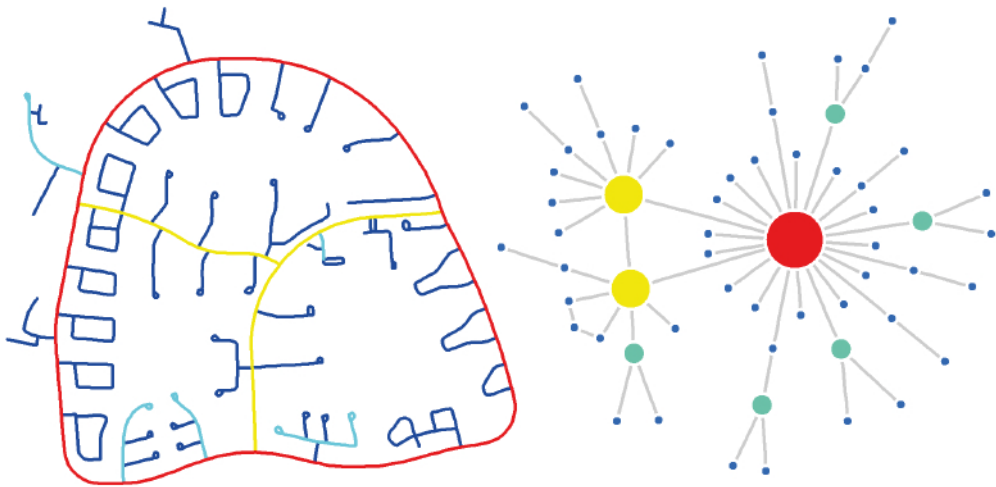


# Topological and Scaling Analysis of Geospatial Big Data

Ding Ma





STUDIES IN THE RESEARCH PROFILE BUILT ENVIRONMENT  
DOCTORAL THESIS NO. 7

# Topological and Scaling Analysis of Geospatial Big Data

Ding Ma




Gävle University Press

© Ding Ma 2018

The front cover illustration describes the transformation of geometric representation into topological representation of Sättra, Gävle. Source: Jiang and Claramunt (2004)

Gävle University Press  
ISBN 978-91-88145-24-6  
ISBN 978-91-88145-25-3 (pdf)  
urn:nbn:se:hig:diva-26197

Distribution:  
University of Gävle  
Faculty of Engineering and Sustainable Development  
SE-801 76 Gävle, Sweden  
+46 26 64 85 00  
[www.hig.se](http://www.hig.se)

 Svanenmärkt trycksak, 3041 0736 Kph Trycksaksbolaget 2018



*To the memory of my grandfathers*  
*Boyuan Ma, 1923 – 2016*  
*Yuanying Luo, 1931 – 2014*



# Abstract

Geographic information science and systems face challenges related to understanding the instinctive heterogeneity of geographic space, since conventional geospatial analysis is mainly founded on Euclidean geometry and Gaussian statistics. This thesis adopts a new paradigm, based on fractal geometry and Paretian statistics for geospatial analysis. The thesis relies on the third definition of fractal geometry: A set or pattern is fractal if the scaling of far more small things than large ones recurs multiple times. Therefore, the terms *fractal* and *scaling* are used interchangeably in this thesis. The new definition of fractal is well-described by Paretian statistics, which is mathematically defined as heavy-tailed distributions. The topology of geographic features is the key prerequisite that enables us to see the fractal or scaling structure of geographic space. In this thesis, *topology* refers to the relationship among meaningful geographic features (such as natural streets and natural cities).

The thesis conducts topological and scaling analyses of geographic space and its involved human activities in the context of geospatial big data. The thesis utilizes the massive volunteered geographic information coming from location-based social media platforms, which are the global OpenStreetMap database and countrywide, geo-referenced tweets and check-in locations. The thesis develops geospatial big-data processing and modeling techniques, and employs complexity science methods, including heavy-tailed distribution detection and head/tail breaks, along with complex network analysis. Head/tail breaks and the induced ht-index are a powerful tool for geospatial big-data analytics and visualization. The derived scaling hierarchies, power-law metrics, and network measures provide quantitative insights into the heterogeneity of geographic space and help us understand how it shapes human activities at city, country, and world scales.

**Keywords:** third definition of fractal, scaling, topology, power law, head/tail breaks, ht-index, complex network, geospatial big data, natural cities, natural streets



# Sammanfattning

Geografisk informationsvetenskap och geografiska informationssystem står inför utmaningar kopplat till förståelsen av det geografiska rummets inneboende heterogenitet, i och med att den konventionella geospaciala analysen huvudsakligen är grundad på euklidisk geometri och Gaussisk statistik. Den här avhandlingen antar ett nytt paradig för geospacial analys baserad på fraktalgeometri och Pareto-statistik. Avhandlingen bygger på fraktalgeometris tredje egenskap: en uppsättning eller ett mönster kommer att uppvisa fraktala egenskaper om flertalet små saker är större, men med samma form, uppträder multipla gånger vid skalförskjutning (s.k. skalning). De två termerna fraktal och skalning används därmed på ett utbytbar sätt i avhandlingen. Den nya definitionen av fraktal kan väl beskrivas av Pareto-statistik, vilket matematiskt definieras som s.k. "tung svans-fördelning" (eng. heavy-tailed distribution). Topologin hos geografiska företeelser fungerar som den viktigaste förutsättningen som möjliggör för oss att se det geografiska rummets fraktal- eller skalningsstruktur. I den här avhandlingen avses topologi som sambandet emellan geografiska formelement av betydelse (t.ex. s.k. naturliga städer och naturliga gator).

Avhandlingen bedriver analyser av topologi och skalning för att undersöka det geografiska rummet och de mänskliga aktiviteter som berörs i kontexten av geospaciala stordata (eng. big data). Avhandlingen nyttjar den enorma mängden frivilligt given geografisk information som härrör från platsbaserade sociala mediaplattformar, som den globala databasen OpenStreetMap, samt landsomfattande georefererade tweets och incheckningsplatser. Avhandlingen utvecklar geospacial stordatabehandling och modelleringstekniker, och tar sig an metoder i komplexitetsvetenskap, vilka inkluderar detektion av tungsvansfördelning och s.k. "huvud/svans-brytpunkter" (eng. head/tail breaks), tillsammans med komplex nätverksanalys. Head/tail breaks och det föranledda ht-indexet är ett kraftfullt verktyg för analys och visualisering av geospaciala stordata. Den erhållna skalningshierarkin, power law-metrika och nätverksåtgärder ger kvantitativa insikter till det geografiska rummets heterogenitet, och hjälper oss att förstå hur det formar mänskliga aktiviteter på olika skalnivåer; såsom stad-, land-, och världsskala.

**Keywords:** tredje fraktalegenskapen, skalning, topologi, power law, head/tail breaks, ht-index, komplexa nätverk, geospaciala stordata, naturliga städer, naturliga gator



# Acknowledgements

It has been five years since I came to Sweden, and how time flies! It is really difficult for me to summarize these years with a few lines, but I would like to express my gratitude to everyone who has supported me.

First and foremost, I would like to thank my primary supervisor, Prof. Bin Jiang. It has been a great honor to be his student. He always kept me on the right track, encouraged me to be creative, and motivated me to overcome difficulties. He was available when I need supervision, and I could easily comprehend theory from his clear instructions and explanations. He taught me the right way of thinking and doing research, from which I can benefit for a lifetime.

I would like to particularly acknowledge Prof. Mats Sandberg, who provided important comments and suggestions during my study, and supported me when I confronted research challenges. I would like to sincerely express my gratitude to Prof. Stefan Seipel and Dr. Anders Brandt, as well as our Geospatial Information Science research group for their guidance and support. I would like to say thanks to Prof. Itzhak Omer, Prof. Toshihiro Osaragi, and Dr. Junjun Yin for their help and guidance for some of my studies, which substantially improved my work. I also appreciate Prof. Lars Harrie, who carefully examined my work at mid-term and gave me very useful comments. Besides, I should say thanks to Prof. Shaowen Wang, Prof. Danny Czamanski, Prof. Thomas Blaschke, Prof. Tiina Sarjakoski, and Prof. Lars Bengtsson for the willingness of being disputation committee members and constructive comments that improve the thesis.

My colleagues in the Division of GIScience at University of Gävle helped and taught me a lot throughout these years. In particular, I would like to thank Jonas Boustedt for his support and encouragement throughout my PhD study. Appreciations go to Eva Sahlin for her kindly help with the Swedish translation. My thanks also go to Markku Pyykönen, Jakob Nobuoka, Nancy Joy Lim, Zheng Ren, and Fei Liu for the time we had together and the insightful discussion on my research work.

My deepest thanks go to my parents for their unconditional love and care. I love them so much! The past five years have not been an easy ride. Thanks to all of my family for sticking by my side.

Gävle, March 2018

Ding Ma





## List of papers

This thesis is based on the following papers, which are referenced in the text by Roman numerals.

### Paper I

**Ma D.**, Sandberg M., and Jiang B. (2015), Characterizing the heterogeneity of the OpenStreetMap data and community, *ISPRS International Journal of Geo-Information*, 4(2), 535–550.

### Paper II

**Ma D.**, Sandberg M., and Jiang B. (2016), A socio-geographic perspective on human activities in social media, *Geographical Analysis*, 49(3), 328–342.

### Paper III

**Ma D.** and Jiang B. (In press), A smooth curve as a fractal under the third definition, *Cartographica*.

### Paper IV

**Ma D.**, Omer I., Osaragi T., and Jiang B. (Submitted), Why topology matters in predicting human activities?

### Paper V

Jiang B. and **Ma D.** (2015), Defining least community as a homogeneous group in complex networks, *Physica A: Statistical Mechanics and its Applications*, 428, 154–160.

### Paper VI

Jiang B. **Ma D.**, Yin J., and Sandberg M. (2016), Spatial distribution of city tweets and their densities, *Geographical Analysis*, 48, 337–351.

### Paper VII

Jiang B. and **Ma D.** (In press), How complex is a fractal? Head/tail breaks and fractional hierarchy, *Journal of Geovisualization and Spatial Analysis*.

Reprints were made with permission from the respective publishers.



## Abbreviations

API	Application Programming Interface
CC	Clustering Coefficient
CDF	Cumulative Distribution Function
CRG	Cumulative Rate of Growth
GIScience	Geographic Information Science
GISystem	Geographic Information System
GoF	Goodness of Fit
GPS	Global Positioning System
JOSM	Java OpenStreetMap Editor
JSON	JavaScript Object Notation
KS	Kolmogorov-Smirnov
LBSM	Location-Based Social Media
LiDAR	Light Detect and Ranging
MLE	Maximum Likelihood Estimation
OSM	OpenStreetMap
PDF	Probability Distribution Function
RA	Ratio of Areas
REST	Representational State Transfer
RS	Remote Sensing
TIGER	Topologically Integrated Geographic Encoding and Referencing
TIN	Triangular Irregular Network
VGI	Volunteered Geographic Information
XML	eXtensible Markup Language



## Glossary of terms

Fractal:	<i>Fractal</i> means broken or irregular shapes. The shapes of geographic features, such as rivers and mountains, are essentially irregular. Therefore, fractal geometry well-describes them. Fractal geometry is self-similar, meaning that a part of the shape is similar to the whole. Based on self-similarity, there are three definitions of fractal: The first refers to strict self-similarity; the second relaxes the first definition to statistical self-similarity; and the third further relaxed self-similarity to the recurrence of a pattern of far more small things than large ones.
Scaling:	<i>Scaling</i> refers to the notion of far more small things than large ones. In other words, if a pattern or phenomenon is scaling, the majority of things must be small, whereas the minority of things are large. According to the third definition of fractal, scaling and fractal are interchangeable in this thesis.
Topology:	<i>Topology</i> focuses on the relationships among meaningful geographic features (such as a continuous natural street, rather than its contained, meaningless segments) and neglects their geometric details (such as location, length, and sinuosity).
Head/tail breaks:	<i>Head/tail breaks</i> define a data-classification scheme with the scaling pattern. It separates the values of data into two imbalanced parts: The head (a minority of values greater than the arithmetic mean) and the tail (the rest of the values). This process can be iteratively applied to the head until the new head is no longer the minority. The derived heads are the resulting classes.
Ht-index:	The <i>ht-index</i> is equal to one plus the number of derived heads during the process of head/tail breaks. It measures how fractal a geographic feature or phenomenon is.
Natural streets:	<i>Natural streets</i> are a collection of individual street segments with good continuity.
Natural cities:	<i>Natural cities</i> refer to spatially clustered human activities, such as agglomerated patches grouped from

night-time image pixels, street blocks and junctions, and social media users' individual tweet and check-in locations.

- Power law: Mathematically, a *power law* indicates the probabilities of a value ( $y$ ) being proportional to some power of a quantity ( $x$ ). It is the most typical member in the heavy-tailed distribution family. If data is a power law, it possesses a very strikingly fractal or scaling property.
- Complex network: A network comprises nodes and links. A *complex network* indicates that the nodes of a network hold a scaling pattern of far more less-connected nodes than well-connected ones, or can connect within a small number of steps.
- Geospatial big data: *Geospatial big data* is a massive, geo-referenced dataset that is hard to store, process, analyze, and visualize in the current geographic information system (GISystem).

# Table of contents

<b>1. Introduction</b>	<b>1</b>
1.1. Background	1
1.2. Research objectives	4
1.3. Thesis organization	4
<b>2. Literature review</b>	<b>7</b>
2.1. Overview	7
2.2. Tobler's Law versus scaling law	7
2.3. Euclidean versus fractal geometry	9
2.4. Geometric versus topological representation	12
2.5. Geospatial small versus big data	18
2.5.1. OpenStreetMap	21
2.5.2. Twitter, Brightkite, and Gowalla	23
<b>3. Experimental design</b>	<b>25</b>
3.1. Overview	25
3.2. The study areas	25
3.3. Data processing	26
3.3.1. Geospatial data extraction	26
3.3.2. Quadtree indexing	29
3.3.3. Data cleaning	31
3.4. Data modeling	31
3.4.1. Natural streets and street blocks from street segments	31
3.4.2. Natural cities from individual locations and street blocks	33
3.4.3. Network construction from human activities	34
<b>4. Methodology</b>	<b>37</b>
4.1. Overview	37
4.2. Mathematical detection of heavy-tailed distributions	38
4.2.1. Different types of heavy-tailed distributions	38
4.2.2. Mathematical detection	38
4.3. Head/tail breaks and its induced ht-index	42
4.3.1. Concept and definitions	42
4.3.2. Applications for geospatial big data	44
4.4. Complex network analysis	44
4.4.1. Network topological measures	45
4.4.2. Community detection using head/tail breaks	47
<b>5. Results and discussion</b>	<b>49</b>
5.1. Overview	49
5.2. Paper I: The heterogeneity of OSM data and community	49
5.3. Paper II: A socio-geographic perspective on human activities	50
5.4. Paper IV: Why topology matters in predicting human activities?	51
5.5. Paper VI: Spatial distribution of city tweets and their densities	53
5.6. Paper III: A smooth curve as a fractal under the third definition	55
5.7. Paper VII: How complex is a fractal?	57
5.8. Paper V: Least community as a homogeneous group in complex networks	59

<b>6. Conclusions and future work</b>	<b>61</b>
6.1. Conclusions	61
6.2. Future work	62
<b>References</b>	<b>65</b>



## List of tables

Table 1.1: Overview of seven papers in this thesis .....	5
Table 2.1: The comparison between Tobler's Law and scaling law .....	9
Table 2.2: The comparison between geospatial small data and big data.....	20
Table 2.3: OSM statistics in November 2017 .....	21
Table 3.1: VGI datasets and their applied study areas .....	26
Table 4.1: The estimated constant $k$ for four heavy-tailed distributions .....	40
Table 5.1: The results of scaling analysis of OSM elements .....	50
Table 5.2: The related metrics of location-location and city-city network .....	51
Table 5.3: The correlation results in central London .....	53
Table 5.4: Power law metrics and ht-index for smooth curves .....	57
Table 5.5: The head/tail breaks statistics of 10 and 15 values respectively .....	58
Table 5.6: Fht-index versus ht-index of a data series using FHTcalculator .....	59
Table 5.7: Scaling analysis of derived communities from 8 networks.....	60



## List of figures

Figure 2.1: Histograms of (a) a city's temperature and (b) population of cities .....	8
Figure 2.2: First and second definitions of fractal .....	11
Figure 2.3: The third definition of fractal .....	12
Figure 2.4: Two representations of the London underground map .....	14
Figure 2.5: Transformation of geometric representation into topological one.....	15
Figure 2.6: Different examples of natural city derivation .....	17
Figure 2.7: Two types of spatial configuration of cities .....	18
Figure 2.8: The overview of OSM components .....	22
Figure 2.9: An example of three OSM data types in XML .....	23
Figure 3.1: The framework of data processing and modeling in this thesis .....	25
Figure 3.2: The study areas in the thesis .....	26
Figure 3.3: The workflow of the data extraction of OSM planet history dump .....	27
Figure 3.4: The nested OSM element structure .....	28
Figure 3.5: Quadtree indexing on the global data set.....	30
Figure 3.6: The workflow for generating natural streets.....	33
Figure 3.7: The workflow for generating natural cities from individual locations ..	34
Figure 3.8: Illustration of a binary network construction.....	35
Figure 3.9: Illustration of a weighted network construction .....	36
Figure 4.1: The overview of the complexity science methodology .....	37
Figure 4.2: The workflow of mathematical detection of heavy-tailed distributions ..	39
Figure 4.3: A working example showing a general power law distribution.....	40
Figure 4.4: Illustration of the process of head/tail breaks using Koch snowflake....	43
Figure 4.5: The overview of the structural analysis of complex networks .....	45
Figure 4.6: Illustration of community-detection algorithm using head/tail breaks ..	48
Figure 5.1: The rank-size plot of degree distribution of co-contribution network ....	50
Figure 5.2: Distributions of connectivity values in different street representations..	52
Figure 5.3: Spatial distribution of tweet numbers and densities in London .....	55
Figure 5.4: Illustration of the new definition of fractal.....	56



# 1. Introduction

## 1.1. Background

Geographic space, or the Earth's surface, encompasses many of the factors that shape how people behave. Research in understanding geographic space from physical and human aspects has become a hot topic in the fields of geographic information science (GIScience) over the past few decades (Langlois 2013). Issues such as the environment, climate change, consumption of resources, and public health have been triggered by the increasingly urbanized world and accelerating socioeconomic development (Knox 1994, Bettencourt and West 2010). Therefore, it is extremely important to study the geographic forms and functions to develop a reasonable, quantitative theory for sustainable development. In this thesis, geographic forms refer to how geographic space looks; that is, the geometry of geographic space at various levels of resolution, including individual geographic features, cities, countries, and the entire world. Geographic functions refer to how geographic space works, indicating the influence of geographic forms on human activities within the space.

Geographic space is inherently heterogeneous and diverse, represented by its containing geographic features regarding their geometric and statistical aspects (Jiang and Yin 2014). Geometrically, the shapes of geographic features, such as mountains and rivers, look neither simple nor regular. Statistically, small geographic features constitute the majority, whereas big geographic features only constitute the minority. One good example of this is city size, which follows Zipf's Law (Zipf 1949): The size of a city is equal to the reciprocal of its rank. In other words, the largest city is twice as big as the second largest city, and so on. Human activities are very complex over the geographic space because it is affected by this inherent heterogeneity (e.g. Penn 2003, Brockmann et al. 2006, Song et al. 2010, Jiang and Jia 2011a, 2012, Osaragi 2013, Omer and Jiang 2015). However, conventional geospatial analysis is insufficient for dealing with such heterogeneity, as it is mainly founded on Euclidean geometry and Gaussian statistics (Jiang 2015a, Jiang and Brandt 2016, Jiang 2016). Euclidean geometry refers to simple, regular shapes such as triangles, rectangles, and circles. Gaussian statistics indicate that things are more or less similar around a well-defined mean and always follow a normal-like distribution. People tend to investigate geographic space and human activities from a local perspective through conventional geospatial analysis. Many things might look regular and more or less similar at a local level. However, geographic space may exhibit great heterogeneity from a holistic or global perspective because there is no average place on the Earth's surface (Goodchild 2004).

To better understand the geographic forms and functions, we must adopt a novel paradigm to carry out in-depth studies that characterize the heterogeneous geographic space and its involved human activities (Jiang 2015b). Jiang (2015b) suggested a new paradigm based on fractal geometry and Paretian statistics to morphologically and statistically better understand such heterogeneity. Fractal means broken or irregular shapes. Fractal geometry, coined by Benoit Mandelbrot, denotes rough, infinitely heterogeneous shapes. The concept of fractal geometry has been adopted to study the urban layout (Batty and Longley 1994, Frankhauser 1994, Chen 2009, 2011, Jiang 2015a). Mandelbrot (1967, 1982, and 2004) used the power-law relationship between the measurement scale and the measured size to describe how fractal geometry works. However, this thesis is not limited to this framework, but relies on a new definition of fractal (Jiang and Yin 2014): *A set or pattern is fractal if the scaling of far more small things than large ones recurs multiple times*. Therefore, the terms fractal and scaling are used interchangeably in this thesis. The new definition of fractal indicates the scaling hierarchy of numerous smallest things, a few largest things, and some things that are between the smallest and the largest things. This definition of fractal is well-described by Paretian statistics. This type of statistics refers to the Pareto distribution: The well-known 80/20 principle (Koch 1998) and long-tail distribution. Mathematically, the Pareto distribution should be defined as heavy-tailed distributions. Typical heavy-tailed distributions are power-law distribution, exponential distribution, and lognormal distribution. Therefore, fractal or scaling structure of geographic space implies heavy-tailed distributions of the geographic features.

Topology of geographic features is the necessary, sufficient condition for us to see the fractal or scaling structure. Topology focuses on the relationships among geographic features, yet neglects geometric details such as location, length, and sinuosity. Unlike topology in conventional GIS systems, which examines how graphic primitives (points, polylines, polygons, and pixels) interconnect, topology in this study refers to the relationship among meaningful geographic features (such as a continuous named street, rather than its contained, meaningless segments or vertices). This refined topology leads us to perceive the fractal or scaling structure of the geographic space. As is the case with the road network at a city or country scale, the street-street topology reveals the underlying scaling pattern of far more less-connected streets than well-connected ones (Jiang and Claramunt 2004). In this connection, topology, and fractal or scaling, link closely with each other and form a theoretical foundation.

The rapid development at the end of the 20th century of geospatial technologies such as satellites, light detection and ranging (LiDAR), and Remote Sensing (RS) have generated very detailed information on the physical aspect of geographic space. Recently, GIScience has considerably benefitted from massive amounts of data from location-based social media (LBSM) such as Twitter, OpenStreetMap (OSM), Gowalla, and Brightkite (Jiang 2013c, Gao and Liu 2014). Data from LBSM has been developed into diverse forms supported by web 2.0 technologies, ranging from check-in locations to various location-embedded media (such as video, photos, and text) and has gradually

become a tool for people to communicate with each other. Technological advancements open up a new horizon to study geographic space because of the notable shift from authoritative, to crowd-sourced data (Crooks et al. 2015), or from geospatial small data to big data in general. We are now at the stage of the fourth scientific paradigm, which is also called data-intensive science (Watts 2007, Gray and Szalay 2007, Bell et al. 2009, Ball 2012). Unlike traditional geospatial data, which was only up to the size of kilobyte or only covered a small scale of geographic space (such as a neighborhood), today's geospatial big data exceeds gigabytes or terabytes, and is country- or worldwide. In OpenStreetMap (OSM), billions of geographic features (nearly all types) across the globe have been created and shared through the Internet. Rather than traditional data, which was contributed by authorities and often at the aggregated level, geospatial big data stores each individual's information at a very fine, spatio-temporal level that leads to enormous observed data about human activities. The wide coverage of the datasets enables people to view geographic space globally, rather than locally. The fine-grained level of data creates new geographic units to organically decompose geographic space from the bottom up, so that people can inspect the diversity of geographic space instead of the homogeneity. Geospatial big data offers a powerful basis for studying the scaling structure of geographic space, which helps further develop new insights into general human activities. The emergence of geospatial big data provides not only an effective means of studying geographic space, but also an invaluable opportunity to shift our paradigm for geospatial analysis.

Through the topological and scaling way of thinking in the big-data context, some intriguing findings have emerged about the structure and patterns of the geographic space. The scaling structure not only describes what the geographic space looks like, but is vital in explaining how it shapes human activities. The most prominent examples are natural cities, which refer to the basic clustering unit of human activities on the Earth's surface, which is represented by social media data (Jiang and Jia 2011b, Jiang and Liu 2012, Jiang and Miao 2015, Long 2016, Long et al. 2016). The extraction of natural cities is neither a subjective view, nor is human intervention included. The derived natural cities exhibit a strikingly scaling pattern of far more small cities than large ones, and the boundary shape of each city is extremely irregular. In this way, natural cities provide a completely new scientific perspective for better understanding urban forms and dynamics. Furthermore, studies have found that a majority of human activities (up to 80 percent) are shaped by the underlying scaling structure of far more less-connected streets than well-connected ones (Jiang 2009a, 2009b). This topological perspective is based on the notion of natural streets that are generated from individual street segments. Eventually, the street-street topological relationship forms the foundation for the scaling analysis. This street-street topology is a *de facto* complex network (Jiang and Claramunt 2004, Gao et al. 2013), bearing not only the scaling property, but many other clustering properties of complex networks such as community structure (Newman 2003, 2010), and assortativity (or disassortativity) (Zhou and Mondragón 2007). These clustering properties, as well as the scaling property, may sub-

stantially shape human activities over geographic space at different scales. Besides, detecting cities' scaling patterns can trigger a sense of beauty in peoples' deep psyche (Jiang and Sui 2014). This beauty is the structural or objective beauty. Therefore, the topological characteristics and scaling pattern of geographic space have far-reaching implications for geospatial research. Given the growing complexity of big data, it becomes increasingly essential to adopt this paradigm to analyze geospatial big data to understand geographic forms and their relationships with human behavior.

## 1.2. Research objectives

Driven by big data, the overall goal of this thesis is to investigate the scaling structure of geographic space and examine how this structure influences human activities at city, country, and world scales. To achieve this goal, it is necessary to follow and adopt a complexity science methodology (Miller and Page 2007, Newman 2011), including both topological and scaling analyses for characterizing the heterogeneity of geographic space and its involved human activities. Additionally, plausible technique solutions for data processing and modeling should be proposed, given the size and complexity of datasets that are difficult for the conventional GISystem to handle.

To achieve the overall goal, the thesis must conduct in-depth research to obtain new insights on the fractal or scaling of geographic space and human-movement behavior from both theoretical and practical perspectives. Therefore, there are three specific aims, as follows:

**Aim A:** Develop new understandings of fractal geometry under the third definition

**Aim B:** Design effective data-processing and modeling techniques to cope with geospatial big-data computing

**Aim C:** Adopt complexity science methods to explore and better understand the heterogeneity of geographic space and how it further shapes human behavior or activities

## 1.3. Thesis organization

The thesis is based on the papers listed below. The Roman numerals will be used to refer to the corresponding papers in the text.

I: Ma D., Sandberg M., and Jiang B. (2015), Characterizing the heterogeneity of the OpenStreetMap data and community, *ISPRS International Journal of Geo-Information*, 4(2), 535–550.

II: Ma D., Sandberg M., and Jiang B. (2016), A socio-geographic perspective on human activities in social media, *Geographical Analysis*, 49(3), 328–342.

III: Ma D. and Jiang B. (In press), A smooth curve as a fractal under the third definition, *Cartographica*.



IV: Ma D., Omer I., Osaragi T., and Jiang B. (Submitted), Why topology matters in predicting human activities?

V: Jiang B. and Ma D. (2015), Defining least community as a homogeneous group in complex networks, *Physica A: Statistical Mechanics and its Applications*, 428, 154–160.

VI: Jiang B., Ma D., Yin J., and Sandberg M. (2016), Spatial distribution of city tweets and their densities, *Geographical Analysis*, 48, 337–351.

VII: Jiang B. and Ma D. (In press), How complex is a fractal? Head/tail breaks and fractional hierarchy, *Journal of Geovisualization and Spatial Analysis*.

Table 1.1: Overview of seven papers in this thesis

	Analysis				Scope				Research focus			
	Fractal or scaling		Topological		World	Countries	Cities	Individual features	Theory		Application	
	Power law metrics	Head/tail breaks (ht-index)	Network measures	Community structure					Fractal geometry	Scaling law	Geographic space	Human activities
Paper I	x	x	x		◊					o	o	o
Paper II		x	x			◊						o
Paper III	x	x						◊	o		o	
Paper IV	x	x	x				◊			o	o	o
Paper V	x	x	x	x						o		
Paper VI		x					◊		o		o	o
Paper VII		x					◊	◊	o		o	

Table 1.1 summarizes these listed papers from three aspects: analysis, scope, and research focus. As the table shows, seven papers conducted both topological and scaling analysis on geospatial big data at four different levels, from local to global. It should be stressed that it is difficult to have a sharp boundary between theory and application when conducting studies. Most studies have an integrated focus on both theory and application. Additionally, head/tail breaks and ht-index are the primary means for data analytics throughout the studies.

Specifically, Paper III reviewed three definitions of fractal and develops a novel perspective on viewing individual smooth curves as a fractal under the third definition through power-law metrics and the ht-index. Paper VII extended the previous integral ht-index to a fractional one, in order to more accurately measure fractal degrees. Therefore, Papers III and VII developed a new understanding of fractal geometry (Aim A) and further consolidated the theoretical framework of this thesis.

A set of techniques were designed and implemented for data processing and modeling (Aim B) to effectively handle the massive amounts of geospatial data from LBSM platforms such as OSM and Twitter. Paper I designed an OSM data-processing workflow, which successfully took out related attributes of more than 2 billion OSM elements and modeled co-contribution relationships, based on roughly 1 million users. Paper II constructed big socio-geographic networks by mapping the social connections of 50,000 users into their 6 million check-in locations. Paper VI obtained natural cities by using millions of street blocks from six European countries.

Based on processed data, implemented geographic representations and constructed networks, the thesis further employs several complexity science methods, including power-law detection, head/tail breaks, and complex network analysis to investigate the scaling property of geographic space and its influence on human-movement behavior (Aim C). Papers I, II, III, V, IV, and VI demonstrated the existence of fractal or scaling structure of the geographic space at individual, urban, national, and global levels. The second issue was covered by Papers I, II, IV, and VI, from which we can see how the spatial scaling structure shapes human activities.

Among seven papers, Ding Ma was responsible for research design, programming, data collection, processing, analysis, and manuscript writing for Papers I, II, III, and IV. Ding Ma contributed to data processing, modeling, and programming for Papers V, VI, and VII.

The remainder of this thesis is organized as follows. Chapter 2 reviews the literature and argues for a new paradigm for geospatial analysis by comparing Tobler's Law and scaling law, Gaussian and Paretian statistics, Euclidean and fractal geometry, geometric and topological representation, and geospatial small and big data. Chapter 3 introduces the datasets and study areas applied in the thesis and details each technique and algorithm concerning data processing and modeling, respectively. Chapter 4 describes the complexity science methods used for geospatial big-data analytics and visualization, including heavy-tailed distribution, mathematical detection, head/tail breaks and ht-index, and complex network analysis. Chapter 5 presents the results from the seven papers and discusses the implications of each study. Finally, Chapter 6 concludes the thesis and points to future work.

## 2. Literature review

### 2.1. Overview

This chapter reviews the literature and argues for a new paradigm for geospatial analysis in the big-data context. To be more specific, the chapter lists fundamental differences in geospatial analysis between Tobler's Law and scaling law, Gaussian and Paretian statistics, Euclidean and fractal geometry, geometric and topological representation, and geospatial small data and big data. These differences are emphasized in order to call for changes in ways of thinking about, and performing, geospatial analysis. To sum up: scaling law, Paretian thinking, fractal geometry, and topology-based geographic representations should be employed for developing in-depth insights and better understanding of the inherent heterogeneity of geographic space and human activities.

### 2.2. Tobler's Law versus scaling law

Tobler's Law (1970) echoes strongly in the geography field (Miller 2004) and has been one of the most important principles in GIScience research. It states: *Everything is related to everything else, but near things are more related than distant things*. Under Tobler's Law, things are more or less similar. More precisely, nearby things tend to be similar and more related to each other. The law further states that spatial phenomena are not random, but auto-correlated, or dependent. From a statistical perspective, the law states that things in geographic space tend to follow a Gaussian-like distribution, which is well-known as a 50/50 principle and shaped as a bell curve. Following a Gaussian-like distribution, things can be characterized by using a well-defined mean and a very small deviation.

Tobler's Law denotes a fact in geographic space, but only appears locally or at local scales (Sui 2004). As is the case with housing prices, spatial auto-correlation can effectively characterize the housing market of a local area (Dubin 1988). The market for geographically adjacent houses behaves in a homogeneous manner, indicating that prices are more or less similar. At a larger scale, such as at a city or country scale, whether or not this law holds become questionable. Specifically, the housing prices at a city or country scale would likely to show that the majority of housing prices are low, while only a few prices are very high. The majority and minority underlie the Paretian-like distribution, which is well-known as the 80/20 principle and long-tail distribution. The values of a long-tail distribution do not center on the mean. In this regard, the mean value and deviation are no longer effective.

The Paretian-like distribution describes a global fact that is the existence of scaling or scaling hierarchy over the geographic space. Geographic space, or the Earth's surface, cannot be described using the concept of an average place (Goodchild 2004). Scaling or scaling hierarchy denotes that: *There are far more small things than large ones across all scales, ranging from the smallest to the largest*. Numerous studies have found that geographic features follow this type of uneven distribution. For example, there are far more small bends

than large ones in an individual coastline feature (Jiang et al. 2013, Paper III); far more short, poorly connected streets/axial lines than long, well-connected ones (Kalapala et al. 2006, Jiang 2007, Jiang and Liu 2009, Jiang 2013b, Paper IV); far more small street blocks than large ones (Jiang and Liu 2012, Lämmer et al. 2006, Paper VI); and far more small cities than large ones (Zipf 1949, Jiang and Miao 2015, Jiang 2016, Paper II).

Let us further examine statistical difference between Gaussian- and Paretian-like distributions using two variables: temperature of Stockholm city and population of Swedish cities. The former is based on raw individual temperature observations from 2013 to 2016 at a daily basis (Bolin Centre Database 2018), while the latter is according to the statistics from world population review (2018). As shown in the histograms (Figure 2.1), temperature well-obeyes the Gaussian-like distribution since its values center around the mean (8.4), however, the population follows Paretian-like distribution where the values are highly right-skewed, indicating that the mean value (29,721) makes little sense for description of urban population in Sweden. The difference between Gaussian- and Paretian-like distributions also lies on the lengths on their “tails”. Gaussian-like distributions possess short or light tails, and the values at the tail drop exponentially. Paretian-like distributions have long or heavy tails, and the values at the tail continuously approach zero but never reach it.

Even more, Gaussian- and Paretian-like distributions fundamentally differ in their ways of thinking. McKelvey and Andriani (2005) noted that the Gaussian way of thinking treats the world as simple, static, and equilibrium, while the Paretian mindset views the world as complex, dynamic, and non-equilibrium. Since many studies have validated that geographic phenomena are the result of complex processes through interactions and interdependence among human and geographic features, rather than just a simple series of actions (e.g., Benguigui and Czamanski 2004, Jiang and Jia 2011a), Paretian thinking should be very suitable for characterizing the heterogeneity of geographic space.

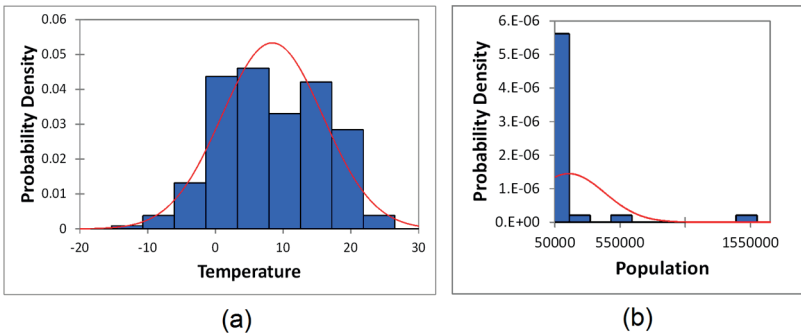


Figure 2.1: (Color online) Histograms of (a) a city's temperature and (b) population of cities

(Note: The city's temperature follows a Gaussian-like distribution, whereas population of cities follows a Paretian-like distribution.)

Scaling is not limited to geographic studies, but has long resonated strongly in many other sciences, including biology (e.g. Jungers 1984) and physics (e.g.

Bak 1996, Brockmann et al. 2006, Song et al. 2010, Chen 2015). Furthermore, numerous laws in different disciplines are rooted in scaling, such as Zipf’s Law (1949) for city sizes, Pareto’s Law for people’s wealth, Korcak’s Law for sizes of lakes and islands, Gutenberg-Richter’s Law for earthquakes, and Benford’s Law for anomalous numbers. In line with these examples, scaling should be formulated as another law in geography and deserves attention equal to Tobler’s Law. However, scaling law, or Paretian thinking, has not yet been sufficiently embraced for geospatial analysis (Jiang 2015b).

Table 2.1: The comparison between Tobler’s Law and scaling law  
(Source: Jiang 2015b)

<b>Tobler’s law</b>	<b>Scaling law</b>
Local	Global
Short-tailed	Long-tailed
Equilibrium	Non-equilibrium
Spatial homogeneity	Spatial heterogeneity
Gaussian statistics (50/50)	Paretian statistics (80/20)

Tobler’s Law and scaling law complement each other. Table 2.1 summarizes the comparisons between Tobler’s Law and scaling law. Tobler’s Law portrays geographic space at the local scale, and scaling law depicts the space across all scales. The former describes things at the local extent, while the latter looks at them globally. Tobler’s Law focuses on homogeneity, while scaling law focuses on on heterogeneity. Tobler’s Law is characterized by Gaussian statistics (50/50), while scaling law is characterized by Paretian statistics (80/20). Through these connections, we must utilize Tobler’s Law and scaling law together to more comprehensively understand geographic space. Section 2.3 will illustrate the association of two laws with Euclidean and fractal geometry, respectively.

### 2.3. Euclidean versus fractal geometry

Euclidean geometry was founded approximately 2,000 years ago. It is the basis of geometry math, through which people understand its theories and axioms that are used to measure objects with different dimensionalities, such as points (one-dimensional), lines (two-dimensional), and surfaces (three-dimensional). In GIScience, or geography in general, Euclidean geometry is also the most important tool for gauging and describing geographic space. One of its most important applications is projecting the surface of the Earth globe into 2D maps.

Euclidean geometry uses geometric primitives, such as points, polylines, and polygons to describe geographic features. Using an example of a cartographic line, the elements of Euclidean geometry include length, orientation, and sinuosity. In this way, Euclidean geometry concentrates on the measurements of geographic features and its effectiveness at this was universally acknowledged until the second half of the 20<sup>th</sup> century. Scientists gradually

realized that Euclidean geometry has limitations in describing geographic features or phenomena, since they are neither regular nor simple, but inherently irregular and heterogeneous. Mandelbrot (1982) claimed that Euclidean geometry treats shapes and patterns of nature in a simple manner: *Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line.*

Mandelbrot (1967) coined the term *fractal geometry* for irregular and complex shapes in nature. The word *fractals* originally comes from Latin word *fractus*, meaning *broken*. The Koch curve is a classic example of a fractal curve (Von Koch 1904). The Koch curve starts from one straight line with a scale (length) of 1, which is then replaced by four segments of an equal scale of 1/3. Each segment is further substituted by four subsequent segments of an equal scale of 1/9, and so on and so forth. This evolution from a simple, regular shape to a complex, irregular one can be deemed as the first definition of a fractal.

The essence of the Koch curve lies in its self-similarity. In other words, the shape of a part can be similar to that of a whole (Irving and Segerman 2013). Because of its self-similarity, the Koch curve has no characteristic length. This curve is not measurable because the total length is always increasing as there are more and more segments of smaller scales. Although such self-similarity cannot be measured, it can be characterized. Mandelbrot (1967) introduced the concept of *fractal dimension* to describe the self-similarity in a quantitative manner. To be specific, fractal dimension refers to the exponent of a power-law relationship between measurement scales and details (number of segments; Figure 2.2). For example, the scale ( $x$ ) of the Koch curve decreases at a power of 1/3, and the number of segments ( $y$ ) increases at a power of 4, so these two variables would constitute a power-law relationship:  $y = x^{-1.26}$ . The exponent is fractal dimension.

Under the first definition of a fractal, two variables strictly follow the equation for the fractal construction. Mandelbrot (1967) noted that there is no need to follow such a strict way of fractal construction, but added some randomness to both scales ( $1 + \varepsilon_1$ ,  $1/3 + \varepsilon_2 \dots$ ) and a number of segments ( $1 + d_1$ ,  $4 + d_2 \dots$ ), as well as maintained the same power-law relationship (Figures 2.2b and 2.2d). In this way, a Koch curve can resemble the British coastline. This is the second definition of fractal, which extends or relaxes the concept of fractal from strict to statistical. The second definition grew in popularity because it helped address the famous question: “How long is a coastline,” known as *the coastline paradox* (Richardson 1961, Steinhaus 1983).

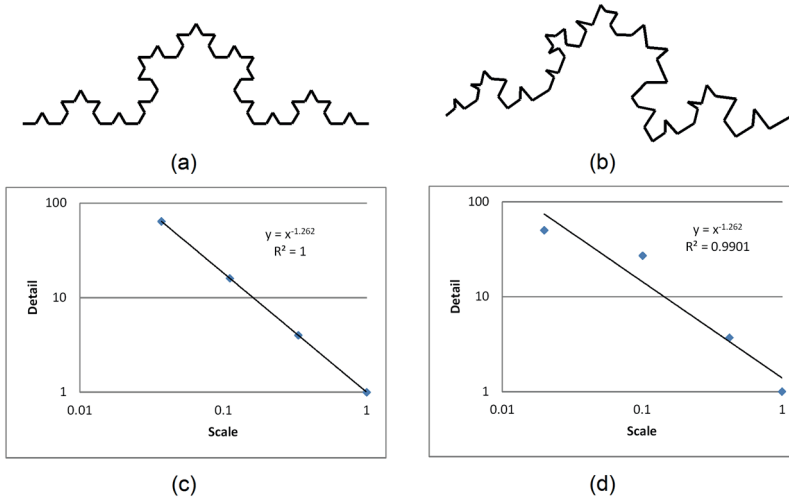


Figure 2.2: (Color online) First and second definitions of fractal  
 (Note: The plots use Koch curve as an example in relation to strict self-similarity (the left column) and statistical self-similarity (the right column))

However, the second definition may still confine us when characterizing how a geographic feature is fractal. To a certain extent, the power-law relationship between scales and details is an idealistic model to fit with real-world fractals. Additionally, a power-law relationship could be too difficult for a fractal pattern at earlier stages. In this respect, Jiang and Yin (2014) further relaxed the second definition. Instead of a power-law relationship between scales and details, they proposed that: *A set or pattern is fractal if the scaling of far more small things than large ones recurs multiple times*. This is further recognized as the third definition of fractal (Jiang 2015, Gao et al. 2017). Fractal and scaling are then interchangeable according to the third definition.

This new, third definition (Jiang and Yin 2014) provides a holistic view of all scales of the Koch curve during its evolution and introduces a new metric ht-index for fractal characterization. The ht-index is derived based on head/tail breaks (Jiang 2013a), which is a novel classification scheme for data with a heavy-tailed distribution (see Chapter 4). To illustrate, as shown in Figure 2.3a, the Koch curve at phase three has 21 segments (1 + 4 + 16). The head/tail breaks then recursively use defined mean values of segment-length to partition those segments into heads (lengths above or equal to the mean) and tails (lengths below the mean), through which we can know how many times the fractal or scaling pattern occurs (Figure 2.3b). The number of occurring times plus one is the value of the ht-index. In this working example, the ht-index is 3, since the scaling pattern occurred twice. The higher the ht-index value, the more fractal a feature will be. In addition, the ht-index can help characterize the degrees of fractal at different phases of fractal construction or development. In Figure 2.3, the ht-index at the second iteration is smaller than that at the third iteration. In this connection, this new definition is from the bottom-up,

whereas the first and second definitions are top-down. Top-down means the fractal construction moves from a simple Euclidean shape to a complex, irregular one by iteratively following a defined rule iteratively, such as a Koch curve.

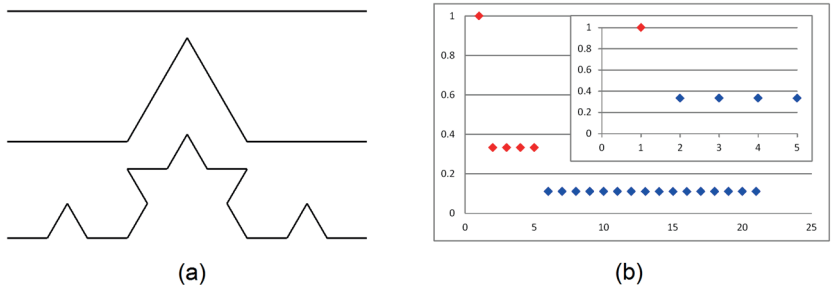


Figure 2.3: (Color online) The third definition of fractal  
 (Note: Three iterations of a Koch curve (a) and the nested, rank-size plot (b) showing two iterations of the recurring pattern of far more small segments (blue points) than large segments (red points))

It is interesting to note that these three definitions of fractal have inheritance relationships. As illustrated, the second definition is relaxed from the first, and the third definition is further relaxed from the second. The first definition refers to the classic fractal, under which only a strict mathematical model, such as a Koch curve, is fractal. Under the second definition, the strict model changes to statistical one (Cattani and Ciancio 2016), so that both the Koch curve and a coastline are fractal. The third definition takes other long-tail statistics into consideration. Under the third definition, other regular curves, whose geometric shapes were previously considered Euclidean (such as a highway) are fractal.

The third definition brings the scaling law or Paretian statistics into geometry and helps us shift our thinking of geographic features from Euclidean to fractal geometric. The comparison between Euclidean and fractal geometry is very much in line with Tobler’s Law and scaling law, and Gaussian distribution and Paretian distribution. The third definition recognizes that the shape of a geographic feature comprises far more small things than large ones. This is different from Euclidean geometric thinking, which decomposes the shape into more or less similar geometric primitives. It should be stressed that, under the third definition, all geographic features are fractal, given the right scope and perspective. Section 2.2 demonstrated that a bigger scope (such as a country or city, rather than neighborhood) is essential for seeing fractal. Section 2.4 will illustrate that topology is the right perspective for seeing the fractal or scaling structure of geographic space.

### 2.4. Geometric versus topological representation

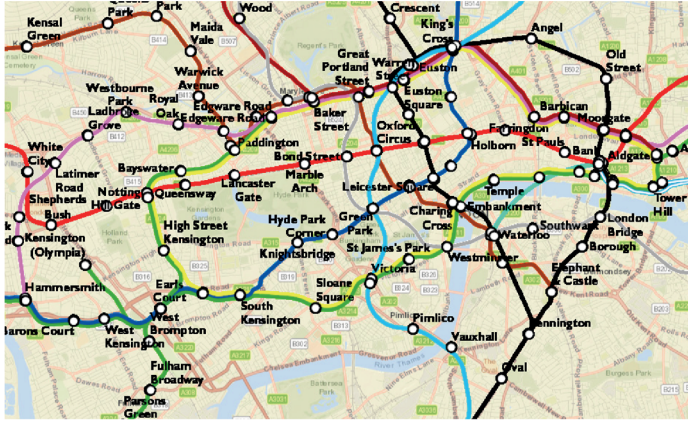
Geographic data representation is vital for both GISystem and GIScience because it stores and displays the geographic features or real-world entities in a computer. Current GIS mainly relies on geometric representation for the spa-



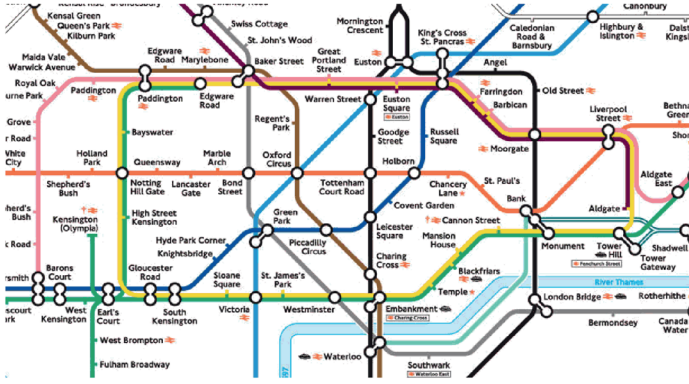
tial component of geographic features. The most common geometric representations are vector and raster, which contain different types of geometric elements, such as points, polylines, polygons, and pixels (Longley et al. 2015). In the past few decades, people have made extensive use of geometric representations to describe and analyze geometric details of different geographic features. The counterpart of geometric representation is topological representation. Topology in mathematics refers to the properties that remain unchanged under the distortion of geometric space. In the GIS literature (Corbett 1979, Egenhofer and Herring 1990), topology is employed for topological data structures and cartograms. Without paying attention to geometric details (such as lengths and angles), topological representation only stores the relationships (0 and 1).

The lack of geometric information in a topological representation of geographic information does not lose information on a map, but makes it more informative. One clear example is the London Underground map, created in 1933 by Harry Beck. Figure 2.4 shows both geometric and topological representations of the London Underground map. The geometric map keeps all the tube lines' geometric details correct, as well as other geographic features, such as the Thames River and residential areas. The topological map distorts the geometric shape of the tube lines and ignores other geographic features, but retains the intersectional relationships among tube lines. It is apparent that the topological representation conveys more information on route plans for travelers. The reason why the topological representation is effective is that it reveals the structure of the tube system by omitting unrelated geometric aspects.

However, this kind of topological representation is still not enough to see a more in-depth structure because the adopted topology is established at the geometric level. More specifically, such a topological relationship is built upon individual geometric primitives, such as points and lines, which only contain metric information such as location and length. Furthermore, geometric primitives are mechanistic, as they are only the spatial component of geographic features, but have little meaning in our perception. In the most well-known GIS data structure, TIGER (topologically integrated geographic encoding and referencing), topological relationships among geometric primitives (such as disjoint, within, overlap, and covers) are used to organize geospatial data (Egenhofer and Herring 1990). Additionally, the type of topology is also used for geospatial analysis. For example, a street network in the ArcGIS network analysis extension (ESRI 2017c) is constructed using a node-node or segment-segment topology, in which a single line must be separated at a street junction node (Figure 2.5a). It is a good model for calculating distances in navigation, but it does not capture street-street topology, so lacks the ability to reveal the underlying street structure.



(a)



(b)

Figure 2.4: (Color online) Two representations of the London underground map (Note: (a) Traditional map without geometric distortion and (b) Topological relationship of stations with relatively distorted geometric details)

To establish street-street topology, we must transform the individual, meaningless segments to meaningful streets. Jiang et al. (2008) proposed the concept of natural streets as joined street segments that have good continuity based on the Gestalt principle. There are three principles for joining the street segments: Every-best-fit, self-best-fit, and self-fit. Three principles represent three choices of a small deflection angle when a segment meets other candidate segments at a junction node. Axwoman 6.3 (Jiang 2015e) is a research prototype for automatic generation of natural streets. The resulting natural streets, compared to street segments, possess a more organic structure. Both lengths and connectivities among segments are more or less similar, but those among nat-

ural streets are rather different. Their connectivity graph exhibits a scaling pattern of far more less-connected streets than well-connected ones. As we can see in Figure 2.5, the street-street topology (without geometric details such as locations, lengths, and directions) can help us detect the scaling structure.

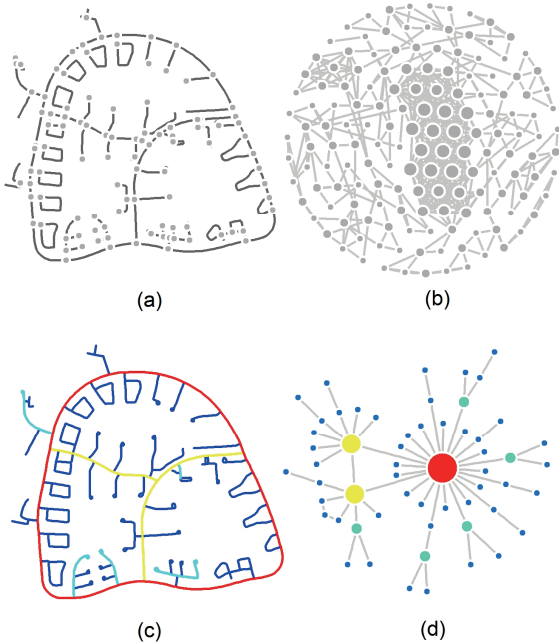


Figure 2.5: (Color online) Transformation of geometric representation into topological one (Source: Jiang and Claramunt 2004)

(Note: Geometric representation is represented by street segments (a) and segment-segment connectivity graph (b), whereas topological representation is represented by natural streets (c) and street-street connectivity graph (d))

This kind of topology is the true topology for geospatial analysis (Jiang et al. 2008) because it considers spatially coherent entities, such as morphologically continuous streets (rather than segments), as the basic unit (Jiang and Ren 2018) and enables us to perceive the scaling structure of geographic space. Such topology is fundamentally different from the existing topology applied in GIS. As mentioned, conventional GIS views geographic space mechanically through geometry-based geographic representations. Unlike conventional GIS, the topological representations apply the introduced topology to better understand geographic forms or city structures, and their nonlinearity (Jiang 2015d).

The structure of urban space is complex and hard to directly understand (Brelsford et al. 2015). Topological representations are effective means for us to overcome this difficulty. Street-based topological representation depicts mostly the structure inside the city. To view the structure among cities, another topological representation is developed, which takes cities in a country as a

coherent whole (Jiang 2018). This representation enables us to conduct a country-wide, spatial, configurational analysis so we can clearly examine how individual cities relate to each other both visually and statistically for better understanding cities' structure. Before illustrating how this topological representation works, it is important to introduce the concept of *natural cities*, which are the basic unit for the representation.

A city can be regarded as a large area with a high concentration of human settlements (Lynch 1960). However, this definition does not give any qualification or criterion for determining what a city should look like. Traditionally, the boundary of a city is decided by local authorities or administration (Eeckhout 2004). This leads to the notion that cities vary tremendously from one country to another. For example, a middle-sized city in Sweden is not comparable to a town in China. The conventional way of delimiting a city area is not very objective or natural. Therefore, an interesting research question arises: Can the areas of cities or urban spaces be naturally and objectively defined (regardless of administrative boundaries) across a country, or even all over the world?

Using massive geospatial datasets, we applied the head/tail division rule to successfully derive cities that objectively and naturally represent human activities or settlements (Jiang and Jia 2011b, Jiang and Liu 2012, Jiang and Miao 2015, Jiang et al. 2015). These derived cities are so-called *natural cities*. The head/tail division rule is: *Given a variable X, if its values x follow a heavy-tailed distribution, then the mean (m) of the values can divide all the values into two parts: A high percentage in the tail, and a low percentage in the head.* To elaborate how the head/tail division rule works for obtaining natural cities, Figure 2.6 presents four examples related to previous studies. Jiang and Jia (2011b) used the nearest-neighborhood algorithm to cluster 7 million street nodes and grouped nearby nodes into natural cities. Jiang and Miao (2015) constructed a huge TIN model of 3 million check-in locations and converted short TIN edges (lengths smaller than the average length of all edges) into city patches.<sup>1</sup> Jiang et al. (2015) used the mean pixel value as the cutoff value to delineate natural cities' boundaries. Jiang and Liu (2012) extracted natural cities by clustering small blocks (block sizes smaller than the mean of all block sizes). There was a universal mean effect for extracting the natural cities. Moreover, all obtained natural cities at either national or cross-national levels possessed striking scaling property of far more small cities than large ones.

---

<sup>1</sup> An online video on using a TIN-based method to extract natural cities:  
<https://www.youtube.com/watch?v=DzeDFULHaEs>

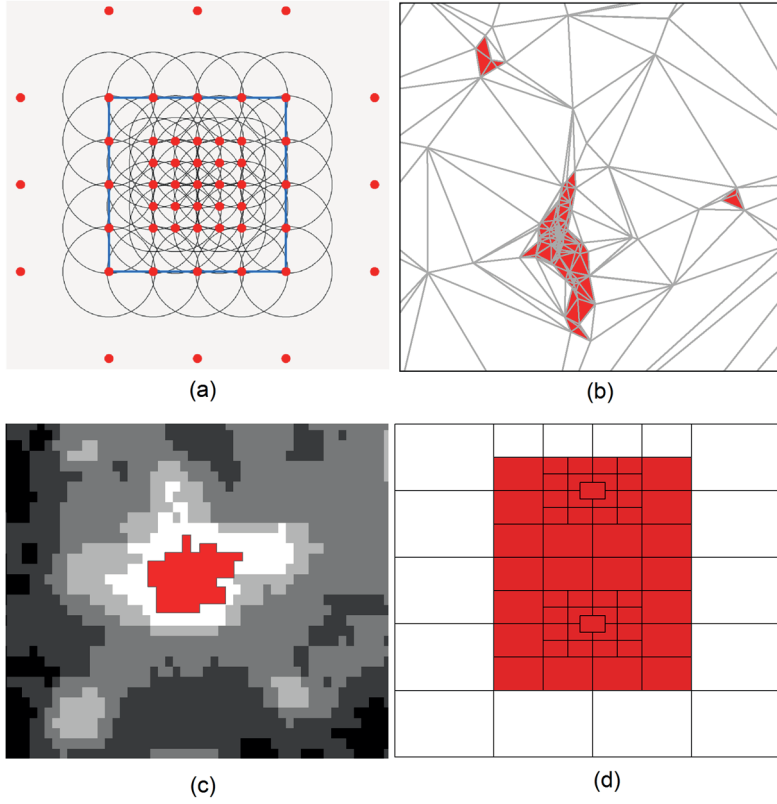


Figure 2.6: (Color online) Different examples of natural city derivation  
 (Note: The natural city can be extracted from (a) street nodes (Jiang and Jia 2011b), (b) check-in locations (Jiang and Miao 2015), (c) night light image (Jiang et al. 2015), and (d) street blocks (Jiang and Liu 2012))

The topological representation is built upon the constructed natural cities and inspired by central place theory (CPT, Christaller 1933). CPT describes a hierarchical urban layout in which a central place is surrounded hexagonally by its next lowest-level place, and so on (Figure 2.7a). This representation puts theory into practice. To do so, the hierarchical levels of cities are first obtained through the head/tail breaks method (Jiang 2013a, 2015a). Thiessen polygons are then created, according to cities' locations and hierarchical levels, respectively (Figure 2.7b). Polygon-polygon relationships are then used to construct a complex network. There are two types of polygon-polygon relationships: Small polygons that point to large ones at the same level; and contained polygons that point to containing polygons between two consecutive levels. The network model can help detect how cities are spatially adaptive with each other, according to their sizes. The model can also be further applied to understand how human activities are shaped by space in the big-data context (Jiang and Ren 2018).

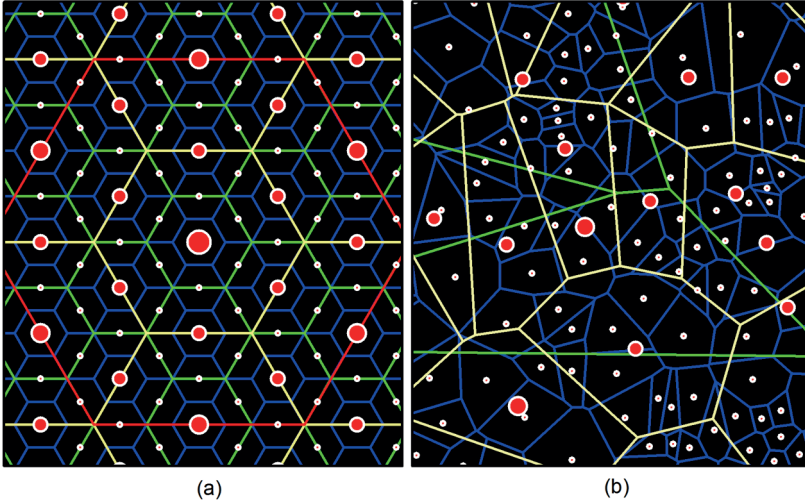


Figure 2.7: (Color online) Two types of spatial configuration of cities  
(Note: (a) CPT and (b) topological representation)

Geographic space inherently bears a diverse and heterogeneous structure. Geographic representation is designed and implemented for geospatial analysis over the geographic space to give us deep insights into geographic entities and phenomena. Conventional geographic representations, such as raster and vector, are good for modeling the geometric aspects of geographic objects and phenomena, but fall short of representing geographically meaningful features. Therefore, such models fail to reveal the relationships between every meaningful feature to every other one. This drawback largely prevents us from seeing the underlying spatial heterogeneity – the scaling structure of far more small things than large ones – of geographic space.

We have entered the geospatial big-data era, in which geographic data has been disruptively changed. Massive geographic datasets with rich attributes are generated at a very fast rate and freely available (see Section 2.5 for more details). On the one hand, geospatial analysis is confronting the big challenges of dealing with such data using basic geographic representations. On the other hand, big data offers an unprecedented opportunity to comprehensively investigate the diverse, heterogeneous geographic space and involved human activities. Triggered by geospatial big data, topological representations, based on natural streets and natural cities, will be very useful for developing new insights into geographic forms and functions, and better understanding the inter-relationship between space and society.

## 2.5. Geospatial small versus big data

In the 20th century, people were accustomed to relying on small data for analysis. Small data literally means *data with a limited volume and that is accessible, informative, and actionable* (TechTarget 2017). In the mid-1990s,

the term *big data* was just a fresh idea in the scientific communities. About five years ago, the concept of big data went viral and drew worldwide attention. Nowadays, big data has become part of our life, since we experience and contribute big data on a daily basis through mobile sensors, messaging applications, and social-network software. There is no clear-cut boundary of how to define big data. It cuts across many fields such as technology, industry, and academics (Chen et al. 2014). Generally, big data refers to large, diversely sourced, unstructured datasets that are difficult to manage by conventional technologies (Li et al. 2016). Basically, big data can be characterized by four Vs: Volume, variety, velocity, and veracity (Lazer et al. 2009). *Volume* refers to the vast amount of data; *variety* refers to diverse formats and sources of data; *velocity* refers to the speed of data generation and processing; and *veracity* refers to data uncertainty, integrity, and quality. There are many other Vs to make understanding big data more comprehensive such as *value*, *validity*, and *visibility*.

Geospatial data is data that is geo-referenced with *xy* coordinates in a spatial referencing system. Geospatial data was previously collected through ground surveying, supported by GPS, RS, photogrammetry, and LiDAR. Now there is another big repository of geospatial data in LBSM, supported by the Internet and Web 2.0 technology. The rapid growth of geospatial data in both size and diversity lets us step into the geospatial big-data era. We can use the four Vs to specifically characterize geospatial big data. *Volume* can now refer to gigatype, terabyte, petabyte, or even exabyte for geospatial datasets. *Variety* means not only the numerous types of geospatial datasets (such as remotely sensing imagery data and LBSM data), but also complex structures (Li et al. 2016). *Velocity* refers to fast geospatial data generation and processing via the Internet (Dasgupta 2013). Finally, *veracity* means varying degrees of quality of geospatial data from diverse sources.

The most prominent type of geospatial big data is volunteered geographic information (VGI; Goodchild 2007). This refers to geographic information created, assembled, and voluntarily disseminated by individual people through the web. VGI has developed into diverse forms, ranging from check-in locations, to geo-tagged web content (videos, photos, and texts), to online mapping. One of the most famous VGI projects is OpenStreetMap, in which the site receives a large amount of geo-related data creation and edits from users all over the world. It is now the most popular VGI platform, with billions of geographic elements. In recent years, the widespread use of social media, such as Twitter and Facebook (Boyd and Ellison 2008), enables us to create enormous amount of location-related information from hundreds of millions of users (Kaplan and Haenlein 2010, Cho et al. 2011, Sui and Goodchild 2011, Ferrari et al. 2011, Wakamiya et al. 2011, Takhteyev et al. 2012, Kulshrestha et al. 2012, Cranshaw et al. 2012, Hawelka et al. 2014, Li et al. 2014). VGI has the potential to make great contributions to social decision-making and problem-solving, such as disaster surveillance and response (Zook et al. 2010), transportation planning (Wang et al. 2018), and environmental monitoring (Gouveia and Fonseca 2008).

In essence, the difference between geospatial big data and small data is much more than size. Table 2.2 lists five aspects, three of which are fundamental (Mayer-Schonberger and Cukier 2013, Jiang and Thill 2015). For more details, given a big population, big data refers to the whole population while small data refers to the sampled part. Big data is measured with timestamps and coordinates, while small data estimates these things. Big data is collected at an individual level, while small data is collected at an aggregated level. These three characteristics make big data more capable than small data of reflecting the diversity and heterogeneity of the Earth’s surface. Therefore, geospatial data analytics must be updated accordingly, in both geometric and statistical aspects. As mentioned in Sections 2.2 and 2.3, Gaussian statistics and Euclidean geometry were widely adopted for conventional geospatial analysis in the small-data era. In the context of geospatial big data, fractal geometry (especially the third definition) should be employed for understanding the complexity of geographic forms (Jiang 2015). Statistically, the heterogeneous geographic space is likely to bear a scale-free or scaling effect and exhibit a long-tail distribution, which is better captured by Paretian statistics. Big data is a new paradigm for geospatial analysis, which fundamentally differs from analysis in the small-data era (Hey et al. 2009, Jiang and Thill 2015).

Table 2.2: The comparison between geospatial small data and big data  
(Source: Jiang and Thill 2015)

<b>Small data</b>	<b>Big data</b>
Simple	Complex
Structured	Unstructured
Sampled	All
Estimated	Measured
Aggregated	Individual

Handling geospatial big data is another important issue that requires broad scientific and technological advances. With respect to underlying data-intensive computing (Yang et al. 2011b, Jiang 2013c, Hey et al. 2009), Apache Hadoop (Apache 2017a, White 2012) is one of the most popular open-source software packages for scalable, distributed computing from a single desktop to thousands of computers. It contains two major components: MapReduce and Hadoop Distributed File System (HDFS). MapReduce works as a parallel data-processing and computing paradigm for big data and has been used by Google (Maitrey and Jha 2015). HDFS is a distributed file system written in Java and comprises a NameNode and multiple DataNodes for distributed storage of Hadoop applications (Apache 2017b). Emerging cloud computing and CyberGIS (Wang 2010, Wright and Wang 2011, Wang 2013), which is a synthesis of cyberinfrastructure, GIScience, and spatial analysis and modeling, offer a promising direction of high-performance and distributed computing for knowledge discovery, collaborative problem-solving and decision-making. There are also a large number of studies in big-data analytics and visualization, quality assessment, and other topics. Regarding big-data analytics and visualization, existing geometry-based data models for raster and vector are insufficient. The topological models developed based on natural streets or



natural cities capture or predict human activities in the big-data context (Jiang and Ren 2018, Paper IV). Cheshire and Batty (2012) visualized the transportation network in reference to individual and collective travel trips using the London Oyster and provided valuable information to planners for optimizing travel scheduling. For quality assessment, Neis et al. (2012) studied the completeness of the street-network data in OSM and found that OSM data differs significantly between city and rural areas. The next part of this section briefly introduces some of the most well-known LBSMs and describes the characteristics of geospatial big data from these platforms.

### 2.5.1. OpenStreetMap

OpenStreetMap (OSM) is the most famous and successful VGI platform in the world. It follows the philosophy of Wikipedia by letting users freely create and edit geographic objects on the platform and collaborate with each other, using their own knowledge. The OSM project was started in August 2004 by Steve Coast at University College London (Bennett 2010). From that point until now, there has been an enormous surge in the amount of users and geographic elements. During this period, the OSM community also received continuous help from different organizations. Yahoo! donated digital images in December 2006 to facilitate direct mapping. In April 2007, Automotive Navigation Data donated a complete dataset of Dutch roads. In October 2007, the US Census TIGER road dataset was transformed into the OSM database. This data is freely available for anyone, without restriction. As Table 1 shows, there were more than 4.6 billion geographic elements contributed by 4.35 million registered users in OSM’s global database, as of November 2017 (OSM 2017).

Table 2.3: OSM statistics in November 2017  
(Note: # = number)

# of users	4,355,579
# of uploaded GPS points	5,938,674,424
# of nodes	4,176,936,518
# of ways	451,452,429
# of relations	5,363,304

OSM is essentially not just a map that visualizes geographic features, but more importantly, is also a system with advanced architecture. The OSM system has five general components: Geodata, map editing, backend (database), map rendering, and visualization (Figure 2.8). Popular map-editing applications are Potlatch, Java OpenStreetMap Editor (JOSM), Merkaartor, and plugins in GIS software such as ArcGIS and QGIS. The most often used applications are Potlatch and JOSM. The former is a web-based online editor for beginners, written in Flash. The latter is a desktop application offering features and tools for professional editing styles, and targeted at advanced users. Map-rendering applications are categorized into either server- or client-side. The most popular server-side map renderer is Mapnik, which needs PostgreSQL and C++ as prerequisites. Mapnik aims for fast generating, high-quality map tiles at high-end servers. On the client side, users can also choose applications to make on-the-

fly renderings in 2D or 3D from their own OSM data, as is the case with Ken-  
dzi3d, Kosmtik, and TileMill. Many of these client-side tools are based on  
Mapnik. The following content will focus on the database part.

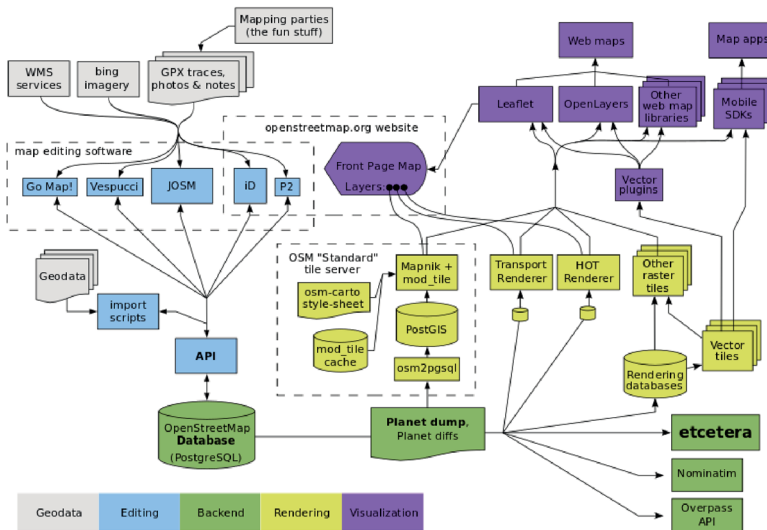


Figure 2.8: (Color online) The overview of OSM components  
(Source: [http://wiki.openstreetmap.org/wiki/Component\\_overview](http://wiki.openstreetmap.org/wiki/Component_overview))

The database controls terabytes of dynamic geographic data. Furthermore, the schema of the database must maintain and support wiki-like operations by allowing users to edit a geographic object and track the edit history on the object (Hakley 2008). The database is hosted by a PostgreSQL server. The OSM database uses a data model comprising three principal data elements: Nodes, ways, and relations. A node  $N$  is a geographic point or an  $XY$  coordinate pair. A way represents a polyline or a polygon in the form of a series of nodes with a sequence. A relation is used to determine the relationship between at least two geographic elements. Therefore, these three data types have a nested structure. OSM data can be stored in eXtensible Markup Language (XML) format. As the Figure 2.9 shows, each way contains nodes, and each relation contains ways and nodes. In addition to storing information about the shape of each element, the database stores other information, such as ID, timestamp, user ID, and version (Figure 2.9). Each element can contain a set of tags, which are demonstrated as key-value pairs. For example, if a way element is a motorway in reality, it should be tagged `<tag k="highway" v="motorway"/>`. Users can also define their own tags to describe the elements.



Figure 2.9: (Color online) An example of three OSM data types in XML

There are numerous applications developed for people to use OSM data for different purposes. Generally, the applications range from information browsing, to map comparing, to track collecting, to routing, to data-quality assessment and control (OSM Wiki 2017). Software applications have been developed for data-quality assessment and control. Keep Right is a widespread OSM street-data assurance tool. It automatically detects errors for all OSM data, such as dead-ends, one-way streets, ways without nodes, and missing tags. There are also applications for specific OSM data elements. For example, OSM Relation Analyzer (<http://ra.osmsurround.org/>) is a web-based application for rating OSM relations in terms of several criteria, including tags, type of ways, and existence of gaps.

### 2.5.2. Twitter, Brightkite, and Gowalla

Twitter is a social-media site providing social-networking and microblogging services. Microblogging means that users can post short texts (the so-called tweet, which has a limit of 140 characters) in real time. Twitter was founded in March 2006 in San Francisco. Just a few years later, it developed a huge global user base. Currently, Twitter has an average of 330 million monthly active users (Statista 2017), who produce tens of millions of tweets every day (Kumar et al. 2013). Generally, users can post tweets in two ways. One is an original tweet a user can create. The other way is by retweeting another user's original tweet (Kwak et al. 2010). Tweets can be in different forms, such as text, image, and video. In addition, each user can add a geo-tag to their original tweet. The geo-tag is formatted as longitude and latitude. Approximately 1 percent of tweets are geo-tagged (Twitter Help Center 2017).

Brightkite and Gowalla both started in 2007 and ended in 2012 when Facebook bought them. These social media sites were similar to the most popular,

location-based, social-media site Foursquare, but without gaming features. Britekite and Gowalla were primarily intended for networking registered users all over the world via places they visited; namely, check-in locations from mobile devices. Users could establish mutual friendship connections, share locations and photos, and leave comments for each other. The location was a useful tool through which users could check for other nearby users and see who had been there before. In Gowalla, users could also check a user's recent history in a given place (Scellato et al. 2011).

The most direct way to fetch LBSM data is by a web search on the official webpage. However, the data is not directly downloadable this way, so it is difficult to keep it for later use. To effectively collect data, such as users' friends list, tweet content and check-in records, one can use a public application program interface (API). In Twitter, there are two types of APIs (Twitter Developer 2017): Representational state transfer (REST) APIs and streaming APIs. REST APIs allow users to use script to read someone else's profile and collected tweets. The results response to users can be in XML or JavaScript Object Notation (JSON) format. The streaming APIs mainly include public, user, and site streams. Public streams are suitable for following trending topics and data mining. User streams are for tracking one user's tweeting activities. Site streams are multi-user streams for obtaining tweet data corresponding to a group of users. REST APIs are different from streaming APIs in three aspects: REST APIs help find the historical tweets (up to the past week), while streaming APIs show new tweets in real time; REST APIs do not require a solid infrastructure, while the streaming APIs do; and REST APIs are limited in the number of calls that users can make to the server, whereas the limitation for streaming APIs is the number of tweets that can be delivered.

## 3. Experimental design

### 3.1. Overview

The experimental design plays a very important role, since this study is situated in the big-data context, which requires data-intensive geo-computation. This chapter presents some technical details for processing and modeling the VGI datasets. More specifically, the datasets mainly come from four sources: OSM, Twitter, Brightkite, and Gowalla. Each dataset comprises at least hundreds of thousands geographic features and tens of thousands of users. Data is collected from each individual user in a bottom-up manner. As the framework (Figure 3.1) shows, this chapter will illustrate the data-processing method, including how to extract relevant information, build up the data structure, and remove redundant parts. This chapter will also show the data modeling method for converting the raw datasets into different geographic and network representations. It should be noted that not all the implementation details are included in this chapter, due to space limitations.

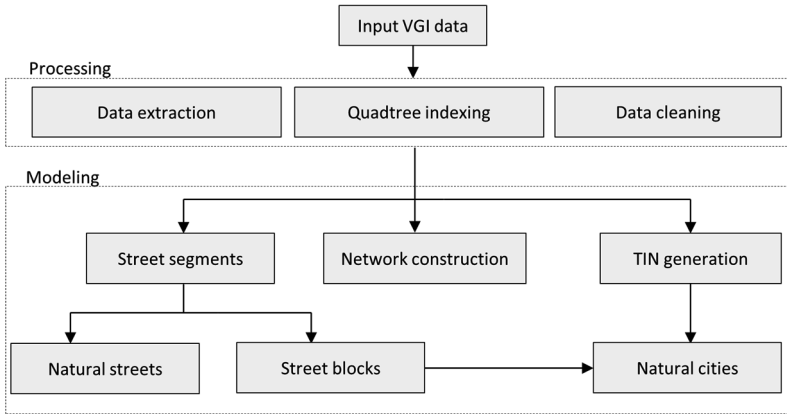


Figure 3.1: The framework of data processing and modeling in this thesis

### 3.2. The study areas

The study areas in this thesis were selected at a wide range of scales, ranging from the entire world, to country, city, and individual geographic features. The reason we chose such a range of study areas was because we think that fractal or scaling is universal across the geographic space. Apart from the study at the global level, we chose the US and six big European big cities (London, Birmingham, Paris, Toulouse, Berlin and Munich) (Figure 3.2). The VGI datasets mainly came from OSM, Twitter, Brightkite, and Gowalla, of which OSM data was relied on the most. To a large extent, the selection of study areas was directed by the quality and availability of datasets, since there was an imbalance of data from one area to another, e.g. the popularity of Twitter in Europe is

totally different from that in China. Table 3.1 shows which social-media data was used at which scale of the study area. Specifically, the study was stimulated by the global coverage of OSM data to explore the scaling property of the entire world. The high concentration of Brightkite and Gowalla check-in locations in the US are good for analyzing human activities at the country level. The completeness of street data and active tweeting activities among European capital cities were suitable for studying human intra-city movements.

Table 3.1: VGI datasets and their applied study areas

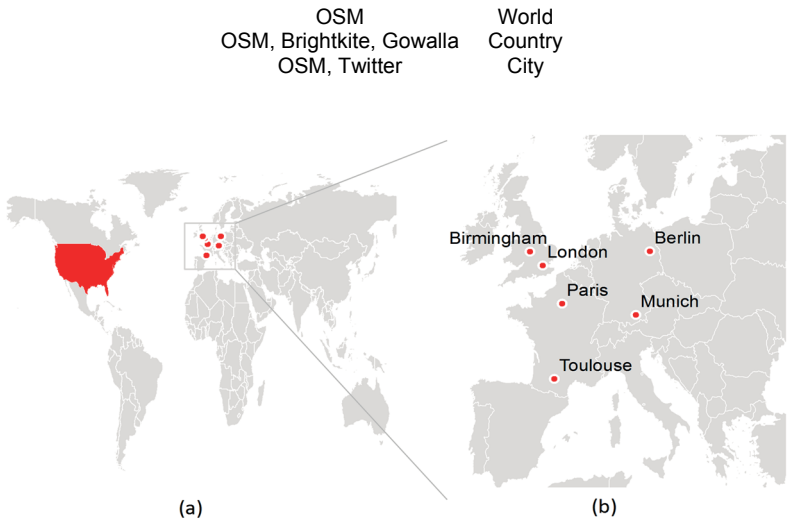


Figure 3.2: (Color online) The study areas in the thesis

### 3.3. Data processing

#### 3.3.1. Geospatial data extraction

##### *Extraction of OSM element*

The OSM full-history dataset is a huge XML file (692 GB after being decompressed by February 2013) and can be directly downloaded at the official OSM website (<http://planet.openstreetmap.org/>). The XML file is well-organized. Specifically, the file starts with node content, moves on to way content, and then relation content. Every element is in sequence of its element ID. By knowing this, we created a workflow that easily and efficiently extracted all required information for each element (Figure 3.3). As the flowchart shows, two types of attributes were saved for each element (The latest version of the element attributes, including its shape information in terms of one, or a set of, ordered  $xy$  coordinate pairs and a list of its member ID (if applicable); and the historical attributes including time, user ID, and version number. For the practical part, the workflow was based on .Net environment in Visual Studio 2010 and the

C# XML library was used to process the file. Since there were billions of elements to be read, it was impossible for us to put all elements and their relation attributes into memory using an ordinary desktop. Instead, we wrote every element onto the hard disk with a structure that could be recognized and loaded by our program for later use. The resulting file contained attributes of approximately 2.1 billion elements and used approximately 130 GB in .txt format, including both the latest and historic attributes for all elements. Thanks to the sequentially organized element, there was no need to sort the elements for searching after loading the content. We applied the binary search for the inherently sorted array for any element using its ID, which normally takes less than one second to pinpoint the element and get its related information.

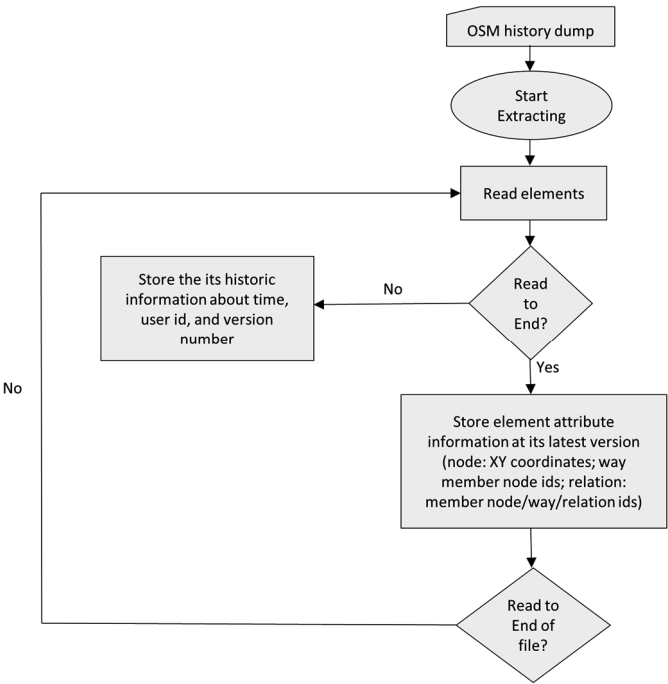


Figure 3.3: The workflow of the data extraction of OSM planet history dump

We refined the data structure for the OSM element extraction. As introduced in the previous chapter, OSM elements have a hierarchical structure that enables them to embed within each other. Figure 3.4 shows this nested relationship. A relation can have nodes, ways, and another relation that contains other ways, nodes, and possible relations. In fact, there are numerous complex relations containing some member relations that are also the parents of other relations in the OSM database. Therefore, it would be insufficient to only rely on the member ID to extract the precise geometric information of a complex relation element such as the number of all containing nodes. To solve this problem, we refined the data structure of the three OSM elements by adding their parent

id(s). Namely, we added a list of way IDs and a list of relation IDs to which that node pertained for each node element. We applied the same strategy to each way and relation element as well. By using this data structure, a set of simple and straightforward algorithms could be used to iteratively find the shape information. As was the case with computing the size of an element in terms of the number of nodes, the recursive algorithm as pseudo-code is presented as:

```

Recursive function CountOSMElementNode (OSM element):
  Foreach member element in the input element:
    If (member element is node):
      NodeCount = NodeCount + 1;
    Else:
      CountOSMElementNode (member element);
End Function

```

We found a beneficial byproduct from this refined data structure – the worldwide street junctions could be directly extracted if a node satisfied the condition of having more than one highway element (the tag value of the way element was “highway”) as parents. This condition could be improved using tag information to avoid the intersection node between a highway bridge and its underlying highway(s). Because there were no spatial operations performed (such as a spatial query), and each node element was well-structured in terms of its node ID and parent highway ids, the extraction could be extremely fast and avoid intensive computing. It could also save a lot of storage in keeping the junction nodes, since it only required updating the “street junction flag” of a node to the value of true in the predefined data structure, rather than storing all the attributes again in a separate file.

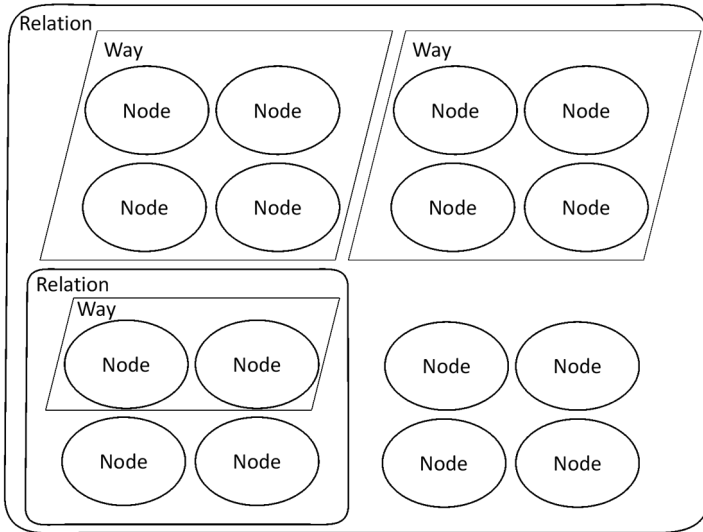


Figure 3.4: The nested OSM element structure



### *Extraction of human activities from LBSM data*

For the study of LBSM, the first step is obtaining the geo-related data. The study used two types of geo-related data: Check-in locations and geo-tagged tweets. The check-in location data was from Brightkite and Gowalla, respectively. The data could be directly downloaded from the official site of Stanford Network Analysis Project (<http://snap.stanford.edu/index.html>). Both datasets had several millions check-in locations and tens of thousands of users over approximately a two-year time span. Each check-in location was recorded with a user ID,  $x$ ,  $y$ , and timestamp. For the geo-tagged tweets, we extensively used Twitter's streaming APIs that allowed us to partially access Twitter's database through an HTTP request. We wrote some Python scripts to do the crawling and deployed the crawler remotely to have multiple instances in the Linux system running at the same time. Finally, it took several days for us to get one week of tweets from June 1 to June 8 in Western Europe. We did not fetch the data for a longer period of time because of the data stream limitation. If the data stream exceeded the 1-percent limit, it would randomly drop some of the messages. Furthermore, since only the spatial-temporal information is of interest, we filter out the tweet content and only keep the user ID, location, and time stamp (the same as the check-in location format). These were some statistics about the tweet location data for the three biggest countries: 4,127,159 in France, 837,627 in Germany, and 3,704,351 in the UK.

The next step was to extract the human activities from the obtained geo-data. Since each location was associated with a user ID and timestamp, it was very easy to extract each user's movement trajectory during the time period. The trajectory can be denoted as a set of time-stamped locations:

$$Trajectory(u) = \{(x_1, y_1, t_1), (x_2, y_2, t_2) \dots (x_i, y_i, t_i)\} \quad (3.1)$$

in which  $u$  is a user;  $x$  and  $y$  are the geographic coordinates; and  $t$  is the timestamp. The pseudo-code below describes how the extraction was implemented. After extracting the movement trajectories, we recorded not only the location history for each user, but also the list of users at each location. Both types of information can be useful for conducting research in human dynamics in either an urban or country space. Furthermore, this data can also be used to construct a socio-geographic network (more details in Section 3.4.3).

```
Function ExtractMovementTrajectory (Check-in locations):  
  While (Not the end of file):  
    Store XY coordinates, user id, and timestamp into list;  
    Sort the list by user id then sort by timestamp;  
    Generate each user's trajectory by grouping the list by id;  
End Function
```

### **3.3.2. Quadtree indexing**

Quadtree (Samet 1990) is a common approach in GISystem that is widely applied for hierarchically structuring spatial data. For this study, it was imperative to build up such a structure to recursively decompose the input VGI data. Taking an example of an OSM element, if we used a standard linear index to

conduct a simple spatial point-in-polygon query to know how billions of elements spatially distribute in each country, it might take years to get the answer. Although R-tree is recommended nowadays for spatial indexing at multi-dimensions (Murray 2003), quadtree was suitable for this study since all datasets to be applied were only 2D geometric objects and quadtree supports a faster, more stable structure. Building quadtree can also help organically partition the geographic space and give a better overview of the geospatial data because the size of the grid can indicate spatial density (Figure 3.5).

Two points must be clarified for building up a quadtree in this study. First, the tree was not a full tree, as each quadtree subdivision (quadrant) stopped when the resolution condition was met. The resolution of the quadrant (or to which detail level) depended on the number of the containing geo-data. For example, the resolution was 100 for point data because an ordinary spatial query can instantly return the result for 100 points with little computing resources. Second, each quadrant was indexed based on its depth. Neighboring quadrants were always saved and iteratively updated. The process of quadtree implementation is described by the pseudo-code shown below. The introduced quadtree structure greatly facilitated processing and modeling the big datasets in terms of computing speed and capacity.

```
Recursive function CreateQuadtreeIndex(Input Geo-data):
  Create envelope polygon for the input geo-data;
  Divide the envelope into 4 sub-regions
  Foreach sub-region:
    Assign the sub-region an index based on the recursion depth;
    Let sub-data = the geospatial data within the sub-region;
    If (number of sub-data > threshold):
      CreateQuadtreeIndex(sub-data)
    Else:
      Assign each feature in the sub-dataset the index of the
      sub-region
End Function
```

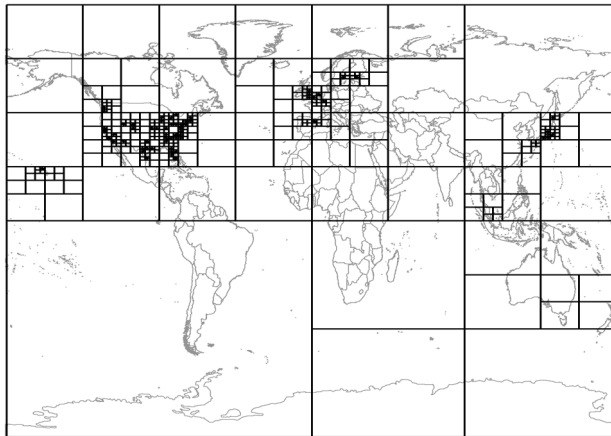


Figure 3.5: Quadtree indexing on the global data set

### 3.3.3. Data cleaning

Cleaning the input VGI removed topological errors. Three of the most common topological errors among the geospatial datasets were duplication, self-overlapping, and self-intersection. Duplication exists in three vector types (point, polyline and polygon), indicating that more than one feature shared an identical shape. Self-overlapping only exists in the polyline type of data, meaning that a single line overlapped itself, or comprises segments that are partially identical. Self-intersection occurs in both polyline and polygon types of data, in which one or more parts of a polyline or polygon cross another part. Since the data quality could probably affect the final results, it was necessary to perform an error check and resolve any errors before doing further analysis.

Current GIS software provides mature solutions for detecting and correcting the listed topological errors for vector data. For example, the tools in ArcMap, such as *Delete Identical Features* and *Repair Geometry*, can help correct errors that were previously mentioned. However, these software solutions are generally unable to cope with big data. In this regard, the study designed a batch-processing procedure to perform the topological correction. The input data is first partitioned through its quadtree structure. Namely, the features are grouped with the same quadrant index. In this way, the data is effectively split into numerous operable batches for the processing program. For each data slice, the program uses the *ITopologicalOperator* interface of ArcObjects library (ESRI 2017a) to identify and correct the topological errors. The *ITopologicalOperator.Simplify* method can effectively deal with self-intersection, self-overlapping, and duplicated parts removal.

## 3.4. Data modeling

### 3.4.1. Natural streets and street blocks from street segments

The natural streets can be automatically generated by Axwoman 6.3 (Jiang 2015e). Jiang et al. (2008) provided the related algorithms for three types of joining principles (self-fit, self-best-fit, and every-best-fit). In practice, Axwoman can calculate the deflection angles among street segments, do the connections in real time, and provide good results. However, there are two deficiencies. First, this solution fails to deal with the complex junction situation. For example, the natural-road tracking process would stop when meeting a roundabout. It made the natural roads less connected than what they should be. Another problem was that the program did not consider the importance of each segment and always started the joining process from a random segment. This could lead to a suboptimal natural road result (Jing et al. 2015).

In this regard, this study made related improvements using a workflow (Figure 3.6). The first step was to generate street segments at street junctions. This step can be done directly through the transformation from inputted street data to street segments using the *Data Interoperability* extension of ArcGIS. The resulting file format was Esri ArcInfo Coverage, and the segment file was in arc type. In addition, the segment-segment topology was correctly created for the data. The street blocks, which were the closed areas of street segments,

could subsequently be generated using the *Feature to Polygon* tool of ArcMap with street segments as input. We then detected the roundabout from the street-block data by calculating the circularity using the equation  $C = \frac{4\pi a}{p^2}$  for each block, in which  $a$  was the block area and  $p$  was the block perimeter. The closer the circularity value was to 1, the more circular a shape the polygon would have. Since roundabouts are usually circle-like shapes, we selected street blocks with a big circularity value (for example,  $C > 0.95$ ) and an acceptable area (that could not be too big) as the roundabout. The segments that intersected with each roundabout were then locally connected, based on the good-continuity principle through the maximum-matching solution (Edmonds 1965), as recommended by Yang et al. (2011a). Each roundabout was deleted from the dataset after the connection of its intersected segments was finished.

The following steps were designed to produce the optimal set of natural roads by assigning the importance of each segment. In order to know how important each segment was, we needed to build up the connectivity graph of the segments. At this point, the isolated segments could easily be detected by checking if they were outside the largest component of the graph (equal to the graph itself if there was no isolated line). After the isolated lines were removed, the length of each segment and its network centrality measures (degree, closeness, and betweenness) were used to determine the segment's importance. The importance  $I$  of each segment could be obtained by integrating four measures, as the equation denotes:

$$I = w1 * length + w2 * degree + w3 * closeness + w4 * betweenness \quad (3.2)$$

Four weights of  $w1, w2, w3$ , and  $w4$  were objectively determined using the CRITIC method (Diakoulakli et al. 1995), based on each parameter's own characteristics (variance), and the relationship between the four parameters (correlation coefficient). The resulting importance values ranged from 0 (unimportant) to 1 (most important). Finally, the sorted segments with a descending order were the input for generating the natural streets. Under this condition, the sequence of tracking natural streets is preset. Specifically, the process started tracking the segments to be connected with the most important segments, and then the segments for the second-most important segment, and so on. This ensured the unique, optimal solution of natural roads. Furthermore, the most important segments tended to be located in the central part of the street network, so it was also very effective and efficient to prioritize them in the tracking algorithm for the sake of reducing computation time.

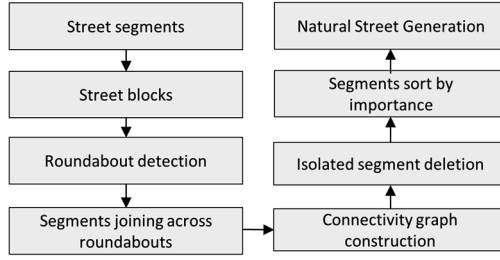


Figure 3.6: The workflow for generating natural streets

### 3.4.2. Natural cities from individual locations and street blocks

The natural cities objectively depict human activities or settlements. Two types of natural cities were generated in this thesis: Individual locations and street blocks. Different techniques were applied to generate different types of natural cities. To be more specific, natural cities generated from individual locations were based on the triangulated irregular network (TIN) (Jiang and Miao 2015). Natural cities generated from street blocks were created by iterative clustering of small street blocks (Jiang and Liu 2012). We developed some improvements to existing solutions to handle the big geospatial datasets.

As mentioned previously, the individual locations from different VGI sources (such as tweet locations and OSM nodes) can be tens, or even hundreds, of millions of points at the country level. It becomes a problem to build up a TIN based on such a large number of points using GIS software. As was the case with ArcMap, 10 million to 15 million nodes are the largest sizes to process (ESRI 2017b). We designed a workflow (Figure 3.7) based on the batch processing method to conduct the TIN-based clustering of a huge point set to derive the natural cities. Based on the quadtree structure, the individual locations were first partitioned into operable batches, each of which contained a limited number of points. The TIN was then locally constructed for a set of points in each quadrant. The convex hull for each point set can be subsequently derived by dissolving all the triangles of each local TIN. The points touching the boundary of a convex hull were then marked as the border points for each point set. Each local TIN connected to its neighbor's TINs using only the border points to form the global TIN. Since the process of TIN generation was a bottom-up approach, it could be used for any number of points. Each TIN edge was stored each time it was created during each local process so it would not create a problem with computer memory. Finally, the points were grouped using the mean edge length of edges from all local TINs and the global TIN. Natural city patches were then produced via *Feature to Polygon* and *Dissolve* tools in ArcMap. See more details about generating natural cities in Jiang and Miao (2015).

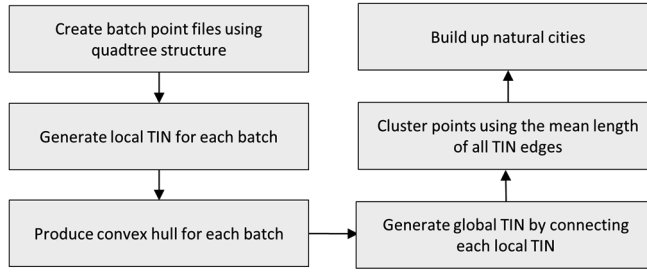


Figure 3.7: The workflow for generating natural cities from individual locations

A street block refers to a minimum area closed by neighboring street segments (Jiang and Liu 2012). The street blocks could be directly produced, based on street segments, by using ArcMap (see Section 3.4.1). If the size of input segments exceeded the limitation of the software, the street blocks could also be created programmatically by following the previous work (Jiang and Liu 2012). The street blocks were categorized into two types: City blocks and field blocks. City blocks were those smaller than the mean size of all street blocks, whereas field blocks were those larger than the mean size. The natural cities could then be generated through an iterative clustering algorithm for grouping the city blocks. The pseudo-code describes how the algorithm worked. This algorithm could be very slow for countrywide searching because it consumed a lot of memory when the recursion went very deep. One way to efficiently alleviate this problem was to first traverse all city blocks to mark those whose neighboring blocks were also all city blocks. This saved a lot of computation time and resources.

```

Recursive Function NaturalCityGeneration (Street Block)
  If (this block is a city block):
    Add this block into Blocklist;
    Get its neighboring blocks;
    Foreach neighboring blocks:
      If (all its neighboring blocks are city blocks):
        NaturalCityGeneration (this block);
  Return Blocklist as a Natural City;
End Function
  
```

### 3.4.3. Network construction from human activities

We constructed two types of networks for modeling human activities in different LBSMs: A binary network and a weighted network. Both of these were undirected networks. In a binary network, the nodes represented individual users and links for relationships. There are two kinds of relationships based on: Users' interactions about their co-contributions to OSM elements; and co-location of users' movement trajectories, respectively. The network was built using Equation 3.3:

$$link_{ij} = \begin{cases} 1, & \text{if } E_i \cap E_j \neq \emptyset \text{ or } T_i \cap T_j \neq \emptyset \\ 0, & \text{else} \end{cases} \quad (3.3)$$

in which  $i$  and  $j$  refer to a pair of individual users,  $E$  stands for a user's edit history, and  $T$  is a user's movement trajectory (location history). Figure 3.8 illustrates a co-contribution network. For example, element  $b$  exists in the edit history of users 1, 3, and 4. Therefore, these three users have a co-contribution relationship between every two of them. This same rule applies to the co-location network.

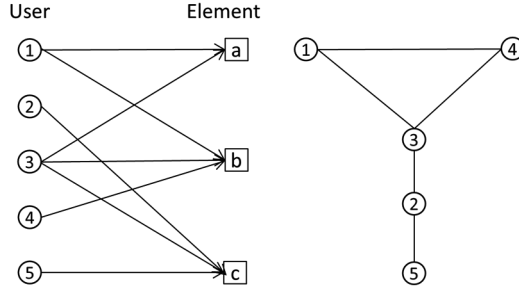


Figure 3.8: Illustration of a binary network construction  
(Source: Paper I)

A weighted network is used to establish socio-geographic networks (Zheng 2011) based upon a user's location and social connections. The user location can be a single location or a natural city, either of which can be shared by other users. The network can be described by Equation 3.4:

$$link_{ij} = |S|, \quad S = U_i \cap U_j \quad (3.4)$$

in which  $i$  and  $j$  refer to a pair of individual locations or natural cities;  $U_i$  or  $U_j$  stands for the users who visited the location/city  $i$  or  $j$ ; and  $S$  stands for the pairs of users who are friends in their social network. The link between a pair of locations/cities is weighted by the number of pairs of socially connected users. As Figure 3.9 illustrates, the socio-geographic network maps user social connections into locations/cities.

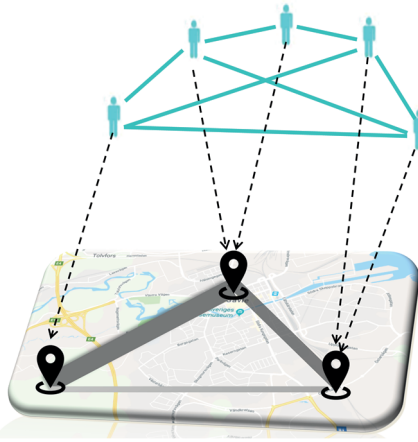


Figure 3.9: (Color online) Illustration of a weighted network construction



## 4. Methodology

### 4.1. Overview

Following the topological and scaling way of thinking, the thesis employs a set of complexity science methods, including heavy-tailed distribution statistics (e.g., Clauset et al. 2009), head/tail breaks (Jiang 2013a), ht-index (Jiang and Yin 2014), topological analysis (Jiang and Claramunt 2004), and complex network analysis (Figure 4.1). This methodology helps uncover the scaling property of geospatial big data in a quantitative manner, and, more importantly, illustrate the underlying fractal or scaling pattern of geographic space and its involved human activity. Overall, the methodology can be seen from both scaling and topology. Scaling analysis comprises heavy-tailed distribution statistics (or power-law metrics in particular), and head/tail breaks and its induced ht-index. Topological analysis mainly utilizes the complex network structural parameters and community structure. Note that the two parts were not separated. Instead, they complemented each other to provide deep insights into scaling pattern and topological properties while performing the analysis. For example, the topological parameters helped detect the underlying scaling pattern of a network, and vice versa, the scaling hierarchy and power-law metrics of network measures help uncover the network community structure. Through a series of studies, we found that this methodology could cope with geospatial big-data analytics.

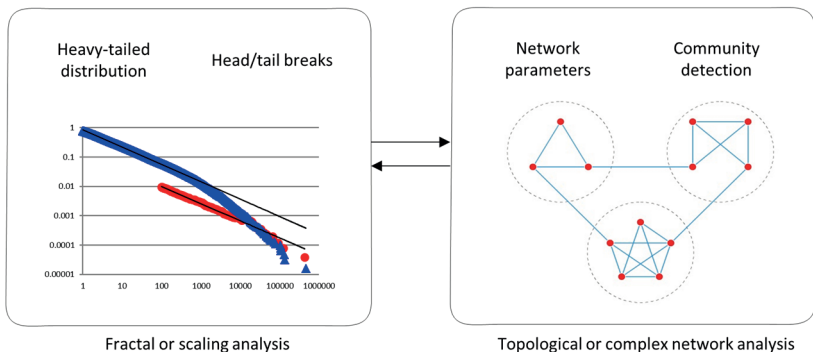


Figure 4.1: (Color online) The overview of the complexity science methodology

## 4.2. Mathematical detection of heavy-tailed distributions

### 4.2.1. Different types of heavy-tailed distributions

Heavy-tailed distribution was one of the theoretical foundations in this study. Compared to a normal, or normal-like distribution, in which data tended to distribute evenly around the mean, the data of a heavy-tailed distribution tended to have much more data to the right side of the mean (rightly skewed). In general, data with a heavy-tailed distribution refers to a nonlinear relationship between quantity  $x$  and its probability (Clauset, Shalizi and Newman 2009). The most common types of heavy-tailed distributions are power-law distribution, exponential distribution, lognormal distribution, and their variations: power law with an exponential cutoff, and stretched exponential. The distributions are as follows:

A power law distribution indicates the probabilities of a value ( $y$ ) being proportional to some power of a quantity ( $x$ ), which is denoted as follows:

$$y = kx^{-\alpha} \quad (4.1)$$

in which  $x_{min}$  is the smallest value from which the power law is obeyed, and  $\alpha$  is the power-law exponent.

A power law with an exponential cutoff is a degenerated form of the power law. Simply put, it is not an ideal power law, but a hybrid between a power law and an exponential, which can be expressed as:

$$y = kx^{-\alpha}e^{-\lambda x} \quad (4.2)$$

An exponential distribution simply denotes the exponential relationship between  $y$  and  $x$ , given by:

$$y = ke^{-\lambda x} \quad (4.3)$$

The degenerated version of an exponential distribution is the stretched exponential, which can be described as:

$$y = kx^{\alpha-1}e^{-\lambda x^\alpha} \quad (4.4)$$

The lognormal distribution is another common type of heavy-tailed distribution. *Lognormal* means that if we take the logarithm of quantity  $x$ , it results in a normal distribution. The lognormal distribution can be formatted as:

$$y = k \frac{1}{x} e^{\left[ \frac{(\ln x - \mu^2)}{2\sigma^2} \right]} \quad (4.5)$$

### 4.2.2. Mathematical detection

The mathematical detection of the heavy-tailed distribution in this study adopted the method from previous studies (Clauset et al. 2009, Jiang and Jia

2011b). The procedures for mathematical detection included two main steps: (1) Examine if quantity  $x$  is power-law distributed and, if not; (2) see if quantity  $x$  follows any of the four alternative distributions. Figure 4.1 shows the workflow of the mathematical detection. The workflow primarily focuses on if the data fits the power-law distribution with an acceptable  $\alpha$  and  $p$  value. The power-law detection is based on a rigorous method suggested by Clauset et al. (2009) that combines the maximum likelihood estimation (MLE) (Shanbhag and Rao 2001) and Kolmogorov-Smirnov (KS) test. The other four types of heavy-tailed distribution (as presented in Equations 4.2, 4.3, 4.4, and 4.5) were examined by the likelihood ratio to see which model had the largest similarity to the input quantity  $x$ . Since the power-law distribution was the mainstay throughout the study, the following content in this section will only concentrate on power-law detection. For the other four members, Table 4.1 presents their estimated constant  $k$  based on MLE method. Interested readers can refer to Clauset et al. (2009) and Jiang and Jia (2012) for more details on how the estimated parameters were calculated.

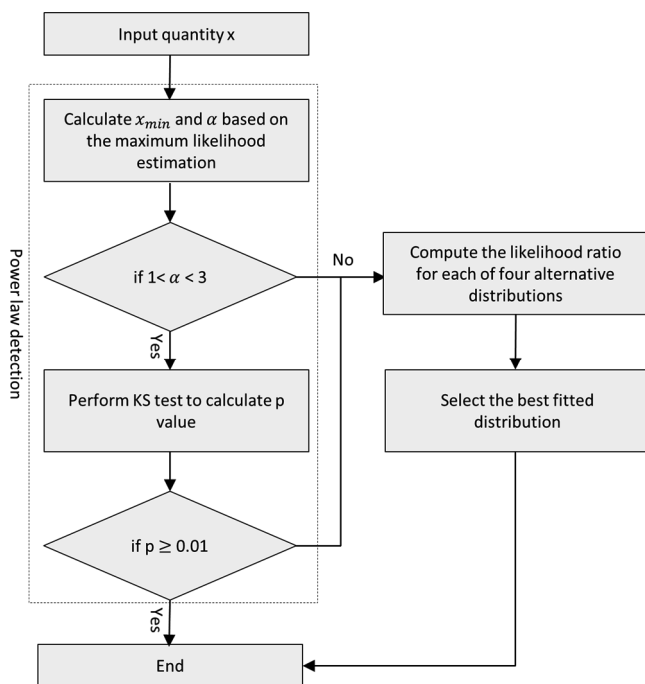


Figure 4.2: The workflow of the mathematical detection of heavy-tailed distributions

Table 4.1: The estimated constant  $k$  for four heavy-tailed distributions

Exponential	$k = \lambda e^{\lambda x_{min}}$
Stretched exponential	$k = \alpha \lambda e^{\lambda x_{min}^\alpha}$
Power law with an exponential cutoff	$k = \frac{\lambda^{1-\alpha}}{\tau(1-\alpha, \lambda x_{min})}$
Lognormal	$k = \sqrt{\frac{2}{\pi \sigma^2}} \left[ \frac{(\ln x_{min} - \mu^2)}{2\sigma^2} \right]^{-1}$

For power-law detection, the simplest way was to plot quantity  $x$  from the largest to smallest, and then take logarithms on the  $x$ - and  $y$ -axes. According to Equation 4.6, the distribution line should become a straight line if the quantity  $x$  is a power law. However, it only suited the data that was perfectly power-law distributed. For more general data, taking logarithms can lead to very noisy results (numerous fluctuations) in the tail (Newman 2005) (Figure 4.2). Furthermore, the assessment of fitness is based on the linear regression that cannot further provide probability distribution estimations.

$$\ln(y) = -\alpha \ln(x) + \ln(k) \quad (4.6)$$

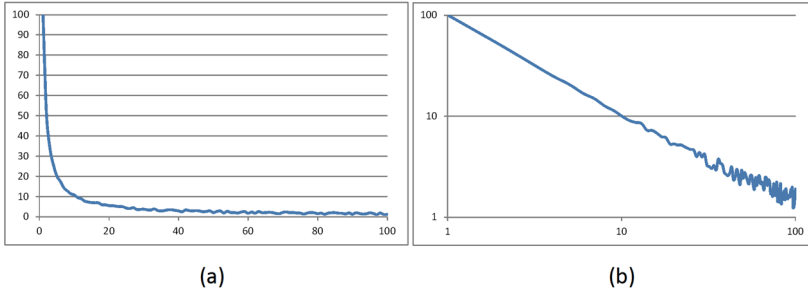


Figure 4.3: (Color online) A working example showing a general power law distribution (Note: The rank-size plot on a fake dataset (a) and its log-log plot (b). Panel b exhibits the noisy tail of the plot line after taking logarithm)

Instead of a rough estimation, Newman (2005) and Clauset et al. (2009) suggested using the MLE method to accurately estimate parameters, such as  $x_{min}$  and  $\alpha$ , in order to calculate the probability distribution. It advances traditional methods in many aspects, since it makes full use of each value in quantity  $x$  by avoiding binning. To elaborate, assuming that a quantity  $x$  is continuous and follows a power-law distribution, and its probability distribution function (PDF) according to Equation 4.1 can be formatted as:

$$P(x) = p(x)dx = kx^{-\alpha}dx \quad (4.7)$$

Since the integral of  $P(x)$  is 1 of the PDF, the following equation can be derived:

$$1 = \int_{x_{min}}^{\infty} kx^{-\alpha} dx = \frac{k}{1-\alpha} [x^{-\alpha+1}]_{x_{min}}^{\infty} \quad (4.8)$$

The constant  $k$  can be obtained as:

$$k = (\alpha - 1)x_{min}^{\alpha-1} \quad (4.9)$$

In this case, the power-law distribution is:

$$p(x) = \frac{\alpha - 1}{x_{min}} \left( \frac{x}{x_{min}} \right)^{-\alpha} \quad (4.10)$$

Considering that there are  $n$  values of quantity  $x$ , the probability of quantity  $x$  is proportional to:

$$p(x|\alpha) = \prod_{i=1}^n p(x_i) = \prod_{i=1}^n \frac{\alpha - 1}{x_{min}} \left( \frac{x}{x_{min}} \right)^{-\alpha} \quad (4.11)$$

Equation 4.11 denotes the likelihood of quantity  $x$ . If we use the logarithmic form, likelihood  $\mathcal{L}$  is:

$$\mathcal{L} = \ln p(x|\alpha) = n \ln(\alpha - 1) - n \ln x_{min} - \alpha \sum_{i=1}^n \ln \frac{x_i}{x_{min}} \quad (4.12)$$

To get the maximum likelihood, we set  $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$ , so the exponent  $\alpha$  could then be expressed by:

$$\alpha = 1 + n \left[ \sum_{i=1}^n \ln \frac{x_i}{x_{min}} \right]^{-1} \quad (4.13)$$

It should be noted that the acceptable range of the exponent  $\alpha$  is normally from 1 to 3 to formulate a power-law distribution.

The next step was to use a KS test to evaluate to what extent the quantity  $x$  fits a power-law distribution based on the estimated parameters (goodness of fit, GoF). The KS statistics used the maximum distance between the cumulative density functions (CDF) of the estimated power-law model and the CDF of the quantity  $x$ .

$$D_e = \max_{x \geq x_{min}} |f(x) - g(x)| \quad (4.14)$$

in which  $D_e$  was the maximum distance,  $f(x)$  is the CDF of the quantity  $x$  with a value of at least  $x_{min}$ , and  $g(x)$  is the CDF for the power-law model that best fits the data in which  $x \geq x_{min}$ .

Since we know that  $\alpha$  was dependent with  $x_{min}$ , the fitted power-law model would change every time a new  $x_{min}$  was employed. Therefore, before assessing the GoF,  $x_{min}$  can be refined through a loop process of the KS statistics. To be specific, we selected a set of  $x_{min}$  candidates as inputs and calculated a series of maximum distances  $D$  between each estimated model and real data, using Equation 4.14. Finally, we found the proper  $x_{min}$  through the estimated model with the smallest  $D$ .

After identifying the suitable  $x_{min}$ , we made 1,000 artificial datasets by using the estimated model  $g(x)$  to conduct a solid assessment of GoF. Each of the 1,000 datasets could be separated into two parts: The values above  $x_{min}$  completely followed a power-law distribution, while the values below  $x_{min}$  were non-power-law distributed. We employed Equation 4.14 again for computing maximum distances  $D_i$  for each artificial dataset and the estimated model  $g(x)$ . The GoF  $p$  value could be derived using the following equation:

$$p = \frac{\# \text{ of } (D_i > D_e)}{1000} \quad (4.15)$$

Since  $D_e$  denotes the maximum distance between the estimated power law model and the real dataset, Equation 4.15 illustrates the idea that the more artificial datasets than the real dataset do not fit well with the estimated model ( $D_i > D_e$ ), the higher the chance for the real dataset being a power law distribution. The  $p$  value ranges from 0 to 1, where 0 means rejection of the hypothesis of a dataset being a power law distribution. Empirically, if the  $p$  value is greater than 0.01, the dataset follows a power law distribution (Marta et al. 2008).

### 4.3. Head/tail breaks and its induced ht-index

#### 4.3.1. Concept and definitions

*Head/tail breaks* (Jiang 2013a) is a classification scheme for data with a heavy-tailed distribution. Data with a heavy-tailed distribution inherently possess a scaling pattern or hierarchy. It developed from the head/tail division rule (Jiang and Liu 2012): *Given data with a heavy-tailed distribution, the arithmetic mean, or average, can split all the data values into two unbalanced parts: A minority of big values above the mean, called the head(for example, < 40%); and a majority of small values below the mean, called the tail.* The unbalance between the head and tail parts refers to the scaling pattern of far more small things than large ones. This process recursively continued for the head part until the head part bore no scaling property or was no longer heavy-tailed distributed. Figure 4.4 uses a working example of a classic fractal – a Koch Snowflake – to clearly explain the process of head/tail breaks.

Figure 4.4a shows the original snowflake which contains 64 equilateral triangles of different sizes. More specifically, there are 48, 12, 3, and 1 triangles, with side lengths of  $1/27$ ,  $1/9$ ,  $1/3$  and 1, respectively. The area of equilateral triangles, according to the formula  $A = \frac{\sqrt{3}}{4} l^2$  ( $l$  is side length), are approximately 0.43, 0.14, 0.05, and 0.016, respectively. The areas of the triangles obviously followed a heavy-tailed distribution, in which there were far more small triangles than big ones. Therefore, we conducted head/tail breaks. The first mean was  $m_1 = (1 \times 0.43 + 3 \times 0.14 + 12 \times 0.05 + 48 \times 0.016)/64 = 0.03$ . This split the triangles into 16 triangles above  $m_1$ , and 48 triangles below  $m_1$ . Those 16 triangles were the head part ( $16/64 = 0.25$ ,  $< 40\%$ ) and selected to be the first class (Figure 4.4b). The rest could be determined in the same manner. The second mean  $m_2 = (1 \times 0.43 + 3 \times 0.14 + 12 \times 0.05)/16 = 0.09$  helped us to obtain the new head with four triangles (Figure 4.4c). Finally there was only one triangle above the third mean  $m_3 = (1 \times 0.43 + 3 \times 0.14)/4 = 0.21$ , which was the third class (Figure 4.4d). We observed that three levels of snowflakes were derived during the head/tail breaks process; namely the scaling hierarchy of all triangles (Figure 4.4e).

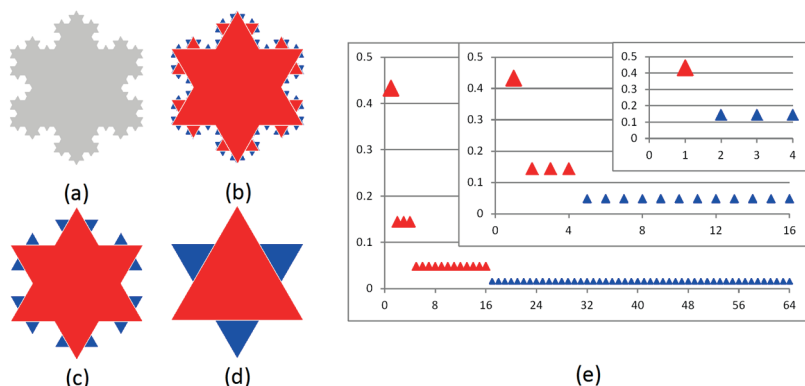


Figure 4.4: (Color online) Illustration of the process of head/tail breaks using Koch snowflake  
(Note: Red triangles indicate the head part, and blue triangles the tail part at each recursion)

The number of arithmetic means that were iteratively derived during the process resulted in the number of classes or hierarchical levels. The number of classes of the data + 1 is the ht-index (Equation 4.16; Jiang 2014). It captured how many times the scaling pattern recurred in the data. It quantified the fractal or scaling characteristic of the data. In other words, the higher the ht-index was, the more fractal or scaling the data became. Note that data with a power-law distribution always has a relatively high ht-index value. In this respect, power-law detection and ht-index complement each other. Head/tail breaks and its induced ht-index are the underlying foundation of the third definition of fractal. Unlike mostly widely used classification K-means (MacQueen 1967) and natural breaks (Jenks 1967) in traditional geospatial analysis,

head/tail breaks classification can effectively capture the underlying scaling hierarchy of geographic data. They have also been applied in many other fields, such as urban planning and transportation (Long et al. 2016), and biodiversity (Ontoy and Padua 2014).

$$ht = \#classes + 1 \quad (4.16)$$

#### **4.3.2. Applications for geospatial big data**

In most cases, geospatial big data comes from diverse sources and has very fine spatio-temporal granularity. These characteristics make big data full of nonlinearity. Previous studies have proven that such nonlinearity is exactly the right picture of the reality, as is the case with the structure of the natural cities from the massive check-in locations (Jiang and Miao 2015, Jiang 2015a). Since the reality is fractal (Bak 1996, Mandelbrot and Hudson 2004), there is an inherently fractal or scaling structure inside geospatial big data. In this regard, head/tail breaks can be used as an effective, efficient tool to analyze and visualize big data.

As mentioned earlier, geospatial big data sometimes appears too big to handle. The strategy behind applying head/tail breaks to big data is that we can always recursively take the head part until the head part is small enough to analyze and visualize. The reason why we can use this strategy is because of the self-similarity property of fractals; that is, the head is self-similar to the whole dataset. Similarly, the red triangles in Figure 4.4 can geometrically and statistically represent the whole at different levels. In other words, dropping out the tail part(s) would not affect much about the scaling pattern of the whole dataset. In this regard, head/tail breaks help us greatly reduce the size of big data (where the tail part is the majority), but without distortion. In this way, big data becomes manageable for both analytics and visualization. Additionally, the whole process of head/tail breaks only calculates the arithmetic means, which requires little computing capacity and a short waiting time for results.

#### **4.4. Complex network analysis**

Complex network analysis is an interdisciplinary research method shared across mathematics, physics, biology and geography. It analyzes nodes for individual features and links for their relationships (Cohen and Havlin 2010). This complex network fundamentally differs from its simple counterparts of regular and random networks. The major difference between complex and simple networks lies in their topological parameters (Helbing 2007, Flack and Krakauer 2011, Jiang et al. 2014). For example, the degrees of nodes in a random or regular network vary little from one to the other. In contrast, the degrees of nodes in a complex network tend to be very heterogeneous or scaling. We conducted a structural analysis of complex networks of various types, based on a set of network parameters that will be introduced in the following sections. Figure 4.5 presents the workflow of the complex network analysis in this thesis.



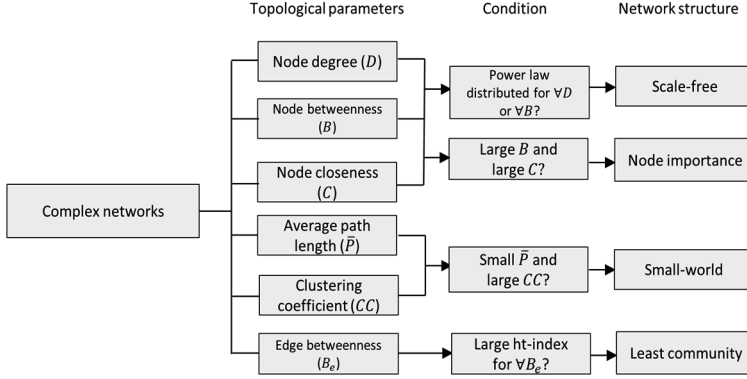


Figure 4.5: The overview of the structural analysis of complex networks

#### 4.4.1. Network topological measures

A network can be denoted as a graph  $G = (N, E)$ , in which  $N$  stands for the set of  $k$  nodes  $\{n_1, n_2, n_3 \dots n_k\}$ ,  $E$  is a set of edges or links among the nodes, and  $\forall E \in N \times N$  (Harary 1969). In this study, two types of graphs are generated – a binary graph and a weighted network graph. We can use the adjacency matrix  $A(G)$  to store and represent both types of graphs. In a binary network, nodes  $i$  and  $j$  are connected if they have a relationship, and the weight of that edge is  $A_{ij} = 1$ . A weighted network is generated by appending the weight value  $A_{ij} \geq 1$  to the edge.

$$A_{ij} = \begin{cases} \geq 1, & \text{if node } i \text{ and } j \text{ are linked} \\ 0, & \text{else} \end{cases} \quad (4.17)$$

##### Node degree

*Degree*, also known as *connectivity*, is the simplest parameter for evaluating a network structure, as Equation 4.18 expresses. The degree of a node is equal to the number of directly linked nodes in a binary network. In a weighted network, a node's degree value equals the sum of the weights of the edges that directly link to that node.

$$D_i = \sum_{j=1}^k A_{ij} \quad (4.18)$$

##### Node betweenness

Node betweenness measures the extent to which a node acts as a bridge in a network (Freeman 1979, Barthélemy 2004). For a node  $i$ , its betweenness score

is calculated by the ratio of the number of shortest paths between any two node  $m$  and  $j$  passing through itself.

$$B_i = \sum_{m,j \in N, m \neq i \neq j}^k \frac{\# \text{ of } Path_{m,i,j}}{\# \text{ of } Path_{m,j}} \quad (4.19)$$

### Node closeness

Node closeness calculates how close a node is to every other node in a network, which is denoted by:

$$C_i = \frac{k-1}{\sum_{j=1, i \neq j}^k d_{ij}} \quad (4.20)$$

in which  $k$  is the number of nodes in a network, and  $d_{ij}$  is the shortest topological distance between nodes  $i$  and  $j$ .

### Path length

The path length of a node describes how far the node is from all other nodes in a graph, which is formatted by:

$$L_i = \sum_{j=1}^k d_{ij} \quad (4.21)$$

It should be noted that  $d_{ij}$  is the shortest path between nodes  $i$  and  $j$ . In a binary graph, the *shortest path* means that it contains the minimum number of edges from node  $i$  to  $j$ . In a weighted graph, the *shortest path* means the smallest sum of the edge weights among all possible paths.

### Clustering coefficient

The *clustering coefficient* (CC; Watts and Strogatz 1998) is an important parameter, together with *average path length* to determine if a network possesses small-world property. As Equation 4.22 shows,  $k$  is the number of nodes,  $\#E_i$  is the number of actual edges between node  $i$  and its neighbors, and  $D_i(D_i - 1)/2$  calculates the maximum number of edges between node  $i$  and its neighbors. The ratio of each node's actual edges to their maximum edges can indicate the extent to which nodes clustered with each other in a network.

$$CC(G) = \frac{1}{k} \sum_{i=1}^k \frac{\# \text{ of } E_i}{D_i(D_i - 1)/2} \quad (4.22)$$

### Edge betweenness

For an edge in a network, its *edge betweenness* value equals the number of shortest paths between any two nodes that pass along it (Girvan and Newman 2002).

$$B_e = \sum_{i,j \in N, i \neq j}^k \frac{\# \text{ of } Path_{i,e,j}}{\# \text{ of } Path_{i,j}} \quad (4.23)$$

#### 4.4.2. Community detection using head/tail breaks

A *community* in a network, also called a *cluster* or a *module*, means a group of nodes in which the links are denser than in nodes outside the group (Newman 2004, Fortunato 2010). Since the degrees of nodes in a random or regular network vary little from one node to another, it is difficult to have many communities or clusters in such a network. In contrast, the degrees of nodes in a complex network tend to be very heterogeneous, which leads to a network with a community structure (Girvan and Newman 2002, Newman 2003, 2004). Community structure is a very critical aspect of a complex network because it reflects the internal relationship among nodes and a network's organization (Lancichinetti et al. 2008).

Many algorithms or methods have been proposed over the past 20 years to detect communities. One of the most fundamental works is by Girvan and Newman (2002). It introduces edge betweenness (Equation 4.23) as a key measurement for detecting communities because each edge of high betweenness value is probably the only connection between one community and another. In other words, edge betweenness reflects how important a role one edge plays as a bridge between different groups in the network. Removing such edges can reveal the network's community structure. Girvan and Newman (2002) implemented a hierarchical decomposition process by deleting edges, one at a time, in the descending order of their edge betweenness values. The resulting community structure is a dendrogram, from which we can see that clusters have an overlapped or nested relationship.

For a complex network, we tend to think not just in terms of degrees of nodes, but also that the nested clusters are inclined to exhibit a scaling hierarchy. This means that there are far more small clusters than large ones. To effectively discover such scaling structures of complex networks, we propose a community-detection method using head/tail breaks. It applies iterative head/tail breaks on edge betweenness scores. It will keep retrieving the head part of the edges as the sub-network until the head percentage of the sub-network disobeys the preset threshold, as the pseudo-code shows.

```
Recursive Function Head/tailCommunity (network, head)
  Extract all subnetworks of the input network;
  Foreach subnetwork
    Calculate edge betweenness for each edge;
    Calculate head percentage in this subnetwork;
    If (head percentage >= head) //this subnetwork is homogenous
```

```

Add subnetwork into EdgeList;
Else
  Head/tailCommunity (subnetwork, head);
Return EdgeList;
End Function

```

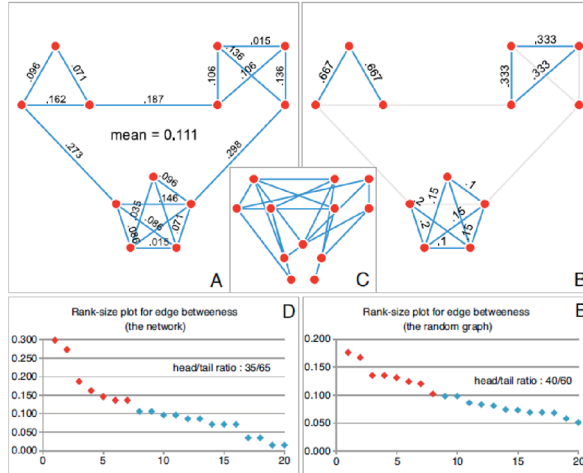


Figure 4.6: (Color online) Illustration of the community-detection algorithm using head/tail breaks  
(Source: Paper V)

To objectively determine the preset threshold, we used an equivalent random graph which contained the same number of nodes and edges as the original network. In this algorithm, the preset threshold was the head percentage of the random graph. We also introduced the concept of least, or homogenous, community, in which the head percentage was larger than, or equal to, one of its equivalent random graphs. As Figure 4.6 presents, a fictive social network comprised 12 nodes and 20 links (Figure 4.6A), its head percentage of edge betweenness was 35 (Figure 4.6D), and the head percentage in its random counterpart was 40 (Figure 4.6C and 4.6E). Community detection using head/tail breaks could then be applied to the fictive network. After removing the edges below the mean value, three sub-networks and an isolated node remained (Figure 4.6B). This procedure was then repeated for each sub-network until we found all of them were least communities. As a result, we found four groups with sizes 5, 3, 3, and 1. In this way, we revealed the community structure of a complex network using head/tail breaks.

## 5. Results and discussion

### 5.1. Overview

This chapter summarizes the outcome of the seven papers and discusses their scientific contributions. The main results concentrate on the topological and scaling analysis of geospatial big data for better understanding scaling structure of the geographic space and how it influences human activities. This section presents results and discussion from the applications of complexity science methods; to geographic space at different scales (global, country, and city levels); to related human activities, as reflected by several LBSM datasets; to development of theories of fractal geometry and scaling law in terms of smooth curves, fractional ht-index, and community structure. The following content is organized according to this sequence, and the results of each paper are subsequently presented and discussed.

### 5.2. Paper I: The heterogeneity of OSM data and community

This paper investigated the global OSM historical database to study the underlying scaling property of geographic space and involved human activities. This XML-format database stores approximately 2 billion geographic features, 1 million users, and 2.7 billion contributions. Therefore the dataset in this study is very big (692 GB when uncompressed, with an eight-year time span from April 9, 2005 to February 5, 2013). This study extracted the related historical and attribute information of all the elements of the entire OSM history dump (see details in Section 3.3.1). The paper then employed head/tail breaks and power-law detection to characterize the scaling pattern of the global OSM database from both the user and the element perspectives.

The findings from the element perspective were as follows: (1) The number of edits, users, and size of each element greatly varied from one to another. (2) The scaling patterns among each attribute were remarkable, indicated by very high ht-index values and power-law metrics (Table 5.1). The statistical results reflected a great heterogeneity of the Earth's surface, or far more small elements than large ones. (3) The global spatial distribution of elements (at the country level) demonstrated a notable power law. (4) Head/tail breaks and the induced ht-index effectively complemented the mathematical heavy-tailed distribution characterization, especially when the data was too big to handle.

Table 5.1: The results of scaling analysis of OSM elements  
 (Note: The scaling analysis is conducted in terms of users, edits, and sizes, respectively, #examined = number of elements at top hierarchical levels for power law detection)

	ht-index	#examined	alpha	p	max
#users	15	745,943	4.95	0	197
#edits	15	548,914	3.39	0.006	3,084
#size	12	479,004	2.37	0.13	5,118,276

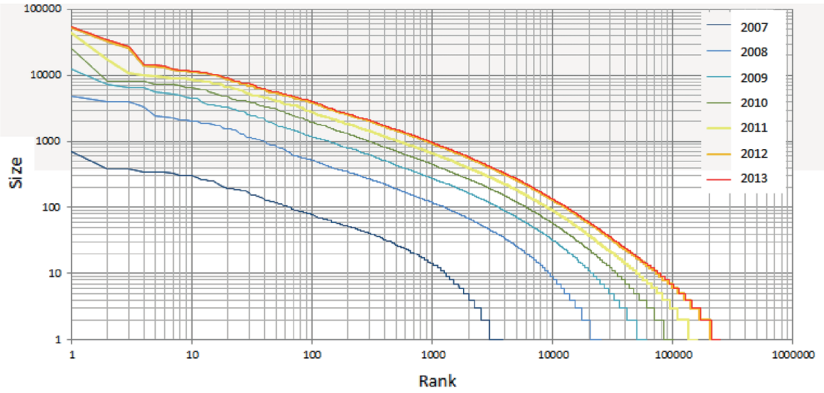


Figure 5.1: (Color online) The rank-size plot of degree distribution of co-contribution network from 2007 to 2013

The nonlinearity also existed from the user perspective. Approximately 30 percent of users made contributions to OSM elements. A clear scaling pattern was also detected in regard to the number of contributions among those users, indicating that there were far more inactive users than active ones. Furthermore, we built up co-contribution networks per annum from 2007 to 2013. We found that online mapping participation grew at a nonlinear pace, and the degree distribution of the co-contribution network also grew more power law-distributed annually (Figure 5.1). This was also seen by the increasing ht-index values of node degrees of networks year by year.

### 5.3. Paper II: A socio-geographic perspective on human activities

This paper investigates the relationship between social and geographic aspects of human activities in social media. To be more specific, it seeks correlation between social connections and check-in locations, through the scaling analysis of human movement behavior and socio-geographic networks. The data came from the former LBSM platforms of Brightkite and Gowalla in the US.

To accomplish the task, we extracted each user's first, or most recent, check-in locations as user locations, and then used them to build up natural cities through a TIN-based clustering method (Figure 2.6b). The study constructed three types of socio-geographic networks for each social media

platform: A people-people network, a location-location network, and a city-city network. The socio-geographic networks contained up to tens of thousands of nodes and tens of millions of links, based on social-media information from approximately 50,000 users and their 6 million check-in locations. The details of network construction are illustrated in Section 3.4.3. The constructed networks were then utilized for correlation analysis between social and geographic aspects of human activities.

Table 5.2: The related metrics of location-location and city-city network  
(Note: # = number, ht-index = ht-index value of edge weight,  $R^2$  (population) = the R-square value between node weighted degree and population,  $R^2$  (location #) = the R-square value between node weighted degree and location number)

	Location-location	City-city
# of nodes	19,450	451
#of edges	3,261,608	9,450
ht-index	11	6
$R^2$ (population)	0.66	0.90
$R^2$ (location #)	0.39	0.91

The study illustrates the underlying complexity and scaling hierarchy of human activities in social media from a socio-geographic perspective. Relying on the head/tail breaks, the study found that there clearly were scaling patterns in the user check-in patterns, edge weights of socio-geographic networks, and the population at user locations and natural cities. As Table 5.2 presents, it was significant that the node degree of the constructed socio-geographic networks was highly correlated with the population at locations (mostly with R-square = 0.7) or cities (greater than 0.9). The correlation results led to two new insights about human activities in social media: Either the first check-in or the most frequent check-in could possibly be a good proxy of the user location; and at the country level, the number of social connections did not correlate well with geographic proximity at the country level, but depended on the characteristics of a city.

This paper is another showcase of how effective head/tail breaks method is for big-data analytics and visualization. Because the head part was important and was self-similar to the data with a scaling property, we could select the top few heads to measure and visualize the nonlinearity of human activity. Moreover, identifying the head and tail parts had large implication for geospatial big-data mining and analytics. In the case of Brightkite, the “head” of unique check-in locations was only a very small part (5 percent) but accounted for the majority of all check-in locations. It then led us to successfully capture the image between people’s social connections and physical locations.

#### 5.4. Paper IV: Why topology matters in predicting human activities?

This paper studies why the topology of space, or the topological relationship of natural streets at the city scale, matters for predicting human activities. This

paper also made a comprehensive comparison between geometric and topological representations to see their effectiveness in capturing or predicting human activities, respectively. In this study, the geometric representations at the city level particularly refer to segment-based models, while topological representations refer to natural streets or axial lines. This paper showed that segment-analysis methods are ineffective to predict human activities because they are essentially geometric, focusing on details such as segment length and the turning angle of a pair of intersected segments, which hardly have scaling property. In comparison, topological models are effective, as they enable us to see the underlying scaling of far more less-connected streets than well-connected ones.

To support this argument, we conducted a series of experiments using London streets and one week of tweet location data, based on the related concepts of natural streets and natural street segments (or street segments for short), axial lines and axial line segments (or line segments for short). As shown in Figure 5.2, we found that the distributions of connectivity values for natural streets and axial lines possessed striking scaling properties, while the street and line segments did not, which was also indicated by some power-law fitting metrics. The scaling property of street connectivity was further seen for predicting human activity.

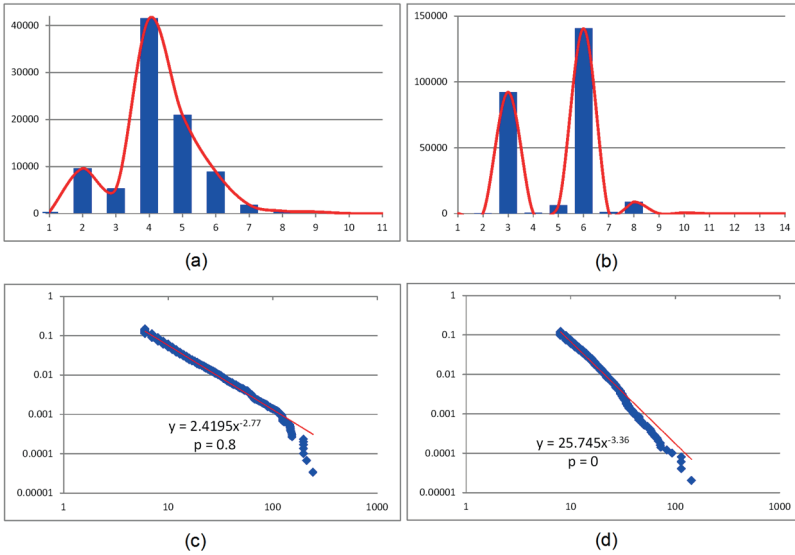


Figure 5.2: (Color online) Distributions of connectivity values in different street representations

(Note: (a) street segment, (b) line segment, (c) natural streets, and (d) axial lines)

We chose nine study areas in London, of which five were from the city center to the periphery. The other four represented high concentrations of geo-located tweets that were scattered around the city. We subsequently found that natural streets were the best representation in terms of predicting human activity,



followed by axial lines. Neither street segments nor line segments bore good correlation between network parameters and tweet locations. To further support our argument that topology matters in predicting human activities, rather than do geometric factors, we conducted a segment analysis based on centrality measures of degree, betweenness, and closeness. Betweenness and closeness were computed under different kinds of segment-segment geometric relationships in terms of minimum length, least angle, and fewest turns (Hillier and Lida 2005). We select one study area that was located in the center of London to show the correlation results (Table 5.3). All correlations between segment metrics and the number of tweet locations were remarkably low ( $R^2 < 0.1$ ). However, correlations significantly improved ( $R^2 > 0.5$ ) when we added the centrality measures of street segments and line segments into individual natural streets. Note that the results of other study areas also followed this trend.

Table 5.3: The correlation results in central London  
(Note: The correlation analysis is conducted between the centrality metrics and the number of tweet locations. M = Metric, F = Fewest turns, A = Angular)

	Degree	Betweenness			Closeness		
		M	F	A	M	F	A
Street segment	0.003	0.022	0.041	0.052	0.014	0.111	0.083
Line segment	0.02	0.009	0.009	0.031	0.001	0.077	0.073
Natural streets	0.85	0.61	0.66	0.64	0.85	0.86	0.86
Axial lines	0.31	0.37	0.07	0.28	0.42	0.41	0.42

Based on these findings, we conclude that natural street- or axial line-based space syntax, or these general types of topological models, significantly coincide with human travel behavior or how humans conceptualize distances or spaces. Topology among natural streets or axial lines make it possible for us to perceive the underlying scaling hierarchy of streets, with numerous least-connected streets, a very few most-connected streets, and some streets that are somewhere between the least- and most-connected. This scaling or fractal structure makes human activities or urban traffic predictable, but in the sense of collective behavior, rather than individual, human, moving behavior.

### 5.5. Paper VI: Spatial distribution of city tweets and their densities

The paper investigates urban space structures and how they shape the spatial distribution of human activities. More specifically, the paper describes the spatial distribution of the number of geo-tagged tweets and their densities at the city level over street blocks, which are used to form the area of natural city (red patches in Figure 2.6d). Unlike previous studies using top-down geographic units such as administrative city boundaries and census tracts (e.g. Gehlke and Biehl 1934, Clark 1951, Wang and Zhou 1999), geographic units in this study were totally bottom-up (the OSM street network). Top-down units are imposed by government and may be outdated and subjective, whereas bottom-up ones are up-to-date (dynamic) and objective, and have very fine spatial-temporal

resolution (Figure 5.3 a and b). Therefore, this paper attempts to find the city forms and how they affect human activities in the geospatial big-data context.

The study selected six cities (Paris, Toulouse, Berlin, Munich, London, and Birmingham) from three European countries (France, Germany, and the United Kingdom). City boundaries were extracted from the street blocks of the entire country (see Section 3.4.2 for details). Each city contains tens of thousands of street blocks. The study also crawled approximately 5 million geo-referenced tweets over three countries, including user ID, latitude, longitude, and timestamp. We then assigned a number of tweet locations to each city block. We calculated each city's topological center as the city center. The topological center was determined by the adjacency relationship among street blocks. We started from the border blocks marked with zero, found their neighboring blocks, and marked those with 1, and so on until all blocks were traversed. The blocks with the highest numbers were the topological center. We used the topological center, rather than the geometric one, because the topological center considers the spatial heterogeneity of street blocks. After pinpointing the city center, we derived the spatial distribution of tweets and their densities from the city center to the periphery.

The findings of the study are multi-fold. Using much finer geographic units and the topological center, the tweet numbers first rose and then descended from the city centers to the borders (Figure 5.3c), and tweet densities followed an overall decreasing trend (Figure 5.3e). The plot lines of both tweet numbers and densities noticeably fluctuated. However, such fluctuations did not exist when using the authorized geographic units (Figure 5.3d and 5.3f). Furthermore, the decreasing trend of tweet densities could disappear if city boundaries were arbitrarily delimited. These observations gave us deep insight into geographic research in the geospatial big-data era. The remarkable ups and downs on the distribution line could be the real picture of human activities based on the scaling structure of urban space, given the assumption that geo-tagged tweets are a good proxy of population in cities. Additionally, the objectively defined natural cities provide a new, effective channel for better understanding the urban space.

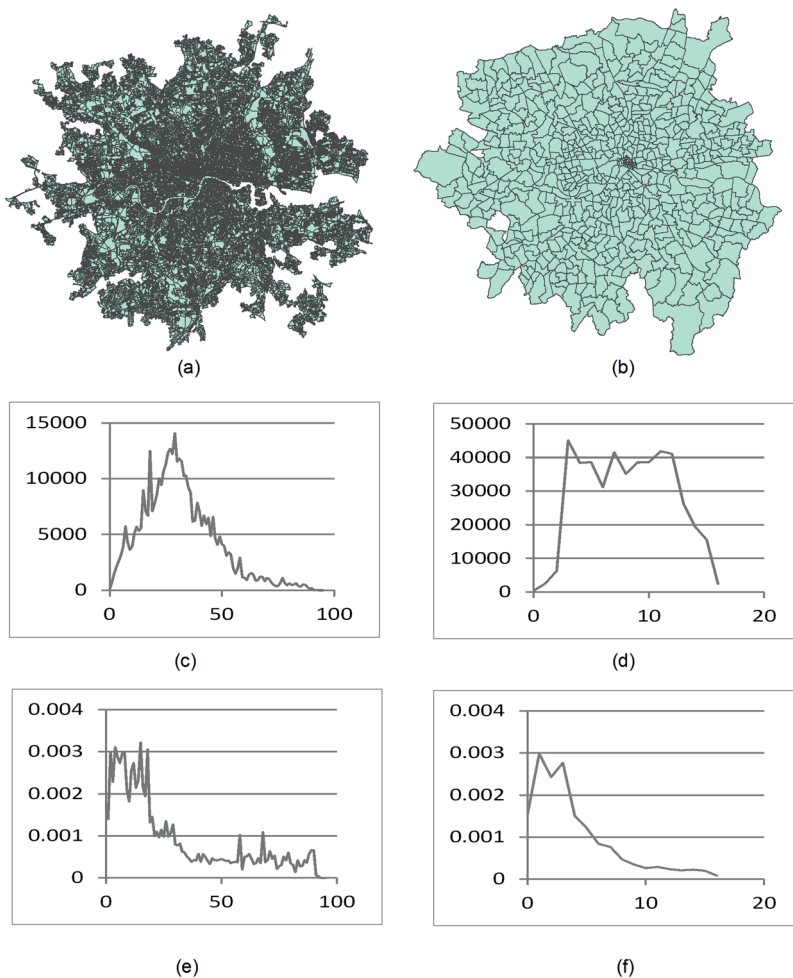


Figure 5.3: (Color online) Spatial distribution of tweet numbers and densities in London (Note: The left column shows the results based on the natural city boundary and natural street blocks. The right column shows the administrative boundary and census tracts)

### 5.6. Paper III: A smooth curve as a fractal under the third definition

Previous empirical studies showed the fractal or scaling pattern of geographic space at different scales. This paper studies the fractal property of individual geographic features. Traditional fractal geometry claims that a curve is fractal if its shape bears the property of self-similarity by following the power-law distribution either strictly (definition 1) or statistically (definition 2) (Mandelbrot 1982). Based on the two traditional definitions, many smooth curves, such as circles and smoothed cartographic curves, are not fractal. To make

fractals more accessible and universal, Jiang and Yin (2014) proposed a third definition of fractal (see Section 2.3 for more details).

Figure 5.4 demonstrates how this new definition applies for a simple polyline. The polyline consists of 10 bends, which are recursively calculated. Furthermore, the scaling of far more small bends than large ones recurs twice (ht-index = 3):  $x_1 + x_2 + x_3 > x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$ , and  $x_1 > x_2 + x_3$ , so the polyline is a fractal. Under the third definition, almost all smooth curves are fractal, so long as the scaling of far more small bends than large ones recurs multiple times. To further verify this viewpoint, the work presents related analyses of four types of smooth curves: A half-circle, a half-ellipse, the logarithmic spiral (Thompson 1917, Bader 2013), and the British coastline (after being smoothed).

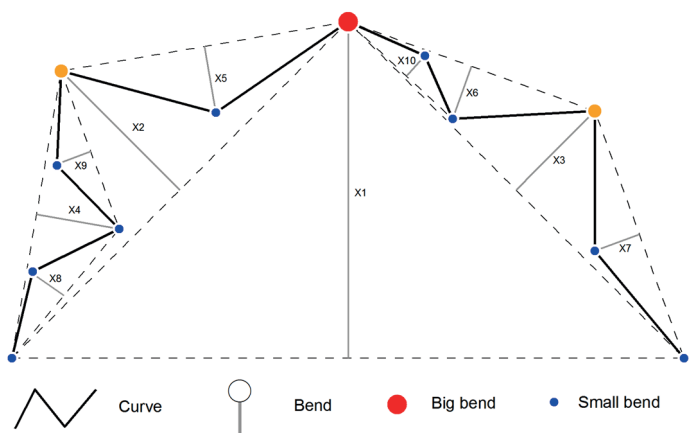


Figure 5.4: Illustration of the new definition of fractal  
(Source: Paper III)

For each smooth curve, we first partitioned it into numerous bends, as illustrated in Figure 5.4, and then examined if the sizes of bends were power-law distributed and the scaling pattern of far more small bends than large ones recurred multiple times (as indicated by the ht-index). This power-law detection method was based on maximum likelihood, which is the most rigorous detection method (Clauset et al. 2009). As a result, we found that these curves were all fractal under the new, relaxed, third definition. Table 5.4 shows that the sizes of bends for nearly each curve were power-law distributed, as the exponent (alpha) and p-value indicate, and had an ht-index value greater than 3.

Table 5.4: Power law metrics and ht-index for smooth curves

Curves	Ht-index	Power law	
		Alpha	p
Half-circle(128)	4	1.61	0.65
Half-circle(250)	4	1.61	1
Half-circle(500)	5	1.61	1
Half-circle(1000)	5	1.61	1
Half-circle(6000)	5	1.61	1
Half-ellipse(upper, 5998)	5	1.51	0.01
Half-ellipse(left, 5998)	4	1.51	0.05
Logarithmic spiral (37)	NA	NA	NA
Logarithmic spiral (74)	4	1.67	0.12
Logarithmic spiral (138)	4	1.71	0.46
Logarithmic spiral (300)	5	1.58	0.25
Logarithmic spiral (720)	5	1.61	0.52
Coastline (10,859)	7	2.15	0.83
Smoothed coastline (62,550)	7	1.56	0.68
Coastline (14)	3	2.52	0.02
Smoothed coastline (25,612)	5	1.57	0.01

This paper develops novel understandings of fractal geometry by demonstrating the universality of the fractal nature. The third definition gives not only a rule to determine if a shape is fractal, but also a completely new perspective on how to look at it, which is fundamentally different from the viewpoint of Euclidean geometry. Fractal geometric thinking leads us to decompose a line into a number of dissimilar bends, whose sizes are long-tail distributed, whereas Euclidean geometric thinking leads us to split it into a number of similar segments, whose lengths follow a normal distribution. This finding has far-reaching implications for understanding geographic features and helps us place geospatial big-data research into the introduced theoretical framework.

### 5.7. Paper VII: How complex is a fractal?

The third definition of fractal is based upon head/tail breaks, or its induced ht-index. The ht-index ranges from one to positive infinity and measures effectively how fractal a geographic feature is (the larger the ht-index value, the more fractal the geographic feature is). To make this section self-contained, here is a working example of calculating ht-index of a data series of 10 values following Zipf's Law (Zipf 1949):  $\{1, 1/2, 1/3, \dots, 1/10\}$ . There are two mean values obtained using head/tail breaks (Table 5.2). The first mean of 0.28 split the data series into a head  $\{1, 1/2, 1/3\}$  and a tail  $\{1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10\}$ , and the second mean of 0.61 separated the head into a new head  $\{1\}$  and a new tail  $\{1/2, 1/3\}$ . Therefore, the ht-index of this data series is 3. More than that, each single value in the series can be assigned to an ht-index value during the head/tail breaks process. For example,  $\{1\}$  has an ht-index of 3,  $\{1/2, 1/3\}$  has an ht-index of 2, and the rest has an ht-index of 1. In this regard, the data series is hierarchically classified, and we know how fractal each single value is. However, ht-index is not sensitive in capturing the small changes from one data series to another. If we add five further values into this data series of

$\{1/11, 1/12, 1/13, 1/14, 1/15\}$ , the ht-index value of the updated data series is still 3 (Table 5.5). This is a sensitivity issue with the ht-index.

Table 5.5: The head/tail breaks statistics of 10 and 15 values respectively

#Data	#head	%head	#tail	%tail	mean
10	3	30%	7	70%	0.28
3	1	33%	2	67%	0.61

#Data	#head	%head	#tail	%tail	mean
15	4	26%	11	74%	0.22
4	1	25%	3	75%	0.52

Some studies have noted the lack of sensitivity of the ht-index. Gao et al. (2016a, 2016b, 2017) proposed three different indexes for improvements: Cumulative rate of growth (CRG), ratio of areas (RA) in a rank-size plot, and unified metrics. All three indexes can well handle the sensitivity issue, as the data series becomes more complex than in its previous state. However, there are certain disadvantages for each of these indexes. To be specific, both CRG and RA suffer from their interpretability (Gao et al 2017), and all three only calculate how fractal the entire data series is, but fail to capture the contribution of each single value to the fractal.

This paper develops the so-called *fractional ht-index* (fht-index) as an improvement of ht-index to address these above-mentioned problems. Given a data series, fht-index provides decimals after the integral ht-index to overcome the sensitivity problem. Moreover, fht-index assigns a continuously hierarchical level to each value of the data series. The idea behind fht-index is the concept of a whole and sub-wholes. To determine if a data series containing  $k$  values is a whole, we first sorted the list from the largest value to the lowest and calculated the ht-index for each single value. We then examined if it met the condition  $ht-index(k) - ht-index(k-1) = 1$ . If the condition was satisfied, the data series could be considered as a whole. If not, we conducted the heavy-tailed distribution detection, as Section 4.2 introduced, to find its most-fitted function. This was done so that we could keep appending small numbers indicated by its fitting function until the data series became a whole. Within the whole, all sub-wholes could then be determined using the same condition, so we assigned the corresponding integral ht-index to the border value of each sub-whole. For those values between any pair of the border values, we calculated an equal interval using  $\frac{1}{\text{number of values}}$ . The fht-indexes were assigned to each value using the function  $f(\text{interval}_j) = (j * \text{interval})^2$ , in which  $j$  is the index of each interval. The fht-index of the data series was assigned to the largest value. FHTcalculator (2017) was developed for conducting the above processes. Table 5.6 shows both ht- and fht-indexes for each of the 10-number arrays. It is worth noting that the fht-index increases from 3.004 to 3.05 with the addition of five small values.

Table 5.6: Fht-index versus ht-index of a data series using FHTcalculator

Size	Ht-index	FHt-index
1	3	3.004
1/2	2	3.000004
1/3	2	3
1/4	1	2.694
1/5	1	2.444
1/6	1	2.25
1/7	1	2.111
1/8	1	2.028
1/9	1	2
1/10	1	1.925

This paper contributes to both understanding and development of fractal geometry theory from several aspects. Fht-index transforms the formerly discrete ht-index values into continuous ones. Therefore, fht-index provides a more accurate way of quantifying the fractal or scaling structure of geographic features. The increased accuracy of fht-index conquers the sensitivity issue and, more importantly, uncovers a continuous-scaling hierarchy that is compatible with ht-index. The fht-index is not only accurate, it can also help detect different levels of fractal of a dataset from its immature stage (not whole) to a mature one (whole). In this way, we can investigate the fractal geometry from a dynamic perspective.

## 5.8. Paper V: Least community as a homogeneous group in complex networks

Community is an important aspect of the topological structure of a complex network. As mentioned earlier, a complex network tends to have communities because the distribution of links can vary significantly from one node to another. Bearing such heterogeneity of links, fractals can emerge from the communities of a complex network. This paper attempts to extend the application of third definition of fractal and scaling law to the network space, by uncovering the fractal or scaling pattern of the community structure of a complex network.

As Girvan and Newman (2002) suggested, edge betweenness is vital for finding communities. In this work, the heterogeneity of links is characterized by the scale-free property of edge-betweenness scores. In order to describe precisely such heterogeneity, this paper relies on a new concept of least community, whose links are as homogeneous as a random graph (see section 4.4.2 for details). For any real-world network, we found its random graph counterpart that maintains the same number of nodes and links. Next, we derived edge-betweenness scores of all links in both the real-world network and its random counterpart, respectively. By applying the head/tail division rule, the heterogeneity of the real-world network was determined by whether or not the network's head percentage was smaller than that of its random graph.

Table 5.7: Scaling analysis of derived communities from 8 networks  
(Note: #comm = number of all communities, head/tail = head/tail ratio of the corresponding random graphs with respect to their edge betweenness, NA = not available.)

	#comm	alpha	p	ht-index	head/tail
Internet	8,398	3.3	0.05	5	43/57
Scientist	7,222	2.82	0.11	5	43/57
Protein	3,767	2.45	0.03	5	43/57
Brightkite	206	2.21	0.35	5	48/52
Erdos	3,262	2.69	0.16	5	39/61
Street	2,642	2.05	0.16	4	41/59
Gowalla	453	1.92	0.12	4	49/51
WWW	6	NA	NA	2	48/52

We developed a new community detection algorithm, as Section 4.4.2 illustrated, by recursively applying the head/tail breaks method on edge betweenness in the entire network and its sub-networks. The study applied the new algorithm on eight complex networks, ranging from social networks in Brightkite and Gowalla, to biological networks (protein interaction), to informational networks (the World Wide Web), to a technological network (the Internet). Furthermore, we also applied the algorithm to each network's random (Erdős and Rényi 1959), scale-free (Barabási and Albert 1999), small-world (Watts and Strogatz 1998) counterparts using the same numbers of nodes and edges. Table 5.7 shows the results. Interestingly, the study found: The edge-betweenness values in those networks were very heterogeneous, as indicated by power-law metrics and large ht-index values; and that there were nested relationships among heterogeneous and homogenous communities. These nested relationships were akin to the human brain whose structure is organic and fractal. To statistically describe the complex structure, we also found that the sizes of the derived communities of each network followed a power-law distribution, indicating that there were far more small communities than large ones. In other words, the fractal or scaling structure was revealed in each network's communities.



## 6. Conclusions and future work

### 6.1. Conclusions

Today's global issue of urbanization has created a great interest in better understanding the underlying mechanism of geographic space and its relationship with human activities. Conventional geospatial analysis is geometrically and statistically limited in describing the heterogeneous nature of the geographic space, as it is based on Euclidean geometry and Gaussian statistics. Meanwhile, it has become increasingly difficult for traditional GIS tools and methods to tackle the massive geospatial big data. In this respect, we should explore geographic space by a different way of thinking and doing.

The thesis was initially motivated by the shortage of conventional geospatial analysis, and was further triggered by emerging geospatial big data. Geospatial big data offers not only a large, fine-grained data source, but also a new paradigm for geospatial analysis. The new paradigm is fundamentally different from the analysis in the small-data era in both geometric and statistical manners. Geometrically, big data obviously exhibits a fractal structure that is hard to capture by Euclidean geometry. Statistically, big data is likely to show a heavy-tailed distribution, in which there is no well-functioning mean value for characterization. Therefore, a new theoretic framework, based on fractal geometry and Paretian statistics, should be designed for geospatial analysis in the big-data era. Together with the new paradigm, big-data analytics also must be equipped with data-intensive computing techniques. In line with these requirements, the three objectives of this thesis (Section 1.2) are established to better understand geographic space in terms of how it looks and works.

The first objective of this thesis was to develop new understanding of fractal geometry. Based on the third definition of fractal, Paper III observed that not only do irregular shapes tend to be fractal, but regular shapes, like smooth curves, are also fractal, given the right perspective and scope. Conventionally speaking, a smooth curve is just a collection of vertices or segments, so it has no chance of being fractal. However, changing the perspective to a set of recursively defined bends can lead us to see the fractal property of the line feature. This bottom-up examination of fractal opened up a new horizon of looking at geographic features. Furthermore, Paper VII improved the quantification of fractal by proposing the fht-index. This novel index extended the integral ht-index to fractional to provide a more accurate, continuous scaling hierarchy of a fractal structure. This improvement greatly helped overcome the sensitivity issue. Moreover, fht-index can measure the fractal of every single part of geographic features or space.

To meet the second objective, the thesis developed techniques of geospatial big-data processing and modeling, and designed a complexity science methodology, including heavy-tailed distribution detection and head/tail breaks, along with some complex network analysis. On the one hand, massive amounts of geospatial data from LBSM platforms, such as OSM and Twitter, were handled by the developed techniques. Paper I processed and extracted billions of rec-

ords of element-attributed information and construct a network of approximately 1 million users from the global OpenStreetMap history database (approximately 700 GB). Paper II built up big socio-geographic networks, which contain up to tens of thousands of nodes and tens of millions of links. Paper VI derived natural cities from countrywide street blocks. On the other hand, the complexity science methodology was effective to obtain the scaling hierarchy and examine the scaling properties of geographic features, network space, urban space, and human activities (Papers I, II, III, IV, V, and VI). Relying on head/tail breaks, the methodology fit very well to the big-data analytics. Through the studies, head/tail breaks proved to be a very powerful analysis and visualization tool, and its induced ht-index complemented mathematical heavy-tailed distribution characterization.

The third objective was accomplished by investigating two major issues: How geographic space looks and how it shapes human activities. The first issue was comprehensively examined by six papers. Paper III revealed the fractal of individual line features such as the UK coastline. Papers IV and VI examined the scaling structure of a city in terms of natural streets and street blocks. Paper II found the scaling pattern of natural cities in the US mainland. Paper I found the global scaling property through the sizes of OSM elements. Paper V identified the striking scaling property of communities in the network space.

The second issue was covered by four papers. Paper I illustrated that both user contribution to the OSM database and the degree of the co-contribution network showed a clear power-law distribution, due to the heterogeneous nature of the Earth's surface. Paper II detected a high correlation between users' social connections and spatial distribution based on natural cities across the country. Paper IV employed the scaling structure of natural streets to capture and predict human activities. Paper VI explored the spatial distribution of human activities shaped by the scaling structure of street blocks.

The thesis accomplished the tasks of understanding the scaling structure of geographic space and its involved human activities in the big-data context. First, the new paradigm of geospatial analysis based on fractal geometry, particularly the third definition of fractal and power-law statistics, is widely used throughout this study. Second, this thesis devised a number of effective and efficient data-processing and modeling techniques, which were applied to massive LBSM data for uncovering knowledge of geographic space and its involved human activities. Third, the derived scaling hierarchies, power-law metrics, and network measures provide new insights into the heterogeneity of the geographic space and help us to understand how it shapes human behavior and activities. Furthermore, this work generated valuable data and corresponding source codes, which could be useful in the field of GIScience in the geospatial big-data era.

## **6.2. Future work**

Although the thesis provided useful knowledge of the underlying scaling structure of geographic space and its involved human activities, with the help of geospatial big data, it is far from enough to say the findings are definitive or

exhaustive. There should be more effort to extend or consolidate the findings. Additional studies should address several latent research problems in the future.

The primary focus of the thesis was studying geographic forms or city structures. Geospatial big data provides us with not only enormous location-related information, but also vast amounts of temporal information. In addition to the structure of geographic space, endeavors in the future can examine geographic processes and urban dynamics. For example, we can study the nonlinear dynamics of the evolution of global user-mapping activities in OSM, apply the  $h$ - $t$ -index to capture the subtle change of the unfolding process of urbanization, or identify the change of the scaling pattern of human movement behavior from time to time at different temporal granularities (day, week, and month).

This thesis relies heavily on head/tail breaks and  $h$ - $t$ -index in two aspects: To determine if a set or pattern is fractal; and to derive the scaling hierarchy or hierarchical levels of data with a heavy-tail distribution. Although this method is straightforward and well-known, there is a sensitivity issue when setting the head percentage. Throughout the studies, we employed 40 as the default head percentage, since it made good sense for the minority. However, it needs more experiments to find statistical support for the default setting of head percentage. Additionally, it is worthy of considering the sensitivity issue on the obtained arithmetic mean values. For example, a sensitivity study on how different mean values lead to different sets of natural cities would develop new and deep insights on scaling analysis.

The thesis showed some data-intensive computing on geospatial big data at various scales. Further improvements concerning big-data computation can be made in two areas. First, there is still some space to improve algorithm performance in terms of speed and memory consumption. Second, it is imperative to develop a scalable, distributed system that supports parallel computing, such as a system based on Hadoop. These improvements can greatly cope with a larger dataset, such as the construction of worldwide natural cities and natural streets. Additionally, we will integrate machine-learning algorithms into geospatial big-data modeling to more effectively detect the human mobility patterns or urban clusters.

In addition to extending present findings, it is also necessary to bring the developed methodology and procured knowledge to other geospatial research fields. One potential field is map generalization. The research community in cartography has long paid too much attention to generalizing geographic features based on Euclidean geometry, but ignored the inherent fractal or scaling structure (Jiang et al. 2013, Jiang 2015c). Future studies can focus on this area to implement map generalization based on the scaling law and fractal geometry. Another promising direction should be the morphological design at the individual level (a geographic object) or the collective level (the layout of a building complex) by adopting the characterized heterogeneity or scaling property.



## References

- Apache (2017a), What is Apache Hadoop, <http://hadoop.apache.org/>, accessed November 2017
- Apache (2017b), HDFS Users Guide, <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html>, accessed November 2017
- Bader M. (2013), *Space-Filling Curves: An introduction with applications in scientific computing*, Springer: Berlin.
- Bak P. (1996), *How Nature Works: The science of self-organized criticality*, Springer-Verlag: New York.
- Ball P. (2012), *Why Society Is a Complex Matter: Meeting twenty-first century challenges with a new kind of science*, Springer Science and Business Media: New York.
- Barabási A. L. and Albert R. (1999), Emergence of scaling in random networks, *Science*, 286(5439), 509–512.
- Barthelemy M. (2004). Betweenness centrality in large complex networks. *The European Physical Journal B*, 38, 163–168.
- Batty M. and Longley P. (1994), *Fractal Cities: A geometry of form and function*, Academic Press: London.
- Benguigui L. and Czamanski D. (2004), Simulation analysis of the fractality of cities, *Geographical Analysis*, 36(1), 69–84.
- Bennett J. (2010), *OpenStreetMap: Be your own cartographer*, PCKT Publishing: Birmingham.
- Bettencourt L and West G (2010) A unified theory of urban living, *Nature* 467(7318): 912–913.
- Bolin Centre Database (2018), Stockholm temperature from 2013 to 2016, [https://bolin.su.se/data/stockholm/raw\\_individual\\_temperature\\_observations.php](https://bolin.su.se/data/stockholm/raw_individual_temperature_observations.php), accessed January 2018.
- Boyd D. M. and Ellison N. B. (2008), Social network sites: Definition, history, and scholarship, *Journal of Computer-Mediated Communication*, 13, 210–230.
- Brelsford C, Martin T, Hand J, Bettencourt L.M.A. (2015), The topology of cities, *SFI Working Paper*, 15–06–021.
- Brockmann D., Hufnagel L. and Geisel T. (2006), The scaling laws of human travel, *Nature*, 439, 462–465.
- Cattani C. and Ciancio A. (2016), On the fractal distribution of primes and prime-indexed primes by the binary image, *Physica A*, 460, 222–229.
- Chen Y. (2009), A new model of urban population density indicating latent fractal structure, *International Journal of Urban Sustainable Development*, 1(1–2), 89–110.
- Chen Y. (2011), Modelling fractal structure of city-size distributions using correlation functions, *PLOS ONE*, 6(9): e24791. doi:10.1371/journal.pone.0024791.

- Chen Y. (2015), Power-law distributions based on exponential distributions: Latent scaling, spurious Zipf's law, and fractal rabbits, *Fractals*, 23(2): 1550009.
- Christaller W. (1933), *Central Places in Southern Germany*, Englewood Cliffs, NJ: Prentice Hall.
- Cho E., Myers S. A., and Leskovec J. (2011), Friendship and mobility: user movement in location-based social networks, *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, U.S.A, 1082–1090.
- Clark C. (1951), Urban population densities, *Journal of the Royal Statistical Society: Series A (General)*, 114(4), 490–496.
- Clauset A., Shalizi C. R., and Newman M. E. J. (2009), Power-law distributions in empirical data, *SIAM Review*, 51, 661–703.
- Cohen R. and Havlin S. (2010), *Complex Networks: Structure, Robustness and Function*, Cambridge University Press: Cambridge.
- Corbett J. P. (1979), *Topological Principles in Cartography*, Technical paper 48, U.S. Dept. of Commerce, Bureau of the Census: Washington D. C.
- Cranshaw J., Schwartz R., Hong J. I., and Sadeh N. (2012), The livelihoods project: Utilizing social media to understand the dynamics of a city, *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 58–65.
- Diakoulakli D., Mavrotas G. and Papayannakis L. (1995), Determining objective weights in multiple criteria problems: The critic method, *Computers and Operations Research* 22(7), 763–770.
- Dubin R. (1988), Estimation of regression coefficients in the presence of spatially autocorrelated error terms, *The Review of Economics and Statistics*, 70(3), 466–474.
- Edmonds J. (1965), Paths, trees, and flowers, *Canadian Journal of Mathematics*, 17, 449–467.
- Egenhofer M. J. and Herring J. R. (1990), A mathematical framework for the definition of topological relationships, *Proceedings of the Fourth International Symposium on Spatial Data Handling*, International Geographical Union, Zurich 1990, 803–813.
- Erdős P. and Rényi A. (1959), On random graphs I, *Publicationes Mathematicae*, 6, 290–297.
- ESRI (2017a), ArcObjects Library Reference (Geometry), <http://resources.esri.com/>, accessed November 2017.
- ESRI (2017b), ArcMap Documentation, <http://desktop.arcgis.com/en/arcmap/10.3/manage-data/tin/fundamentals-of-tin-surfaces.htm>, accessed November 2017.
- ESRI (2017c), ArcGIS Network Analyst documentation, <http://www.esri.com/software/arcgis/extensions/networkanalyst>, accessed November 2017.
- Ferrari L., Rosi A., Mamei M., and Zambonelli F. (2011), Extracting urban patterns from location-based social networks, *3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, November 1, 2011, Chicago, IL, USA.

- FHTCalculator (2017), <https://github.com/dingmartin/FHTCalculator>.
- Flack J. C. and Krakauer D. C. (2011), Challenges for complexity measures: A perspective from social dynamics and collective social computation, *Chaos*, 21(3), 037108.
- Fortunato S. (2010), Community detection in graphs, *Physics Reports*, 486(3–5), 75–174.
- Frankhauser P. (1994), *La fractalit'e des structures urbaines [The fractals of urban structure]*, Economica: Paris.
- Freeman L. C. (1979), Centrality in social networks: conceptual clarification, *Social Networks*, 1, 215–239.
- Gao H. and Liu H. (2014), Data analysis on location-based social networks, *Mobile Social Networking: An innovative approach*, Springer, in: Chin A. and Zhang D. (editors, 2014), 165–194.
- Gao P. C., Liu Z., Liu G., Zhao H., and Xie X. (2017), Unified metrics for characterizing the fractal nature of geographic features, *Annals of American Association of Geographers*, 1–17.
- Gao P. C., Liu Z., Tian K., and Liu G. (2016a), Characterizing traffic conditions from the perspective of spatial-temporal heterogeneity, *ISPRS International Journal Geo-Information*, 5(3), 34.
- Gao P. C., Liu Z., Xie M. H., Tian K. and Liu G. (2016b), CRG index: A more sensitive ht-index for enabling dynamic views of geographic features, *The Professional Geographer*, 68(4), 533–545.
- Gao S., Wang Y., Gao Y., and Liu Y. (2013), Understanding urban traffic flow characteristics: a rethinking of betweenness centrality, *Environment and Planning B: Planning and Design*, 40, 135–153.
- Gehlke C. E. and Biehl K. (1934), Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association*, 29(185A), 169–170.
- Girvan M. and Newman M. E. J. (2002), Community structure in social and biological networks, *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.
- Goodchild M. (2004), The validity and usefulness of laws in geographic information science and geography, *Annals of the Association of American Geographers*, 94.2, 300–303.
- Goodchild M. F. (2007), Citizens as sensors: The world of volunteered geography, *GeoJournal*, 69(4), 211–221.
- Gouveia C. and Fonseca A. (2008), New approaches to environmental monitoring: the use of ICT to explore volunteered geographic information, *GeoJournal*, 72(3–4), 185–197.
- Harary F. (1969), *Graph Theory*, Addison-Wesley: Reading, Mass.
- Hawelka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P. and Ratti C. (2014), Geo-located Twitter as proxy for global mobility patterns, *Cartography and Geographic Information Science*, 41(3), 260–271.
- Helbing D. (2007), *Managing Complexity*, Springer: New York.
- Hey T., Tansley S. and Tolle K. (2009), *The Fourth Paradigm: data intensive scientific discovery*, Microsoft Research, Redmond, Washington.

- Hillier B. and Iida S. (2005), Network and psychological effects in urban movement, in: A. G. Cohn and D. M. Mark (Eds.): *Proceedings of the International Conference on Spatial Information Theory*, COSIT 2005, Elllicottville, NY, USA, September 14–18, 2005, LNCS 3693, 475–490.
- Irving G. and Segerman H. (2013), Developing fractal curves, *Journal of Mathematics and the Arts*, 7(3–4), 103–121.
- Jenks G. F. (1967), The data model concept in statistical mapping, *International Yearbook of Cartography*, 7, 186–190.
- Jiang B. (2009a), Ranking spaces for predicting human movement in an urban environment, *International Journal of Geographical Information Science*, 23(7), 823–837.
- Jiang B. (2009b), Street hierarchies: a minority of streets account for a majority of traffic flow, *International Journal of Geographical Information Science*, 23(8), 1033–1048.
- Jiang B. (2013a), Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution, *The Professional Geographer*, 65(3), 482–494.
- Jiang B. (2013b), The image of the city out of the underlying scaling of city artifacts or locations, *Annals of the Association of American Geographers*, 103(6), 1552–1566.
- Jiang B. (2013c), Volunteered geographic information and computational geography: New perspective, In *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*, ed. Sui D., Elwood S. and Goodchild M., Dordrecht: Springer.
- Jiang B. (2015a), Head/tail breaks for visualization of city structure and dynamics, *Cities*, 43, 69–77.
- Jiang B. (2015b), Geospatial analysis requires a different way of thinking: The problem of spatial heterogeneity, *GeoJournal*, 80(1), 1–13.
- Jiang B. (2015c), The fractal nature of maps and mapping, *International Journal of Geographical Information Science*, 29(1), 159–174.
- Jiang B. (2015d), Wholeness as a hierarchical graph to capture the nature of space, *International Journal of Geographical Information Science*, 29(9), 1632–1648.
- Jiang B. (2015e), Axwoman 6.3: An ArcGIS extension for urban morphological analysis, <http://giscience.hig.se/binjiang/axwoman/>, University of Gävle, Sweden.
- Jiang B. (2016), Scaling as a design principle for cartography, *Annals of GIS*, 23(1), 67–69.
- Jiang B. (2018), A topological representation for taking cities as a coherent whole, *Geographical Analysis*, xx(x), xx–xx, DOI: 10.1111/gean.12145.
- Jiang B. and Brandt S.A. (2016), A fractal perspective on scale in geography, *ISPRS International Journal of Geo-Information*, 5(6): 95.
- Jiang B. and Claramunt C. (2004), Topological analysis of urban street networks, *Environment and Planning B: Planning and Design*, 31(1), 151–162.



- Jiang B. and Jia T. (2011a), Agent-based simulation of human movement shaped by the underlying street structure, *International Journal of Geographical Information Science*, 25(1), 51–64.
- Jiang B. and Jia T. (2011b), Zipf's law for all the natural cities in the United States: a geospatial perspective, *International Journal of Geographical Information Science*, 25(8), 1269–1281.
- Jiang B. and Jia T. (2012), Exploring human mobility patterns based on location information of US flights, Preprint, arXiv:1104.4578.
- Jiang B. and Liu C. (2009), Street-based topological representations and analyses for predicting traffic flow in GIS, *International Journal of Geographical Information Science*, 23(9), 1119–1137.
- Jiang B. and Liu X. (2012), Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information, *International Journal of Geographical Information Science*, 26(2), 215–229.
- Jiang B. and Miao Y. (2015), The evolution of natural cities from the perspective of location-based social media, *The Professional Geographer*, 67(2), 295–306.
- Jiang B. and Sui D. (2014), A new kind of beauty out of the underlying scaling of geographic space, *The Professional Geographer*, 66(4), 676–686.
- Jiang B. and Thill J. C. (2015), Volunteered Geographic Information: Towards the establishment of a new paradigm, *Computers, Environment and Urban Systems*, 53, 1–3.
- Jiang B. and Yin J. (2014), Ht-index for quantifying the fractal or scaling structure of geographic features, *Annals of the Association of American Geographers*, 104(3), 530–541.
- Jiang B. and Zheng R. (2018), Geographic space as a living structure for predicting human activities using big data, *International Journal of Geographical Information Science*, xx(x), xx–xx, DOI: 10.1080/13658816.2018.1427754.
- Jiang B., Duan Y., Lu F., Yang T. and Zhao J. (2014), Topological structure of urban street networks from the perspective of degree correlations, *Environment and Planning B: Planning and Design*, 41(5), 813–828.
- Jiang B., Liu X. and Jia T. (2013), Scaling of geographic space as a universal rule for map generalization, *Annals of the Association of American Geographers*, 103(4), 844–855.
- Jiang B., Yin J. and Liu Q. (2015), Zipf's Law for all the natural cities around the world, *International Journal of Geographical Information Science*, 29(3), 498–522.
- Jiang B., Zhao S., and Yin J. (2008), Self-organized natural roads for predicting traffic flow: a sensitivity study, *Journal of Statistical Mechanics: Theory and Experiment*, July, P07008.
- Jing T., Xiong F., Lei Y., and Zhan Y. (2015), Revising self-best-fit strategy for stroke generating, *In Advances in Spatial Data Handling and Analysis*, Springer, 183–192.
- Jungers W. L. (1984), *Size and Scaling in Primate Biology*, Springer: Berlin.

- Kaplan A. M. and Haenlein M. (2010), Users of the world, unite! The challenges and opportunities of social media, *Business Horizons*, 53, 59-68.
- Knox P.L. (1994), *Urbanization: introduction to urban geography*, Prentice-Hall: New Jersey, US.
- Koch R. (1998), *The 80/20 Principle: The secret of achieving more with less*, DOUBLEDAY: New York.
- Kulshrestha J., Kooti F., Nikraves A. and Gummadi P. K. (2012), Geographic Dissection of the Twitter Network, *The International AAAI Conference on Web and Social Media (ICWSM)*, the 6<sup>th</sup> International Conference, Dublin, Ireland, 202-209.
- Kumar S., Morstatter F., and Liu H. (2013), *Twitter Data Analytics*, Springer: Berlin.
- Kwak H., Lee C., Park H., and Moon S. (2010), What is Twitter, a social network or a news media? *Proceedings of the 19th International World Wide Web Conference*, 2010.
- Lancichinetti A., Fortunato S., and Radicchi F. (2008), Benchmark graphs for testing community detection algorithms, *Physical review E*, 78(4), 046110.
- Langlois P. (2013). The structure of the geographic space, *Simulation of Complex Systems in GIS*, 1-4.
- Lazer D., Pentland A., Adamic L., Aral S., Barabási A.-L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., and Van Alstyne M. (2009), Computation social science, *Science*, 323, 721-724.
- Li L., Goodchild M. and Xu B. (2014), Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr, *Cartography and Geographic Information Science*, 40(2), 61-77.
- Li S., Dragicevic S., Castro F.A., Sester M., Winter S., Coltekin A., Pettit C., Jiang B., Haworth J., and Stein A. (2016), Geospatial big data handling theory and methods: A review and research challenges, *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119-33.
- Long Y. (2016), Redefining Chinese city system with emerging new data, *Applied Geography*, 75, 36-48.
- Long Y., Shen Y. and Jin X. (2016), Mapping block-level urban areas for all chinese cities, *Annals of the Association of American Geographers*, 106, 96-113.
- Longley P., Goodchild M., Maguire D. and Rhind D. (2015), *Geographic Information Science and Systems*, Wiley: Chichester, England.
- Lynch K. (1960), *The image of the city*, MIT Press: Cambridge, MA
- MacQueen J.(1967), Some methods for classification and analysis of multivariate observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 281-297.
- Mandelbrot B. B. (1967), How long is the coast of Britain? Statistical self-similarity and fractional dimension, *Science*, 156(3775), 636-638.
- Mandelbrot B. B. (1982), *The Fractal Geometry of Nature*, W. H. Freeman and Co.: New York.

- Mandelbrot B. B. (2004), *Fractals and Chaos: The Mandelbrot set and beyond*, Springer: New York.
- Mandelbrot B. B. and Hudson R. L. (2004), *The (Mis)Behavior of Markets: A fractal view of risk, ruin and reward*, Basic Books: New York.
- Marta C. G., Cesar A. H. and Barabási A.-L. (2008), Understanding individual human mobility patterns, *Nature*, 453, 779–782.
- Mayer–Schonberger V. and Cukier K. (2013), *Big Data: A revolution that will transform how we live, work, and think*, Eamon Dolan/Houghton Mifflin Harcourt: New York.
- McKelvey B. and Andriani P. (2005), Why Gaussian statistics are mostly wrong for strategic organization, *Strategic Organization*, 3(2), 219–228.
- Miller J. H. (2004), Tobler’s First Law and spatial analysis, *Annals of the Association of American Geographers*, 94(2), 284–289.
- Miller J. H. and Page S. E. (2007), *Complex Adaptive Systems: An introduction to computational models of social life*, Princeton University Press: Princeton.
- Murray C. (2003), *Oracle Spatial Quadtree Indexing*, <http://www.oracle.com/technetwork/testcontent/qt-128949.pdf>, Retrieved November 2017,
- Newman M. E. J. (2003), The structure and function of complex networks. *SIAM review*, 45(2), 167–256.
- Newman M. E. J. (2004), Detecting community structure in networks, *European Physical Journal B*, 38, 321–330.
- Newman M. E. J. (2005), Power laws, Pareto distributions and Zipf’s law, *Contemporary Physics*, 46(5), 323–351.
- Newman M. E. J. (2010), *Networks: An Introduction*, Oxford University Press: Oxford.
- Newman M. E. J. (2011), Complex Systems: A survey, *American Journal of Physics*, 79, 800–810.
- Omer I. and Jiang B. (2015), Can cognitive inferences be made from aggregate urban flow data? *Computers, Environment and Urban Systems*, 54, 219–229.
- Ontoy D.S. and Padua R.N. (2014), Measuring species diversity for conservation biology: Incorporating social and ecological importance of species, *Biodiversity Journal*, 5, 387–390.
- Osaragi T. (2013), Modeling a spatiotemporal distribution of stranded people returning home on foot in the aftermath of a large-scale earthquake, *Natural Hazards*, 68(3), 1385–1398.
- OSM (2017), OpenStreetMap stats report, [https://www.openstreetmap.org/stats/data\\_stats.html](https://www.openstreetmap.org/stats/data_stats.html), accessed November 2017
- Penn A. (2003), Space syntax and spatial cognition: Or why the axial line? *Environment and Behavior*, 35(1), 30–65.
- Penn A., Hillier B., Banister D. and Xu J. (1998), Configurational modelling of urban movement networks, *Environment and Planning B: Planning and Design*, 25, 59–84.

- Richardson L. F. (1961), The problem of contiguity: An appendix to statistic of deadly quarrels, *General systems: Yearbook of the Society for the Advancement of General Systems Theory*, Society for General Systems Research: Ann Arbor, Mich., 6(139), 139–187.
- Samet H. (1990), Hierarchical spatial data structures, *In Proceedings of the first symposium on Design and implementation of large spatial databases*, New York: Springer–Verlag, 193–212.
- Scellato S., Noulas A., Lambiotte R. and Mascolo C. (2011), Socio-spatial properties of online location-based social networks, *The International AAAI Conference on Web and Social Media (ICWSM)*, the 5<sup>th</sup> International Conference, Barcelona, Spain, 329–336.
- Shanbhag D.N. and Rao C.R. (2001), *Stochastic Processes: Theory and Methods*, Elsevier Science B.V: Amsterdam, Netherlands.
- Song C., Koren T., Wang P. and Barabási A.–L. (2010), Modelling the scaling properties of human mobility, *Nature Physics*, 818–823.
- Statista (2017), Twitter: number of monthly active users 2010–2017, <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>, accessed November 2017.
- Steinhaus H. (1983), *Mathematical Snapshots*, 3rd edition, Oxford University Press: London.
- Sui D. (2004), Tobler's First Law of geography: A big idea for a small world? *Annals of the Association of American Geographers*, 94(2), 269–277.
- Sui D. and Goodchild M. (2011), The convergence of GIS and social media: challenges for GIScience, *International Journal of Geographical Information Science*, 25(11), 1737–1748.
- Takhteyev Y., Gruzd A. and Wellman B. (2012), Geography of Twitter networks, *Social Networks*, 34(1), 73–81.
- TechTarget (2017), Definition of small data, <http://whatis.techtarget.com/definition/small-data>, accessed November 2017.
- Thompson D. W. (1917), *On Growth and Form*, Cambridge University Press: Cambridge.
- Tobler W. (1970), A Computer movie simulating urban growth in the Detroit region, *Economic geography*, 46(2), 234–240.
- Twitter Developer (2017), Docs, <https://developer.twitter.com/en/docs>, accessed November 2017.
- Twitter Help Center (2017), Tweet location FAQs, <https://help.twitter.com/en/safety-and-security/tweet-location-settings>, accessed November 2017.
- Von Koch H. (1904), Sur une courbe continue sans tangente, obtenue par une construction geometrique elementaire, *Arkiv för Matematik*, 1:681–704.
- Wakamiya S., Lee R., and Sumiya K. (2011), Crowd-based urban characterization: Extracting crowd behavioral patterns in urban areas from Twitter, *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 77–84.

- Wang F. and Zhou Y. (1999), Modelling urban population densities in Beijing 1982–90: Suburbanisation and its causes, *Urban Studies*, 36(2), 271–287.
- Wang S. (2010), A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and Spatial Analysis, *Annals of the Association of American Geographers*, 100(3): 535–557.
- Wang S. (2013), CyberGIS: Blueprint for integrated and scalable geospatial software ecosystems, *International Journal of Geographical Information Science*, 27(11): 2119–2121.
- Wang S., Gao S., Feng X., Murray A.T. and Zeng Y. (2018): A context-based geoprocessing framework for optimizing meet-up location of multiple moving objects along road networks, *International Journal of Geographical Information Science*, DOI:10.1080/13658816.2018.1431838
- Watts D. J. (2007), A twenty-first century science, *Nature*, 445(x), 489.
- Watts D. J. and Strogatz S. H. (1998), Collective dynamics of “small-world” networks, *Nature*, 393, 440–442.
- White T. (2012), *Hadoop: The Definitive Guide*, O’Reilly Media.
- World population review (2018), The population of Swedish cities, <http://worldpopulationreview.com/countries/sweden-population/>, accessed January 2018.
- Wright D.J. and Wang S. (2011), The emergence of spatial cyberinfrastructure, *Proceedings of the National Academy of Sciences*, 108 (14): 5488–5491.
- Yang B., Luan X., Li Q. (2011a), Generating hierarchical strokes from urban street networks based on spatial pattern recognition, *International Journal of Geographical Information Science*, 25(12), 2025–2050.
- Yang C., Goodchild M., Huang Q., Nebert D., Raskin R., Xu Y., Bambacus M. and Fay D. (2011b), Spatial cloud computing: How can the geospatial sciences use and help shape cloud computing? *International Journal of Digital Earth*, 4, 305–329.
- Zheng Y. (2011), Location-based social networks: Users, In *Computing with Spatial Trajectories*, Springer (Editors Zheng Y. and Zhou X.): New York, 243–276.
- Zhou S. and Mondragón R. J. (2007), Structural constraints in complex networks, *New Journal of Physics*, 9, 173. doi:10.1088/1367-2630/9/6/173.
- Zipf G. K. (1949), *Human Behavior and the Principles of Least Effort*, Addison Wesley: Cambridge, MA.
- Zook M., Graham M., Shelton T. and Gorman S. (2010), Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake, *World Medical and Health Policy*, 2(2), 7–33.



Associated papers have been removed in the electronic version of this thesis.

For more details about the papers see:

<http://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-26197>

Gävle University Press  
ISBN 978-91-88145-24-6  
ISBN 978-91-88145-25-3 (pdf)

University of Gävle  
Faculty of Engineering  
and Sustainable Development  
SE-801 76 Gävle, Sweden  
+46 26 64 85 00  
[www.hig.se](http://www.hig.se)



## Topological and Scaling Analysis of Geospatial Big Data

The geographic space and phenomena are inherently heterogeneous and diverse (e.g., urban layout and human movement behavior). However, current geospatial analysis lacks ability to reveal such underlying heterogeneity as it relies on Euclidean geometry and normal distribution statistics which focus on geometric details (such as locations, directions, and sizes) that are with a well-defined mean and small variance. Instead, topology without geometric details, fractal geometry, and power-law distribution statistics represent new perspectives for geospatial analysis, particularly in the era of big data, for better understanding geographic space and its involved human activities.

Given the circumstances, it is time for us to shift our paradigm of geospatial analysis towards topological and scaling ways of thinking. In this connection, it is necessary to adopt complexity science methods such as power law detection, head/tail breaks, ht-index, topological analysis, and complex networks in various applications of geospatial modeling and human behavior. The thesis is further motivated by the emerging big data harvested from the Internet and social media such as OpenStreetMap, Twitter, Brightkite, and Gowalla which provide a new instrument for studying space and society.

Ding Ma