2.98	1.08	$7.46{\cdot}10^{-4}$	3.12	1.38	section 2.3.1, Trafiklab 2013
⊞	Karlstadsbuss*	96	222	2.31	2.36	2.68	2.47	1.04	$5.43{\cdot}10^{-4}$	4.05	1.24	section 2.3.1, Trafiklab 2013
⊟	Luleå Lokaltrafik*	126	308	2.44	2.58	2.77	2.74	1.06	$4.40{\cdot}10^{-3}$	3.70	1.29	section 2.3.1, Trafiklab 2013
◫	Stadsbussarna Östersund*	140	344	2.46	2.53	2.89	2.53	1.00	$6.63{\cdot}10^{-4}$	3.19	1.34	section 2.3.1, Trafiklab 2013
⊡	Swebus*	67	236	3.52	3.82	4.51	4.36	1.14	$1.13{\cdot}10^{-2}$	4.40	1.36	section 2.3.1, Trafiklab 2013
◇	Power grid in the USA	4 941	6 594	2.67	2.97	3.48	3.49	1.17	$7.82{\cdot}10^{-4}$	4.24	1.29	Watts et al. 1998
◈	Network of airports in the USA	500	2 980	11.92	12.77	41.22	60.57	4.74	$7.48{\cdot}10^{-4}$			Colizza et al. 2007
◇	Water distribution network Anytown	24	43	1.79		1.92	1.78		$1.09{\cdot}10^{-2}$			Walski et al. 1987
◇	Water distribution network Anytown†	24	43	3.58	4.33	4.23	4.61	1.07	$7.44{\cdot}10^{-3}$	2.14	1.98	Walski et al. 1987
◇	Water distribution network Wolf-Cordera Ranch	1 785	1 983	1.11		1.51	1.54		$1.34{\cdot}10^{-3}$			Lippai 2005
◇	Water distribution network Wolf-Cordera Ranch†	1 785	1 982	2.22	2.19	2.49	2.26	1.03	$3.12{\cdot}10^{-3}$	2.98	1.3	Lippai 2005
☆	Pizza Napoletana	2 291	3 687	1.61	1.75	2.46	2.16	1.24	$5.64{\cdot}10^{-3}$	4.59	1.13	section 2.3.2, European Comission 2010
△	Tic-tac-toe (2x2 board)	30	44	1.47	1.46	1.55	1.57	1.08	$4.33{\cdot}10^{-3}$	7.95	1.05	section 2.3.3
△	Tic-tac-toe (3 moves, 3x3 board)	3 890	74 169	19.07	23.86	25.63	43.40	1.82	$3.03{\cdot}10^{-3}$			section 2.3.3
△	Rubik's Cube (3 rotations, 2x2x2 size)	1 417	1 644	1.16	1.19	5.80	5.12	4.32	$3.70{\cdot}10^{-2}$			section 2.3.3
△	Rubik's Cube (3 rotations, 3x3x3 size)	4 602	5 364	1.17	1.39	7.13	7.18	5.17	$4.47{\cdot}10^{-2}$			section 2.3.3
▽	p2p Gnutella network 09	8 114	26 013	3.21	2.20	4.93	4.89	2.23	$7.96{\cdot}10^{-3}$			Ripeanu et al. 2002
▽	Met. network of Archaeoglobus fulgidus	1 567	3 631	2.32	2.33	8.32	10.16	4.36	$1.42{\cdot}10^{-2}$			Jeong et al. 2000
▽	Met. network of Caenorhabditis elegans	1 469	3 447	2.35	2.33	12.39	14.15	6.08	$1.45{\cdot}10^{-2}$			Jeong et al. 2000
▽	Met. network of Escherichia coli	2 897	7 104	2.45	2.43	15.51	18.21	7.48	$7.20{\cdot}10^{-2}$			Jeong et al. 2000
▽	Graph of Wikipedia votes	7 115	103 689	14.57	25.59	57.18	82.34	3.22	$4.62{\cdot}10^{-4}$			Leskovec et al. 2010[a, b]
⬠	$\mathcal{G}_{\text{Gilbert, directed}}(150, 0.5)$	150	11 170	74.47	253.78	74.91	249.05	0.98	$1.21{\cdot}10^{-6}$			section 3.3
⬠	Complete graph (100 nodes)	100	9 900	99.00	$1.04 \cdot 10^6$	99.00	215 786.70	0.21	$1.37{\cdot}10^{-29}$			

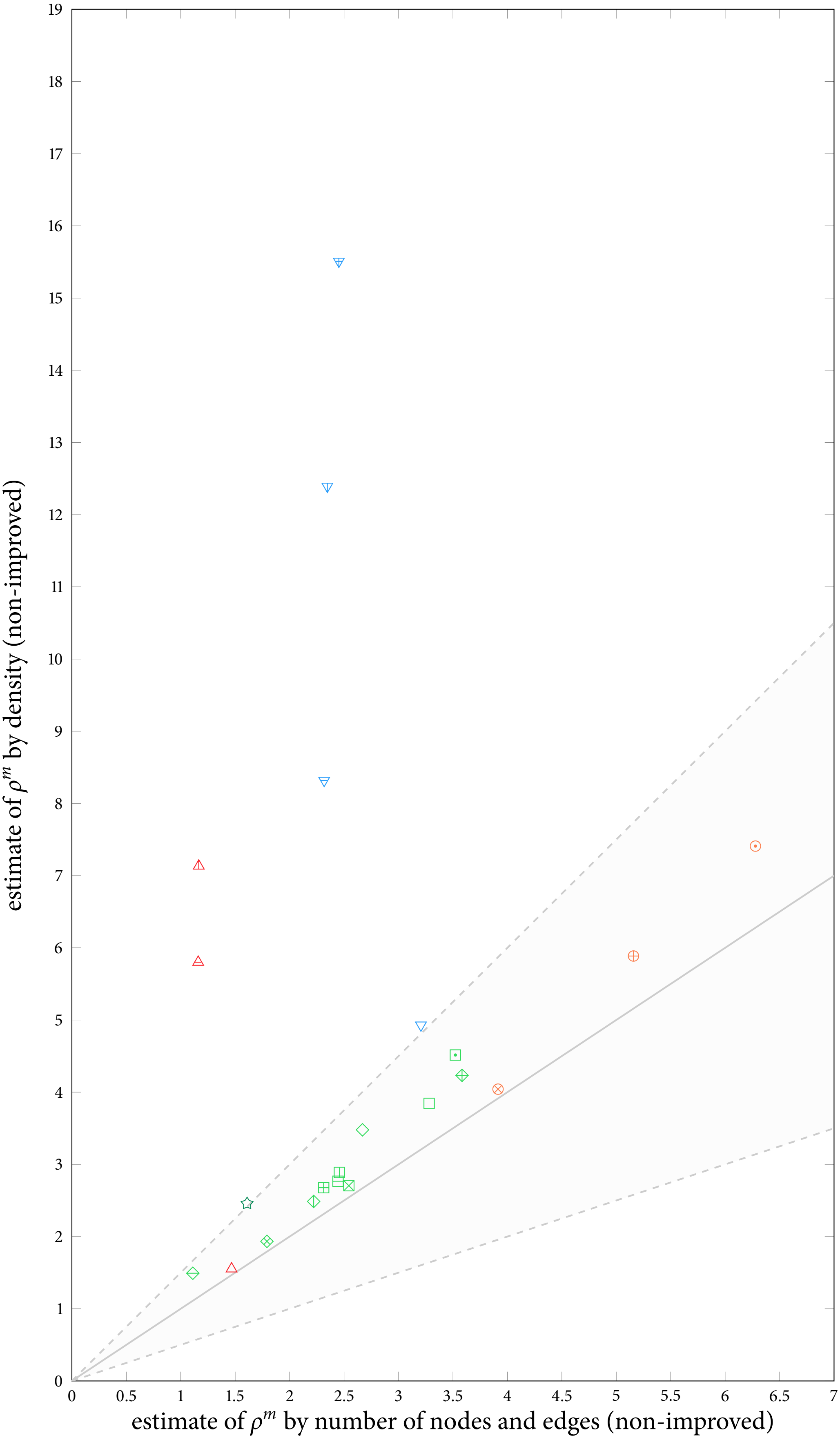


Figure 5.12
Estimation of ρ^m for several data sets (cf. table 5.1);
if a data set has typical properties of spatial information, both estimates coincide;
for the grey area, the estimates differ less than a factor of 1/2;
graphs marked by * are timetable-packed;
graphs marked by † are the associated undirected graphs

The *network of airports in the USA* can be embedded in space by the natural location of the airports, but both estimates are, nevertheless, very different. This effect is caused by the large number of non-spatial aspects which are influencing the network: the importance of a low average number of connections separating two airports, cultural aspects leading to more connections, legal restrictions (night flight restrictions, ban on unsafe airlines, taxes), etc. These aspects cause a number of structural properties that SISG models do not have, because these properties are not typical for spatial data in general: a non-uniform distribution of the airports in space, a large number of long-distance connections, a large number of hubs, a strong hierarchical organization (domestic and intercontinental), communities of strongly related airports, etc. (Barthélemy 2003).

The *recipe of Pizza Napoletana* has a temporal structure due to the partial temporal order[13] of the states. The estimates of ρ^m, accordingly, approximately coincide for the recipe. They are much lower than the ones of the spatial data sets, which was expected, because time is one-dimensional and thus of lower dimension than space. The estimate of the minimal dimension is 7.95, which seems not to be within a reasonable range. A possible reason is the lower precision of the estimation of the minimal dimension, compared with the one of ρ^m, as was discussed in section 5.3.

The *game Tic-tac-toe* is expected to have a temporal structure, because there exists a partial order in the state space: a mark is added in each step, and there exists no possibility to go back to a state with less marks. There are, however, many more aspects which are influencing the structure of the graph representation, and they become more important for a larger board. The two estimations of ρ^m accordingly suggest such a temporal structure in case of a $2x2$ board. For a larger board, however, the temporal order is outbalanced by the fact that many states can occur in two or more different courses of the game. The estimates of ρ^m differ, hence, by a factor of 1.34 for the non-improved and a factor of 1.82 for the improved algorithms. The estimation of the minimal dimension is not within a reasonable range.

The *Rubik's Cube* has a temporal aspect, too. The interrelations between the cube's states are, however, determined by the arrangement of the small cubes and the ways they can be rotated[14]. The temporal structure of the representation is outbalanced by other aspects, because our representation reflects only the colours showing up on each face of the cube and not the point in time when the state is gained. The estimates of ρ^m accordingly differ.

The *peer-to-peer network* has spatial aspects, because the computers are placed in space, but other aspects may be much more important. The estimates of ρ^m accordingly do not coincide but are not very different either. The *metabolic networks* and the *graph of Wikipedia votes* do not have any decisive spatial or temporal structure. The estimates of ρ^m accordingly differ.

The estimates of ρ^m for the *directed Gilbert model* (by the non-improved and improved algorithms) and the *complete graph* (by the non-improved algorithm)

[13] The fact that the transition between most states are irreversible induces an ordering relation on the set of states. We can, for example, mix flour and water, but can hardly separate them.

[14] The rotations can be proven to have the structure of a group with six generators, which is much more complex than the structure of time.

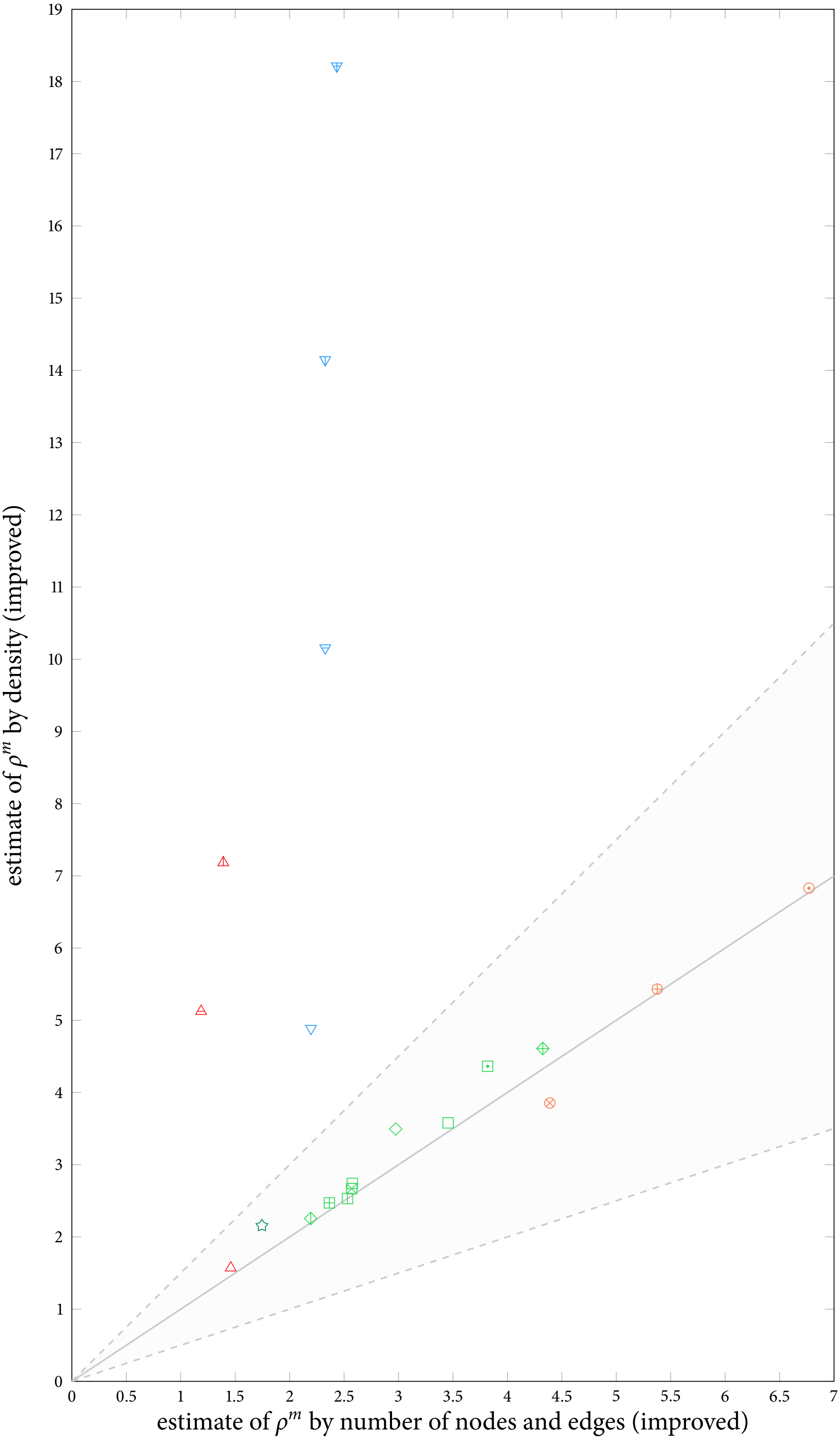
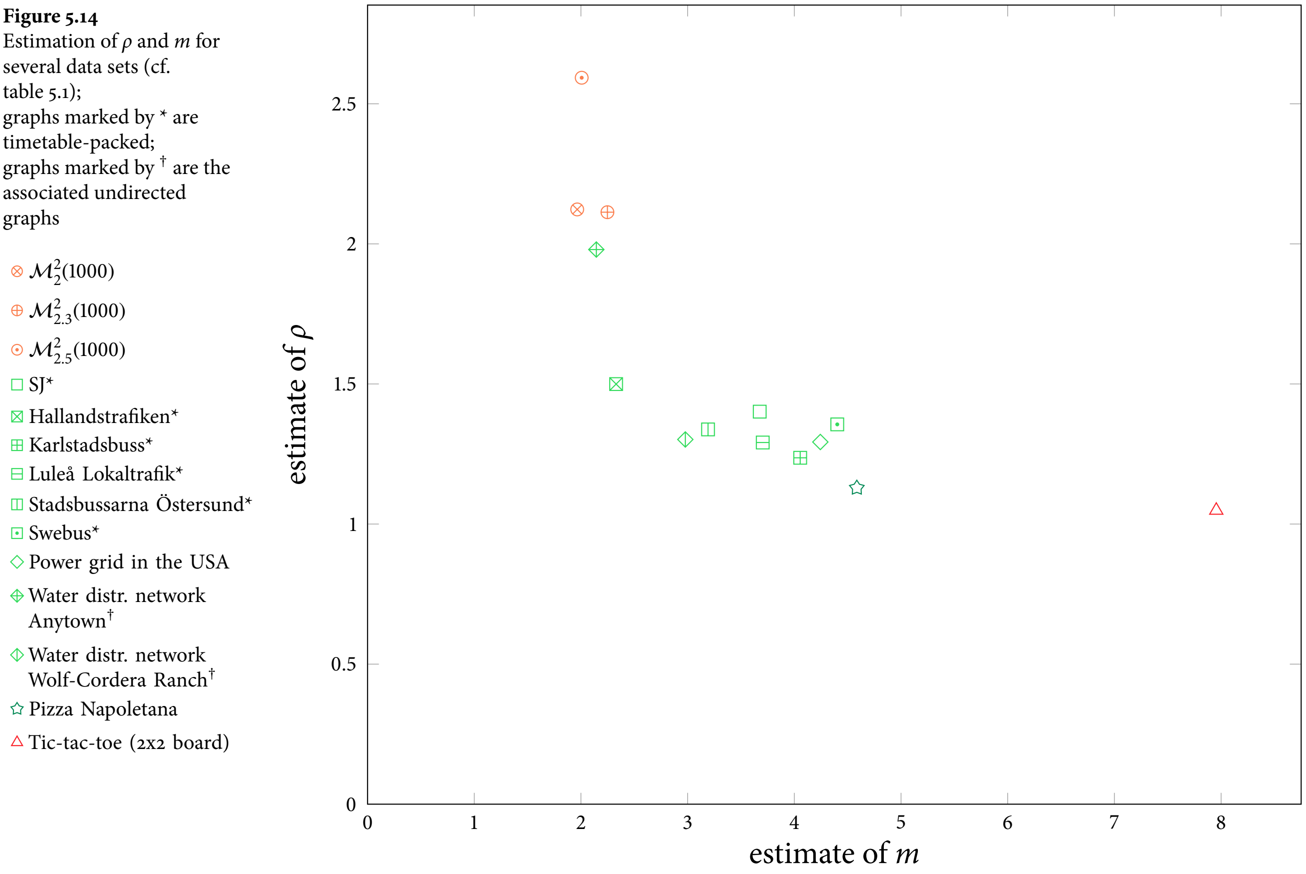


Figure 5.13
Estimation of ρ^m for several data sets (cf. table 5.1);
if a data set has typical properties of spatial information, both estimates coincide;
for the grey area, the estimates differ less than a factor of 1/2;
graphs marked by * are timetable-packed;
graphs marked by † are the associated undirected graphs

approximately coincide but are much higher than expected for spatial data. As can be seen in figure 5.15, the estimates by the non-improved and the improved algorithms are, in case of the complete graph, very different, by an order of magnitude 4 to 5, which indicates that it has no spatial structure. The complete graph is in fact a SISG model with very high density parameter and minimal dimension 1, and the difference between the estimates by the non-improved and the improved algorithms is due to the fact that all parts of the complete graph are outer regions[15] (cf. section 4.2). The complete graph has, in spite of being a SISG model, no spatial structure, because the absolute values of the estimates of ρ^m are meaningless in the context of spatial information.

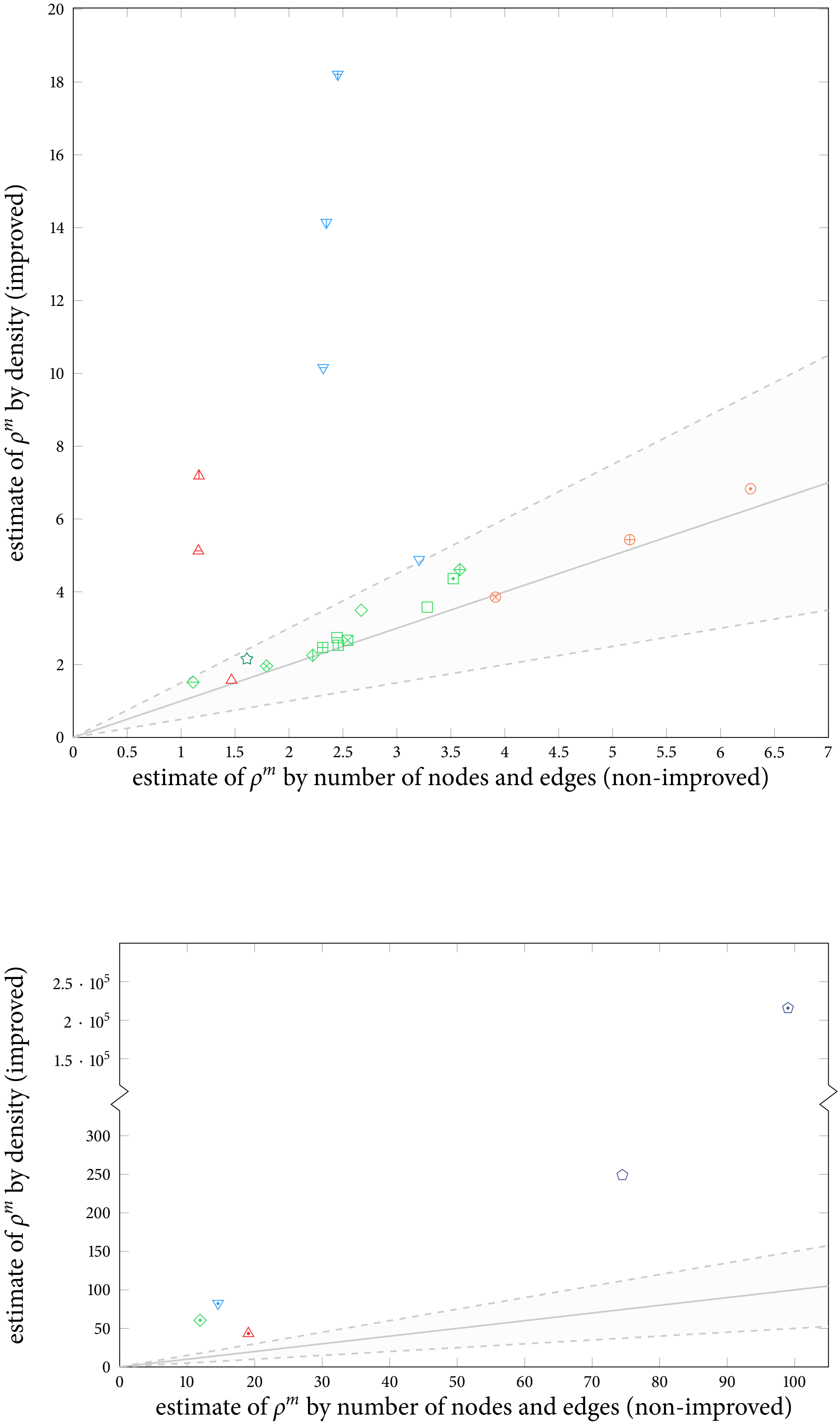
The evaluation shows that the examined data sets can be characterized by the use of the algorithms. The estimates of ρ^m coincide, in particular, for the examined spatial and temporal data sets, and they differ for the examined non-spatial and non-temporal data sets. The difference between the estimates turned out to be much smaller for spatial data than was expected, considering that data sets are usually characterized by more than a spatial structure. A detailed comparison of the estimates is able to characterize each of the examined data sets by its spatial and temporal structure, showing the characterization of spatial information (cf. section 3.2), the concept of spatial structure (cf. section 1.1) and the SISG model (cf. section 3) to be meaningful in the context of the examined data sets.

Figure 5.14
Estimation of ρ and m for several data sets (cf. table 5.1); graphs marked by * are timetable-packed; graphs marked by † are the associated undirected graphs



[15] When the complete graph is enlarged by an additional node, the density of each subgraph changes.

Figure 5.15
Estimation of ρ^m for several data sets (cf. table 5.1);
if a data set has typical properties of spatial information, both estimates coincide;
for the grey area, the estimates differ less than a factor of 1/2;
graphs marked by * are timetable-packed;
graphs marked by † are the associated undirected graphs



The results of the evaluation seem to be meaningful in spite of the random effects and the small size of the graphs. Only for a very small data set, the water distribution network of Anytown, one of the algorithms did not converge. The absolute size of the estimates of the density parameter and the minimal dimension seem not to be meaningful. The high variance of these estimates has already been discussed in section 5.3 and may be one reason for the unexpected absolute sizes of these estimates. Data sets from the same domain however have similar estimates of the density parameter and the minimal dimension, which demonstrates that the results are not random.

The evaluation of the algorithms on real data sets showed that the results match, by and large, our expectations. We were able to characterize data sets in respect to their spatial structure, and similar data sets were characterized similarly. The dimension of space could, in principle, be estimated, but the precision was not high enough to provide meaningful absolute values.

Conclusion

Spatial structure can be modelled by uniform SISG models. A data set can, thus, be tested for a spatial structure by comparing it to uniform SISG models. We argued how such a comparison can be executed by statistical methods, and provided algorithms for the comparison. Evaluations were executed on uniform SISG models and on real data sets. More specifically, the main contributions of this chapter are as follows:

(1) Theoretical foundations to discuss the spatial structure of data sets were laid.

(2) Algorithms were introduced to test data sets for spatial structures, and algorithms to estimate the parameters that generate uniform SISG models similar to a given data set.

(3) The provided algorithms were evaluated and empirically evidenced to be correct for uniform SISG models. The evaluation provides detailed information about the quality of the results.

(4) The concept of spatial structure was proven to be meaningful: the algorithms were able to detect a spatial structure in the data sets which describe spatial information; and similar data sets were characterized similarly in respect to their spatial structure.

(5) We showed that spatial structure reflects the dimension of space, because it is, in principle, possible to reconstruct the dimension of a uniform SISG model.

The results of this chapter demonstrate that the hypotheses of the introduction (cf. section 1.4) are valid. Besides this major result, a characterization of existing data sets was provided.

6

Conclusion

Fremtiden kommer af sig selv,
det gør fremskridtet ikke.

—**Paul Henningsen**
danish designer, architect and author
(1894–1967)

6.1 Summary and Hypotheses

Spatial and temporal aspects of information are central to numerous tasks, because most things exist and happen in space and time. Examples of such spatial tasks are navigation tasks, placement tasks for cell sites or branches of a company, all kinds of spatial analyses, personal planning tasks for activities in space, etc. Solutions to spatial tasks can be found, if we comprehend the role of space and time, and we thus strive for a systematic understanding of space and time.

Laws are a common tool to formalize such a systematic understanding. They aim at hiding context as much as possible and, ideally, at focusing on very few simple relations between very few things. The absence of context and the clear focus on single aspects is advantageous for two reasons: first, it is easy to check whether the preconditions of a law are met and the law applies. Second, we can better understand how laws interact and combine them to theories. Even when laws seem to be advantageous, only very few laws are currently known for geographical space. Tobler's first law of geography is one of the exceptions.

Formal structures are a key issue for laws, because structures capture patterns and configurations of things and describe them formally. A system can be described by different structures. The choice of a convenient structure is crucial for the formulation of laws, and laws assume a simple form in this case, in particular when contextual dependencies are, by and large, eliminated. Euclidean space, for example, is such a convenient structure for the physical concept of spacetime in Newtonian mechanics, and pseudo-Riemannian manifolds, in Einstein's general theory of relativity. The structure of spacetime had however not been discussed with the geographical concept of space in mind. We thus tried, in this thesis, to capture spatial and temporal aspects of information by suitable structures, with the hope to promote the formulation of laws and the development of theories in geographical information science.

Summary. Information about human activities is, in many cases, of spatial nature (chapter 2). Examples are timetable information for public transport; usage information collected by mobile phones, including the position; observation data about movements in space, e. g. walking, cycling, driving; etc. Human activities can strongly influence and constrain each other. We have thus introduced the notion of a human activity system, which denotes a set of related human activities. Such a system can be represented as a graph: states are represented by nodes, and activities, i. e. transitions between these states, by edges. Graph representations of human activity systems provide a basis for the formal discussion of the systems' structures, and structural aspects of such representations can be emphasized by equivalence relations on the node set.

Many spatial data sets share some typical properties, e. g. Tobler's law and scale invariance (chapter 3). These properties lead to a predominance of certain constellations of edges, when data sets are represented by graphs. These constellations give rise to the concept of a spatial structure, which is exposed by many spatial data sets. We introduced a scale-invariant spatial graph (SISG) model of the spatial structure: for a given set of nodes which are embedded in space, edges are introduced such that the resulting graph satisfies Tobler's law, is scale-invariant and has other typical properties of spatial information.

When no assumptions on the nodes' distribution in space are made, it is reasonable to examine uniform SISG models, which assume randomly distributed nodes with a uniform distribution (chapter 4). The examination of the uniform SISG model's properties showed that the number of nodes and edges, the density of the graph and the volume of spheres are statistically characteristic of such a model, and that these properties fundamentally depend on the dimension. The dimension of space and the density parameter of an abstract uniform SISG model can, in principle, be reconstructed, when the number of nodes and edges, the density and the volume of spheres are known for the SISG model and its subgraphs. This is unexpected, because a high dimension of space leads to a larger number of edges, and the dimension could thus be confused with the density of the graph.

The SISG model was evaluated by a comparison to real data sets, which proved the model to be meaningful (chapter 5). This comparison considered the SISG model as an abstract graph, but ignored the properties that it gains by its embedding in space. We showed that all considered spatial data sets are similar to a SISG model and thus expose a spatial structure, and that all considered non-spatial data sets are not. We were moreover able to distinguish spatial and temporal data by the dimension that is reflected by its structure. The results are, interpreted with the semantic knowledge about the data sets in mind, more than reasonable.

Hypotheses. We raised two hypotheses in the introduction: first, the hypothesis that the concept of spatial structure is meaningful; and second, the hypothesis that the spatial structure implicitly reflects the dimension of space. These hypotheses could, by and large, be argued to be valid.

(1) The concept of spatial structure is meaningful, because most spatial data sets share structural properties.

Data sets can be characterized by its structure. Many properties of a data set influence the structure, but some do not. In particular, the same structure can appear for very different data sets. We thus cannot infer from the structure which properties a data set has, e. g. whether the data set exposes references to space. A structure can however suggest the data set to have a certain property, because it statistically occurs more often than other structures for data sets which expose this certain property. The hypothesis claims the existence of a structure that suggests a data set to be spatial, i. e. to expose references to space.

We compared the structure of several spatial and non-spatial data sets in chapter 5. The structures of the considered spatial data sets were shown to be very similar to SISG models, and the structures of the considered non-spatial data sets to be not. This corroborates the hypothesis for the considered data sets.

(2) Spatial structure implicitly reflects the dimension of space.

The uniform SISG model depends on the dimension of the space in which the nodes are placed. We expect the spatial structure to depend on the dimension of space in a similar way. The volume of a sphere is larger for higher dimensions, and as Tobler's law assumes nodes in the neighbourhood to be statistically more often adjacent than others, a uniform SISG model of higher dimension is expected to have more more edges. The hypothesis claims, amongst others, that the effect of a higher dimension is different from the effect of a higher density parameter which influences the number of edges in the graph, and that different dimensions statistically lead to different spatial structures.

We proved that the dimension of space can statistically be reconstructed from an abstract graph representation. This shows that a higher dimension of space does not only cause a larger number of edges but has further effects on the structure of a uniform SISG model, which can be measured by the volume of spheres.

We have, in this thesis, successfully argued that the concept of spatial structure is meaningful and that the spatial structure implicitly reflects the dimension of space. In addition, the thesis sheds light on how to model spatial structure, on the properties of spatial structure and on why spatial structures have these properties. We will discuss in the next section, which questions stay open and point out how answers could be found.

6.2 Future Work

This thesis examines the structure of spatial information. Many issues arise when the formal concept of spatial structure meets real data: possible improvements and modifications of the SISG model become apparent; the role of the parameters in the generation of the SISG model may be explored in more detail; the SISG model may be compared to a larger number of data sets; additional aspects, including non-spatial ones, may be modelled and compared to spatial structure; and possible applications of the SISG model may be examined. We discuss, in the following, these open questions and ideas that were not explored yet.

Improving the SISG Model. Uniform and non-uniform SISG models are not necessarily connected, and the size of the largest connected component grows much slower than the number of nodes, if the model has relatively few edges as in the case of small density parameters and minimal dimensions. Several examples of spatial graph representations, e. g. graph representations of public transport, have however large connected components. It remains an open question whether the connectedness is a property of spatial structure, and in case it is, how to create connected models which are otherwise similar to SISG models.

An undirected SISG model was introduced, but neither were its properties analysed nor were the provided algorithms systematically evaluated for undirected models. The systematic treatment of undirected SISG models makes it possible to address a much larger family of graphs, because many data sets can only be represented by undirected graphs.

Some data sets can be represented by weighted graphs, which is e. g. important to describe the distance between places. The SISG model is not able to capture this aspect. A generalization of SISG models to weighted graphs is yet to be introduced.

Specializations. This thesis, in particular the SISG model and the comparison of real data with the model, aims at discussing spatial structure in general. It does not aim at discussing and characterizing certain types of spatial or temporal data, e. g. public transport or recipes. Models of specific types of spatial data are however needed for different applications. It remains to discuss how the SISG model can be modified, by paying attention to the characteristics of a certain type of data, in order to serve for such a purpose.

Graph representations of public transport, for example, exhibit properties that SISG models do not share: (1) every node usually has at least two edges, (2) the indegree equals the outdegree for most nodes, and (3) the graph has only few connected components, in most cases only one. These properties are not necessarily met by every graph representation of public transport but are actually valid for the data used in this thesis. Future research may discuss how the SISG model can be modified in order to exhibit these three properties.

Discussion of the Role of the Dimension. Data that exhibits the same properties and patterns at any scale is called fractal data. When such data is embedded in space, we may try to compute its fractal dimension by determining to which degree it fills the space. This method of determining the dimension returns, unlike in the case of geometric objects, not necessarily integer values. Two definitions of such a fractal dimension are provided by Song et al. (2005). It remains to discuss whether the minimal dimension of a SISG model and the fractal dimension are related.

The minimal dimension was, for the data sets of the evaluation, systematically higher than expected by the dimension of the space in which the data sets are embedded. The same holds, at least for some data sets, for the fractal dimension (Daqing et al. 2001). It remains to understand this discrepancies and whether both discrepancies share a common reason. An analysis of the density parameter's meaning and the parameter's relation to the minimal dimension may be convenient in this context.

More Extensive Evaluation. The SISG model has been evaluated by a comparison to several data sets. A high number of data sets from different domains, spatial as well as non-spatial ones, has to be considered in order to render the statistical argumentations meaningful. A more extensive evaluation may thus increase the significance of the results. Applications of the SISG model may support the results of the evaluation and provide further insights into the concept of spatial structures.

Comparison with Planar Graphs. Planar graphs are embedded in space and thus have some properties with SISG models in common. In particular, the diameter and the average shortest path lengths of planar graphs are in a similar way related to the number of nodes than the ones of SISG models. It remains to examine whether planar graphs have a spatial structure, and in which aspects they differ.

Modelling Causality and Interdependencies. In this thesis, spatial structure has been modelled without paying attention to relations of higher order, i. e. to how relations relate. A graph representation of public transport activities does, for example, not contain any information about how activities, which are represented by edges, can be concatenated, i. e. how they can be combined to compound activities. In particular, it cannot be modelled that an activity renders another activity possible or even causes another activity to be performed. Without relations of higher order, causality cannot be modelled.

A model of spatial structure presumes relations of higher order, if we assume a more complex understanding of Tobler's law: different kinds of relations exist, for example *left of*, *right of*, etc., and everything is related to (virtually) everything else. In particular, feedback loops arise: one thing relates to another one which itself relates to the former one. In such a feedback loop, the former thing has an indirect effect on itself. It can be argued that a relation of second order is needed to describe a feedback loop formally, because the feedback, i. e. the effect of the loop, cannot be encoded in the representation.

The SISG model does not explicitly reflect the aspects of causality and feedback loops, because relations of higher order are not included in the model; it models spatial structure in a much simpler but yet effective way. A deeper understanding of spatial information can be gained by modelling spatial structure with the use of algebraic methods, in particular by algebraic structures and monoidal homology, because these methods are more suitable to model higher order relations, and thus feedback loops and causality.

Modelling Additional Aspects. The temporal evolution of spatial information is not captures by SISG models. When things move in space, spatial and temporal aspects are inevitably related, and processes alter things and their configurations over time. Future research may model this temporal evolution of spatial information.

Spatial structures can appear in hierarchies. As an example, local, regional and nation-wide transport coexist, all three expose a spatial structure, and they can be interpreted as different levels of the transport system's hierarchy. These different types of transport are related by the stops that they share, and their graph representations can thus be combined to one more extensive graph representation. It has not yet been examined how this compound graph representation relates to the SISG model. Conversely, future research may, for a given compound graph, identify subgraphs that are similar to a SISG model. Such an identification of subgraphs would, for example, provide the possibility to identify and characterize different types of public transport, e. g. local, regional or nation-wide transport.

Modelling Temporal Evolution. An important class of networks evolves over time: users sign up to social networks and add other users as friends, mobile devices steadily move around and connect to different cell sites, new streets are built, and new stops and lines are introduced in public transport networks. Roth et al. (2012) showed that the world's largest metro networks evolve, despite their geographical and economic differences, in a very similar way and thus share some characteristics, e. g. the structure of a core and several branches. It is yet an open question whether the evolution of spatial networks is, at large part, influenced by the properties of space. Future research may model the evolution of spatial networks by iteratively adding nodes to a SISG model, much like in case of other graph models, e. g. the Barabási-Albert model.

Modelling Non-Spatial Structures. Spatial structure is the structure that characterizes spatial information. There is no obvious reason for why there should not exist structures that characterize other types of information, e. g. social, technical or legal information. When structures for such non-spatial types of information have been discovered, a comparison of spatial structure to other ones may provide further insights in how spatial and non-spatial information is structured, and why it is structured the way it is. Knowledge about different structures would even offer the possibility to characterize information more precise, which would render numerous applications possible.

Modelling Interaction Between Several Structures. Several spatial and non-spatial structures can interact. Transport networks for several modes of transport, for example, are in many cases coupled. This coupling leads, in case of the London

Underground, to the existence of an optimal speed to minimize the congestion; an increase of the metro network's speed would even lead to an increase of the congestion (Strano et al. 2015).

Multimodal and multi-layered transport networks often expose similarities, independent of the city in which the transport takes place (Strano et al. 2015). These similarities cannot originate from the specific needs and features of the town, as the needs and features differ for each town but the similarities do not. Future research may show whether these similarities can be explained in terms of the structure of the single layers, and how structures, in general, can interact and lead to new structures.

Improving Algorithms on Spatial Data. When the input data of an algorithm exposes a spatial structure and the algorithm is adapted to this structure, the algorithm can efficiently compute the desired information. If an algorithm does not perform well, it is nearby to ask whether we can improve the adaption of the algorithm to the structure of the data. An algorithm cannot be adapted to the data's structure, if the question to be answered by the algorithm is ill-posed. It is hence important to pose the right question to find well-performing algorithms. Future research can lead to new insights of how algorithms can be designed and optimized in order to take advantage of the spatial structure.

Algorithms can be very slow in the worst case but still perform fast on data with a suitable structure. A depth-first search in a simple graph, for example, considers in the worst case the whole set of edges, which grows quadratically in the number of nodes. The number of edges incident to an edge however can, for graphs with a spatial structure, be expected to be statistically independent of the size of the graph. A depth-first search can hence under certain circumstances be expected to execute much faster, possibly even in linear time. Similar considerations could argue other algorithms to be fast for spatial information.

When data exposes a structure only statistically, algorithms can be improved by heuristics. Dijkstra's algorithm, for example, systematically searches a shortest path in a weighted graph, and the A* algorithm, which is an extension of Dijkstra's algorithm, searches the same path by the use of a heuristic. If the heuristic estimate function for the distance between two nodes is not exact but overestimates the costs in rare cases, the resulting path of the A* algorithm is not necessarily a shortest path, but it is in most cases close to. As Tobler's law, for example, is only statistically true for real data sets, it can only be used for heuristic improvements. Future research may identify suitable aspects of the spatial structure that can be used to heuristically improve algorithms.

Parallelization of Algorithms. The structure of spatial information facilitates the parallelization of algorithms. The problem of parallelization becomes, due to the multicore architecture of current processors, increasingly important. An algorithm can be successfully parallelized by the MapReduce principle, if it can be divided into subproblems. The spatial structure renders a division of the data set and, in consequence, a division of a problem into subproblems possible: first, Tobler's law predicts the existence of neighbourhoods, and the partition of space into neighbourhoods can lead to a partition of the spatial data set. Second, the scale-invariance of data leads to similar structures at different scales, and problems can under certain circumstances be solved at several scales in parallel. A preliminary discussion of the MapReduce principle in respect to Tobler's law and neighbourhoods has been published (Mocnik 2014) but more practical requirements need to be formulated, and scale-invariance needs to be discussed in the context of parallelization.

New Indices for Spatial Data. Indices provide a quick way of accessing data, when the data has a known structure. R-trees (Guttman 1984) and R* trees (Beckmann et al. 1990), for example, require explicit references to space, and as the concept of a neighbourhood is meaningful for space, the data can be indexed by such neighbourhoods. When data does not expose explicit references to space, these indices cannot be used. Data with a spatial structure is however expected to reflect the concept of neighbourhoods, independent of whether the data exposes explicit references to space. Future research may introduce concepts to index spatial data by its spatial structure, even when explicit references to space are missing.

Theory Building. Geographical information science has evolved over the years from the technical science of GIS systems into a more complex science. This more complex science focusses on the information aspect of geography, amongst others on algorithms and computational aspects, spatial statistics, ontological aspects, linked data, volunteered geographic information, the field of cognitive science, etc. Contemporary geographical information science addresses many of the global challenges as well as every-day challenges, which are both complex and multifaceted. This diversity of aspects and challenges leads to a very active and interdisciplinary field of research, but a comprehensive theory which covers these different aspects of spatial information is yet missing.

Goodchild (2003) formulated the long-term goal of finding universal laws to describe how things exist and happen in space and time. Tobler's first law of geography is one of the few known laws and the only one which has gained broad reputation. One possible reason may be that the principles behind space and spatial information, which at bottom apply to most problems of geographical information science, are yet only explored in parts. A promising approach to this long-term goal of finding universal laws is the exploration of the structure of geographical information scientific topics.

Future research may build comprehensive theories about spatial information by continuing the thesis' discussion at a structural level. The concept of graph representations and of spatial structure are fundamental for such a theory building, because the theory is expected to capture general aspects of spatial information. The structural understanding of spatial information and related algorithms can, for example, provide insights into which meaningful transformations of spatial data sets exist and which invariants they have. Such invariants may be unexpected and may not correspond to any intuitive concept, but history has proven many invariants to be extremely useful, for example energy and momenta in physics. The discussion of spatial information at a structural level facilitates the even more general reuse of mathematical and physical theories and thus to find new answers to geographical information scientific questions.

Using The Model Instead Of Real Data. Hypotheses are often evaluated on data sets in order to corroborate them. In many cases, there are not sufficiently many suitable data sets available: data can be hard to collect; unsuitable assumptions may have been made during the data collection; the data may be of poor quality; data sets may be not extensive enough; etc. When a hypothesis is to be tested on spatial data sets and not enough spatial data sets are available or the provision of suitable data sets is to extensive, the SISG model may be used for the evaluation. An evaluation on SISG models is advantageous, because arbitrarily many models with the desired density parameter and minimal dimension can be generated. The evaluation on SISG models cannot substitute evaluations on real data sets, but it can render statistical evaluations on a large number of data sets with a spatial structure possible.

Application: To Which Degree Is Information Spatial? It is widely claimed that information is of spatial nature in large part (Franklin 1992), but evidence is very rare (Hahmann et al. 2011). Attempts to prove or disprove this claim suffer from the fact that one tried to count information, which is not possible. This problem could be circumvented by counting the number of spatial and non-spatial data sets (Hahmann et al. 2011, 2013). A more sophisticated solution would be to measure how spatial data sets are, because the question of whether a data set is spatial by interpretation is a gradual one. This approach of measuring to which degree a data set is spatial may be used to put the attempts to verify the claim on a firm statistical footing.

Application: Improving Search Engines. When data sets are used to solve tasks, they usually need to be equipped with semantic information. If too little semantic information is available, we may try to use data sets with incomplete or missing semantics. The possibility to guess whether data is spatial, to reconstruct the dimension and to compute other properties related to the spatial structure can contribute to this goal of using data with incomplete or missing semantics. Complex search engines[1], for example, require extensive semantics and can yet only access small parts of the world wide web, whilst more simple search engines, which only match for strings, need only very restricted semantics but can access large parts of the world wide web. If data with incomplete or missing semantics can be used to answer complex questions because the semantics can, at least in parts, be guessed, complex search engines can access larger parts of the world wide web. Future research may identify further structural properties that can be used to guess semantic properties.

[1] for example www.wolframalpha.com

The quality of theories and models can, amongst others, be judged by the verifiability of its predictions, its explanatory power, its simplicity and its elegance. The evaluation of our considerations showed that the statistical predictions are applicable, at least for the considered data sets, and the high number of questions raised by the considerations of this thesis demonstrates the considerations to have a broad scope and a high explanatory power. The argumentation is, at the same time, very simple: the SISG model can be defined in a nutshell; the model can be explained in terms of space, objects and relations, i. e. in terms of the entities which constitute spatial information; and no statistical considerations had to be made unless the intended conclusions were of statistical nature. These factors suggest that the proposed concept of spatial structure and the SISG model are viable. The real value of the considerations remains though to be demonstrated in applications and in further theoretical use.

A

Mathematics

We provide, in this appendix, a résumé on the mathematical fundament that this thesis is based on. We present all necessary definitions and some propositions but omit their proofs. The provided résumé may help the reader to agree on definitions, even if more than one definition of a notation may be found in literature. This chapter is however not intended to provide an introduction to the reader who is unfamiliar with the topic. The interested reader may, for such purposes, be referred to the books written by Lang (2002) and Dieudonné (1969) for further details on algebra and metric spaces as well as to the book written by Diestel (2005) for details on graph theory.

A.1 Linear Algebra

Linear algebra is the area of mathematics that focuses on vector spaces and linear transformations between them. Vector spaces have been shown to capture fundamental concepts of the space and time.

Basic Algebraic Structures. A *monoid* $(M, \odot)$ is a set M equipped with an associative operation $\odot\colon M \times M \to M$, such that M is closed under the operation $\odot$ and an identity element exists. A *group* $(G, \odot)$ is a monoid such that inverse elements exist. A group is called *abelian* or *commutative*, if the operation is commutative. A *ring* $(R, \oplus, \odot)$ is a set R equipped with two operations, such that $(R, \oplus)$ is an abelian group, $(R, \odot)$ is a monoid and distributivity laws in respect to the operations $\oplus$ and $\odot$ hold. A ring is called *field*, if $(R, \odot)$ is an abelian group.

A *vector space* $(V, \oplus, \odot)$ *over a field F* is an abelian group $(V, \oplus)$ equipped with an operation $\odot\colon F \times V \to V$, called *scalar multiplication*, such that the scalar multiplication is compatible with the field multiplication, there exists an identity element of the scalar multiplication and distributivity laws in respect to the operations $\oplus$ and $\odot$ hold. The scalar multiplication can alternatively be regarded as a ring homomorphism from the field F into the endomorphism ring of $(V, \oplus)$. A vector space over $\mathbb{C}$ or $\mathbb{R}$ is called *complex* or *real*, respectively.

Affine Transformations, Eigenvectors and Eigenvalues. A *linear transformation* $f\colon (V, \oplus, \odot) \to (V', \oplus', \odot')$ is a homogeneous and additive map $f\colon V \to V'$. A map $g\colon (V, \oplus, \odot) \to (V', \oplus', \odot')$ is called an *affine transformation*, if there exists a linear transformation f and a vector $v' \in V'$, such that $g(v) = v' \oplus' f(v)$ for every $v \in V$.

An *eigenvector* of a linear transformation $f\colon (V, \oplus, \odot) \to (V, \oplus, \odot)$ is a vector $v \in V$, such that an *eigenvalue* $\lambda \in F$ exists and $f(v) = \lambda \cdot v$.

Metric Spaces and Normed Vector Spaces. A *metric space* is a set M equipped with a non-negative and symmetric operation $d\colon M \times M \to \mathbb{R}$, such that the identity of indiscernibles and the triangle equality hold. A common example of metric spaces are metric vector spaces, e. g. real vector spaces equipped with the Euclidean metric or the french railway metric. The french railway metric $d_x(v, w)$ equals the Euclidean metric in case that v and w define a line through the origin, and $d(v, x) + d(x, w)$ for d the Euclidean metric and a fixed point x otherwise.

A *normed vector space* is a vector space $(V, \oplus, \odot)$ equipped with an absolut homogeneous, subadditive and definite operation $\|\cdot\|\colon V \to \mathbb{R}$. Every norm induces a metric by $m(v, w) = \|w \ominus v\|$. An example of normed vector spaces is a real vector spaces equipped with a p-norm (for given $p > 0$)

$$\|v\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

where (x_i) are the components of the vector x for some basis. The norm $\|\cdot\|_1$ is called the *Manhattan norm*, $\|\cdot\|_2$ the *Euclidean norm* and the limit of $p \to \infty$ the *maximum norm* or *Chebyshev norm*.

Euclidean Vector Space. An *inner, dot or scalar product* on a complex or real vector space is a positive definite symmetric sesquilinear form. A *Euclidean vector space* is a real vector space equipped with an inner product. An inner product $\langle \cdot, \cdot \rangle$ defines a norm by $\|x\| = \sqrt{\langle x, x \rangle}$.

Regions and the Uniform Distribution. A *region* of a vector space is an open, connected and non-empty subset of a Euclidean vector space. A set $S \subset U$ of randomly placed points is said to be *randomly distributed in a region U with uniform distribution*, if the chance of a point $s \in S$ to be placed at a point p is equally high for every $p \in U$.

A.2 Graphs

A graph can be used to represent things and their relations. Graphs are, as abstract representations, suitable for the examination of the structure of complex systems.

Graphs. An *undirected graph* $G = (N, E)$ consists of a set of *nodes* N and a set of *undirected edges* E, i. e. unordered pairs of nodes. A *directed graph* is a graph with *directed edges*, i. e. ordered pairs of nodes. We can associate, to every directed graph $G = (N, E)$, a simple undirected graph consisting of the nodes N and an undirected edge (p, q), if either a directed edge (p, q) or a directed edge (q, p) exists in G.

Graphs are even called *networks*, depending on the semantics of a graph, the context it appears in and the author's preferences. We will, in this thesis, not make any distinction between both terms and use the term *network* only if it is part of a widely used compound term.

Relations Between Nodes and Edges. An undirected edge e is said to *join a node* n, if the edge is of the form $(n, \cdot)$ or $(\cdot, n)$. A directed edge (p, q) is said to *start* at p and to *end* at q. Two nodes p and q are called *adjacent* in an undirected graph if an edge (p, q) exists. Two nodes are called adjacent in a directed graph if they are adjacent in the associated undirected graph. Adjacency is a equivalence relation. Two edges are called *incident*, if they share a node. Two edges of a directed graph are called *consecutive*, if one of them ends at the node that the other is starting at.

Types of Graphs. A graph is sometimes called *abstract* to emphasize that the nodes and edges have no meaning apart from the fact that we can check whether two nodes or edges are equal. A graph is called *simple*, if it has no *loops*, i. e. edges starting and ending at the same node, and not more than one edge between each, in case of directed graphs ordered and otherwise unordered, pair of nodes. A graph that is not simple is called a *hypergraph*. A *weighted graph* is a graph where a number is associated to each edge, and a *named graph* is a graph where an object called *name* is associated to each edge.

Degree. The *node degree* of a node n is the number of edges which join the node n. In a directed graph, the number of nodes ending at a node n is called the *indegree* of the node n, and the number of nodes starting at a node n, the *outdegree*.

Subgraphs. A *subgraph* H of a graph G is a graph whose nodes are also nodes of G and whose edges are also edges of G. We even write $H \subset G$, if H is a subgraph of G. A subgraph $H \subset G$ is called *induced*, if every edge of G which only joins nodes of H is also an edge of H. An induced subgraph $H = (N, E) \subset G$ is said to be *induced by the nodes* N.

Complements. The *complement* of a graph $G = (N, E)$ is a graph consisting of the same nodes N such that an edge (p, q) with $p, q \in N$ is contained in the complement if and only if it is not in E.

Complete Graph. A simple graph is called *complete*, if there exists an edge between each, in case of directed graphs ordered and otherwise unordered, pair of nodes.

Walks, Cycles and Holes. A *walk or path of length* k in an undirected or directed graph consists of a number of nodes $p_0, \dots, p_k$ and edges (p_i, p_{i+1}) for all $0 \leq i < k$. The nodes $p_1, \dots, p_{k-1}$ are called *inner nodes*. A closed walk, where the inner nodes are pairwise non-equal, is called a *cycle*. A cycle is called a *hole*, if the nodes of the cycle are only connected by edges which belong to the cycle.

Distance in a Graph. The *distance* $\delta(p, q)$ between two nodes p and q is defined as the minimal length of walks between p and q, and a walk with minimal length is called a *shortest path*. In a directed graph, the *undirected distance* is the distance in the associated undirected graph.

Eccentricity and Centre Nodes. The *eccentricity* of a node p is the maximal distance from p to any other node of the graph. A node p is called a *centre node*, if it is a node with minimal eccentricity.

Maximal Subgraphs. A subgraph $H \subset G$ is called *maximal for a property* π, if the property π is met for H but for no subgraph $H' \supsetneq H$.

Connected Component. An undirected graph is called *connected*, if there exists a walk from each node to each other node of the graph. A directed graph is called connected, if the associated undirected graph is. Maximal connected subgraphs are called *connected components*.

Trees. An (undirected) *tree* is a connected graph without cycles, and a directed tree is a graph whose associated undirected graph is an undirected tree. A *rooted tree* is a directed tree, such that every node of the tree can be reached by a walk starting at a distinguished point, called the *root node*. The *height* of a rooted tree is the maximal distance between the root node and any other node in the graph.

B

Computational Aspects

Several algorithms were presented in the thesis, and some of them reuse existing algorithms. This chapter of the appendix provides an overview on existing algorithms which either were used in the thesis or are of general interest in the context of the thesis. This overview does not aim at explaining the algorithms in detail but at providing a short and concise description and references to literature.

B.1 Root-Finding Problem

A very common mathematical problem is to find a value x such that $f(x) = 0$ for a function $f\colon \mathbb{R}^n \to \mathbb{R}$. Such a value x is called *root* of the function f. Several methods have been developed to iteratively improve a given estimation of a root, e. g. the bisection method, Newton's method and the secant method.

Secant Method. A simple algorithm to find a root of a function is the secant method. The algorithm does, in contrast to Newton's method, not assume the function f to be differentiable but only to be continuous. In every step of the iteration, a new estimation x_{n+1} is computed, based on the last estimation x_n and the last but one estimation x_{n-1}:

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \cdot f(x_n).$$

The computation does, in general, not converge. If the function f is twice continuously differentiable and has a simple root, the computation however converges. The secant method is very old, and its historical development has been discussed by Papakonstantinou et al. (2013).

B.2 Shortest Path Problem

The computation of a shortest path between two nodes of a graph is a common problem. Existing algorithms that solve the shortest path problem have been reviewed by Delling et al. (2009) and Cherkassky et al. (1996). Both authors characterize the algorithms, in particular, by their complexity and by the running time of their implementations. The algorithms can also be used to heuristically compute the average shortest path length. We only discuss one algorithm, because the algorithm seems, at least for the considered data sets, to be much faster than most others.

Partial Shortest Path Trees. Agarwal et al. (2012, 2013) propose the concept of partial shortest path trees (PSPT). The PSPT of a node n is a certain subtree of the original graph, such that it contains the node n, the PSPT of n intersects a high number of PSPTs of other nodes, and at least one inner node of a shortest path between n and an arbitrary node m lies in the intersection of the PSPTs of n and m. An approximation to a shortest path between two nodes n and m can be computed by finding a shortest path in the intersection of the PSPTs of n and m. In some cases, the length of a shortest path between n and m in the intersection of the PSPTs is longer than the shortest path between these nodes in the original graph, but it has been proven by Agarwal et al. (2013) that the shortest path in the intersection is at most one edge longer.

B.3 Centre Node

The centre node of a graph is a node with minimal eccentricity. It is of interest, in the context of this thesis, to estimate inner and outer regions (cf. section 4.2).

2-Sweep and 2-dSweep Algorithms. It is time-consuming to compute the centre node of a graph. Heuristic algorithms like the 2-Sweep algorithm for undirected graphs and the 2-dSweep algorithm for directed graphs can be used to compute a centre node much faster (Corneil et al. 2001, Crescenzi et al. 2012). They are based on the following idea: in an undirected graph, we choose a node n_0 and repeatedly find a node n_{i+1} with maximal distance to n_i. This approach results in a series $(n_i)_i$ of nodes, whose distance increases. The 2-Sweep algorithm stops at $i = 2$, and the distance $\delta(n_1, n_2)$ is a lower bound for the diameter. An estimate of the centre node is the middle node on a shortest path between n_1 and n_2, i. e. the node $m_{\lfloor k/2 \rfloor}$ for a shortest path $(n_1, m_0, \dots, m_k, n_2)$.

In case of a directed graph, a similar algorithm is executed twice: the first time with a forward search to find n_1 and a backward search to find n_2, and the second time with a backward search to find n_1 and a forward search to find n_2. The results are compared and the one with the higher estimate of the graph's diameter is chosen.

B.4 Diameter

The computation of the diameter requires the computation of the distance for each pair of nodes in a graph $G = (N, E)$. A breadth-first search computes the distance for one pair of nodes in $O(|N| + |E|)$ time. As the number of edges $|E|$

varies between $O(|N|)$ and $O(|N|^2)$ in a connected single graph, we expect the computation of the distance for one pair of nodes to take $O(|N|^2)$ time in the worst case, and the computation of the distance for all pairs of nodes to take $O(|N|^4)$ time in the worst case. The algorithms discussed in the following are faster.

Floyd-Warshall Algorithm. Based on the idea that every subpath of a shortest path is a shortest paths again, the *Floyd-Warshall algorithm* computes the distance for each pairs of nodes in a graph simultaneously in $O(|N|^3)$ time (Floyd 1962, Warshall 1962). Assume that all nodes are enumerated from 1 to $|N|$. We define $\delta(i,j,k)$ to be the distance between the nodes i and j in the subgraph which is induced by the nodes $\{1,\dots,k\}$, and we define $\delta(i,j,0) = 1$. We then compute, recursively for $k = 1,\dots,|N|$, the distance in the subgraph induced by the nodes $\{1,\dots,k\}$ as

$$\delta(i,j,k+1) = \min(\delta(i,j,k), \delta(i,k+1,k) + \delta(k+1,j,k)).$$

As $\delta(i,j,|N|)$ is the distance in the original graph G, the distance is computed for all pairs when $k = |N|$. The running time arises from the recursions in k, j and i. The algorithm has, in this formulation, been published by Ingerman (1962).

iFub and DiFub Algorithms. The diameter of a graph can be computed much more efficiently by heuristic algorithms, e. g. by the iFub algorithm for undirected graphs (Crescenzi et al. 2013) and the DiFub algorithm for directed graphs (Crescenzi et al. 2012). Both variants of the iFub algorithms compute the diameter in $O(|N| \cdot |E|)$ time in the worst case. Crescenzi et al. (2012) proved that the diameter can, in many cases, even be computed in $O(|E|)$ time.

Both variants of the algorithms start with a node n_0 that is considered to be central, which can, for example, be computed by the 2-dSweep algorithm[1] (cf. section B.3). A forward and a backward breadth-first search is performed from n_0. Beginning with the nodes of largest distance to n_0, the height of forward and backward breadth-first search trees are computed for all nodes. Considerations about the height of breadth-first search trees lead to a break condition at which the diameter can be computed by the heights of the trees.

2-Sweep and 2-dSweep Algorithms. The 2-Sweep and the 2-dSweep algorithm return an estimate of a centre node and a lower bound for the diameter. The repeated execution of the algorithm for different start nodes results in a number of lower bounds, whose maximum can be used as an estimate of the diameter.

B.5 Cliques

The concept of cliques and complete subgraphs has been used in mathematics since the begin of the 19th century. The problem of computing maximal cliques has been proven to be NP-complete. Algorithms thus have exponential running time, unless they presume certain types of graphs, such as e. g. planar graphs.

[1] The overall result does not depend on the choice of the starting node, but can result in longer running times.

Bron-Kerbosch Algorithm. The maximal cliques in an undirected graph can be computed by the Bron-Kerbosch algorithm. It was introduced by Bron et al. (1973) and published in a simpler formulation by Akkoyunlu (1973). The algorithm recursively collects nodes $n_0, \dots, n_k$ such that each nodes lies in the neighbourhoods of all their preceding nodes. The collected nodes form, by definition, a clique. After adding a node to the collection, it is tested whether the clique is maximal. Using recursive backtracking, all possible choices of nodes are considered, and the cliques that are maximal are returned. The algorithm has exponential running time.

B.6 Dominant Eigenvalue

Eigenvalues have numerous applications, because they characterize matrices and the associated linear transformations well. Numerous algorithms for the computation of the eigenvalue with the largest absolute value exist. We discuss, in the following, a simple algorithm that applies to general matrices. Algorithms for specific types of matrices may have much faster convergence.

Power Iteration Algorithm. Von Mises et al. (1929) introduced an algorithm that estimates the eigenvalue with the largest absolute value. The algorithm inductively computes in each step an improved estimation. For an estimation of an eigenvector b_k which corresponds to the eigenvalue with largest absolute value of a matrix A, we compute an improved estimation of the eigenvector as $b_{k+1} = Ab_k \big/ \|Ab_k\|$. The iteration converges linearly, and each step can be computed in $O(|N|)$ time.

References

Literature that is of special interest for the understanding of the thesis and its results is marked by the symbol ◇.

(Abdalla et al. 2012) Abdalla, A. and Frank, A. U.: *Towards a spatialization of PIM tools*. Proceedings of the Young Researchers Forum on Geographic Information Science at the GIZeitgeist, 2012

(Abdalla et al. 2014) Abdalla, A. and Frank, A. U.: *Designing spatia-temporal PIM tools for prospective memory support*. Proceedings of the 10th International Symposium on Location Based Services, 227–242, 2014

(Abdalla et al. 2013) Abdalla, A.; Weiser, P. and Frank, A. U.: *Design principles for spatio-temporally enabled PIM tools: A qualitative analysis of trip planning*. Proceedings of the 16th AGILE Conference on Geographic Information Science, 323–336, 2013

◇ (Agarwal et al. 2012) Agarwal, R.; Caesar, M.; Godfrey, P. B. et al.: *Shortest paths in less than a millisecond*. Proceedings of the ACM Workshop on Online Social Networks (WOSN), 37–42, 2012

◇ (Agarwal et al. 2013) Agarwal, R.; Caesar, M.; Godfrey, P. B. et al.: *Shortest paths in microseconds*. arxiv/1309.0874v1 [cs.DC], 2013

◇ (Akkoyunlu 1973) Akkoyunlu, E. A.: *The enumeration of maximal cliques of large graphs*. SIAM Journal on Computation, 2(1), 1–6, 1973

◇ (Albert et al. 2002) Albert, R. and Barabási, A.-L.: *Statistical mechanics of complex networks*. Reviews of Modern Physics, 74(1), 47–97, 2002

(Albert et al. 2000) Albert, R.; Jeoing, H. and Barabási, A.-L.: *Error and attack tolerance of complex networks*. Nature, 406, 378–382, 2000

◇ (Aldous et al. 2013) Aldous, D. J. and Ganesan, K.: *True scale-invariant random spatial networks*. Proceedings of the National Academy of Sciences of the United States of America, 110(22), 8782–8785, 2013

◇ (Aldous et al. 2010) Aldous, D. J. and Shun, J.: *Connected spatial networks over random points and a route-length statistic*. Statistical Science, 25(3), 275–288, 2010

(Allen et al. 2014) Allen, M. R.; Barros, V. R.; Broome, J. et al.: *Climate change 2014. Synthesis report*. Intergovernmental Panel on Climate Change (IPCC), 2014

(Anguita et al. 2012) Anguita, D.; Ghio, A.; Oneto, L. et al.: *Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine*. Proceedings of the 4th International Workshop on Ambient Assisted Living (IWAAL), 216–223, 2012

(Anguita et al. 2013) Anguita, D.; Ghio, A.; Oneto, L. et al.: *A public domain dataset for human activity recognition using smartphones*. Proceedings of the 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 437–442, 2013

(Archdeacon et al. 1995) Archdeacon, D.; Bonnington, C. P. and Little, C. H. C.: *An algebraic characterization of planar graphs*. Journal of Graph Theory, 19(2), 237–250, 1995

(Arnborg et al. 1987) Arnborg, S.; Corneil, D. G. and Proskurowski, A.: *Complexity of finding embeddings in a k-tree*. SIAM Journal on Algebraic Discrete Methods, 8(2), 277–284, 1987

(Balassa et al. 1987) Balassa, B. and Bauwens, L.: *Intra-industry specialisation in a multi-country and multi-industry framework*. The Economic Journal, 97(388), 923–939, 1987

(Barabási 2002) Barabási, A.-L.: *Linked. The new science of networks*. Cambridge, MA: Perseus, 2002

◇ (Barabási et al. 1999) Barabási, A.-L. and Albert, R.: *Emergence of scaling in random networks*. Science, 286, 509–512, 1999

(Barabási et al. 2000) Barabási, A.-L.; Albert, R. and Jeon, H.: *Scale-free characteristics of random networks: The topology of the world-wide web*. Physica A, 281(1–4), 69–77, 2000

◇ (Barabási et al. 2001) Barabási, A.-L.; Ravasz, E. and Vicsek, T.: *Deterministic scale-free networks*. Physica A, 299(3–4), 559–564, 2001

(Barabási et al. 2002) Barabási, A.-L.; Jeong, H.; Néda, Z. et al.: *Evolution of the social network of scientific collaborations*. Physica A, 311, 590–614, 2002

◇ (Barrat et al. 2000) Barrat, A. and Weigt, M.: *On the properties of small-world network models*. European Physical Journal B, 13(3), 547–560, 2000

◇ (Barthélemy 2003) Barthélemy, M.: *Crossover from scale-free to spatial networks*. Europhysics Letters, 63(6), 915–921, 2003

◇ (Barthélemy 2011) Barthélemy, M.: *Spatial networks*. Physics Reports, 499(1–3), 1–101, 2011

(Basri 1966) Basri, S. A.: *A deductive theory of space and time*. Amsterdam: North-Holland Publishing, 1966

(Beckmann et al. 1990) Beckmann, N.; Kriegel, H.-P.; Schneider, R. et al.: *The R*-tree: An efficient and robust access method for points and rectangles*. Proceedings of the International Conference on Management of Data (SIGMOD), 322–331, 1990

(Blanchard et al. 2009) Blanchard, P. and Volchenkov, D. (eds.): *Mathematical analysis of urban spatial networks*. Heidelberg: Springer, 2009

(BonJour 2013) BonJour, L.: *Epistemological problems of perception*. In: Zalta, E. N. (ed.), *The Stanford encyclopedia of philosophy*. Stanford University, 2013, spring 2013 edn.

(Bredon 1993) Bredon, G. E.: *Computer methods for mathematical computations*. New York: Springer, 1993

(Brezmes et al. 2009) Brezmes, T.; Gorricho, J.-L. and Cotrina, J.: *Activity recognition from accelerometer data on mobile phone*. Proceedings of the 10th International Work-Conference on Artificial Neural Networks (IWANN), 2, 796–799, 2009

◇ (Bron et al. 1973) Bron, C. and Kerbosch, J.: *Algorithm 457: Finding all cliques of an undirected graph*. Communications of the ACM, 16(9), 575–577, 1973

(Brooks 1941) Brooks, R. L.: *On colouring the nodes of a network*. Mathematical Proceedings of the Cambridge Philosophical Society, 37(2), 194–197, 1941

◇ (Brouwer et al. 2012) Brouwer, A. E. and Haemers, W. H.: *Spectra of graphs*. New York: Springer, 2012

(Chavarriaga et al. 2013) Chavarriaga, R.; Sagha, H.; Calatroni, A. et al.: *The Opportunity challange: A benchmark database for on-body sensor-based activity recognition*. Pattern Recognition Letters, 34(15), 2033–2042, 2013

(Chen et al. 2011) Chen, J.; Shaw, S.-L.; Yu, H. et al.: *Exploratory data analysis of activity diary data: A space-time GIS approach*. Journal of Transport Geography, 19(3), 394–404, 2011

(Cherkassky et al. 1996) Cherkassky, B. V.; Goldberg, A. V. and Radzik, T.: *Shortest paths algorithms: Theory and experimental evaluation*. Mathematical Programming, 73(2), 129–174, 1996

(Choudhury et al. 2006) Choudhury, T.; Philipose, M. and Wyatt, D.: *Towards activity databases: Using sensors and statistical models to summarize people's lives*. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 29(1), 49–58, 2006

◇ (Christaller 1933) Christaller, W.: *Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*. Jena: Fischer, 1933

(Chrobak et al. 1991) Chrobak, M. and Eppstein, D.: *Planar orientations with low out-degree and compaction of adjacency matrices*. Theoretical Computer Science, 86, 243–266, 1991

(Cohen et al. 2003) Cohen, R. and Havlin, S.: *Scale-free networks are ultrasmall*. Physical Review Letters, 90(5), 058701, 2003

◇ (Coleman et al. 1983) Coleman, T. F. and Moré, J. J.: *Estimation of sparse Jacobian matrices and graph coloring problems*. SIAM Journal on Numerical Analysis, 20(1), 187–209, 1983

(Colizza et al. 2007) Colizza, V.; Pastor-Satorras, R. and Vespignani, A.: *Reaction-diffusion processes and metapopulation models in heterogeneous networks*. Nature, 3, 276–282, 2007

◇ (Constantine et al. 1993) Constantine, A. G.; Gower, J. C. and Zielman, B.: *A Cappadocian tablet problem*. Proceedings of the 20th Computer Applications and Quantitative Methods in Archaeology Conference (CAA 1992), 303–315, 1993

◇ (Corneil et al. 2001) Corneil, D. G.; Dragan, F. F.; Habib, M. et al.: *Diameter determination on restricted graph families*. Discrete Applied Mathematics, 113(2–3), 143–166, 2001

(Corry 2004) Corry, L.: *Modern algebra and the rise of mathematical structures*. Basel: Birkhäuser, 2004, second edn.

(Couclelis 2005) Couclelis, H.: *Space, time, geography*. In: Longley, P. A.; Goodchild, M. F.; Maguire, D. J. et al. (eds.), *Geographical information systems: Principles, techniques, management and applications*. New York: Wiley, 2005, 29–38, second edn.

◇ (Crescenzi et al. 2012) Crescenzi, P.; Grossi, R.; Lanzi, L. et al.: *On computing the diameter of real-world directed (weighted) graphs*. Proceedings of the 11th International Symposium on Experimental Algorithms (SEA), 99–110, 2012

◇ (Crescenzi et al. 2013) Crescenzi, P.; Grossi, R.; Habib, M. et al.: *On computing the diameter of real-world undirected graphs*. Theoretical Computer Science, 514, 84–95, 2013

(Daqing et al. 2001) Daqing, L.; Kosmidis, K.; Bunde, A. et al.: *Dimension of spatially embedded networks*. Nature Physics, 7, 481–484, 2001

(Delling et al. 2009) Delling, D.; Sanders, P.; Schultes, D. et al.: *Engineering route planning algorithms*. In: Lerner, J.; Wagner, D. and Zweig, K. (eds.), *Algorithms of large and complex networks. Design, analysis, and simulation*. Berlin: Springer, 2009, 117–139

◇ (Denise et al. 1996) Denise, A.; Vasconcellos, M. and Welsh, D. J. A.: *The random planar graph*. Congressus Numerantium, 113, 61–79, 1996

(Diestel 2005) Diestel, R.: *Graph theory*. Heidelberg: Springer, 2005

(Dieudonné 1969) Dieudonné, J.: *Treatise on analysis. Foundations of modern analysis*, vol. 1. New York: Academic Press, 1969

(Dufour-Lussier et al. 2012) Dufour-Lussier, V.; Le Ber, F.; Lieber, J. et al.: *Semi-automatic annotation process for procedural texts: An application on cooking recipes*. Proceedings of the 1st Cooking with Computers Workshop (CwC), 35–40, 2012

◇ (Erdős et al. 1959) Erdős, P. and Rényi, A.: *On random graphs I*. Publicationes Mathematicae Debrecen, 6, 290–297, 1959

(Esfeld et al. 2011) Esfeld, M. and Lam, V.: *Ontic structural realism as a metaphysics of objects*. In: Bokulich, A. and Bokulich, P. (eds.), *Scientific structuralism*. Dordrecht: Springer, 2011, 143–159

(Euler 1741) Euler, L.: *Solutio problematis ad geometriam situs pertinentis*. Commen-tarii academiae scientiarum Petropolitanae, 8, 128–140, 1741

(European Comission 2010) European Comission: *Commission Regulation (EU) No 97/2010 of 4 February 2010 entering a name in the register of traditional specialities guaranteed [Pizza Napoletana (TSG)]*. Official Journal of the European Union, 53(L34), 7–16, 2010

(Fisher et al. 2000) Fisher, K.; Gershuny, J.; Gauthier, A. et al.: *Exploring new ground for using the Multinational Time Use Study*. Institute for Social and Economic Research Working Paper Series, 2000-28, 2000

◇ (Floyd 1962) Floyd, R. W.: *Algorithm 97: Shortest path*. Communications of the ACM, 5(6), 345, 1962

(Fogliaroni 2013) Fogliaroni, P.: *Qualitative spatial configuration queries. Towards next generation access methods for GIS*. Amsterdam: IOS, 2013

(Forsythe et al. 1977) Forsythe, G. E.; Malcolm, M. A. and Moler, C. B.: *Computer methods for mathematical computations*. Englewood Cliffs, NJ: Prentice-Hall, 1977

(Fortunato 2010) Fortunato, S.: *Community detection in graphs*. Physics Report, 486(3–5), 75–174, 2010

◇ (Fotheringham et al. 1980) Fotheringham, A. S. and Webber, M. J.: *Spatial structure and the parameters of spatial interaction models*. Geographical Analysis, 12(1), 33–46, 1980

(van Fraassen 2006) van Fraassen, B. C.: *Structure: Its shadow and substance*. British Journal for the Philosophy of Science, 57(2), 275–307, 2006

◇ (Frank 2007) Frank, A. U.: *Wayfinding for public transportation users as navigation in a product of graphs*. VGI, 95(2), 195–200, 2007

◇ (Frank 2008) Frank, A. U.: *Shortest path in a combined street and public transportation network: Cognitive agent simulation in a product of two state-transition networks*. Künstliche Intelligenz, 3, 14–18, 2008

(Franklin 1992) Franklin, C.: *An introduction to geographic information systems: Linking maps to databases*. Database, 15(2), 12–21, 1992

(de Fraysseix et al. 1982) de Fraysseix, H. and Rosenstiehl, P.: *A depth-first-search characterization of planarity*. Annals of Discrete Mathematics, 13, 75–80, 1982

◇ (Freeman 1977) Freeman, L. C.: *A set of measures of centrality based on betweenness*. Sociometry, 40(1), 35–41, 1977

◇ (Freeman 1979) Freeman, L. C.: *Centrality in social networks. Conceptual clarification*. Social Networks, 1(3), 215–239, 1979

◇ (Frigg et al. 2011) Frigg, R. and Votsis, I.: *Everything you always wanted to know about structural realism but were afraid to ask*. European Journal or Philosophy of Science, 1(2), 227–276, 2011

◇ (Fruchterman et al. 1991) Fruchterman, T. J. and Reingold, E. M.: *Graph drawing by force-directed placement*. Software: Practice and Experience, 21(11), 1129–1164, 1991

(Gibson et al. 2005) Gibson, D.; Kumar, R. and Tomkins, A.: *Discovering large dense subgraphs in massive graphs*. Proceedings of the 31st International Conference on Very Large Data Bases (VLDB), 721–732, 2005

◇ (Gilbert 1959) Gilbert, E. N.: *Random graphs*. Annals of Mathematical Statistics, 30(4), 1141–1144, 1959

(Glaser 1989) Glaser, R.: *Expertise and learning: How do we think about instructional processes now that we have discovered knowledge structures?* In: Klahr, D. and Kotovsky, K. (eds.), *Complex information processing: The impact of Herbert A. Simon*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1989, 269–282

(Golledge et al. 1997) Golledge, R. G. and Stimson, R. J.: *Spatial behavior: A geographic perspective*. New York: Guilford Press, 1997

(Goodchild 2003) Goodchild, M. F.: *The fundamental laws of GIScience*. Keynote address, Annual Assembly, University Consortium for Geographic Information Science, Monterey, CA, 2003

(Goudie 2013) Goudie, A. S. (ed.): *The human impact on the natural environment*. Oxford: Wiley-Blackwell, 2013, 7th edn.

(Grujičić et al. 2014) Grujičić, I.; Raidl, G.; Schobel, A. et al.: *A metaheuristic approach for integrated timetable based design of railway infrastructure*. Proceedings of the 3rd International Conference on Road and Rail Infrastructure (CETRA), 691–696, 2014

(Guttman 1984) Guttman, A.: *R-trees: A dynamic index structure for spatial searching*. Proceedings of the International Conference on Management of Data (SIGMOD), 47–57, 1984

(Gärdenfors 2000) Gärdenfors, P.: *Conceptual spaces. The geometry of thought*. Cambridge, MA: MIT Press, 2000

(Haggett et al. 1969) Haggett, P. and Chorley, R. J.: *Network analysis in geography*. London: Edward Arnold, 1969

(Hahmann et al. 2013) Hahmann, S. and Burghardt, D.: *How much information is geospatially referenced? Networks and cognition*. International Journal of Geographical Information Science, 27(6), 1171–1189, 2013

(Hahmann et al. 2011) Hahmann, S.; Burghardt, D. and Weber, B.: *80% of all information is geospatially referenced? Towards a research framework: Using the semantic web for (in)validating this famous geo assertion*. Proceedings of the 14th AGILE Conference on Geographic Information Science, 2011

(Halás et al. 2014) Halás, M.; Klapka, P. and Kladivo, P.: *Distance-decay functions for daily travel-to-work flows*. Journal of Transport Geography, 35, 107–119, 2014

(Haufe et al. 2012) Haufe, S.; Schiffel, S. and Thielscher, M.: *Automated verification of state sequence invariants in general game playing*. Artificial Intelligence, 187–188, 1–30, 2012

(Hayashi 2006) Hayashi, Y.: *A review of recent studies of geographical scale-free networks*. IPSJ Digital Courier, 2, 155–164, 2006

(Hecht et al. 2009) Hecht, B. and Moxley, E.: *Terabytes of Tobler: Evaluating the first law in a massive, domain-neutral representation of world knowledge*. Proceedings of the 9th International Conference on spatial information theory (COSIT), 88–105, 2009

(Hemmens 1970) Hemmens, G. C.: *Analysis and simulation of urban activity patterns*. Socio-Economic Planning Sciences, 4(1), 53–66, 1970

(Hildebrandt et al. 2010) Hildebrandt, T.; Frommberger, L.; Wolter, D. et al.: *Towards optimization of manufacturing systems using autonomous robotic observers*. Proceedings of the 7th CIRP International Conference on Intelligent Computation in Manufacturing Engineering, 2010

◇ (Holland et al. 1981) Holland, P. W. and Leinhardt, S.: *An exponential family of probability distributions for directed graphs*. Journal of the American Statistical Association, 76(373), 33–50, 1981

◇ (Holme et al. 2012) Holme, P. and Saramäki, J.: *Temporal networks*. Physics Reports, 519(3), 97–125, 2012

(Hossmann et al. 2012) Hossmann, T.; Efstratiou, C. and Mascolo, C.: *Collecting big datasets of human activity one checkin at a time*. Proceedings of the 4th ACM International Workshop on Hot Topics in Planet-Scale Measurement (HotPlanet), 15–20, 2012

(Houba et al. 2000) Houba, I. H. G.; Hartog, R. J. M.; Top, J. L. et al.: *Using recipe classes for supporting detailed planning in food industry: A case study*. European Journal of Operational Research, 122(2), 367–373, 2000

(Hunter et al. 2008) Hunter, D. R.; Goodreau, S. M. and Handcock, M. S.: *Goodness of fit of social network models*. Journal of the American Statistical Association, 103(481), 248–258, 2008

◇ (Huson et al. 1995) Huson, M. L. and Sen, A.: *Broadcast scheduling algorithms for radio networks*. Proceedings of the Military Communications Conference (MILCOM), 2, 647–651, 1995

(Huynh 2008) Huynh, D. T. G.: *Human activity recognition with wearable sensors*. PhD thesis, Darmstadt University of Technology, 2008

◇ (Ingerman 1962) Ingerman, P. Z.: *Algorithm 141: Path matrix*. Communications of the ACM, 5(11), 556, 1962

(Jacomy et al. 2014) Jacomy, M.; Venturini, T.; Heymann, S. et al.: *ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software*. PLoS ONE, 9(6), e98679, 2014

(Jenkins et al. 1992) Jenkins, P. L.; Phillips, T. J.; Mulberg, E. J. et al.: *Activity patterns of Californians: Use of and proximity to indoor pollutant sources*. Atmospheric Environment, 26A(12), 2141–2148, 1992

(Jeong et al. 2000) Jeong, H.; Tombor, B.; Albert, R. et al.: *The large-scale organization of metabolic networks*. Nature, 407, 651–654, 2000

◇ (Jespersen et al. 2000) Jespersen, S. N. and Blumen, A.: *Small-world networks: Links with long-tailed distributions*. Physical Review E, 62(5), 6270–6274, 2000

(Kahl 2002) Kahl, W.: *A relation-algebraic approach to graph structure transformation*. Habilitation thesis, Bundeswehr University Munich, 2002

◇ (Kalapala et al. 2006) Kalapala, V.; Sanwalani, V.; Clauset, A. et al.: *Scale invariance in road networks*. Physical Review E, 73, 026130, 2006

(Kaltenbrunner et al. 2012) Kaltenbrunner, A.; Scellato, S.; Volkovich, J. et al.: *Far from eyes, close on the web: Impact of geographic distance on online social interactions*. Proceedings of the ACM Workshop on Online Social Networks (WOSN), 19–24, 2012

(Kammer et al. 2015) Kammer, F. and Tholey, T.: *Approximate tree decompositions of planar graphs in linear time*. arxiv/1104.2275v4 [cs.DS], 2015

(Kansky et al. 1989) Kansky, K. and Danscoine, P.: *Measures of network structure*. Flux, 5, 89–121, 1989

(Karantonis et al. 2006) Karantonis, D. M.; Narayanan, M. R.; Mathie, M. et al.: *Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring*. IEEE Transactions on Information Technology in Biomedicine, 10(1), 156–167, 2006

(Karg et al. 2014) Karg, M. and Kirsch, A.: *A human morning routine dataset*. Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 1351–1352, 2014

(Kirchhoff 1845) Kirchhoff, G.: *Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesondere durch eine kreisförmige*. Annalen der Physik und Chemie, 64(4), 497–514, 1845

(Klepeis et al. 2001) Klepeis, N. E.; Nelson, W. C.; Ott, W. R. et al.: *The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants*. Journal of Exposure Analysis and Environmental Epidemiology, 11(3), 231–252, 2001

◇ (Kobourov 2013) Kobourov, S. G.: *Force-directed drawing algorithms*. In: Tamassia, R. (ed.), *Handbook of graph drawing and visualization*. Boca Raton, FL: CRC Press, 2013, 383–408

(Kopczewska 2013) Kopczewska, K.: *Roads as channels of centrifugal policy transfer: A spatial interaction model revised*. Contemporary Economics, 7(3), 39–50, 2013

◇ (Kosmidis et al. 2008) Kosmidis, K.; Havlin, S. and Bunde, A.: *Structural properties of spatially embedded networks*. Europhysics Letters, 82, 48005, 2008

(Krömer 2007) Krömer, R.: *Tool and object. A history and philosophy of category theory*. Basel: Birkhäuser, 2007

◇ (Kuhn 2012) Kuhn, W.: *Core concepts of spatial information for transdisciplinary research*. International Journal of Geographical Information Science, 26(12), 2267–2276, 2012

◇ (Kuratowski 1930) Kuratowski, C.: *Sur le problème des courbes gauches en Topologie*. Fundamenta Mathematicae, 15(1), 271–283, 1930

(Kwapisz et al. 2010) Kwapisz, J. R.; Weiss, G. M. and Moore, S. A.: *Activity recognition using cell phone accelerometers*. Explorations of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 12(2), 74–82, 2010

(Ladyman 1998) Ladyman, J.: *What is structural realism?* Studies in History and Philosophy of Science, Part A, 29(3), 409–424, 1998

(Lang 2002) Lang, S.: *Algebra*. New York: Springer, 2002, third edn.

(Lee et al. 2010) Lee, V. E.; Ruan, N.; Jin, R. et al.: *A survey of algorithms for dense subgraph discovery*. In: Aggarwal, C. C. and Wang, H. (eds.), *Managing and mining graph data*. New York: Springer, 2010, 303–336

(Leech et al. 1996) Leech, J. A.; Wilby, K.; McMullen, E. et al.: *The Canadian Human Activity Pattern Survey: Report of methods and population surveyed*. Chronic Diseases in Canada, 17(3/4), 118–123, 1996

(Leskovec et al. 2010^{a}) Leskovec, J.; Huttenlocher, D. and Kleinberg, J.: *Predicting positive and negative links in online social networks*. Proceedings of the 19th International Conference on World Wide Web (WWW), 641–650, 2010

(Leskovec et al. 2010^{b}) Leskovec, J.; Huttenlocher, D. and Kleinberg, J.: *Signed networks in social media*. Proceedings of the 28th Conference on Human Factors in Computing Systems (CHI), 1361–1370, 2010

(Liebchen 2003) Liebchen, C.: *Symmetry for periodic railway timetables*. Proceedings of the Workshop on Algorithmic Methods and Models for Optimization of Railways (ATMOS), 34–51, 2003

(Lippai 2005) Lippai, I.: *Water system design by optimization: Colorado Springs Utilities case studies*. Proceedings of the ASCE Pipeline Division Specialty Conference (Pipelines), 1058–1070, 2005

(Los Alamos National Laboratory et al. 1998) Los Alamos National Laboratory and Texas Transportation Institute: *Transportation Analysis Simulation System (TRANSIMS). The Dallas case study*. Washington, DC: U.S. Department of Transportation, 1998

◇ (Louf et al. 2014) Louf, R.; Roth, C. and Barthelemy, M.: *Scaling in transportation networks*. PLoS ONE, 9(7), e102007, 2014

◇ (Lovasz 1993) Lovasz, L.: *Random walks on graphs: A survey*. In: Miklós, D.; Sós, V. T. and Szõnyi, T. (eds.), *Combinatorics, Paul Erdős is eighty*. Budapest: János Bolyai Mathematical Society, 1993, vol. 2, 1–46

(Love et al. 2006) Love, N.; Hinrichs, T. and Genesereth, M.: *General game playing: Game Description Language Specification*. Standford University, 2006

◇ (Luce et al. 1949) Luce, R. D. and Perry, A. D.: *A method of matrix analysis of group structure*. Psychometrika, 14(2), 95–116, 1949

(Lukowicz et al. 2004) Lukowicz, P.; Ward, J. A.; Junker, H. et al.: *Recognizing workshop activity using body worn microphones and accelerometers*. Proceedings of the 2nd International Conference on Pervasive Computing (PERVASIVE), 18–32, 2004

(Lukowicz et al. 2007) Lukowicz, P.; Timm-Giel, A.; Lawo, M. et al.: *WearIT@work: Toward real-world industrial wearable computing*. IEEE Pervasive Computing, 6(4), 8–13, 2007

(Lymberopoulos et al. 2008) Lymberopoulos, D.; Bamis, A. and Savvides, A.: *Extracting spatiotemporal human activity patterns in assisted living using a home sensor network*. Proceedings of the 1st International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), 2008

(Machens et al. 2013) Machens, A.; Gesualdo, F.; Rizzo, C. et al.: *An infectious disease model on empirical networks of human contact: Bridging the gap between dynamic network data and contact matrices*. BMC Infectious Diseases, 13, 185, 2013

◇ (Macindoe et al. 2010) Macindoe, O. and Richards, W.: *Graph comparison using fine structure analysis*. Proceedings of the 2nd International Conference on Social Computing (SocialCom), 193–200, 2010

◇ (MacLane 1937) MacLane, S.: *A combinatorial condition for planar graphs*. Fundamenta Mathematicae, 28(1), 22–31, 1937

(Maslov et al. 2004) Maslov, S.; Sneppen, K. and Zaliznyak, A.: *Detection of topological patterns in complex networks: Correlation profile of the internet*. Physica A, 333, 529–540, 2004

(Mazur et al. 1932) Mazur, S. and Ulam, S.: *Sur les transformations isometriques d'espaces vectoriels normes*. Comptes rendus hebdomadaires des séances de l'Académie des sciences, 194, 946–948, 1932

(McCurdy et al. 2000) McCurdy, T.; Glen, G.; Smith, L. et al.: *The National Exposure Research Laboratory's consolidated human activity database*. Journal of Exposure Analysis and Environmental Epidemiology, 10(6), 566–578, 2000

◇ (McDiarmid 2008) McDiarmid, C.: *Random graphs on surfaces*. Journal of Combinatorial Theory, Series B, 98(4), 778–797, 2008

◇ (McDiarmid et al. 2005) McDiarmid, C.; Steger, A. and Welsh, D. J. A.: *Random planar graphs*. Journal of Combinatorial Theory, Series B, 93(2), 187–205, 2005

(Miller et al. 2003) Miller, E. J. and Roorda, M. J.: *Prototype model of household activity/travel scheduling*. Transportation Research Record: Journal of the Transportation Research Board, 1831, 114–121, 2003

(Miller 2004^{a}) Miller, H. J.: *Activities in space and time*. In: Stopher, P.; Button, K. J.; Haynes, K. E. et al. (eds.), *Handbook of transport 5: Transport geography and spatial systems*. Oxford: Pergamon/Elsevier, 2004, 647–660

◇ (Miller 2004^{b}) Miller, H. J.: *Tobler's first law and spatial analysis*. Annals of the Association of American Geographers, 94(2), 284–289, 2004

◇ (von Mises et al. 1929) von Mises, R. and Pollaczek-Geiringer, H.: *Praktische Verfahren der Gleichungsauflösung*. Zeitschrift für Angewandte Mathematik und Mechanik, 9(1/2), 58–77/152–164, 1929

(Mocnik 2014) Mocnik, F.-B.: *MapReduce principle for spatial data*. Extended Abstract Proceedings of the 8th International Conference on Geographic Information Science (GIScience), 100–103, 2014

(Mocnik 2015) Mocnik, F.-B.: *Modelling spatial information*. Proceedings of the 1st Vienna Young Scientists Symposium (VSS), 46–47, 2015

◇ (Mocnik et al. 2015) Mocnik, F.-B. and Frank, A. U.: *Modelling spatial structures*. Proceedings of the 12th Conference on Spatial Information Theory (COSIT), 2015

(Moeslund et al. 2006) Moeslund, T. B.; Hilton, A. and Krüger, V.: *A survey of advances in vision-based human motion capture and analysis*. Computer Vision and Image Understanding, 104(2–3), 90–126, 2006

(İkizler Nazlı et al. 2008) İkizler Nazlı and Forsyth, D. A.: *Searching for complex human activities with no visual examples*. International Journal of Computer Vision, 80(3), 337–357, 2008

(Newman 2002) Newman, M. E. J.: *Assortative mixing in networks*. Physical Review Letters, 89(20), 208701, 2002

(Newman 2003^{a}) Newman, M. E. J.: *Mixing patterns in networks*. Physical Review E, 67, 026126, 2003

◇ (Newman 2003^{b}) Newman, M. E. J.: *The structure and function of complex networks*. SIAM Review, 45(2), 167–256, 2003

(Nieves et al. 2013) Nieves, J. C.; Guerrero, E. and Lindgren, H.: *Reasoning about human activities: An argumentative approach*. Proceedings of the 12th Scandinavian Conference on Artificial Intelligence (SCAI), 195–204, 2013

◇ (Ogden et al. 1923) Ogden, C. K. and Richards, I. A.: *The meaning of meaning*. New York: Harcourt, Brace and World, 1923

(Ortmann 2014) Ortmann, J.: *Semantic integration of human and technical observations*. PhD thesis, University of Münster, 2014

(Papakonstantinou et al. 2013) Papakonstantinou, J. M. and Tapia, R. A.: *Origin and evolution of the secant method in one dimension*. The American Mathematical Monthly, 120(6), 500–518, 2013

(Paradiso et al. 2005) Paradiso, R.; Loriga, G.; Taccini, N. et al.: *WEALTHY – a wearable healthcare system: New frontier on e-textile*. Journal of Telecommunications and Information Technology, 4, 105–113, 2005

(Pastor-Satorras et al. 2001) Pastor-Satorras, R.; Vázquez, A. and Vespignani, A.: *Dynamical and correlation properties of the internet*. Physical Review Letters, 87(25), 258701, 2001

(Poincaré 1905) Poincaré, H.: *Science and hypothesis*. New York: Walter Scott Publishing, 1905

(Poppe 2007) Poppe, R.: *Vision-based human motion analysis: An overview*. Computer Vision and Image Understanding, 108(1–2), 4–18, 2007

◇ (Polya 1921) Polya, G.: *Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz*. Mathematische Annalen, 84(1–2), 149–160, 1921

(Psillos 2006) Psillos, S.: *The structure, the whole structure, and nothing but the structure?* Philosophy of Science, 73(5), 560–570, 2006

(Putnam 1975) Putnam, H.: *Philosophical papers. Mathematics, matter and method*, vol. 1. Cambridge: Cambridge University Press, 1975

(Raubal 2001) Raubal, M.: *Agent-based simulation of human wayfinding: A perceptual model for unfamiliar buildings*. PhD thesis, Vienna University of Technology, 2001

(Raubal et al. 2008) Raubal, M. and Moratz, R.: *A functional model for affordance-based agents*. In: Rome, E.; Hertzberg, J. and Dorffner, G. (eds.), *Towards affordance-based robot control*. Heidelberg: Springer, 2008, 91–105

(Raubal et al. 2004) Raubal, M.; Miller, H. J. and Bridwell, S.: *User-centred time geography for location-based services*. Geografiska Annaler, Series B, 86(4), 245–265, 2004

(Recker et al. 1986^{a}) Recker, W. W.; McNally, M. G. and Root, G. S.: *A model of complex travel behaviour: Part I, Theoretical development*. Transport Research Part A, 20(4), 307–318, 1986

(Recker et al. 1986^{b}) Recker, W. W.; McNally, M. G. and Root, G. S.: *A model of complex travel behaviour: Part II, An operational model*. Transport Research Part A, 20(4), 319–330, 1986

(Regneri et al. 2013) Regneri, M.; Rohrbach, M.; Wetzel, D. et al.: *Grounding action descriptions in videos*. Transactions of the Association for Computational Linguistics (TACL), 1, 25–36, 2013

(Reiss et al. 2013) Reiss, A.; Hendeby, G. and Stricker, D.: *A competitive approach for human activity recognition on smartphones*. Proceedings of the 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 455–460, 2013

(Ribeiro et al. 2006) Ribeiro, R.; Batista, F.; Pardal, J. P. et al.: *Cooking an ontology*. Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA), 213–221, 2006

◇ (Richards et al. 2009) Richards, W. and Wormald, N.: *Representing small group evolution*. Proceedings of the 12th International Conference on Computational Science and Engineering (CSE), 159–165, 2009

(Rilett et al. 2001) Rilett, L. R. and Zietsmann, J.: *An overview of the TRANSIMS micro-simulation model: Application possibilities for South Africa*. Proceedings of the 20th South African Transport Conference, 2001

(Ripeanu et al. 2002) Ripeanu, M.; Foster, I. and Iamnitchi, A.: *Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design*. IEEE Internet Computing, 6(1), 50–57, 2002

(Robertson et al. 1986) Robertson, N. and Seymour, P. D.: *Graph minors. II. Algorithmic aspects of tree-width*. Journal of Algorithms, 7(3), 309–322, 1986

(Rodrigue et al. 2013) Rodrigue, J.-P.; Comtois, C. and Slack, B.: *The geography of transport systems*. Oxon: Routledge, 2013, third edn.

(Rohrbach et al. 2012^{a}) Rohrbach, M.; Amin, S.; Andriluka, M. et al.: *A database for fine grained activity detection of cooking activities*. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 1194–1201, 2012

(Rohrbach et al. 2012^{b}) Rohrbach, M.; Regneri, M.; Andriluka, M. et al.: *Script data for attribute-based recognition of composite activities*. Proceedings of the 12th European Conference on Computer Vision (ECCV), 1, 144–157, 2012

◇ (Roth et al. 2012) Roth, C.; Kang, S. M.; Batty, M. et al.: *A long-time limit for world subway networks*. Journal of the Royal Society Interface, 9(75), 2540–2550, 2012

(Rozenberg 1997) Rozenberg, G. (ed.): *Handbook of graph grammars and computing by graph transformation*. Singapore: World Scientific, 1997

(Rudin 1991) Rudin, W.: *Functional analysis*. New York: McGraw-Hill, 1991, second edn.

(Ryoo 2008) Ryoo, M. S.: *Semantic representation and recognition of human activities*. PhD thesis, University of Texas at Austin, 2008

(Sala et al. 2010) Sala, A.; Cao, L.; Wilson, C. et al.: *Measurement-calibrated graph models for social network experiments*. Proceedings of the 19th International World Wide Web Conference (WWW), 861–870, 2010

(Salathé et al. 2012) Salathé, M.; Bengtsson, L.; Bodnar, T. J. et al.: *Digital epidemiology*. PLoS Computational Biology, 8(7), e1002616, 2012

◇ (Sanders 1997) Sanders, J. T.: *An ontology of affordances*. Ecological Psychology, 9(1), 97–112, 1997

(Schnyder 1989) Schnyder, W.: *Planar graphs and poset dimension*. Order, 5(4), 323–343, 1989

(Scholz et al. 2013) Scholz, C.; Atzmüller, M. and Stumme, G.: *New insights and methods for predicting face-to-face contacts*. Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM), 563–572, 2013

(Schöbel et al. 2013) Schöbel, A.; Raidl, G.; Grujičić, I. et al.: *An optimization model for integrated timetable based design of railway infrastructure*. Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen), 2013

(Seymour et al. 1994) Seymour, P. D. and Thomas, R.: *Call routing and the ratcatcher*. Combinatorica, 14(2), 217–241, 1994

(Shapiro 1997) Shapiro, S.: *Philosophy of mathematics: Structure and ontology*. Oxford: Oxford University Press, 1997

(Shoaib et al. 2014) Shoaib, M.; Bosch, S.; Incel, O. D. et al.: *Fusion of smartphone motion sensors for physical activity recognition*. Sensors, 14(6), 10146–10176, 2014

(Snow 1854) Snow, J.: *The cholera near Golden-square, and at Deptford*. Medical Times and Gazette, 23. September 1854, 321–322, 1854

(Song et al. 2005) Song, C.; Havlin, S. and Makse, H. A.: *Self-similarity of complex networks*. Nature, 433, 392–395, 2005

◇ (Spielman 2012) Spielman, D.: *Spectral graph theory*. In: Naumann, U. and Schenk, O. (eds.), *Combinatorial scientific computing*. Boca Raton, FL: Chapman and Hall, 2012, 495–524

◇ (Stewart 1948) Stewart, J. Q.: *Demographic gravitation: Evidence and applications*. Sociometry, 11(1/2), 31–58, 1948

◇ (Strano et al. 2015) Strano, E.; Shai, S.; Dobson, S. et al.: *Multiplex networks in metropolitan areas: Generic features and local effects*. Journal of the Royal Society Interface, 12(111), 2015

◇ (Strobach 1998) Strobach, N.: *The moment of change. A systematic history in the philosophy of space and time*. Dordrecht: Kluwer, 1998

◇ (Strogatz 2001) Strogatz, S. H.: *Exploring complex networks*. Nature, 410, 268–276, 2001

(Taegi et al. 2001) Taegi, K. and Oh, K.-Y.: *Country size, income level and intra-industry trade*. Applied Economics, 33(3), 401–406, 2001

(Tarjan et al. 1977) Tarjan, R. E. and Trojanowsi, A. E.: *Finding a maximum independent set*. SIAM Journal on Computing, 6(3), 537–546, 1977

(Tentori et al. 2008) Tentori, M. and Favela, J.: *Activity-aware computing for healthcare*. IEEE Pervasive Computing, 7(2), 51–57, 2008

(Thielscher 2010) Thielscher, M.: *A General Game Description Language for incomplete information games*. Proceedings of the 24th AAAI Conference on Artificial Intelligence, 994–999, 2010

(Thielscher 2011) Thielscher, M.: *The General Game Description Language is universal*. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 1107–1112, 2011

◇ (Tobler 2004) Tobler, W.: *On the first law of geography: A reply*. Annals of the Association of American Geographers, 94(2), 304–310, 2004

◇ (Tobler et al. 1971) Tobler, W. and Wineberg, S.: *A Cappadocian speculation*. Nature, 231, 39–41, 1971

◇ (Tobler 1970) Tobler, W. R.: *A computer movie simulating urban growth in the detroit region*. Economic Geography, 46, 234–240, 1970

(De la Torre et al. 2008) De la Torre, F.; Hodgins, J.; Bargteil, A. et al.: *Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database*. Robotics Institute, Carnegie Mellon University, 2008

(Trafiklab 2013) Trafiklab: *GTFS Sverige*. https://www.trafiklab.se, accessed at 28th April 2013

(Tunca et al. 2014) Tunca, C.; Alemdar, H.; Ertan, H. et al.: *Multimodal wireless sensor network-based ambient assisted living in real homes with multiple residents*. Sensors, 14(6), 9692–9719, 2014

◇ (Turvey 1992) Turvey, M. T.: *Affordance and prospective control: An outline of the ontology*. Ecological Psychology, 4(3), 173–187, 1992

(Uttal 2008) Uttal, W. R.: *Time, space, and number in physics and psychology*. Cornwall-on-Hudson, NY: Sloan Publishing, 2008

(Van Laerhoven et al. 2004) Van Laerhoven, K.; Lo, B. P. L.; Ng, J. W. P. et al.: *Medical healthcare monitoring with wearable and implantable sensors*. Proceedings of the 3rd International Workshop on Ubiquitous Computing for Healthcare Applications (UbiHealth), 2004

(de Verdière 1990) de Verdière, Y. C.: *Sur un nouvel invariant des graphes et un critère de planarité*. Journal of Combinatorial Theory, Series B, 50(1), 11–21, 1990

(Vázquez et al. 2002) Vázquez, A.; Pastor-Satorras, R. and Vespignani, A.: *Large-scale topological and dynamical properties of the internet*. Physical Review E, 65, 066130, 2002

(Wagner 1937) Wagner, K.: *Über eine Eigenschaft der ebenen Komplexe*. Mathematische Annalen, 114(1), 570–590, 1937

(Walski et al. 1987) Walski, T. M.; Brill, E. D.; Gessler, J. et al.: *Battle of the network models: Epilogue*. Journal of Water Resources Planning and Management, 113(2), 191–203, 1987

◇ (Warshall 1962) Warshall, S.: *A theorem on boolean matrices*. Journal of the ACM, 9(1), 11–12, 1962

(Wasserman et al. 1994) Wasserman, S. and Faust, K.: *Social network analysis: Methods and applications*, 1994

(Watanabe 1977) Watanabe, H. (ed.): *Human activity system. Its spatiotemporal structure*. Tokyo: University of Tokyo Press, 1977

◇ (Watts et al. 1998) Watts, D. J. and Strogatz, S. H.: *Collective dynamics of small-world networks*. Nature, 393, 440–442, 1998

◇ (Waxman 1988) Waxman, B. M.: *Routing of multipoint connections*. IEEE Journal on Selected Areas in Communications, 6(9), 1617–1622, 1988

◇ (Whitney 1931) Whitney, H.: *Non-separable and planar graphs*. Proceedings of the National Academy of Sciences in the United States of America, 17(2), 125–127, 1931

(Wilson et al. 2012) Wilson, G. and Shpall, S.: *Action*. In: Zalta, E. N. (ed.), *The Stanford encyclopedia of philosophy*. Stanford University, 2012, summer 2012 edn.

(Wittgenstein 1967) Wittgenstein, L.: *Philosophical investigations*. Oxford: Basil Blackwell, 1967, third edn.

(Wobst 1975) Wobst, R.: *Isometrien in metrischen Vektorräumen*. Studia Mathematica, 54(1), 41–54, 1975

(Wormald 1999) Wormald, N. C.: *Models of random regular graphs*. In: Lamb, J. D. and Preece, D. A. (eds.), *Surveys in combinatorics 1999*. Cambridge: Cambridge University Press, 1999, 239–298

◇ (Worral 1989) Worral, J.: *Structure realism: The best of two worlds?* Dialectica, 43(1–2), 99–124, 1989

(Xie et al. 2007) Xie, F. and Levinson, D. M.: *Measuring the structure of road networks*. Geographical Analysis, 39(3), 336–356, 2007

◇ (Xulvi-Brunet et al. 2002) Xulvi-Brunet, R. and Sokolov, I. M.: *Evolving networks with disadvantaged long-range connections*. Physical Review E, 66, 026118, 2002

◇ (Yook et al. 2002) Yook, S.-H.; Jeong, H. and Barabási, A.-L.: *Modeling the internet's large-scale topology*. Proceedings of the National Academy of Sciences of the United States of America, 99(21), 13382–13386, 2002

◇ (Zipf 1947) Zipf, G. K.: *The hypothesis of the minimum equation as a unifying social principle: With attempted synthesis*. American Sociological Review, 12(6), 627–650, 1947

Curriculum Vitae

Franz-Benjamin Mocnik

Date of birth: 8th of October 1983 — Nationality: German

Education

1990–2002	School education (skipped the 10th grade; final grade: *sehr gut*)
2000–2001	University courses for pupils at the University of Essen in the fields of physics, mathematics and informatics
7/2001	Summer school of the Deutsche SchülerAkademie Title: *Mathematical Structures of Theoretical Physics*
8–10/2002	Preparatory class in mathematics and physics, University of Bonn
10/2002	Begin of studies, University of Bonn
12/2004	Pre-diploma in physics (final grade: *sehr gut*)
3/2011	Diploma in mathematics (final grade: *sehr gut*) Thesis: *Bogomolovs Zerlegungssatz für Calabi-Yau-Mannigfaltigkeiten* Advisor: Daniel Huybrechts
9/2011–2/2013	Doctoral studies, University of Münster Advisor: Werner Kuhn
3/2013–	Doctoral studies, Vienna University of Technology Advisor: Andrew U. Frank

Employment

2005–2010	Tutor Mathematical Department, University of Bonn
2011–2012	Scientific Researcher Institute for Geoinformatics, University of Münster
3/2013–9/2014	Project Assistant Research Group Geoinformatics, Vienna University of Technology
10/2014–8/2015	University Assistant Research Group Geoinformatics, Vienna University of Technology