Modelling Spatial Structures

Franz-Benjamin Mocnik and Andrew U. Frank

Vienna University of Technology, 1040 Vienna, Austria
mail@mocnik-science.net
www.mocnik-science.net

Abstract Data is spatial if it contains references to space. We can easily detect explicit references, for example coordinates, but we cannot detect whether data implicitly contains references to space, and whether it has properties of spatial data, if additional semantic information is missing. In this paper, we propose a graph model that meets typical properties of spatial data. We can, by the comparison of a graph representation of a data set to the graph model, decide whether the data set (implicitly or explicitly) has these typical properties of spatial data.

Keywords: space, spatial structure, spatial data, spatial information, time, Tobler's law, principle of least effort, graph model, spatial network, scale invariance

1 Introduction

It is widely assumed that information is in large part of a spatial nature [17]. Evidence for exact percentages is rare [21], but the large number of spatial data sets demonstrates the importance of spatial information, e. g. data about public transport, cadastres, maps, and weather data.

Tobler claimed that "everything is related to everything else, but near things are more related than distant things" [43], which is known as Tobler's first law of geography. This law is not universally true but has been proven to be statistically valid for many spatial data sets, e. g. for spatially referenced Wikipedia articles [22]. Tobler's law is, in case of human activities, motivated by the principle of least effort [52]: it claims that people choose the path of least effort, and as movement in space usually requires more effort for longer distances, human activities more often relate near than distant things.

Physical properties of tangible space are very similar at different scales. Classical mechanics holds, for example, for everyday items as well as for solar systems. Representations of human activities and processes that depend on physical properties are thus often independent of scale. This is not true in general, but it applies, for instance, to many transport networks [28, 34]. Both, Tobler's law and scale invariance, are characteristics of spatial data in many cases.

Properties of data can be influenced by the data's relation to space, but also by other reasons: for example, the location of cities depends on properties of space (since transport costs are related to distance, etc.) but also on landscape

morphology (shape and structure of water bodies, natural resources, etc.) and others. If the properties of data originating from the properties of space (called spatial structure) predominate, the data is called spatial. This classification is not a binary but rather a fuzzy classification: a data set can expose a spatial structure to a certain extent, and spatial and non-spatial structures may coexist.

Explicit references to space enable us to check whether a data set has typical properties of spatial data, e. g. whether Tobler's law is met [22]. When a data set does not expose explicit references to space, we cannot check in the same way whether Tobler's law is met or whether it has typical properties of spatial data. However, data can meet Tobler's law without explicitly including references to space: some things are more related than others, and relations (within the data set) may have the same structure as relations in spatial information. The issue lies with how to detect a spatial structure without explicit references to space. A very short overview over this topic has been provided by the author at the Vienna Young Scientists Symposium [38].

In this paper, we propose a spatial graph model that has some typical properties of spatial data sets, e. g. Tobler's law and scale invariance. The comparison of a data set with the proposed graph model enables us to determine whether the data set exhibits these typical properties or not.

In the next section, we discuss typical properties of spatial information, including Tobler's law and scale invariance (section 2). Then we outline existing graph models as well as related work and argue why these models are not suitable as general models of spatial data (section 3). We propose a spatial graph model (section 4) and show that it meets the properties discussed in section 2 (section 5). The comparison between spatial data sets and the proposed graph model is discussed and evaluated on several data sets, including data sets about public transport, water distribution networks, formalizations of games, and biological networks (section 6).

2 Typical Properties of Spatial Information

Information which exposes a number of references to space is, by definition, called *spatial information*. These references relate the underlaying data with things that are placed in 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 spatial data; and non-uniform distributions of objects in space lead to not necessarily uniform but in many cases bounded distributions of relations. We call such a structure of data in this paper a *spatial structure*[1], and we say that data *has a spatial structure* (in which case we also

[1] Time has a similar effect on data, because it can be modelled by one-dimensional Euclidean vector spaces.

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 in such a way.

In this section, we discuss three typical properties of spatial information that the proposed graph model meets. For this discussion, we assume data sets to have representations that explicitly expose references to space as well as relations of objects within the data set. Such representations of data as a graph are called *graph representations* in this paper. Graph representations can be stored in triple stores or graph databases to practically verify properties of the data sets.

2.1 Tobler's First Law

A topological core concept of space is neighbourhood [29]: when things are near, we say that they are located in the same neighbourhood. The concept is relative in the sense that the meanings of near and neighbourhood depend on context, and do not necessarily relate to Euclidean space but to some concept of distance, e. g. travelling time or fuel consumption.

The existence of distance and the concept of neighbourhood (both properties of space), as well as the existence of "costs" to relate distant things, lead to spatial autocorrelation, *Tobler's first law of geography*: "Everything is related to everything else, but near things are more related than distant things" [43]. Many data sets reveal that more distant things are statistically less related for different reasons, e. g. due to transport or communication costs. For example, Tobler's law has been proven true in large part for spatially referenced Wikipedia articles [22].

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 physical processes of our tangible world remain (nearly) the same when rescaled. 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 the rescaling of space. The system of objects and interrelations is, in this case, called *scale-invariant*[2]. The effect of scale invariance can be observed in different data sets, e. g. for metro and railway networks [34], and for road networks [28].

2.3 Bound Outdegree

The average edge degree in a planar graph can be proven to be strictly less than 6, which can be seen by Euler's formula and the fact that a face has at least

[2] The concept of *scale invariance* of a graph embedded in space should not be confused with the concept of *scale-freeness* of a graph, which is characterized by a power law distribution of the nodes' edge degrees and hence by the invariance of the distribution's shape under rescaling of the total number of edges.

three edges and each edge has at most two faces. The edges of a planar graph can be oriented such that the outdegree is bounded by 3 [11]. We expect that the outdegree of a graph embedded in space behaves similar, even if it is not completely planar, and we hence expect the outdegree 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-uniform 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.

We have discussed three typical properties of spatial information, as well as the structure that spatial data in consequence has. In the next section, we discuss existing graph models and argue why they are not suitable as general models of spatial data.

3 Existing Graph Models

Existing graph models have been developed in order to model different aspects of information. An overview of space-related graphs and their properties has been provided by Barthélemy [8]. In this section, we review well-known graph models and argue why they are not suitable to model spatial data in general.

- The *Erdős-Rényi model* is a randomly chosen graph with a given number of nodes and edges [15]. The *Gilbert model* is a graph where edges between pairs of nodes exist with a given probability [19]. Both models are not suitable for modelling spatial data in general, because spatial data is not completely random: the structure of spatial data is influenced by space, and some configurations of edges are expected to occur more often than others.
- *Barabási and Albert* proposed a graph model where nodes are added incrementally [4]. Each time a node is added, edges are more likely to be introduced between the new node and nodes that already have a high number of edges. Thus, the majority of nodes is joined by a very low number of edges, whereas only a very low number of nodes is joined by a high number of edges, resulting in a power-law degree distribution. Graphs with power-law distributions are called *scale-free*, because the distribution of the edge degree is scale-invariant. This model and similar ones have been used to model internet links [5, 42], citation networks [3], and social networks [40]. The construction of Barabási-Albert models however does not reflect Tobler's law.
- The family of *exponential random graph models* consist of graphs whose edges follow the distribution of the exponential family [23]. Exponential random graph models have been used to model social networks [24]. These graphs are tailored to fit statistical properties, but they do not refer to spatial properties.
- *Hierarchical network models* relate duplicates of small graphs at different hierarchies [6]. These models are suitable for 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.

- *Watts and Strogatz* proposed a model with a very short typical path length and high clustering [48]. Spatial graphs usually have longer path length than this model, because relations tend to exist only in neighbourhoods.
- *Planar graphs*, i. e. graphs that can be embedded in two-dimensional space such that their edges do not intersect, have been studied widely [2, 18, 30, 35, 41, 45, 46, 50]. Graphs have been proven to be planar if and only if they neither contain the complete graph K_5 with 5 nodes nor the complete bipartite graph $K_{3,3}$ with 6 nodes as a subgraph after the contraction of edges [46]. Spatial data sets usually cannot be represented by planar graphs, because they remain spatial after local modifications (in particular after the introduction of K_5 or $K_{3,3}$ as subgraphs).
- *Spatial generalizations of existing models* have been discussed, e. g. by considering only planar graphs during construction [14, 37], or other modifications [27, 36]. For example, the Barabási-Albert model has been modified by taking distance between nodes into account [7, 42, 51]. These generalizations share aspects of spatial data, but as most of their characteristics originate from the non-generalized models, they are not suitable as models for spatial data in general.
- A class of *geometric graph models* assumes nodes to have explicit locations in space, and edges to be modelled by the distance between nodes: an edge between two nodes with distance $d(p, q)$ is introduced with probability $f(d(p, q))$, where $f\colon \mathbb{R} \to [0, 1]$ is some probability function. For example, radio transmitters (with constant transmitting power) have been modelled [25] using the function

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

  A similar model was discussed by Waxman [49], who proposed a smoothened, continuous probability function

$$f(l) = \beta \exp\left(-\frac{l}{r}\right). \tag{2}$$

  Both models depend on the absolute distance between points and thus are not scale-invariant. A scale-invariant variant of this model was discussed by Aldous [1]: edges to the k nodes with minimal distance are introduced for each node (for a given $k > 0$). This model does not reflect the fact that for a spatial data set, the distribution of the relations, in particular the number of relations per node, usually depends on the locations of the nodes in space.

As argued in this section, existing graph models are tailored to model specific types of data but not spatial data in general. In the next section, we introduce a model of spatial data in general. The construction of the model is motivated by the three typical properties of spatial information of section 2.

4 A Scale-Invariant Spatial Graph Model

In this section, we introduce a graph model that has numerous properties of spatial data, including the three ones discussed in the last section. The graph model does not aim at modelling particular types of spatial data, e. g. data about public transport or communication data. It aims, instead, at having typical properties of spatial data, and thus at sharing similarities with many spatial data sets. In the following, we motivate that the proposed model meets the properties of section 2. The proof is given later in section 5.

For the construction of a graph model, we ask which edges have to be introduced for a given set of nodes in space in order to model spatial data. To be more exact, we ask which configuration of edges would produce the properties of section 2.

Taking the second part of Tobler's law "near things are more related than distant things" literally means that, as soon as a point p is related to a point q by an edge, every edge q' with a shorter distance, i. e. $d(p,q') < d(p,q)$, is with a high probability also related to p by an edge. In the proposed model, an edge between p and q' does not only exist with a high probability, but with a probability of 1.[3] This means that edges to a number of nearest points are introduced, and the number can vary for different nodes.

As Tobler's law describes things in space, the number of edges depends on the distribution of the nodes, in particular on the distance between nodes. If the number depends on the absolute distance between points, the model is not scale-invariant. The following model only uses the relative distance between nodes, and is therefore scale-invariant:

Definition 1 (Scale-invariant spatial graph model – SISG model).
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[4] *(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

$$d(p,q) \leq \rho \cdot \min_{q_0 \in S \setminus \{p\}} d(p,q_0)$$

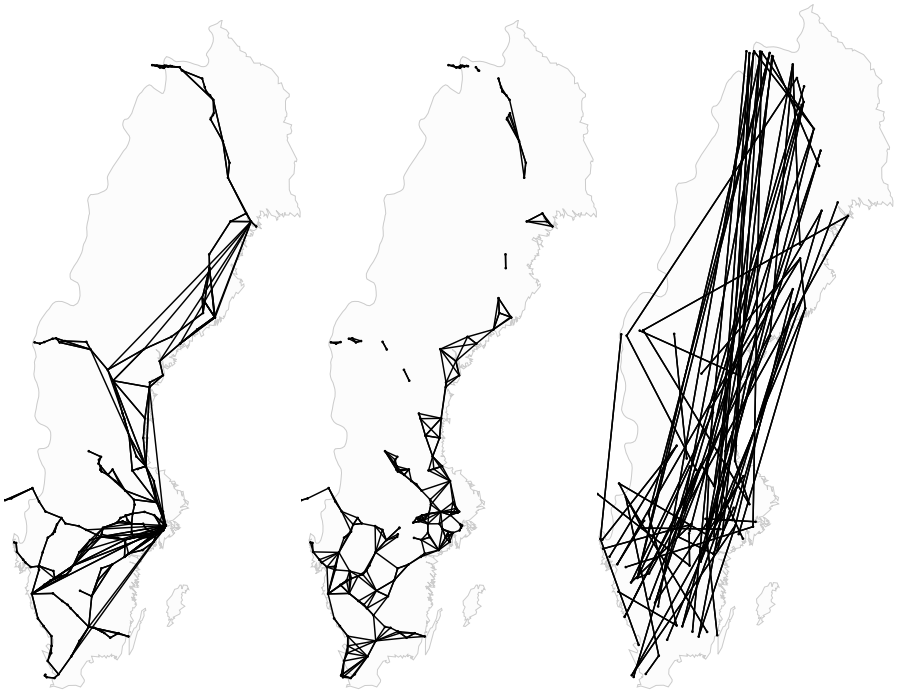
where q_0 denotes the nearest neighbour of p.

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.

[3] This choice is made because it enables us to analytically compute some properties of the model. A variant of the model which introduces edges only with a certain probability is left to future work. As long as this effect is less dominant than other ones, we expect the properties of the proposed model and its variant to be similar.

[4] A graph is called *abstract* if its nodes and edges contain no additional semantics. An abstract graph is, in particular, not embedded in space, and the nodes have no location.

In figure 1(a), a graph representation of data from the national railway operator in Sweden is depicted[5]. Nodes in the graph representation relate to stops, and edges to pairs of successive stops, i. e. stops p and q such that at least one train travels from p to q without stopping in between. Figure 1(b) shows a SISG model with the stops of the data set used in (a) as generating set. As expected, edges exist only between near nodes. The SISG model is not expected to model the graph representation exactly, because it is not a model of public transport but of spatial data in general. We hence expect the model to share the properties discussed in section 2 with the graph representation. The Gilbert model (a random graph) with the stops of the data set used in (a) does not share significant spatial properties with the graph representation in (a). In particular, edges exist between nodes independent of their distance.



(a) graph representation (b) $\mathcal{M}_2(S)$ (c) $\mathcal{G}_{\text{Gilbert}}(S, 6 \cdot 10^{-3})$

Figure 1: Graphs whose nodes S are the stops of the national railway operator SJ in Sweden; (a) graph representation of the data set, (b) a SISG model and (c) a Gilbert model; the parameters are chosen such that similarities and dissimilarities visually stand out

[5] The data is publicly accessible in the General Transit Feed Specification (GTFS) format [44].

The definition of a SISG model is based on a generating set of points placed in space. If a model is to be computed without prior choice of a generating set, a set of randomly placed points can be used. Without further assumptions, however, it is not clear which distribution of the points should be used, and as physical space is uniform, there is no preferred distribution. A uniform distribution[6] of points is hence a convenient choice:

Definition 2 (Uniform scale-invariant spatial graph model).
For a given dimension m, we denote a (uniform) SISG model with a generating set consisting of s randomly distributed points with uniform distribution in an m-dimensional ball by $\mathcal{M}_\rho^m(s)$.

The SISG model as well as the uniform SISG model are constructed to have some typical properties of spatial data. In the next section, we discuss some basic properties of the model and formally prove that the SISG model has the typical properties of spatial data that were discussed in section 2.

5 The Model has Typical Properties of Spatial Data

We have constructed SISG models such that they have some typical properties of spatial data. In this section, we examine the properties that were discussed in section 2, and we formally prove that SISG models have these properties.

5.1 Tobler's First Law

Tobler claimed that "near things are more related than distant things". In accordance with this, we are able to prove:

Theorem 1. *If there is an edge between two nodes p and q in a SISG model $\mathcal{M}_\rho(S, V)$, then there exists an edge between p and any node q' that has distance smaller than $d(p, q)$.*

Proof. We have

$$d(p, q') \overset{(a)}{\leq} d(p, q) \overset{(b)}{\leq} \rho \cdot \min_{q_0 \in S \setminus \{p\}} d(p, q_0)$$

where (a) is due to the presumption of the theorem and (b) due to the definition of a SISG model. The equation proves that p and q' are connected by an edge according to the definition of a SISG model. □

[6] A uniform distribution is a distribution where a point is placed at each location in space with the same probability.

5.2 Scale Invariance

The change of scale in a vector space V can be described by transformations $\tau\colon V \to V$ that scale distances between arbitrary points p and q by a constant factor $\sigma > 0$:

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

As such transformations change scale, they are called *scale transformations* of relative scale σ. SISG models can be proven to be invariant under such transformations:

Theorem 2. *SISG models are invariant under scale transformations, i. e.*

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

(as abstract graphs) for every scale transformation $\tau\colon V \to V$.

Proof. The definition of the SISG model does not explicitly include the location of nodes, just their distances. In particular, the inequality in definition 1 stays invariant because both sides are multiplied by the relative scale $\sigma > 0$. □

5.3 Bound Outdegree

The outdegree of a node has a lower bound:

Theorem 3. *If a SISG model has at least two nodes, each node has outdegree of at least* 1*.*

Proof. As at least two nodes exist, each node p has a nearest node p_0, and thus an outgoing edge to p_0. □

As SISG models are simple by definition (i. e. there exists not more than one edge for each pair of nodes), the outdegree of a node is bound by $(s - 1)$ where s is the number of nodes. This upper bound however is meaningless, because it is met for every simple graph and not only for SISG models. It can be shown that the expected outdegree of a node is much lower than this upper bound, when the number of nodes approaches infinity[7]:

Theorem 4. *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 $m = \dim V$. For an arbitrarily chosen point p and a real number $L > 0$, let S be the set of all points in the m-dimensional ball $B_m(p, L)$ of radius L centred

[7] Analytical results are much easier to derive when the number of nodes approaches infinity and hence only "inner regions" of the graphs have to be considered. The results can, however, be expected to approximately hold for finite graphs as well, when the number of nodes is sufficiently high.

in p. We denote the minimal distance between p and the remaining points by $r = \min_{q \in S \setminus \{p\}} d(p, q)$.

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

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

points in $B(\rho R, R)$, namely the one at minimal distance r (the cases where more than one point is at minimal distance is a null set) and the ones in the inner of $B(\rho R, R)$. (The second equality is due to the fact that $\mathrm{Vol}_m(\rho R)$ equals $\rho^m \mathrm{Vol}_m(R)$.) 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

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

Inserting equation $(*)$ proves that the probability that at least μ edges are starting at p is

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

The probability that at most μ edges are starting at 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 at p, we first compute

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

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

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

When $s \to \infty$, the second summand vanishes. □

We have proven that SISG models have the typical properties of spatial information which were discussed in section 2. In the next section, we use this fact to test data sets for spatial structures.

6 Application: Testing Data for Spatial Structures

SISG models have typical properties of spatial data and can thus serve as prototypes of spatial data. In this section, we discuss an application of SISG models: the comparison of data sets with the model enables us to discover spatial structures.

6.1 The Problem of Testing Data for Spatial Structures

Space can lead to a spatial structure, i. e. typical properties of spatial data, as was discussed in section 2. If a data set contains explicit references to space, it can be checked which typical properties the data set has. The situation is different when a data set only implicitly contains references to space, e. g. when semantics is missing: the data set contains references to space but we are not aware of them, and the references can, in consequence, not be used to check which typical properties the data set has. This raises the question of *how to check for a spatial structure without explicit references to space.*

When a data set is similar to a uniform SISG model, we can conclude that the data set has, by and large, the properties discussed in section 2 because the SISG model has them. In this case, the data set has a spatial structure, i. e. a structure that data typically has when it can be interpreted as spatial information. We can, by no means, conclude that the data set becomes spatial information when it is interpreted[8], but there is a good chance that the data set *can* become spatial information by interpretation if the data set is similar to a SISG model. The information of the data set is, in this case, a good candidate for spatial information.

Data sets very rarely *equal* a SISG model, because the properties discussed in section 2 are not exact laws but rather loose properties. It is however sufficient that a data set and a uniform SISG model are *similar* in order to conclude that they have similar properties. Spatial data sets can even have properties very different to the ones discussed in section 2, depending on how representations are constructed, and on to what extent non-spatial information is included in the representation. Such a data set cannot be detected when it is compared to SISG models.

Many examples of spatial information have, in addition to a spatial structure, further structures. Examples 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. The

[8] There cannot exist any way to conclude whether a certain interpretation of a data set is spatial without knowledge of the interpretation, because there can exist spatial and non-spatial interpretations of the same data set.

comparison of a data set to the SISG model can gradually determine how similar the data set set is to the model, and thus to what degree the data set has a spatial structure.

We have discussed that we can, with some limitations, approach the question of whether a data set has a spatial structure by comparing it to the SISG model. In the next section, we approach the question of how we can compare a data set with a SISG model.

6.2 The Problem of Comparing Data to the SISG Model

In section 6.1, we discussed the importance of comparing a data set to the SISG model in order to approach the question whether the data set is spatial. The question of how to conduct this comparison is approached in this section.

The uniform SISG model depends on a density parameter ρ, a dimension m and a number of nodes s. For given parameters, we can compute an explicit model $\mathcal{M}_\rho^m(s)$. The SISG model $\mathcal{M}_\rho^m(s)$ can be understood as an abstract graph, i. e. as a set of abstract nodes and edges, ignoring the fact that we know the parameters ρ, m, and s that were used to generate the model. We will discuss two methods (theorems 5 and 6) that enable us to approximately recover ρ^m. As both methods estimate the same value, we expect their results to be approximately equal for any SISG model.

As the two methods of estimating ρ^m do computationally not depend on the fact that the graph is a SISG model, we can apply them to any data set that is represented as a graph. This enables us to check whether a graph representation has a spatial structure: if the computations of both estimates result different values for a data set, it does not have the properties discussed in section 2. If the estimates are approximately equal, we can conclude that the graph representation shares some properties with a SISG model.

Both methods of estimating ρ^m assume that the number of nodes approaches infinity. Data sets are necessarily finite, and the used analytical results are hence, in general, not valid for data sets. If the size of a data set is sufficiently large, we will assume that the estimations are good enough for a reasonable comparison to the SISG model. We will see in section 6.4 that the results are reasonable for the examined data sets.

Theorem 4 provides a first method of estimating ρ^m for a given SISG model (as an abstract graph): we expect each node to have an outdegree of ρ^m, and as the outdegree equals the number of edges divided by the number of nodes, we can immediately conclude:

Theorem 5. *For a graph $\mathcal{M}_\rho^m(s)$ with e edges, we expect $\rho^m = e/s$ for $s \to \infty$.*

A second approach to estimate ρ^m compares the density of subgraphs. The density of a graph is defined as:

Definition 3 (Density of a graph [12]). *The* density[9] *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 theorem 5, we expect the density of $\mathcal{M}_\rho^m(s)$ to be $\rho^m/(s-1)$ for $s \to \infty$. As every subgraph of a SISG model is again a subgraph of the SISG model generated with the same density parameter and the nodes of the subgraph as generating set, we expect an induced subgraph[10] of $\mathcal{M}_\rho^m(s)$ with t nodes to have density $\rho^m/(t-1)$ for $t \to \infty$. For small subgraphs, the estimation may be worse than for large ones as the limit of $t \to \infty$ is not considered. If the computation of a subgraph's density does not only include the edges between nodes of the subgraph but also includes all edges that start (and not necessarily end) in the subgraph, the result is similar to the one of an infinite graph because it cannot be distinguished from a fragment of an infinite graph. We define:

Definition 4 (Total density of a subgraph). *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.

Using this definition, we can conclude:

Theorem 6. *A subgraph H of $\mathcal{M}_\rho^m(s)$ with t nodes is expected to approximately have total density $\rho^m/(t-1)$ for $H \ll \mathcal{M}_\rho^m(s)$, i. e. in case that the number of nodes in the subgraph is much smaller than the number of nodes in the SISG model.*

An estimate of ρ^m can be found by computing the total density of subgraphs and fitting by the function $c_{\text{total density}}(t) = \rho^m/(t-1)$ where t is the number of nodes in the subgraph. This estimate of ρ^m by theorem 6 should approximately equal the estimate by theorem 5 for uniform SISG models. In the next section, we introduce spatial and non-spatial data sets that are used in section 6.4 to evaluate the considerations of this section.

6.3 Data Sets for the Comparison

In this section, we introduce data sets from different domains that have explicit references to space resp. time, and some data sets that have no explicit references to space nor time. These data sets are used in the next section to evaluate the methods proposed in section 6.2.

[9] The notation of density of a graph should not be confused with the notation of density of elements distributed in space.

[10] A subgraph H of a graph G is called *induced* if every edge (p, q) of G with p and q nodes in H is also an edge of H.

Table 1: Estimates of ρ^m for different data sets

	Graph	$\lvert N\rvert$	$\lvert E\rvert$	$\widehat{\rho^m}^N$	$\widehat{\rho^m}^D$	χ^2	Ref.
⊗	$\mathcal{M}_2^2(1000)$	1 000	3 939	3.94	4.69	$4.53{\cdot}10^{-3}$	def. 2
⊕	$\mathcal{M}_{2.3}^2(1000)$	1 000	5 111	5.11	5.73	$2.80{\cdot}10^{-3}$	def. 2
⊙	$\mathcal{M}_{2.5}^2(1000)$	1 000	5 947	5.95	6.84	$6.85{\cdot}10^{-3}$	def. 2
□	SJ (national railway provider)	176	544	3.09	4.30	$9.46{\cdot}10^{-3}$	[44]
	Länstrafiken Sörmland	2 100	4 382	2.09	2.37	$9.32{\cdot}10^{-4}$	[44]
	Östgötatrafiken	2 643	5 960	2.26	2.08	$1.07{\cdot}10^{-3}$	[44]
	Blekingetrafiken	1 215	2 643	2.18	2.15	$8.81{\cdot}10^{-4}$	[44]
⊠	Hallandstrafiken	1 503	3 331	2.22	2.36	$4.92{\cdot}10^{-4}$	[44]
	Värmlandstrafiken	1 682	3 743	2.23	2.59	$2.53{\cdot}10^{-3}$	[44]
	Västmanlands Lokaltrafik	1 491	3 223	2.16	2.41	$5.32{\cdot}10^{-4}$	[44]
	Dalatrafik	3 359	7 366	2.19	2.48	$3.92{\cdot}10^{-3}$	[44]
⊞	Karlstadsbuss	251	530	2.11	2.30	$4.52{\cdot}10^{-4}$	[44]
⊟	Luleå Lokaltrafik	205	459	2.24	2.73	$2.51{\cdot}10^{-3}$	[44]
◫	Stadsbussarna Östersund	247	526	2.13	2.54	$5.68{\cdot}10^{-4}$	[44]
⊡	Swebus	158	435	2.75	3.78	$1.21{\cdot}10^{-2}$	[44]
◇	Power grid in the USA	4 941	13 188	2.67	3.74	$1.62{\cdot}10^{-3}$	[48]
◈	Network of airports in the USA	500	5 960	11.92	45.52	$1.26{\cdot}10^{-3}$	[13]
◈	Water distr. netw. Anytown	24	43	1.79	2.07	$2.31{\cdot}10^{-2}$	[33]
◈	Water distr. netw. W.-C. Ranch	1 785	1 983	1.11	1.88	$4.30{\cdot}10^{-3}$	[33]
☆	Pizza Napoletana	2 291	3 687	1.61	2.16	$1.04{\cdot}10^{-2}$	[16]
△	Tic-tac-toe (2x2 board)	30	44	1.47	1.53	$4.57{\cdot}10^{-3}$	
▲	Tic-tac-toe (3 moves, 3x3 board)	3 890	74 169	19.07	24.85	$4.11{\cdot}10^{-3}$	
▲	Rubik's Cube (3 rot., 2x2x2 size)	1 417	1 644	1.16	5.15	$4.44{\cdot}10^{-2}$	
▲	Rubik's Cube (3 rot., 3x3x3 size)	4 602	5 364	1.17	9.59	$3.63{\cdot}10^{-2}$	
▽	p2p Gnutella network 09	8 114	26 013	3.21	6.03	$9.40{\cdot}10^{-4}$	[39]
▽	Met. network of A. fulgidus	1 567	3 631	2.32	10.92	$2.95{\cdot}10^{-3}$	[26]
▽	Met. network of C. elegans	1 469	3 447	2.35	7.68	$1.97{\cdot}10^{-1}$	[26]
▽	Met. network of E. coli	2 897	7 104	2.45	12.03	$1.32{\cdot}10^{-1}$	[26]
▽	Graph of Wikipedia votes	7 115	103 689	14.57	78.82	$3.35{\cdot}10^{-3}$	[31, 32]

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

$\lvert N\rvert$ number of nodes

$\lvert E\rvert$ number of edges

$\widehat{\rho^m}^N$ estimate of ρ^m by the number of nodes and edges (according to theorem 5)

$\widehat{\rho^m}^D$ estimate of ρ^m by density (fit of the arithmetic mean of the total density of 10 (randomly chosen) connected induced subgraphs consisting of n nodes ($n = 1, \ldots, 50$), by $\rho^m/(t-1)$, according to theorem 6)

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

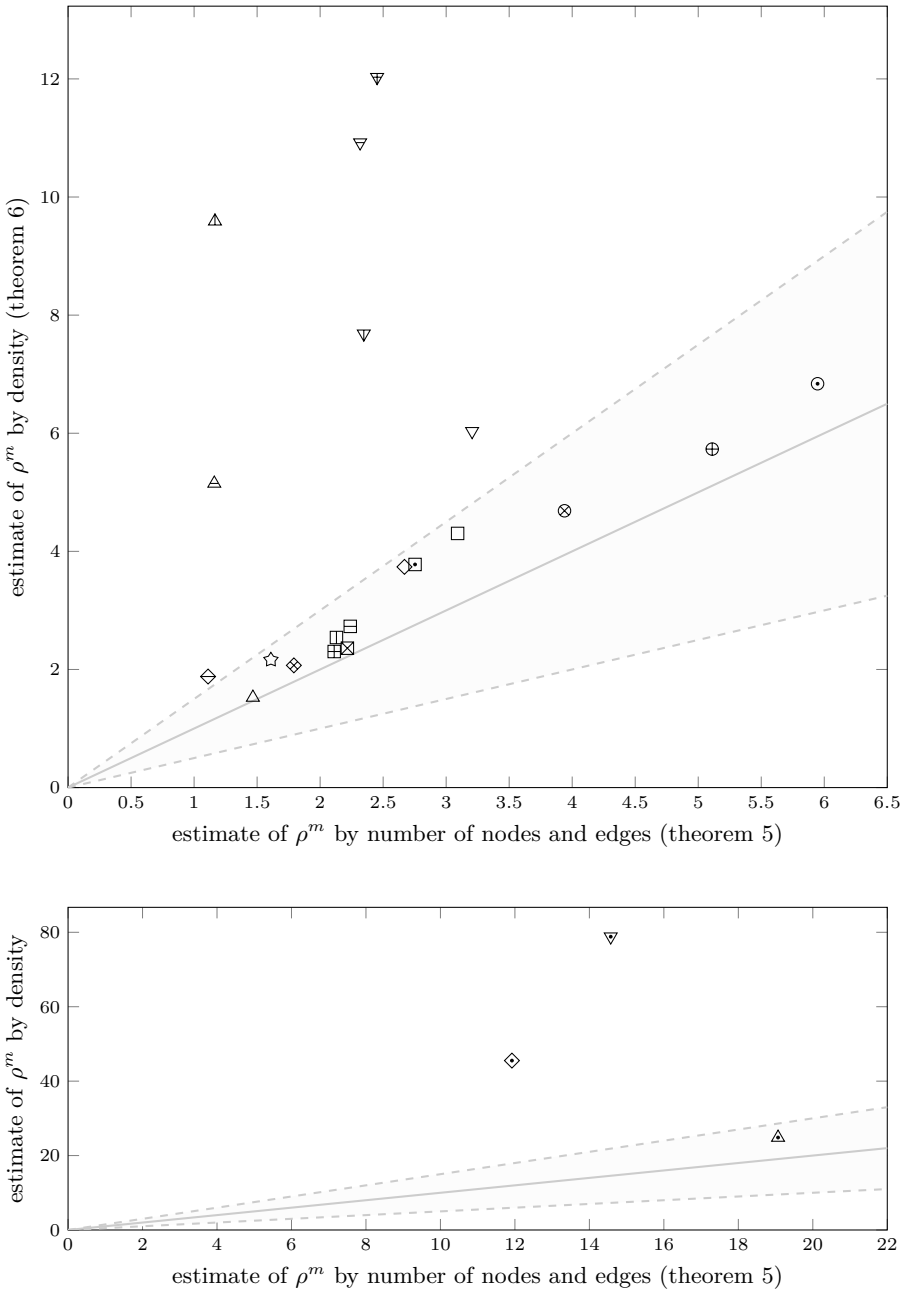


Figure 2: Estimates of ρ^m for different data sets (see table 1); if a data set has the typical properties of spatial data discussed in section 2, both estimates coincide; for the grey area, the estimates differ by less than a factor of 1/2

As examples for spatial data, we examine data about public transport in Sweden which was already used in section 4. Each of these data sets contains data from one transport agency [44]. These networks are explicitly related to space by the coordinates of the stops.

In addition to these spatial data sets, we examine the high-voltage power grid in the Western States of the USA [48], and the network of airports in the USA [13]. In the latter one, airports are represented as nodes, and an edge exists between two nodes if a flight was scheduled between the corresponding airports in 2002.

Water distribution networks are further examples for spatial data sets. They consist of one or more sources and a number of sinks. Pipes are represented by edges, and as the aim of the network is to distribute the water, there is a flow direction and the pipes can hence be represented by directed edges. Walski introduced the hypothetical water distribution network of Anytown which has been used as a prototypical example in many studies [47]. Another example is the water distribution network of the Wolf-Cordera Ranch which distributes water to about 370,000 persons [33].

Recipes and Games are examples of activities that are related to time. We examine a formalization of a pizza recipe [16] where we represent the state of the ingredients as nodes and the actions as edges. Similarly, we examine the Rubik's Cube and Tic-tac-toe games. For both games, we consider different sizes of the cube resp. board, and we restrict the number of rotations resp. moves in the game to restrict the size of the network.

Data about a computer network (p2p Gnutella network 09) [39], metabolic networks of different cellular organisms [26], and a graph of Wikipedia votes [31, 32] have no explicit references to space (in the used representations as abstract graphs).

In addition to these data sets, we consider three uniform SISG models to validate the hypothesis that both estimates coincide.

The data sets introduced in this section are very different in their structure. In section 6.2, the relation between spatial structure and properties of SISG models was discussed. In the next section, we evaluate the considerations about this relation for the introduced data sets and discuss in how far we can distinguish between data with spatial structure, temporal structure, and data exposing none of these structures.

6.4 Results of the Comparison

In section 4, we proposed a graph model for spatial data. As an application of the model, we compare it with data sets in order to evaluate whether the data sets can be categorized by their spatial structure. In particular, we argued in section 6.2 that a data set can only have the properties discussed in section 2 if the two estimates of ρ^m (theorems 5 and 6) are approximately equal. In this section, we compare both estimates for the data sets that we introduced in the previous section.

The estimates[11] of ρ^m for the different data sets can be found in table 1 and figure 2. In case both estimates coincide, the data sets are placed on the diagonal. If a data set has the properties which are discussed in section 2, we expect it to have approximately equal estimate of ρ^m. If a data set has a low number of edges and is only temporal, we expect it to also have approximately equal but lower estimate of ρ^m, because time has one dimension whereas space has more. The estimates are not expected to coincide for data sets without spatial structure, but the values could coincide by chance.

As expected, both estimates and the value ρ^m used during the generation of the model are of similar size for all three SISG models. The differences between the two estimates can be shown to be an effect of the finiteness of the models.

As transport networks have explicit references to space, they have a spatial structure, and both estimates are as expected of similar size, as can be seen in figure 2. The estimates are between about 2 and 4.5. As stops of public transport are placed in two-dimensional space and the density parameter is larger than 1 (but not much larger, because the networks are far from being complete graphs), the estimates are within a reasonable range.

Both estimates approximately coincide for the power grid in the USA, as is expected for a spatial graph. The estimates are between 2.7 and 3.7, which is within a reasonable range compared to the values for transport networks.

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 high 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 high number of long-distance connections, a high number of hubs, a strong hierarchical organization (domestic and intercontinental), communities of strongly related airports, and many more [8].

The formalization of the pizza recipe as well as the game Tic-tac-toe are strongly influenced by time: most steps in pizza baking and all steps in Tic-tac-toe are irreversible, and time therefore induces a partial order on the nodes of the network. As can be seen in figure 2, both estimates are of the same size for each of these examples, which is expected because spatial and temporal data share some typical properties. The estimates for the pizza recipe and for Tic-tac-toe (for a board of size 2x2) are lower than those of transport networks, as is expected because time is one-dimensional (and space usually two- or higher-dimensional). For Tic-tac-toe with a board of size 3x3, there are many more choices in each step (with equal probability) resulting in a much less dominant temporal structure.

Water distribution networks can be embedded in space by the natural locations of the junctions. Both estimates approximately coincide for the water distribution

[11] Note that the estimate by density can slightly differ for each computation because it depends on the random choice of subgraphs.

network of Anytown, as is expected for a spatial graph. The estimates for the water distribution network of the Wolf-Cordera Ranch do not coincide but the values are not very different either. For both data sets, the estimates are between 1.1 and 2.1, which is reasonable due to the existence of a flow direction and the resulting low number of edges compared to the number of nodes.

The estimates for the remaining data sets differ by more than a factor of 1/2.[12] These data sets do not have the typical properties of spatial data that were discussed in section 2, suggesting that neither a spatial nor a temporal structure is *decisive* for these data sets.

We have seen in this section that the argument of section 6.2 is also valid (with minor deviations) for finite SISG models. The comparison of the estimates for different data sets has been shown to provide a meaningful characterization of the examined data sets according to their spatial structure.

7 Conclusion

Information with references to space exposes in many cases some typical properties, including Tobler's law and scale invariance. We introduced the scale-invariant spatial graph (SISG) model and showed that it meets some of the typical properties of spatial data. The model can therefore serve as a model of spatial data.

As an application of the SISG model, we showed how to determine whether a data set shares some typical properties of spatial data with the SISG model, even if the data set does not contain any *explicit* references to space. The evaluation of this consideration showed that spatial and temporal structures could be detected for the examined data sets.

The construction of SISG models introduces edges between two points if their distance is smaller than a certain value, which depends on the distribution of the points in space. This construction is a scale-invariant variant of a model proposed by Huson (equation 1 in section 3). Waxman has introduced a smoothened variant of Huson's model by introducing edges with a probability that depends on the distance between two points (equation 2 in section 3). Future research may, in a similar way, introduce and analyse a smoothened variant of the SISG model.

Properties of the SISG model were analytically computed only for models of infinite size. For finite models, these properties are different, and the influence of the "boundary region" has to be examined. Future research may analytically compute properties of finite models as well as result in algorithms that are taking the influence of the "boundary region" into account, e. g. when testing data sets for spatial structures.

In section 6.2, we discussed two methods that enable us to estimate the value ρ^m for a given data set such that the data set is similar to the uniform SISG model $\mathcal{M}_\rho^m(s)$. Separate estimations of ρ and m would enable a more specific classification of data sets. In particular, it could be examined which

[12] The factor of 1/2 is chosen to visually illustrate how near data sets are depicted to the diagonal in figure 2. This choice is arbitrary and has no relevance for the fact that some data sets are depicted much nearer to the diagonal than others.

influence generalization of data has on spatial structure, and which methods of generalization leave which parameters of the model invariant.

Spatial dependency is only one of the factors that influence the structure of data, and most data sets are characterized by additional aspects. Transport networks, for example, are usually connected (or have very few connected components), but SISG models with low dimension m and low density parameter ρ are in many cases disconnected; the outdegree equals the indegree for most nodes in a transport network; and the number of edges joining a node is usually between 2 and 4 for road networks. Future research may discuss how the SISG model can be modified in order to model specific types of spatial data, e. g. data about public transport or communication, and how other structures of data can be modelled.

Many processes are characterized by a hierarchical system, e. g. in Christaller's central place theory [10]. A combination of SISG models in different hierarchies could be used to model this effect. Such a hierarchical SISG model could be used to model and identify different hierarchies without assuming additional semantics. This would enable us to distinguish between local transport networks and nation-wide ones.

Algorithms are efficient if they take advantage of the data's structure. Future research may show how knowledge gained about the spatial structure of a data set can be used to improve and optimize algorithms, and SISG models, as a prototype for spatial data, could be used to gain insights into how these improvements and optimizations could be carried out. In particular, spatial indexes like R-trees [20] and R*-trees [9] could possibly be generalized to graphs that follow Tobler's law.

References

[1] Aldous, D.J., Shun, J.: Connected spatial networks over random points and a route-length statistic. Statistical Science 25(3), 275–288 (2010)

[2] Archdeacon, D., Bonnington, C.P., Little, C.H.C.: An algebraic characterization of planar graphs. Journal of Graph Theory 19(2), 237–250 (1995)

[3] Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A 311, 590–614 (2002)

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

[5] Barabási, A.L., Albert, R., Jeon, H.: Scale-free characteristics of random networks: the topology of the world-wide web. Physica A 281(1–4), 69–77 (2000)

[6] Barabási, A.L., Ravasz, E., Vicsek, T.: Deterministic scale-free networks. Physica A 299(3–4), 559–564 (2001)

[7] Barthélemy, M.: Crossover from scale-free to spatial networks. Europhysics Letters 63(6), 915–921 (2003)

[8] Barthélemy, M.: Spatial networks. Physics Reports 499(1–3), 1–101 (2011)

[9] Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: an efficient and robust access method for points and rectangles. Proceedings of the International Conference on Management of Data (SIGMOD) p. 322–331 (1990)

[10] Christaller, W.: Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen. Fischer, Jena (1933)

[11] Chrobak, M., Eppstein, D.: Planar orientations with low out-degree and compaction of adjacency matrices. Theoretical Computer Science 86, 243–266 (1991)

[12] Coleman, T.F., Moré, J.J.: Estimation of sparse Jacobian matrices and graph coloring problems. SIAM Journal on Numerical Analysis 20(1), 187–209 (1983)

[13] Colizza, V., Pastor-Satorras, R., Vespignani, A.: Reaction-diffusion processes and metapopulation models in heterogeneous networks. Nature 3, 276–282 (2007)

[14] Denise, A., Vasconcellos, M., Welsh, D.J.A.: The random planar graph. Congressus Numerantium 113, 61–79 (1996)

[15] Erdős, P., Rényi, A.: On random graphs I. Publicationes Mathematicae Debrecen 6, 290–297 (1959)

[16] European Union: 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)

[17] Franklin, C.: An introduction to geographic information systems: linking maps to databases. Database 15(2), 12–21 (1992)
[18] de Fraysseix, H., Rosenstiehl, P.: A depth-first-search characterization of planarity. Annals of Discrete Mathematics 13, 75–80 (1982)
[19] Gilbert, E.N.: Random graphs. Annals of Mathematical Statistics 30(4), 1141–1144 (1959)
[20] Guttman, A.: R-trees: a dynamic index structure for spatial searching. Proceedings of the International Conference on Management of Data (SIGMOD) p. 47–57 (1984)
[21] Hahmann, S., Burghardt, D., 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)
[22] Hecht, B., 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) p. 88–105 (2009)
[23] Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association 76(373), 33–50 (1981)
[24] Hunter, D.R., Goodreau, S.M., Handcock, M.S.: Goodness of fit of social network models. Journal of the American Statistical Association 103(481), 248–258 (2008)
[25] Huson, M.L., Sen, A.: Broadcast scheduling algorithms for radio networks. Proceedings of the Military Communications Conference (MILCOM) 2, 647–651 (1995)
[26] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L.: The large-scale organization of metabolic networks. Nature 407, 651–654 (2000)
[27] Jespersen, S.N., Blumen, A.: Small-world networks: links with long-tailed distributions. Physical Review E 62(5), 6270–6274 (2000)
[28] Kalapala, V., Sanwalani, V., Clauset, A., Moore, C.: Scale invariance in road networks. Physical Review E 73, 026130 (2006)
[29] Kuhn, W.: Core concepts of spatial information for transdisciplinary research. International Journal of Geographical Information Science 26(12), 2267–2276 (2012)
[30] Kuratowski, C.: Sur le problème des courbes gauches en Topologie. Fundamenta Mathematicae 15(1), 271–283 (1930)
[31] Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. Proceedings of the 19th International Conference on World Wide Web (WWW) p. 641–650 (2010)
[32] Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. Proceedings of the 28th Conference on Human Factors in Computing Systems (CHI) p. 1361–1370 (2010)
[33] Lippai, I.: Water system design by optimization: Colorado Springs Utilities case studies. Proceedings of the ASCE Pipeline Division Specialty Conference (Pipelines) p. 1058–1070 (2005)

[34] Louf, R., Roth, C., Barthelemy, M.: Scaling in transportation networks. PLoS ONE 9(7), e102007 (2014)
[35] MacLane, S.: A combinatorial condition for planar graphs. Fundamenta Mathematicae 28(1), 22–31 (1937)
[36] McDiarmid, C.: Random graphs on surfaces. Journal of Combinatorial Theory, Series B 98(4), 778–797 (2008)
[37] McDiarmid, C., Steger, A., Welsh, D.J.A.: Random planar graphs. Journal of Combinatorial Theory, Series B 93(2), 187–205 (2005)
[38] Mocnik, F.-B.: Modelling spatial information. Proceedings of the 1st Vienna Young Scientists Symposium (VSS) (2015)
[39] Ripeanu, M., Foster, I., 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)
[40] Sala, A., Cao, L., Wilson, C., Zablit, R., Zheng, H., Zhao, B.Y.: Measurement-calibrated graph models for social network experiments. Proceedings of the 19th International World Wide Web Conference (WWW) p. 861–870 (2010)
[41] Schnyder, W.: Planar graphs and poset dimension. Order 5(4), 323–343 (1989)
[42] Soon-Hyung, Y., Jeong, H., 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)
[43] Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240 (1970)
[44] Trafiklab: GTFS Sverige. https://www.trafiklab.se, accessed at 2013-04-28 (2013)
[45] 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)
[46] Wagner, K.: Über eine Eigenschaft der ebenen Komplexe. Mathematische Annalen 114(1), 570–590 (1937)
[47] Walski, T.M., Brill, E.D., Gessler, J., Goulter, I.C.: Battle of the network models: epilogue. Journal of Water Resources Planning and Management 113(2), 191–203 (1987)
[48] Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393, 440–442 (1998)
[49] Waxman, B.M.: Routing of multipoint connections. IEEE Journal on Selected Areas in Communications 6(9), 1617–1622 (1988)
[50] 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)
[51] Xulvi-Brunet, R., Sokolov, I.M.: Evolving networks with disadvantaged long-range connections. Physical Review E 66, 026118 (2002)
[52] 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)