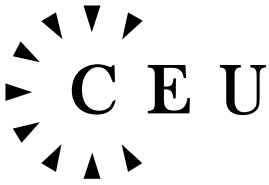
# Multimodality and sustainability in urban networks




**Luis Guillermo Natera Orozco**

Department of Network and Data Science
Central European University

Supervisor: Federico Battiston
External Supervisor: Michael Szell

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Network Science

2021

# RESEARCHER DECLARATION

I Luis Guillermo Natera Orozco certify that I am the author of the work Multimodality and sustainability in urban networks. I certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, the contributions of others. The thesis contains no materials accepted for any other degrees in any other institutions. The copyright of this work rests with its author. Quotation from it is permitted, provided that full acknowledgement is made. This work may not be reproduced without my prior written consent.

## Statement of inclusion of joint work

Signature of PhD Candidate:

January, 2021

Signature of Dr. Federico Battiston, endorsing statement of joint work:

January, 2021

Signature of Dr. Michael Szell, endorsing statement of joint work:

January, 2021

Signature of Dr. Orsyola Vásárhelyi , endorsing statement of joint work:

January, 2021

# Abstract

# Acknowledgements

# Contents

**Contents** i

**List of Tables** iii

**List of Figures** v

**1 Extracting the multimodal fingerprint for transportation networks** 1
- 1.1 Prelude . . . 1
- 1.2 METHODS AND DATA . . . 2
- 1.3 FINDINGS . . . 5

**2 Data-driven strategies for optimal bicycle network growth** 7
- 2.1 Prelude . . . 8
- 2.2 Data acquisition and network construction . . . 9
- 2.3 Defining bicycle network growth strategies and quality metrics . . . 10
- 2.4 Growing bicycle networks shows stark improvements with small investments . . . 16
- 2.5 Different cities have different optimal investment strategies . . . 18
- 2.6 Discussion . . . 20

**3 Life quality as walkability** 25
- 3.1 Prelude . . . 25
- 3.2 Data . . . 26
- 3.3 Quantifying Life Quality . . . 27
  - 3.3.1 The services index: $Q^{\text{services}}$ . . . 29
  - 3.3.2 Safety index: $Q^{\text{safety}}$ . . . 30
  - 3.3.3 Environmental index $Q^{\text{environment}}$ . . . 31
- 3.4 Results . . . 32
  - 3.4.1 Evaluation . . . 33
- 3.5 Discussion . . . 34

**4 Appendices** **37**
4.1 Life quality as walkability . . . . . 37
4.1.1 Secondary data sources . . . . . 37
4.1.2 Weights used in the calculations . . . . . 38
4.2 Data-driven strategies for optimal bicycle network growth . . . . 38
4.2.1 Data . . . . . 38
4.2.2 Bicycle network improvement . . . . . 38
4.2.3 Bicycle network and 30 km/hr streets . . . . . 39

# List of Tables

1.1 Layers measures for analyzed cities . . . . . 3

2.1 Measures for analyzed cities . . . . . 10

# List of Figures


1.1 Manhattan multiplex network . . . . . 2
1.2 Schematic overlap census . . . . . 4
1.3 Overlap Census clusters . . . . . 5

2.1 Multimodal configuration . . . . . 9
2.2 Algorithms schematic representation . . . . . 15
2.3 Algorithmic improvement in Budapest . . . . . 16
2.4 Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$). . . . . 18
2.5 Cities bicycle connectivity improvement 5 new kilometers . . . . 19
2.6 Cities bicycle connectivity improvement 35 new kilometers . . . . 21
2.7 Bicycle infrastructure service area . . . . . 22

3.1 Budapest pedestrian network . . . . . 28
3.2 Budapest neighborhoods life quality index . . . . . 32
3.3 Budapest life quality index correlation with real estate prices . . . 34

4.1 Connected component size distribution for analyzed cities . . . . 39
4.2 Normalized increase in nodes inside the largest connected component ($n_{LCC}$). . . . . 40
4.3 Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$). . . . . 41
4.4 Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$) versus the fraction of extant kilometers to be added. . . . . 42
4.5 Bike-car directness $\Delta$ per invested kilometers. . . . . 43
4.6 Kilometers gain in the largest connected component. . . . . 44
4.7 Normalized increase in nodes inside the largest connected component ($n_{LCC}$). . . . . 45
4.8 Bike-car directness $\Delta$ per invested kilometers. . . . . 46
4.9 Kilometers gain in the largest connected component. . . . . 47

CHAPTER 1

# EXTRACTING THE MULTIMODAL FINGERPRINT FOR TRANSPORTATION NETWORKS

Urban mobility increasingly relies on multimodality, combining the use of bicycle paths, streets, and rail networks. These different modes of transportation are well described by multiplex networks. Here we propose the overlap census method which extracts a multimodal profile from a city's multiplex transportation network. We apply this method to 15 cities, identify clusters of cities with similar profiles, and link this feature to the level of sustainable mobility of each cluster. Our work highlights the importance of evaluating all the transportation systems of a city together to adequately identify and compare its potential for sustainable, multimodal mobility[1].

## 1.1 Prelude

The infrastructure of different modes of transportation can be described as a mathematical object, the multiplex transport network [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. A city's multiplex transport network contains the layer of streets and other co-evolving network layers, such as the bicycle or the rail networks, which together constitute the multimodal transportation backbone of a city. Due to the car-centric development of most cities [**?**], streets form the most developed layers [**?**, **?**] and define or strongly limit other layers: For example, sidewalks are by definition footpaths along the side of a street and make up a substantial part of

[1] A stand alone of this chapter has been published in Transport findings [**?**]

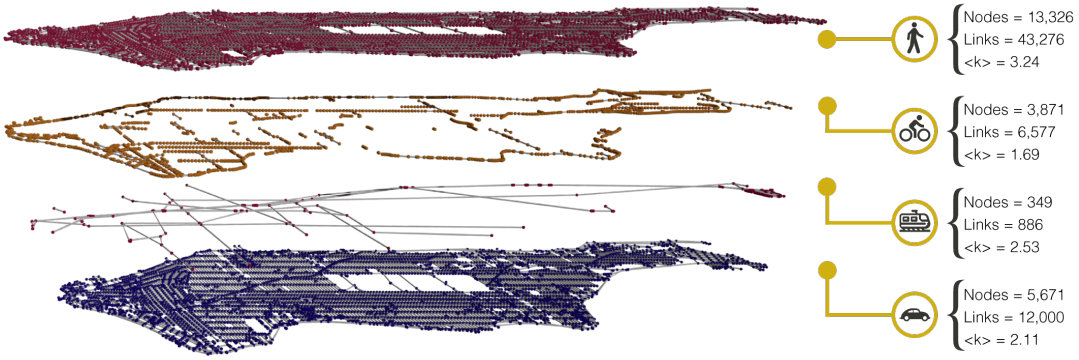



**Figure 1.1.** ***(Map plot left)*** *Multiplex network representation of Manhattan with the four analyzed layers of transport infrastructure (pedestrian paths, bicycle paths, rail lines, and streets), with data from OpenStreetMap.* ***(Right)*** *Network information for each layer, number of nodes, links and average degree* $\langle k \rangle$.

a city's pedestrian space [**?**]. Similarly, most bicycle paths are part of a street or are built along the side. Yet, the different layers of a multimodal network typically serve as diverse channels to permeate a city. Here we consider the transport networks of 15 world cities and develop an urban fingerprinting technique based on multiplex network theory to characterize the various ways in which transport layers can be interconnected, identifying the potential for multimodal transport. Using clustering algorithms on the resulting urban fingerprints, we find distinct classes of cities, reflecting their transport priorities.

## 1.2 METHODS AND DATA

We acquired urban transportation networks from multiple cities around the world, defined by their administrative boundaries, using OSMnx [**?**]. These data sets are of high quality [**?**, **?**] in terms of correspondence with municipal open data [**?**] and completeness [**?**]. The various analyzed urban areas and their properties are reported in Table 1.1. Figure 1.1 shows the different network layers for Manhattan, one of our analyzed cities.

Code to replicate our results is available as Jupyter Notebooks (`https://github.com/nateraluis/Multimodal-Fingerprint`) and data can be downloaded from Harvard Dataverse [**?**].

We characterize each city as a multiplex network [**?**, **?**, **?**] with $M$ layers and $N$ nodes that can be active in one or more layers in the system. Layers follow a primal approach [**?**] where nodes represent intersections (that may be present in one or more layers), and links represent streets (denoted by $s$), bicycle paths

| | Pedestrian | | | Bicycle | | | Rail | | | Street | | | Population |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nodes | Links | $\langle k \rangle$ | Nodes | Links | $\langle k \rangle$ | Nodes | Links | $\langle k \rangle$ | Nodes | Links | $\langle k \rangle$ | |
| Amsterdam | 23,321 | 33,665 | 2.89 | 34,529 | 35,619 | 2.06 | 1,096 | 1,655 | 3.02 | 15,125 | 21,722 | 2.87 | 872,680 |
| Barcelona | 20,203 | 30,267 | 3.00 | 7,553 | 7,647 | 2.02 | 249 | 249 | 2.00 | 10,393 | 15,809 | 3.04 | 1,600,000 |
| Beihai | 2,026 | 2,978 | 2.94 | 0 | 0 | 0.00 | 59 | 62 | 2.10 | 2,192 | 3,209 | 2.93 | 1,539,300 |
| Bogota | 81,814 | 121,038 | 2.96 | 9,760 | 9,651 | 1.98 | 166 | 165 | 1.99 | 62,017 | 91,197 | 2.94 | 7,412,566 |
| Budapest | 73,172 | 106,167 | 2.90 | 10,494 | 10,318 | 1.97 | 1,588 | 1,964 | 2.47 | 37,012 | 52,361 | 2.83 | 1,752,286 |
| Copenhagen | 30,746 | 41,916 | 2.73 | 13,980 | 13,988 | 2.00 | 276 | 369 | 2.67 | 15,822 | 20,451 | 2.59 | 2,557,737 |
| Detroit | 47,828 | 78,391 | 3.28 | 3,663 | 3,626 | 1.98 | 20 | 21 | 2.10 | 28,462 | 45,979 | 3.23 | 672,662 |
| Jakarta | 140,042 | 191,268 | 2.73 | 248 | 231 | 1.86 | 58 | 54 | 1.86 | 138,388 | 188,637 | 2.73 | 10,075,310 |
| LA | 89,543 | 128,757 | 2.88 | 14,577 | 14,428 | 1.98 | 173 | 221 | 2.55 | 71,091 | 101,692 | 2.86 | 3,792,621 |
| London | 270,659 | 351,824 | 2.60 | 62,398 | 60,043 | 1.92 | 2,988 | 3,535 | 2.37 | 179,782 | 219,917 | 2.45 | 8,908,081 |
| Manhattan | 13,326 | 21,447 | 3.22 | 3,871 | 3,777 | 1.95 | 349 | 436 | 2.50 | 5,671 | 9,379 | 3.31 | 1,628,701 |
| Mexico | 108,033 | 158,425 | 2.93 | 5,218 | 5,278 | 2.02 | 370 | 364 | 1.97 | 95,375 | 140,684 | 2.95 | 8,918,653 |
| Phoenix | 111,363 | 157,075 | 2.82 | 35,631 | 35,979 | 2.02 | 105 | 138 | 2.63 | 73,688 | 102,139 | 2.77 | 1,445,632 |
| Portland | 50,878 | 72,958 | 2.87 | 24,252 | 24,325 | 2.01 | 230 | 340 | 2.96 | 35,025 | 49,062 | 2.80 | 583,776 |
| Singapore | 82,808 | 110,612 | 2.67 | 12,981 | 12,947 | 1.99 | 683 | 740 | 2.17 | 50,403 | 66,779 | 2.65 | 5,638,700 |

**Table 1.1.** *Measures for the administrative area of analyzed cities. The number of nodes, links and average degree ($\langle k \rangle$) for each layer in all cities of our dataset are highly diverse due to the varying developmental levels and focus of transport. The range of population in the analyzed cities goes from half million people to ten million people living in Jakarta, this allows to have a range of different sizes and cover different developmental stages.*

and designated bicycle infrastructure ($b$), subways, trams and rail infrastructure ($r$), or pedestrian infrastructure ($p$). Construction of these intersection nodes follows the topological simplification rules of OSMnx **[?]**. This recent approach has been useful to demonstrate how cities grow **[?, ?]**, how efficient **[?]** and dense they are, and to capturethe tendency of travel routes to gravitate towards city centers **[?]**. Each layer $\alpha = 1, \ldots, M$ is described by anadjacency matrix $A^{[\alpha]} = a_{ij}^{[\alpha]}$ where $a_{ij}^{[\alpha]} = 1$ if there is a link between nodes $i$ and $j$ in layer $\alpha$ and 0 otherwise. The multiplex urban system is then specified as a vector of adjacency matrices $A = (A^{[1]}, \ldots, A^{[M]})$.

Whereas one of the simplest features of single layer networks is the degree distribution, in multiplex networks a node can have different degrees in each layer, which inform us about the multimodal potential of a city through the different roles that its intersections play. If a city has nodes that are mainly active in one layer but not in others, there is no potential for multimodality. On the contrary, in a multimodal city we expect to find many transport hubs that connect different layers, such as train stations with bicycle and street access, i.e. nodes that are active in different multiplex configurations. Note that even in a multimodally "optimal" city there will be a high heterogeneity of node activities due to the different speeds and nature of transport modes, implying, for example, a much lower density of nodes necessary for a train network than for a bicycle network. Still, if we had a way to see and compare all combinations of node ac-

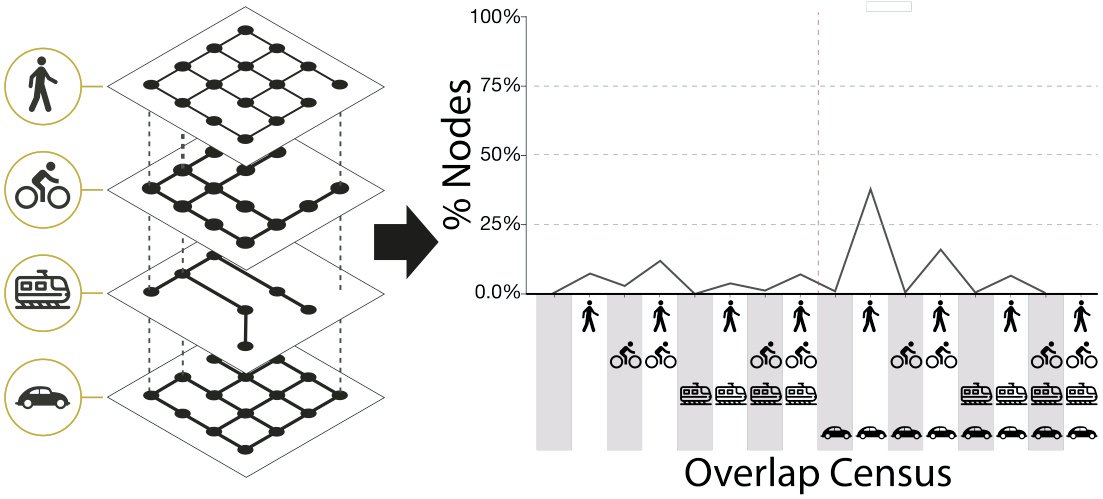



**Figure 1.2.** *Schematic of multiplex layers in a city (left) and its transformation to the overlap census (right). In the overlap census, the vertical red line gives a visual separation of the left from the right half where nodes become active in the street layer. High spikes in the right half indicate car-centricity.*

tivities in the system, we could learn how much focus a city puts on connecting different modes. We can define such a fingerprint using the multiplex network formalism

In a multimodal city, we expect to find many transport hubs that connect different layers, such as train stations with bicycle and street access, i.e. nodes that are active in different multiplex configurations. Here we propose a method to assess all such combinations of node activities in the system, helping us to learn how well connected different modes are. For each city, we build a profile based on the combinations of node activities, and refer to it as *overlap census* (Figure 1.2). The overlap census captures the percentage of nodes that are active in different multiplex configurations and provides an "urban fingerprint" of its multimodality [**?**]. To define the overlap census formally, given a multiplex transport network with $M$ layers the overlap census is a vector of $(2^M) - 1$ components, which accounts for the fractions of nodes that can be reached through at least one layer.

In Fig. 1.2 we show a schematic of how the overlap census is built: taking the multiplex network, and calculating the percentage of nodes that overlap in different configurations. The multiplex approach addresses the multimodality of a city: it not only counts how many nodes or links there are in each layer, but it shows how they are combined, revealing the possible multimodal mobility combinations in the city. Understanding the possibilities for interchange

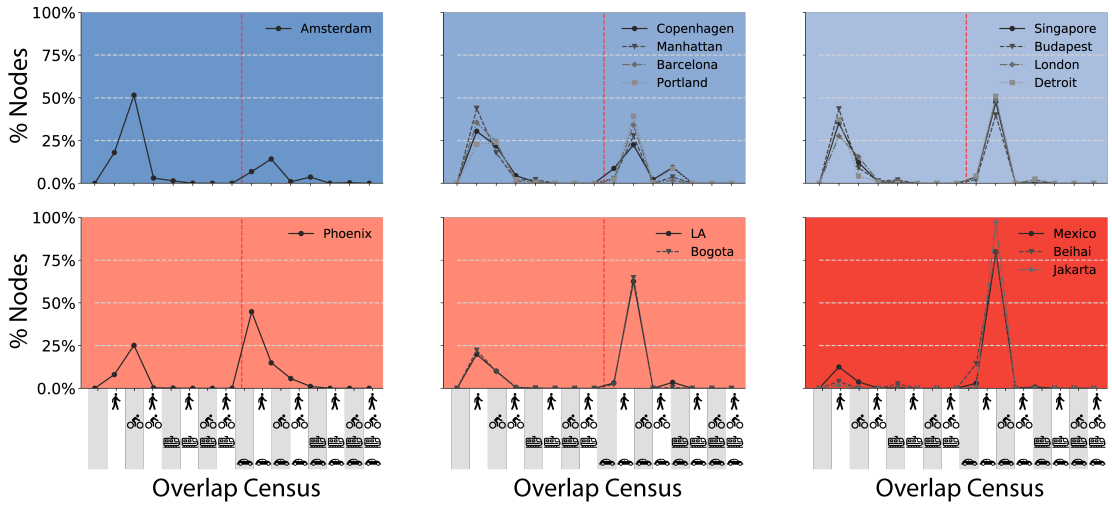



**Figure 1.3.** *Clusters of cities based on similarity of their overlap census. We find six different clusters using a k-means algorithm (coloured areas), which explain more than* 90% *of the variance.*

between mobility layers provides us with a better understanding of urban systems, showing us the complexity and interplay between layers.

## 1.3 FINDINGS

Due to the expected heterogeneity of node activities in different layers, the overlap census of a specific city is also expected to be heterogeneous and hard to assess on its own. Therefore, a good way to assess a city's overlap census is by comparing it with the overlap census of other cities. We find similarities between cities via a k-means algorithm fed with fifteen vectors (one per city), where each vector contains the percentages of nodes active in each possible configuration. The algorithm separates the 15 analyzed cities into six different clusters [Fig. 1.3].

On the left half of the overlap census, we show the configurations in which nodes are not active in the street layer, while the right half contains car-related configurations [Fig. 1.3] . These clusters of cities are useful to explain similarities in infrastructure planning in different transport development paths [**?**, **?**], with clusters of car-centric urbanization (like Mexico, Beihai, and Jakarta) opposed to clusters that show a more multimodal focus in their mobility infrastructure (like Copenhagen, Manhattan, Barcelona, and Portland). In the extreme cluster that contains only Amsterdam, close to 50% of nodes are active in the bicycle layer, whereas in the Mexico-Beihai-Jakarta cluster more than 50% of nodes are active

in the street-pedestrian configuration. The concentration of nodes in just one configuration informs not only about the mobility character of the city, i.e. Amsterdam being a bicycle-friendly city, but unveils the importance of explicitly considering overlooked layers and their interconnections. For example, Singapore, Budapest, London, and Detroit have two main peaks indicating that most of their nodes are either active in the street-pedestrian or only in the pedestrian configuration. This is not the case in Los Angeles and Bogota, where the majority of nodes are active in the car-pedestrian combination, i.e. the pedestrians have to share most of the city with cars. Our multimodal fingerprint unravels how different transport modes are interlaced, helping identifying which layer (or set of layers) could be improved to promote multimodal, sustainable mobility.

To summarize, we propose the new "overlap census" method based on multiplex network theory allowing to rigorously identify and compare the multimodal potential of cities.

CHAPTER 2

# DATA-DRIVEN STRATEGIES FOR OPTIMAL BICYCLE NETWORK GROWTH

Urban transportation networks, from sidewalks and bicycle paths to streets and rails, provide the backbone for movement and socioeconomic life in cities. To make urban transport sustainable, cities are increasingly investing to develop their bicycle networks. However, it is yet unclear how to extend them comprehensively and effectively given a limited budget. Here we investigate the structure of bicycle networks in cities around the world, and find that they consist of hundreds of disconnected patches, even in cycling-friendly cities like Copenhagen. To connect these patches, we develop and apply data-driven, algorithmic network growth strategies, showing that small but focused investments allow to significantly increase the connectedness and directness of urban bicycle networks. We introduce two greedy algorithms to add the most critical missing links in the bicycle network focusing on connectedness, and show that they outmatch both a random approach and a baseline minimum investment strategy. Our computational approach outlines novel pathways from car-centric towards sustainable cities by taking advantage of urban data available on a city-wide scale. It is a first step towards a quantitative consolidation of bicycle infrastructure development that can become valuable for urban planners and stakeholders.[1]

[1] A stand alone version of this chapter has been published in Royal Society Open Science [**?**]

## 2.1 Prelude

Most modern cities have followed a car-centric development in the 20th century [?] and are today allocating a privileged amount of urban space to automobile traffic [?, ?]. From a network perspective, this space can be described as the street layer of a larger mathematical object, the multiplex transport network [?, ?, ?]. A city's multiplex transport network contains other network layers that have co-evolved with the street layer, such as the bicycle layer or the rail network layer, Fig. 2.1. Due to the car-centric development of most cities, street layers are the most developed layers and define or strongly limit other layers: For example, sidewalks are by definition footpaths along the side of a street and make up a substantial part of a city's pedestrian space [?]; similarly most bicycle paths are part of a street or are built along the side.

From an urban sustainability perspective, this situation is suboptimal because the unsustainable mode of automobile transportation dominates sustainable modes like cycling. Consequently, urban planning movements in a number of pioneering cities are increasingly experimenting with drastic policies, such as applying congestion charges (London) [?] and repurposing or removing car parking (Amsterdam, Oslo) [?, ?, ?]. These efforts agree in one common goal, together with the literature on cycling safety [?, ?, ?, ?] and with cost-benefit analysis [?]: Protected bicycle lanes need to be extended considerably to create complete bicycle networks that provide a safe infrastructure for cycling citizens. Although scattered efforts in this direction have shown preliminary success, a quantitative framework for developing and assessing systematic strategies is missing [?].

In this chapter we analyze the bicycle infrastructure network from the 14 world cities we previously analyzed in Chapter 1, the cities included come from leading bicycle-culture countries like Netherlands, to car-centric countries like Great Britain, the USA, or Colombia. We first uncover network fragmentation within the bicycle dedicated infrastructure. Then, to improve a city's vital dedicated bicycle infrastructure [?, ?, ?, ?], we develop algorithms for connecting disconnected graphs based on concrete quality metrics from bicycle network planning [?] and apply them to the empirical bicycle networks via network growth simulations. We find that localized investment into targeted missing links can rapidly consolidate fragmented bicycle networks, allowing to significantly increase their connectedness and directness, with potentially crucial implications for sustainable transport policy planning.

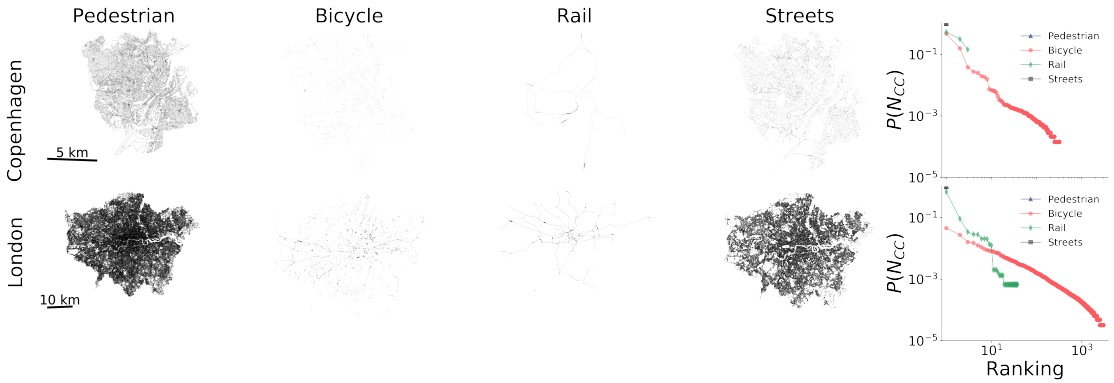



**Figure 2.1.** ***(Map plots, left)*** *Networks representing various layers of transport infrastructure (pedestrian paths, bicycle paths, rail lines, and streets) for Copenhagen and London, with data from OpenStreetMap.* ***(Right)*** *Connected component size distribution $P(N_{cc})$ as a function of the ranking of the component for all considered network layers and cities. All layers are well connected except the bicycle layer: Copenhagen has 321 bicycle network components despite being known as a bicycle-friendly city, while London's bicycle layer is much more fragmented, featuring over 3000 disconnected components. Copenhagen's largest connected bicycle component (leftmost data point) spans 50% of the network, but London's only less than 5%.*

## 2.2 Data acquisition and network construction

We acquired street and bicycle infrastructure networks from multiple cities around the world using OSMnx [**?**], a Python library to download and construct networks from OpenStreetMap (OSM). OSMnx simplifies the OpenStreetMap's raw data to retain only nodes at the intersections and dead ends of streets, and the spatial geometry of the edges, generating a length-weighted nonplanar directed graph [**?**]. These data sets are of high quality [**?**, **?**] in terms of correspondence with municipal open data [**?**] and completeness: More than 80% of the world is covered by OSM [**?**]. In particular, OSM's bicycle layer has better coverage than proprietary alternatives like Google Maps [**?**]. We collect data from a diverse set of cities to capture different development states of bicycle infrastructure networks; from consolidated networks like Amsterdam and Copenhagen, less developed ones like Manhattan and Mexico City, to rapidly developing cities like Jakarta and Singapore. The various analyzed urban areas and their properties (number of nodes $N$, number of connected components $CC$, and population) are reported in Table 2.1. Code to replicate our results is available as Jupyter Notebooks (`https://github.com/nateraluis/bicycle-network-growth`) and data can be downloaded from Harvard Data-

verse [?].

| | walk | | | bike | | | rail | | | drive | | | Population |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $N$ | $CC$ | $\ell(km)$ | $N$ | $CC$ | $\ell(km)$ | $N$ | $CC$ | $\ell(km)$ | $N$ | $CC$ | $\ell(km)$ | |
| Amsterdam | 23,321.0 | 1.0 | 2,075.67 | 34,529.0 | 355.0 | 972.08 | 1,096.0 | 8.0 | 288.72 | 15,125.0 | 1.0 | 2,010.49 | 872,680 |
| Barcelona | 20,203.0 | 1.0 | 2,122.6 | 7,553.0 | 122.0 | 229.19 | 263.0 | 29.0 | 105.98 | 10,393.0 | 1.0 | 1,551.44 | 1,600,000 |
| Bogota | 81,814.0 | 1.0 | 8,686.51 | 9,760.0 | 171.0 | 367.33 | 166.0 | 12.0 | 20.2 | 62,017.0 | 1.0 | 7,383.69 | 7,412,566 |
| Budapest | 73,172.0 | 1.0 | 7,746.12 | 10,494.0 | 257.0 | 336.13 | 1,588.0 | 20.0 | 522.06 | 37,012.0 | 1.0 | 5,332.97 | 1,752,286 |
| Copenhagen | 30,746.0 | 1.0 | 2,286.66 | 13,980.0 | 321.0 | 417.01 | 276.0 | 3.0 | 123.56 | 15,822.0 | 1.0 | 1,547.3 | 2,557,737 |
| Detroit | 47,828.0 | 1.0 | 6,769.46 | 3,663.0 | 53.0 | 141.06 | 20.0 | 3.0 | 11.54 | 28,462.0 | 1.0 | 5,624.49 | 672,662 |
| Jakarta | 140,042.0 | 1.0 | 13,947.96 | 248.0 | 19.0 | 8.44 | 60.0 | 8.0 | 81.24 | 138,388.0 | 1.0 | 14,194.2 | 10,075,310 |
| LA | 89,543.0 | 1.0 | 14,329.92 | 14,577.0 | 230.0 | 653.16 | 173.0 | 9.0 | 90.82 | 71,091.0 | 1.0 | 13,324.46 | 3,792,621 |
| London | 270,659.0 | 1.0 | 23,846.62 | 62,398.0 | 3,023.0 | 1,281.71 | 2,988.0 | 38.0 | 1,045.39 | 179,782.0 | 1.0 | 18,154.52 | 8,908,081 |
| Manhattan | 13,326.0 | 1.0 | 1,320.78 | 3,871.0 | 105.0 | 111.42 | 349.0 | 5.0 | 197.51 | 5,671.0 | 1.0 | 1,022.13 | 1,628,701 |
| Mexico | 108,033.0 | 1.0 | 14,547.18 | 5,218.0 | 52.0 | 332.37 | 371.0 | 18.0 | 253.48 | 95,375.0 | 1.0 | 13,732.39 | 8,918,653 |
| Phoenix | 111,363.0 | 1.0 | 14,314.0 | 35,631.0 | 141.0 | 1,221.18 | 105.0 | 4.0 | 71.64 | 73,688.0 | 1.0 | 11,841.49 | 1,445,632 |
| Portland | 50,878.0 | 1.0 | 5,324.78 | 24,252.0 | 198.0 | 596.36 | 230.0 | 2.0 | 132.36 | 35,025.0 | 1.0 | 4,583.47 | 583,776 |
| Singapore | 82,808.0 | 1.0 | 8,633.13 | 12,981.0 | 104.0 | 339.39 | 683.0 | 14.0 | 428.66 | 50,403.0 | 1.0 | 6,635.37 | 5,638,700 |

**Table 2.1.** *Measures for the administrative area of analyzed cities. The number of connected components (CC) and nodes (N) for each layer in all cities of our dataset are highly diverse due to the varying developmental levels and focus of transport.*

We characterize each city street and bicycle infrastructure as a primal network, [?] in which nodes are intersections, while links represent bicycle paths, and designated bicycle infrastructure. This recent approach has been useful to demonstrate how cities grow [?, ?], how efficient [?] and dense they are, and to capture the tendency of travel routes to gravitate towards city centers [?]. This network is described by an adjacency matrix $A = \{a_{ij}^{[\alpha]}\}$ where $a_{ij} = 1$ if there is a link between nodes $i$ and $j$ and 0 otherwise.

## 2.3 Defining bicycle network growth strategies and quality metrics

Across all cities considered, we find that almost all network layers are made up of one giant component, except for the bicycle layer which is always fragmented into many disconnected components (see Table 2.1). This discovery is remarkable given that the fragmentation occurs also in bicycle-friendly cities like Copenhagen (Fig. 2.1), showing that cycling infrastructure can be suboptimal even in the leading cycling cities on the planet. To quantify such an underdevelopment in the sustainable mobility infrastructure of cycling, we focus on the single layer of bicycle networks and on two well-established metrics in bicycle infrastructure quality assessment [?, ?, ?, ?, ?]: *connectedness* and *directness*. Connectedness indicates "the ease with which people can travel across the transportation system" [?], and it is related to answering the question "can I go where I want to, safely?". Directness addresses the question "how far out of their way do users have to travel to find a facility they can or want to use?",

and can be measured by how easy it is to go from one point to another in a city using bicycle infrastructure versus other mobility options, like car travel.

As our main approach, we choose to measure connectedness and directness over the designated bicycle infrastructure only, without considering travel on streets. Although it is possible to cycle on streets, growing evidence from bicycle infrastructure and safety research is unveiling serious safety issues for cycling when mixed with vehicular traffic [**?**, **?**, **?**]. However, we also tested our algorithms on a combination of bicycle infrastructure plus streets for which the maximum speed is 30 km/h, following common best-practice reasoning that low speed limits can make streets safe for cycling [**?**]. The results of these additional simulations are available in the Supplementary Information; they do not differ significantly from the case of designated bicycle infrastructure presented below, as the developed algorithms follow the same rules to connect the multiple components in both cases of segregated bicycle infrastructure only and of included bikeable streets.

To quantify connectedness, we first measure the number of disconnected components of each city's bicycle network. It is no surprise that car-centric cities have a highly fragmented bicycle infrastructure: for example, London has more than 3,000 disconnected bicycle infrastructure segments. However, even bicycle-friendly cities like Copenhagen have over 300 disconnected bicycle path components – see the connected component size distribution $P(N_{cc})$ in Fig. 2.1. This infrastructure fragmentation in the bicycle layer poses a challenge for a city's multimodal mobility options [**?**] and for the safety of its cycling citizens [**?**, **?**].

There are various approaches in developing automated strategies for bicycle infrastructure planning. Hyodo et al. [**?**] have proposed a bicycle route choice model to plan bicycle lanes taking into account facility characteristics. Other studies have used input data from bicycle share systems [**?**] or origin destination matrices [**?**] to plan bicycle lanes. More recently, taxi trips have been used to identify susceptible clusters for bicycle infrastructure [**?**]. Here we attempt an alternative approach: Since hundreds of bicycle network components already exist in most cities, we aim at consolidating the existing infrastructure by making strategic connections between components rather than starting from scratch.

Our approach takes into account the currently available bicycle infrastructure and uses an algorithmic process to improve the network by finding the most important missing links step by step. This way we focus on optimizing the connectedness metric, growing the bicycle infrastructure by making it more connected, merging parts into fewer and fewer components. We develop two iterative greedy algorithms that we check against a random and a minimum

**Algorithm 1** Largest-to-Second. The algorithm takes the bicycle network $G$ and a list of its weakly connected components *wcc*, then it iterates over the weakly connected components, sorts them by their size (number of nodes inside each component), locates the closest pairs of nodes between the first and the second components. The process is repeated until all the components have been connected.

```
procedure L2S
    G ← bicycle network graph
    wcc ← components of network G
    for i in length(wcc)-1 do
        sort wcc by components size
        cc ← two biggest components from wcc
        i_j ← closest nodes between cc_0 and cc_1
        connect cc_0 and cc_1 in i_j
```

**Algorithm 2** Largest-to-Closest. The algorithm takes the bicycle network $G$ and a list of its weakly connected components *wcc*, then it iterates over the weakly connected components, sorts them by their size (number of nodes inside each component), locates the largest connected component and the closest of the remaining components, the components are connected. The process is repeated until all the components have been connected.

```
procedure L2C
    G ← bicycle network graph
    wcc ← components of network G
    for i in length(wcc)-1 do
        sort wcc by components size
        cc_0 ← biggest component from wcc
        cc_n ← clossest component to cc_0
        i_j ← closest nodes between cc_0 and cc_n
        connect cc_0 and cc_n in i_j
```

investment approach. The first algorithm, *Largest-to-Second* (L2S), identifies in each step the largest connected component in the bicycle infrastructure network and connects it to the second largest (see algorithm 1 for details). The second algorithm, *Largest-to-Closest* (L2C), also identifies the largest connected component, but connects it to the closest of the remaining bicycle infrastructure components (see algorithm 2 for details). In both algorithms, components are connected through a direct link between their two closest nodes. We use this technique as an approximation to the underlying street-shortest path – since the

most relevant shortest 100 connections typically range from 14 to 500 meters, roughly the length of two blocks, this approximation is reasonable. The algorithms repeat this process until there are no more disconnected components in the network.

To have a random baseline, we compare our algorithms with a *Random-to-Closest* (R2C) component approach. In each step of this baseline approach, one component is picked at random and connected with the closest remaining one (see algorithm 3 for details). This baseline allows us to model a scenario where infrastructure is developed following a systematic but random linking approach – in urban development this corresponds to uncoordinated local planning that randomly connects close pieces of bicycle infrastructure. We also implement a second baseline, the extreme case of *Closest-Components* (CC), which prioritizes connecting the closest two components disregarding their size (see algorithm 4 for details). This CC approach is equivalent to an "invest as little as possible" development strategy – it builds up a minimum-spanning-tree-like structure following a modified Kruskal's algorithm [**?**]. All four algorithms connect components optimizing a well-defined criterion, finding the critical missing links in the network, and adding one new link per iteration. See Fig. 2.2 for a schematic of the four algorithms.

**Algorithm 3** Random-to-Closest. The algorithm takes the bicycle network $G$ and a list of its weakly connected components *wcc*, then it iterates over the weakly connected components, randomly picks a component and connects it to the closest of the remaining components. The process is repeated until all the components have been connected.

**procedure** *R2C*
  $G \leftarrow$ bicycle network graph
  $wcc \leftarrow$ components of network G
  **for** i in length(wcc)-1 **do**
    $cc_{ran} \leftarrow$ random component from $wcc$
    $cc_n \leftarrow$ clossest component to $cc_{ran}$
    $i\_j \leftarrow$ closest nodes between $cc_{ran}$ and $cc_n$
    connect $cc_{ran}$ and $cc_n$ in $i\_j$

We apply the algorithms to the bicycle infrastructure inside the political demarcation of the cities, however it is possible to extend the methods and include bicycle highways and cross-city trails, since they use as input a set of spatial network components to connect. We opt to not include cross-city links, since they are a special case only available in a few regions and where adequate intra-urban bicycle infrastructure has already been established [**?**, **?**].

**Algorithm 4** Closest-Components. The algorithm takes the bicycle network $G$ and a list of its weakly connected components $wcc$, then it iterates over the weakly connected components, calculate the distance between available components and connect the two closests ones. The process is repeated until all the components have been connected.

```
procedure CC
    G ← bicycle network graph
    wcc ← components of network G
    for i in length(wcc)-1 do
        Δ_min ← clossest components in wcc
        cc_0 ← first component for Δ_min
        cc_1 ← second component for Δ_min
        i_j ← closest nodes between cc_0 and cc_1
        connect cc_0 and cc_1 in i_j
```

To test how much cities improve their bicycle layers using these four algorithms, we define two metrics on the bicycle layer that operationalize the notion of connectedness: i) $n_{LCC} = \frac{N_{LCC}}{N}$, the fraction of nodes from the bicycle infrastructure inside the largest connected component ($N_{LCC}$) compared to the total number of nodes from the same type of infrastructure ($N$), and ii) $\ell_{LCC} = \frac{L_{LCC}}{L}$, the fraction of link kilometers inside the bicycle infrastructure largest connected component ($L_{LCC}$) compared to the total number of link kilometers in the bicycle network ($L$). Both metrics take values between 0 and 1, where 1 means that there is only one connected component. An intermediate value, for example 0.2, means that the largest connected component contains 20% of all bicycle intersections or path kilometers. Executing our algorithms step by step these metrics can only grow, approaching 1 when the process is complete and they terminate. What distinguishes the algorithms is *how fast* these values grow.

We quantify directness through the metric: iii) bicycle-car directness $\Delta$, which answers the question "how direct are the average routes of bicycles compared to cars?" via the ratio between average distance by car and average distance by bicycle. For example, if the shortest car-route from west to east Manhattan is 4 km and the shortest route on the bicycle network between these two points is 5 km, the bicycle-car directness is $4/5 = 0.8$. Note that if the bicycle network is a subset of the street network, then $\Delta$ cannot be larger than 1. Formally we write $\Delta = \frac{\langle \delta^b_{ij} \rangle_{ij}}{\langle \delta^s_{ij} \rangle_{ij}}$, where $\langle \delta^s_{ij} \rangle_{ij}$ is the average car-route distance, and $\langle \delta^b_{ij} \rangle_{ij}$ is the average length of the shortest bike-route between $i$ and $j$. In each iteration of any of our algorithms, we implement this measure by randomly se-

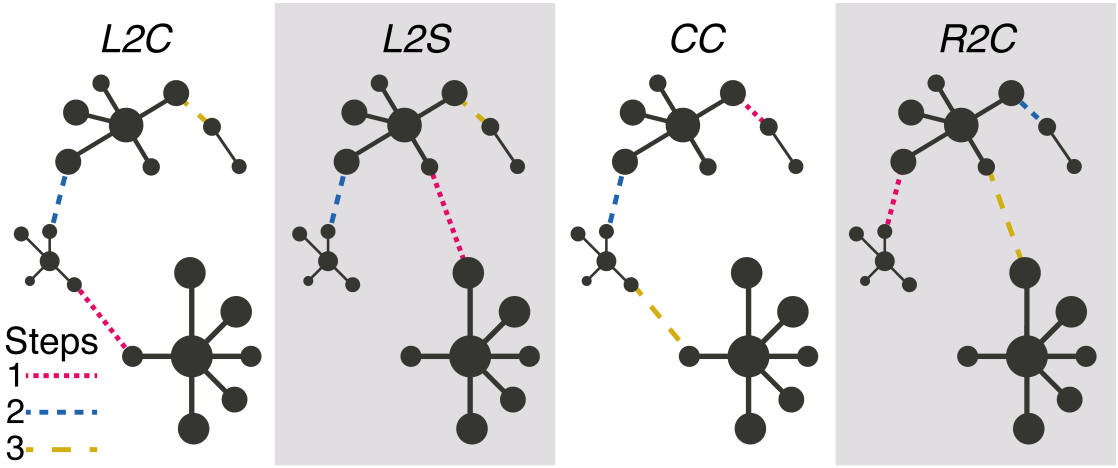



**Figure 2.2.** *Schematic representation of algorithms to improve bicycle network infrastructure: Largest-to-Closest (L2C) finds the largest component and connects it with the closest one; Largest-to-Second (L2S) connects the largest component with the second largest; Closest-Connected (CC) connects the two closest components; and Random-to-Closest (R2C) picks a random component and connects it to the closest.*

lecting one thousand pairs of origin-destinations nodes and then averaging the corresponding street/bicycle distance. To avoid undefined values due to disconnected components in the bicycle layer, we add the following condition: If a node from the pair $i$ and $j$ is in a different component, we assign the value $\delta_{ij}^{b} = 0$. This condition also ensures consistency of growing directness values while the algorithm merges more and more nodes into the same component.

Finally, in order to measure the cumulative efficiency of our algorithms, we define the metric: iv) $G_{LCC}$ as the relative gain of bicycle path kilometers in the largest connected component. For example, $G_{LCC} = 1.5$ means that the algorithm has increased the largest connected component's original size by 150%. Formally, $G_{LCC} = \frac{L_{LCC} - L_{LCC_0}}{L_{LCC_0}}$, where $L_{LCC_0}$ is the sum of kilometers in the largest connected component before the algorithm runs. As with all other metrics, $G_{LCC}$ is monotonically increasing with the growth algorithm, and reaches $\frac{1-\ell_{LCC_0}}{\ell_{LCC_0}}$ at the end of the dynamics.

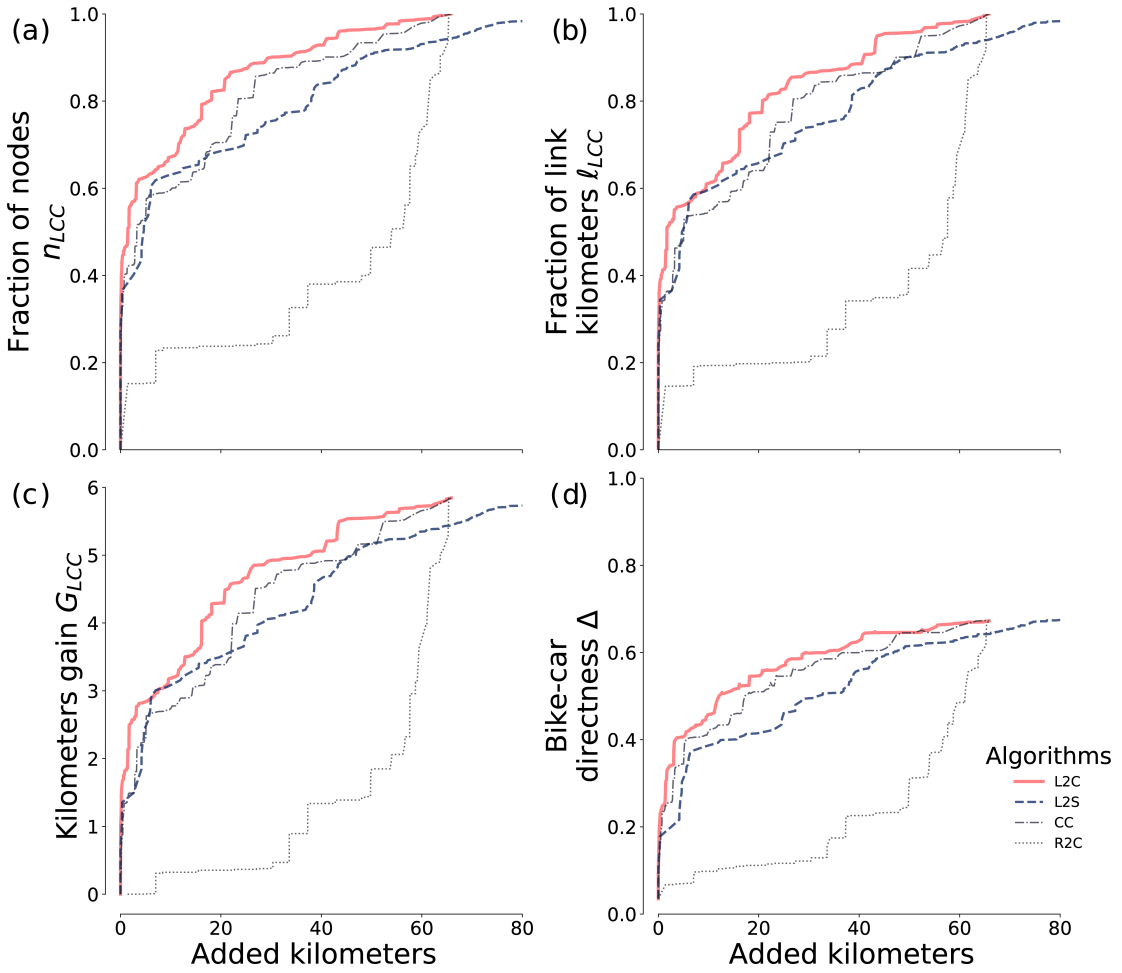



**Figure 2.3.** ***(a)*** *Normalized increase in nodes inside the largest connected component ($n_{LCC}$).* ***(b)*** *Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$).* ***(c)*** *Kilometers gain ($G_{LCC}$).* ***(d)*** *Bicycle-car directness ($\Delta$). Measures in (b-e) are plotted as a function of the sum of added links in kilometers, for the case of Budapest (for all cities see Fig. 4.1)*

## 2.4 Growing bicycle networks shows stark improvements with small investments

We demonstrate in Fig. 2.3 the power of the various growth strategies by showing the initial state of the bicycle layer for the case of Budapest and its state after 85 iterations of the *Largest-to-Closest* algorithm: At this point the network has almost quadrupled the size of its largest connected component (from 82 km to 313 km), with a negligible investment of just less than 5 km (corresponding to

1.4% of the previously existing bicycle infrastructure) in new connecting bicycle paths. In terms of connectedness, it goes from 15% to 56% connected. This rapid increase shows that the city can easily improve its bicycle infrastructure with small investments. For some extreme cases, like Bogota, with the same 5 km investment (an increase of 1.3% to the previously existing infrastructure) the bicycle-car directness increases from 6% to almost 48% and connectedness from 34% to 89%. Similar encouraging results hold for other cities (See Section 4.2.2).

The fraction of nodes inside the largest connected component increases rapidly with newly added links for all considered algorithms except *Random-to-Closest*, Fig. 2.3(a). The *Largest-to-Closest* algorithm performs better than the others, even more than *Closest-Components* which prioritizes minimum investments in the network. Since we are considering bicycle infrastructure, a better practical measure than the number of intersections is the number of kilometers that can be cycled using only designated paths. Figure 2.3(b) shows how this measure improves in a similarly explosive way: with an investment of only 20 km (5.9% of the existing infrastructure), the largest connected component will contain 80% of the original bicycle infrastructure. Results for the kilometer gain $G_{LCC}$ are shown in Fig. 2.3(c). Three of the four algorithms rapidly gain new kilometers, but as the invested new kilometers grow, each algorithm follows a different gain rate. Also for this metric, *Largest-to-Closest* is the algorithm with the best performance.

We also measure the bicycle-car directness ratio, Fig. 2.3(d). The bicycle-car directness $\Delta$ improves as the algorithms consolidate the network. These improvements are, however, indirectly driven by the improvement of connectedness, which boosts the accessibility of bicycles to different areas of the city. The flattening of the curves at a value considerably smaller than 1 (around 0.65) shows that cars will always outperform bicycles in terms of directness, having on average at least 33% shorter paths in the city. This suboptimal flattening is a natural consequence of the algorithms optimizing for connectedness only, not adding "redundant" connections. Nevertheless, the measure shows that, similar to connectedness, with a relatively negligible investment of bicycle path kilometers into the system, the bicycle network's directness improves drastically, even in the greediest case where the shortest possible missing link is added in every iteration. This result holds for all analyzed cities (see SI). The large differences between the baseline *Random-to-Closest* and our two algorithms (*Largest-to-Second* and *Largest-to-Closest*) show the importance of following an approach that consolidates and grows the largest connected component. Similar results are obtained for all the cities when taking into account the combination of bicycle infrastructure and safely bikeable streets ($\leq$ 30 km/h), in Figure 2.4 we

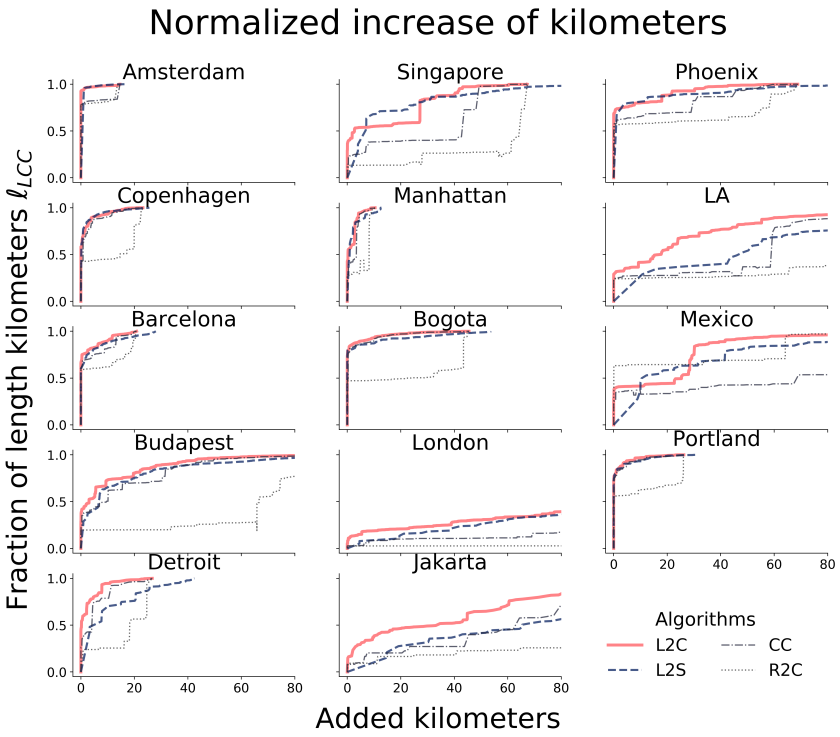



**Figure 2.4.** *Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$).*

show the increase in bikable kilometers when taking into consideration bikable streets and designated bicycle infrastructure combination for all the analyzed cities (see Section 4.2.3 for the rest of the measures).

## 2.5 Different cities have different optimal investment strategies

Differences arise in the state of the bicycle layer and its improvement after applying a growth algorithm. To see this effect, we rank how cities improve using the *Largest-to-Closest* algorithm in two different investment scenarios: investing either i) 5 km, or ii) 35 km. Figure 2.5 shows how cities improve when investing 5 km of bicycle infrastructure. We see that some cities get above 75% of their existing infrastructure connected, meaning that their bicycle layer only needs a small extension. On the other hand, cities like London, Los Angeles, and Jakarta need a larger investment to improve. Concerning bicycle-car directness, cities

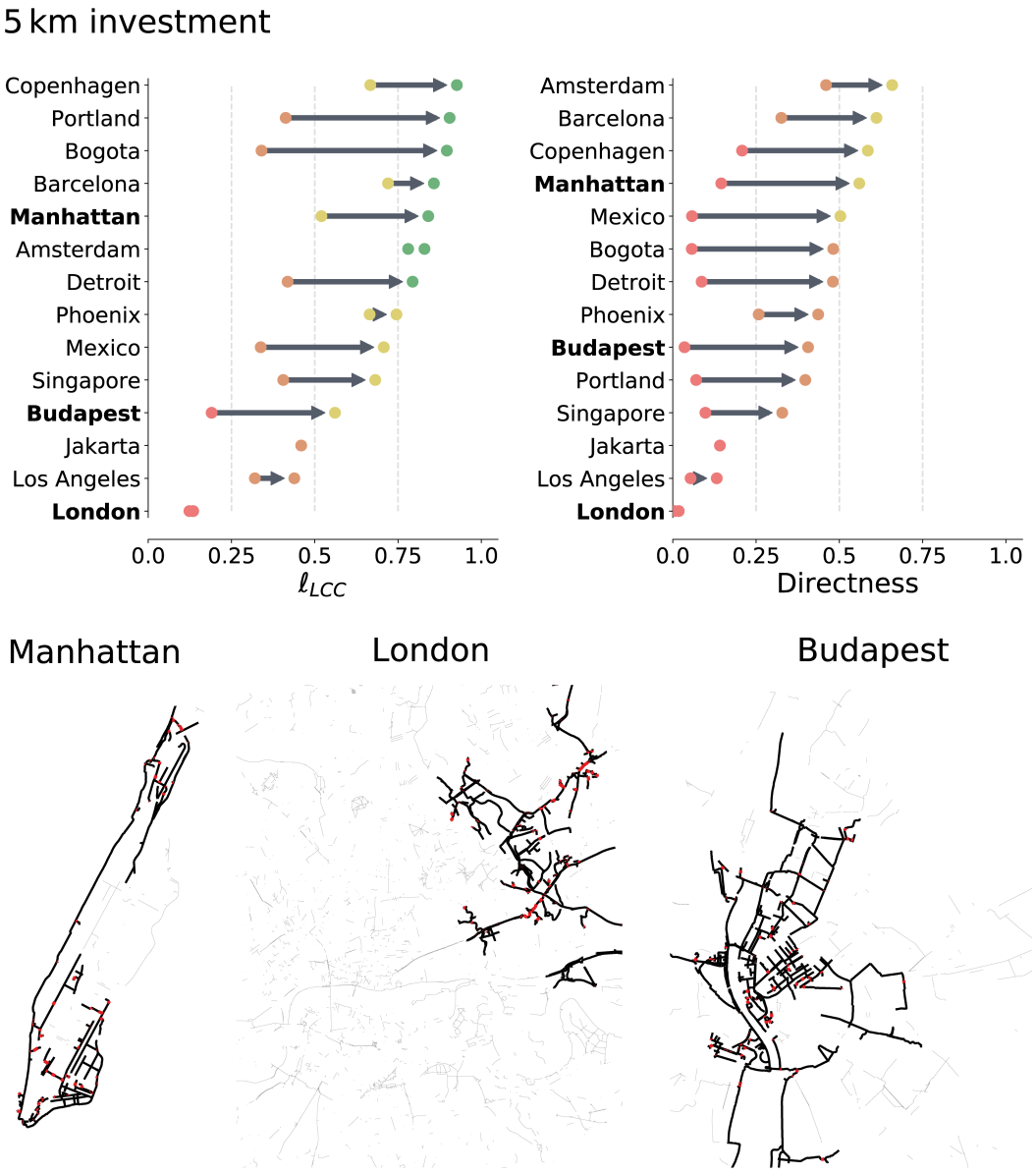



**Figure 2.5.** *Cities improvement and ranking using the* Largest-to-Closest *algorithm. We report the improvement and ranking on the fraction of total kilometers of bicycle infrastructure in the largest connected component ($\ell_{LCC}$) and in the bicycle-car directness ($\Delta$). Dotted lines show thresholds of* 25%, 50%, *and* 75%. *Plots (a-b) show investment strategies of 5 km and 35 km, respectively, the Manhattan, London, and Budapest plots show the suggested new links (red) after adding 5 km and 35 km, the newly created largest connected component (black), and the remaining separated components (grey).*

reach lower values due to the focus of the algorithms on completeness. In the worst performing cities like Los Angeles, a covered length close to 50% can be reached easily, while the bicycle-car directness ratio stays below 20%, showing that it is much harder to gain an acceptable bicycle infrastructure in cities where cars are overprioritized. The 35 km investment strategy shows that most cities can get at least 75% of their bicycle infrastructure connected, Fig. 2.6. The worst performing outlier is London, due to its bicycle layer containing more than 3000 connected components scattered around 1600 $km^2$ (see Table 2.1). In terms of bicycle-car directness London also performs badly, while Amsterdam is the best performing one.

The four proposed metrics capture the impact of newly created connections on the various components of the bicycle network. By linking previously disconnected neighbourhoods with a sustainable mode of transport, our approach focuses on consolidating bicycle infrastructure networks, thus making cities more cohesive and green. It does not, however, focus on growing the bicycle network into large areas of the city not currently served. To test to which extent such connectedness-based algorithms bring an indirect benefit for coverage, we measured the proportion of the city that is covered and reachable by bicycle with an epsilon of 500 meters around the bicycle infrastructure of the largest connected component, and calculated the percentage of nodes in the street layer that are covered by the bikeable area. The results of this measurement show a wide range of effects: Cities with an already high coverage above 80% (Amsterdam, Copenhagen) reach near instantly 100%, cities with an intermediate coverage (Manhattan, Bogota, Budapest) follow a more linear progression per added kilometer, while underdeveloped or sprawling cities (LA, London, Jakarta) show negligible growth (Fig. 2.7).

While our present goal (consolidation of the bike network) is intended to show the potential of our approach, an extension towards the exploration of new city areas will increase further the real-world applicability of our results. We consider this extension an interesting line of future research in the challenge of developing optimal data-driven strategies of transport network growth, potentially informed by theoretical frameworks such as optimal percolation [**?**, **?**]. Besides, the use of other network metrics, such as network efficiency [**?**], might unveil new dimensions characterising the impact of the proposed algorithms on the development of the bicycle infrastructure.

## 2.6 Discussion

Our starting point showed that a common characteristic of cities is the fragmentation of their bicycle networks. We have proposed the use of data-driven algo-

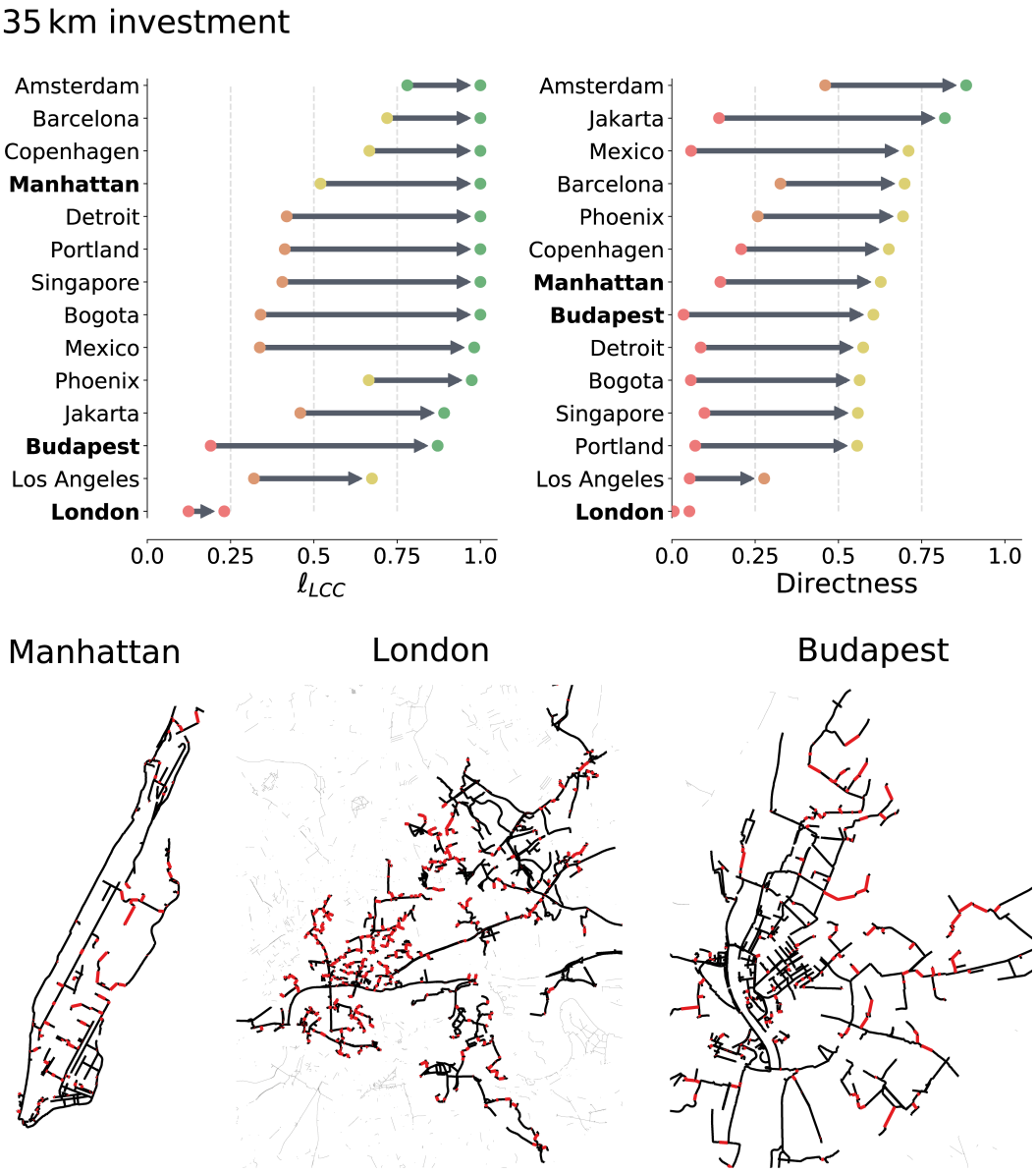



**Figure 2.6.** *Cities improvement and ranking using the* Largest-to-Closest *algorithm. We report the improvement and ranking on the fraction of total kilometers of bicycle infrastructure in the largest connected component ($\ell_{LCC}$) and in the bicycle-car directness ($\Delta$). Dotted lines show thresholds of* 25%, 50%, *and* 75%. *Plots (a-b) show investment strategies of 5 km and 35 km, respectively, the Manhattan, London, and Budapest plots show the suggested new links (red) after adding 5 km and 35 km, the newly created largest connected component (black), and the remaining separated components (grey).*

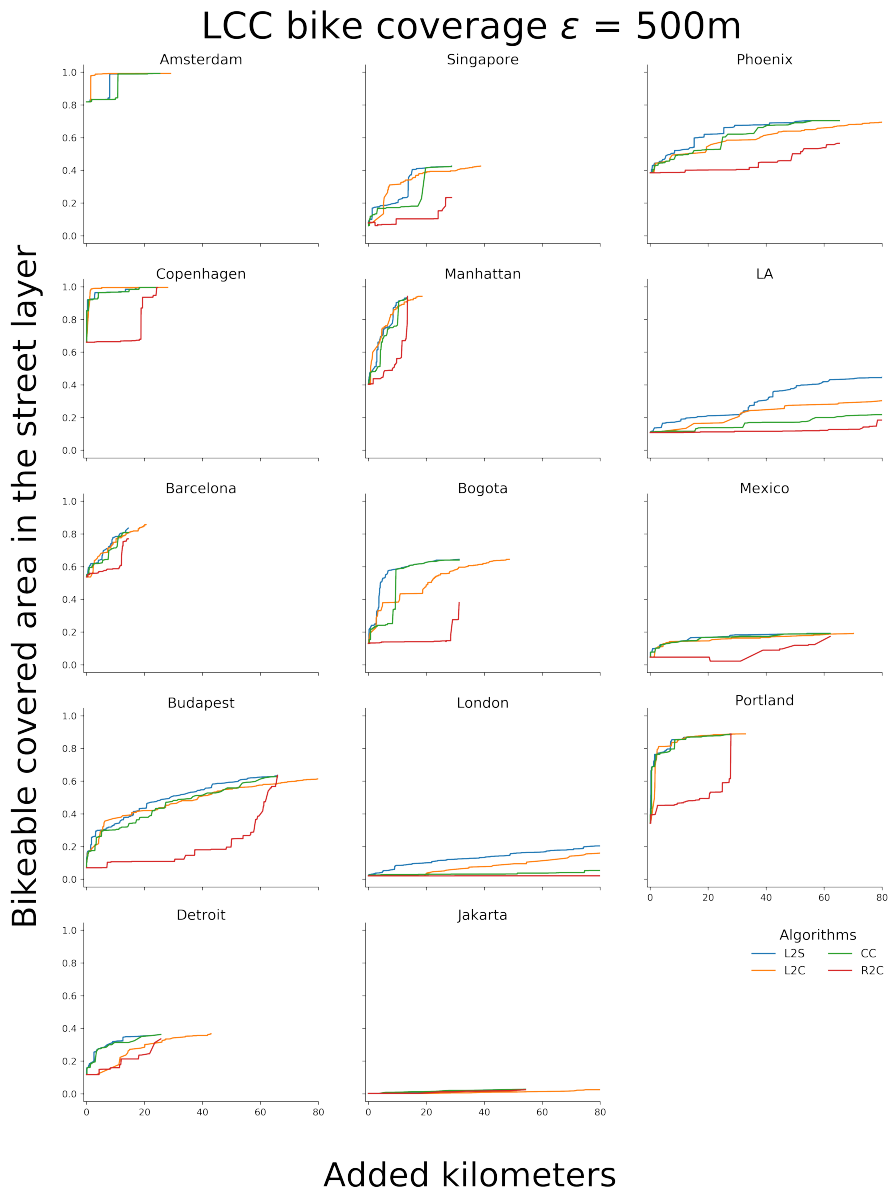



**Figure 2.7.** *Area of influence (coverage) for the bicycle infrastructure in the street layer.*

rithms to consolidate bicycle network components into connected networks to improve efficiently sustainable transport. We have shown that connecting the bicycle infrastructure in an algorithmic way rapidly improves the connectedness and directness of the bicycle layer. These algorithms, when compared with two baselines, highlight the usefulness of growing the bicycle network on a city-wide scale (considering all areas of the city) rather than randomly adding local

bicycle infrastructure. Improving the connectivity of bicycle lanes and paths improves not only the network itself, but also promotes the use of bicycles as means of transportation in a city, improving the health of its inhabitants [**?**].

Improving bicycle infrastructure one link at the time (by identifying suitable components to connect) is only the first step towards a systematic framework for realistic bicycle network growth strategies. Our current approach is not the last word in this development, since it does not yet explicitly optimize for directness and does not account for transport flow. Further, our proposed approach helps implement a more connected transport network which can improve the possibilities for multimodal transport. This could be starting point for implementing truly multimodal strategies, such as integration with public transportation, or bicycle parkings in transportation hubs [**?**].

In our algorithms, each new link works as a bridge between components, potentially having large betweenness centrality. Such high-betweenness segments could become overused and create bottlenecks in practice. To improve this situation, it would be necessary to create links in the network that act as redundant paths. In doing so, directness and coverage would also be improved, along with the network's robustness to interruptions. This is an interesting and possibly demanding task that we leave for future research, as the new links would have to be created in a coherent manner balancing trade-offs between network structure and mobility dynamics. We anticipate that complementing OpenStreetMap data with additional information on the use of traffic flow and movement data, like trips from bike share systems or origin-destination matrices, possibly from alternative sources such as municipalities and transportation agencies, might further improve the algorithms by better detecting underserved and optimal areas in the city where new links should be created. Despite these various possibilities for qualitative updates to the studied growth strategies, our first models have demonstrated the capability to generate substantial improvements with minimal effort.

The use of data-driven algorithms to identify crucially missing links in bicycle infrastructure has the potential to improve the mobility infrastructure of cities efficiently and economically. This approach is not only useful for planning city structure, but could also be used together with simulating mobility flows and to provide insights on how the system will behave after new measures are implemented. Ultimately, planning cultures and processes will also have to be accounted for [**?**]. We anticipate that a future stream of work should include longitudinal studies [**?**] in multiple cities, along with algorithmic simulations to first model and simulate possible changes to the transport network, and then to test those models with ground truth data, to compare the evolution of infrastructure and mobility dynamics between cities with different transport

priorities.

CHAPTER 3

# LIFE QUALITY AS WALKABILITY

## 3.1 Prelude

During the 20th century, most cities have evolved to accommodate a car-centric vision [?], allocating a privileged amount of urban space to motorized traffic [?, ?]. From a liveability perspective, this situation is suboptimal because the automobile infrastructure dominates and defines the walkable area, increasing car traffic, air pollution and deteriorating walkable conditions.

The concept of walkability is an important factor to consider in connection with liveability. Liveability refers to an environment from an individual perspective [?] which includes "a vibrant, attractive and secure environment for people to live, work and play and encompasses good governance, a competitive economy, high quality of living and environment sustainability" [?]. Thus in a liveable city, there must be an emphasis not only on sustainable transportation and built environment to reduce the harm on nature [?, ?] but also encouraging citizens to walk for supporting their physical and mental well-being [?]. However, improving walkability is more complex than we would think. Walking should be an available, safe and well-connected mode of transportation, but as Speck put it well, it should be interesting and comfortable as well, to have a feeling of the streets as 'outdoor living rooms' [?].

The pedestrian infrastructure that sustains walkability in a city can be described as a network [?]. This approach has been useful to identify street patterns [?, ?] and its evolution [?, ?], measure the morphology of cities [?], and how the streets connectivity impacts on pedestrian volume [?].

The various approaches to create a walkability index or so-called walk score consider mainly the following components: safety and security [?, ?]; convenience, attractiveness and public policy [?, ?], connectedness [?], but also reckon

with the land use mix and residential density of the certain area[**?**]. Another approximation rather accents the importance of its effect on air pollution, health problems, travel costs and even on the sense of community[**?**]. Thus measuring walkability not only captures the propensity to walk in a city but also includes the components a liveable city must have and support, under the umbrella of sustainability.

There are good examples of how sustainable city development initiatives tackle growing inequalities with data-driven approaches. Long Island used city data to analyze which amenities are needed to increase the quality of life in a newly built environment [**?**], other cities are investing in smart technologies to develop public transport, connecting spatially discriminated areas [**?**, **?**].

Since the number of components which should be taken into consideration in creating a walkability index is high, the types of data are also mixed and thus difficult to integrate. While the information on connectedness, security, residential density, etc. is quantitative and in general easily available, gaining opinion about attractiveness, convenience, or even about the feeling of security is more complicated. Here we propose to use a data-driven approach as a proxy to quantify life quality, making it reproducible and easily expanded to include different data sources. We apply our methods to Budapest, but as having an emphasis on the online and easily available quantitative data, the methods can be generalized and applied to any city.

## 3.2 Data

We work with three different data sources: networks, points of interest and city attributes. The pedestrian network and points of interest were acquired using OSMnx [**?**], a python library to download and construct networks from OpenStreetMap (OSM). The data contained in OSM is of high quality [**?**, **?**] in terms of correspondence with municipal open data [**?**] and completeness: More than 80% of the world is covered by OSM [**?**].

The majority of points of interest were downloaded from OpenStreetMap, from different classification keys (amenity, tourism, shop, office, leisure) using OSMnx [**?**]. We filtered the points of interest using the districts' demarcation [**?**], to get only the data within Budapest boundaries, having, as a result, more than $39,000$ data points. We complement the data sets with secondary data sources as specialized directories of doctors and childcare facilities (see appendix).

We categorize the points of interest in six main categories: I) Family friendliness (Access to education and daycare, and family support services), II) Access to health care and sport facilities, III) Art and culture (e.g. museums, exhibitions), IV) Nightlife (e.g. bars, restaurants), V) Environment (air quality and ac-

cess to green areas), and VI) Public Safety. The points of interest and secondary data sources are available at `https://github.com/nateraluis/Budapest_LQI`

The district-level data (population and crimes) were obtained from the Hungarian Police's public database, calculated based on the number of crimes committed in public places 100 thousand per capita in 2018 [**?**]. Population data is coming from the 2016 micro-census conducted by the Hungarian Statistical Bureau [**?**]. We took into account the air pollution, this data set coming from National Air Pollution Measurement Network [**?**], containing the geolocation of the air quality stations and different measures (annual median concentration of carbon monoxide, nitrogen dioxide, and PM10 dust).

Accuracy of the Life Quality Index (LQI) model highly depends on how comprehensive the distribution of listed services. We use OSM as our key data source, but to achieve a more comprehensive and country-specific database we collect publicly available data from various Hungarian websites for each category (See Appendix A for databases and sources)

The network contains all the sidewalks and pedestrian designated infrastructure, it is conceptualized as undirected, nonplanar and primal network [**?**]. The pedestrian network is described as a weighted graph, with its adjacency matrix $W = w_{ij}$ where the weight $w_{ij}$ contains the length between $i$ and $j$ if connected, and 0 otherwise.

We assigned properties to the nodes of the network, matching the nodes with their corresponding districts, then assign nodes as attributes based on the district level data (population and crimes, see section 3.3.2). For the pollution data, we calculated the corresponding Voronoi cells, for the air quality stations, and matched the nodes with them, we divided the pollution by the number of nodes in each corresponding cell and assigned the value to the nodes (See section 3.3.3). For the edges, we encoded their length $\ell_{ij}$ along with the traversal time $Tt_{ij}$ between nodes $i$ and $j$ calculated as $Tt_{ij} = \frac{\ell_{ij}}{ps}$ where $ps$ is the pedestrian speed as a constant rate of $5km/h$.

## 3.3 Quantifying Life Quality

The life quality of a person is largely subjective and hard to quantify. However, it is both intuitive and has been scientifically shown that the environment and personal well-being strongly correlate [**?**]. Thus using environmental factors as proxies, life quality and livability becomes quantifiable [**?**].

The main environmental factors we consider in our model are: the availability of services and amenities, the quality of the infrastructure, environmental factors and safety. The goal of our model is to quantitatively characterize the

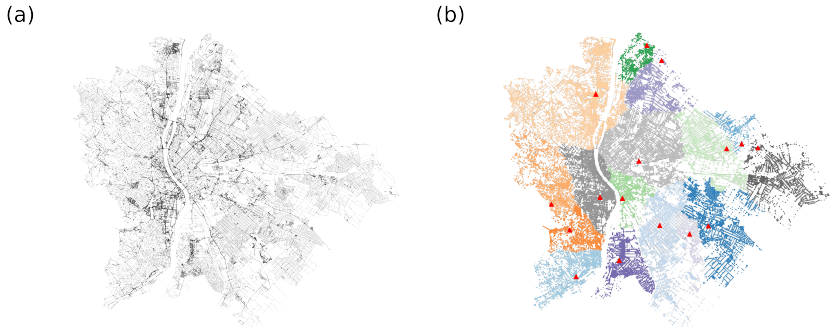



**Figure 3.1.** ***(a)*** *Network representing Budapest pedestrian structure. The network was built following a primal approach, where the edges are sidewalks and pedestrian infrastructure, and nodes are intersections.* ***(b)*** *The graph-Voronoi tessellation of the Budapest network, generated using a subset of 15 parks as seeds. The color of the nodes represents the cell they belong to and the highlighted red dots are the seeds of each cell. The distance measure between two points is defined as the weighted shortest path on the graph, the weights being the average time required to cross a given edge.*

immediate environment of residents in the space of factors that affect life quality.

The fundamental framework of our model and our calculations is the network representation of Budapest's pedestrian infrastructure. The nodes of the network represent intersections, while links are sidewalks and pedestrian infrastructure. The output of our model is an index, that characterizes every node of the Budapest network, giving a high-resolution quality-landscape of the city. The index is ultimately a number aggregated from multiple sub-categories, and its main value is highlighting inequalities and relative deficiencies within the city.

The final value of the index is a weighted sum, characterizing every node (intersection) in the network:

$$Q_i = w^{services}\tilde{Q}_i^{services} + w^{safety}\tilde{Q}_i^{safety} + w^{environment}\tilde{Q}_i^{environment} \tag{3.1}$$

In the equation $i$ represents an individual node in the network. The "tilde" above the $Q$ terms means that the values of the different category indices are normalized within the category. The weights $w$ assigned to every term are arbitrary and are highly context-dependent. We include the weights used for producing the results of this paper in Appendix B. All terms of the equation are discussed in the following sections.

### 3.3.1 The services index: $Q^{\text{services}}$

The number quantifying each node in terms of how well it is connected with amenities and services is a weighted sum of sub-categories as well.

$$Q_i^{services} = \sum_c w^c Q_i^c \tag{3.2}$$

where, $c$ denotes categories (family, culture, health, sport, and nightlife), and $w^c$ the importance (weight, see Appendix B) of category $c$. Some categories, like family, have further subcategories. Even though we have also had data and made a separate analysis on tourism, its effects on life quality of the residents are ambiguous, so we decided to omit it from the index.

What sets categories apart is that they incorporate different sets of amenities, with a few overlaps. The details of the categorization of amenities are included in the Appendix.

For every service/amenity class we have a given set of points of interest (POI) along with where the amenities of that class are available, with exact geolocation. We assign every POI of a given amenity class (e.g supermarket, pharmacy, school, etc.) to the nearest node on the infrastructure network. Each set of POIs organically generates a spatial partitioning of the city with one partition per POI. The partition of a POI is the set of all the nodes from which that particular POI can be reached faster than any other POI of the same class.

Mathematically these partitions are called graph-Voronoi cells [**?**, **?**], where every node of a cell is assigned to its closest seed (POI). Distance, in this case, is not euclidean or geometric distance, but the distance on the network, where we use the weighted shortest path between two nodes as the distance measure. The weight of links is a temporal parameter encoding the average time required to cross the represented street from one end to the other, thus the weight is a simple product of average speed and length of the street. This is in principle very similar to the way navigation systems find routes between points. For an example of a graph-Voronoi partitioning see Figure 1 (b).
To assess how well connected a node is to amenities we consider the following factors:

- How important is an amenity - weight ($w_a$)
- How long does it take to reach the amenity - time to reach ($t_i a$)
- Relatively how many nodes (or people) does the amenity share with - exclusivity ($P_a$)

From the three factors, the latter two are calculated using the city infrastructure network. The index for an amenity class, from the perspective of node $i$, is

proportional with its importance (weight) and it is inversely proportional with the time to reach the closest POI from $i$ and with the degree of exclusivity.

$$q_i^a = \frac{w_a}{(P_a + 1)(t_{ia} + 1)} \tag{3.3}$$

There can be certain singular cases when a Voronoi cell is empty ($P_a = 0$, i.e. no residents in the area) or the node i in question is right at the POI ($t_{ia} = 0$). To avoid anomalies in the index we added 1 to both parameters.

The index of one category is proportional to the sum of its amenity-indices (calculated in (3.3)). To treat this number on the right scale (in practice we can get very large and very small numbers) we take the natural logarithm of the sum across amenities.

$$Q_i^c = log(\sum_a q_i^a) : \tag{3.4}$$

As we have mentioned earlier the final services index is the weighted sum of the indices of the sub-categories.

$$Q_i^{services} = \sum_c w^c Q_i^c$$

Finally we normalize the values of $Q^{services}$ so its values are comparable to the other values of the final $Q$ equation (3.1):

$$\tilde{Q}_i^{services} = \frac{Q_i^{services} + |min(Q^{services})|}{max(Q^{services}) + |min(Q^{services})|} \tag{3.5}$$

### 3.3.2 Safety index: $Q^{\text{safety}}$

The safety index is calculated across districts based on the number of crimes committed per one hundred thousand residents. Since the highest resolution data available to us was on the district level, every node $i$ in the same district will have the same safety index value. The crime index:

$$Q_i^{crime} = \frac{N_{crime}^{district}}{n_i^{district}}$$

Where $N_{crime}^{district}$ is the number of crimes committed in a district in a year, and $n_i^{district}$ is the number of nodes in the district. The safety index is one minus the normalized crime index.

$$\tilde{Q}_i^{safety} = 1 - \frac{Q_i^{crime}}{max(Q^{crime})} \tag{3.6}$$

### 3.3.3 Environmental index $Q^{\text{environment}}$

The environmental index is made up of two components: air pollution ratio and ratio of natural areas.

#### Air pollution ratio

We use the data provided by Budapest's air pollution measuring stations for the year 2018. For this study, we used the yearly median value of three polluters: carbon monoxide, nitrogen dioxide, and PM10 dust-pollution. As an approximation, we project the geometric Voronoi cells of the measuring stations onto the city map and each node will receive the pollution metrics of the geometrically closest station. We divide these values with the yearly upper health limit for the given polluter to assess to what degree do these values approximate the health limit. Thus the air pollution index of one node is formalized as follows:

$$C_i = \frac{c_i^{CO}}{c_{limit}^{CO}} + \frac{c_i^{NO_2}}{c_{limit}^{NO_2}} + \frac{c_i^{Pm10}}{c_{limit}^{Pm10}}$$

Where $c_{limit}^{CO} = 3000g/m^3$, $c_{limit}^{NO_2} = 40g/m^3$ , $c_{limit}^{Pm10} = 40g/m3$ are the yearly upper limits based on data from the Hungarian Air Quality Network [**?**].

#### Ratio of natural areas

For this index, we have data on the neighborhood level, which is a more granular level of administrative partitioning the city than the districts are. In this case, we project the same index onto every node in the same neighborhood. We consider as natural areas forests, parks and water surfaces (ponds, rivers, etc). The index:

$$T_i = \frac{R_{water}^{nh(i)} + R_{forest}^{nh(i)} + R_{park}^{nh(i)}}{max(T)}$$

Where $R_x^{nh(i)}$ is the relative surface area of natural area $x$ within the neighborhood that $i$ belongs to ($nh(i)$). In other words, the surface area of a natural area is divided by the number of nodes in the neighborhood and the surface area of the neighborhood. Thus $R_x^{nh(i)} = \frac{T(x)}{T(nh(i))n_{nh}}$, where $T(x)$ is the surface area of $x$ natural area, $T(nh)$ is the surface area of $nh$ neighborhood and $n_{nh}$ is the number of nodes in neighborhood $nh$. The final environmental index:

$$Q_i^{environment} = \frac{1 + T_i}{1 + C_i}$$

, That after a normalization is:

$$\tilde{Q}_i^{environment} = \frac{Q_i^{environment}}{max(Q^{environment})}$$

## 3.4 Results

We quantify life quality in terms of each category (family support, education healthcare, sport, culture, nightlife, environment), and an overall measurement which contains all 6 categories and crime rate normalized by the population for the city of Budapest. Our method allows us to measure life quality for each intersection of the city, which helps to capture within neighborhood inequalities too. Analysis on the category level is beneficial for targeted policy interventions for better service allocation.

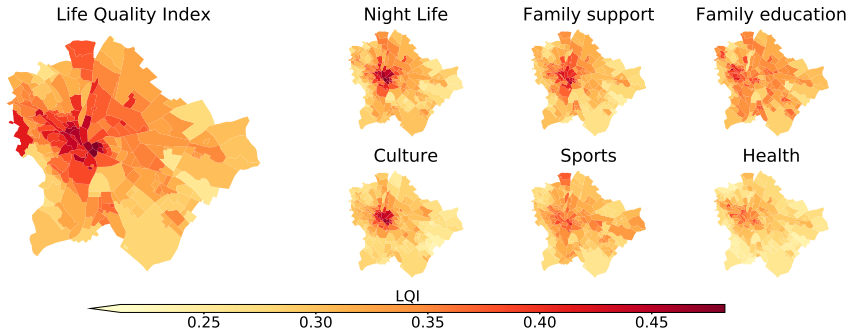



**Figure 3.2.** *Budapest neighborhoods, average life quality by categories and aggregated life quality index.*

Figure 3.2 shows our overall life quality index (LQI) and by categories. Heatmaps reveal important features of Budapest. Similarly to most European cities life quality is much better in the inner districts [**?**, **?**], especially in the case of Night Life and Culture.

Budapest is divided by the Danube river into two main parts: Buda and Pest. The river does not only serve as a geographical border but due to historical reasons, it also divides the citizens by social status. Hilly Buda, on the West side of the river, used to be the capital of the country, with the residence of the former Hungarian king. On the other side, the mainly flat Pest used to be the agricultural supporter of the aristocrats in Buda [**?**]. Even though the city has changed dramatically since the Monarchy, the division of Buda and Pest persists, and our life quality index captures it well. However certain services

are legally guaranteed to be evenly distributed in the city, such as education and healthcare, for precise modeling one should take into account private care too, which highlights inequalities. So, the traditional division of Buda and Pest is even visible in categories where there should not be that much of a difference (Education, Family Support, Healthcare).

Results also highlight that category LQI-s are highly correlated, less liveable neighborhoods are constant regardless of the amenity category, and well-performing neighborhoods do not change either. It is caused by two main factors: the lack of amenities and the relatively high walking distances in the suburbs.

The compact city concept focuses on building more sustainable and livable cities while designing practical neighborhoods where citizens can maintain everyday life without a car [**?**]. Since, the walkability of a neighborhood highly correlates with its liveability [**?**] and the suburbs in Budapest do not show any compact city design features, both long distances and the lack of amenities effects suburban habitats lives negatively.

### 3.4.1 Evaluation

Multiple methods have been developed to evaluate the accuracy of quality of life metrics: Scholars used expert validation with geographic visualization [**?**, **?**], correlations with socioeconomic characteristics [**?**] and surveying citizens' perceptions of the conditions of life [**?**].

Our evaluation is based on the micro-economical *hedonic approach* of estimating the values of public goods. In a capitalist market, real estate prices reflect the recognition of a neighborhood's characteristics: Prices are formed based on demand, more desirable places are more expensive, due to the underlying assumption of providing a higher quality of life. [**?**] Estimating neighborhoods life-quality with real estate prices has a long tradition in urban literature [**?**, **?**, **?**], therefore we adopt this method to evaluate our model.

We collected the average $m^2/EUR$ price for all 23 districts of Budapest in January 2019 [**?**] and correlated each LQI category averaged by district with it. Figure 3.3 shows that our overall LQI correlates the most (R=0.91) with the real-estate prices. Most of its components have a positive correlation with real-estate prices, except the environment which is calculated based on air pollution and green surface proximity. The life quality (LQI) in Budapest is much higher in densely populated downtown districts, which are lack of green surface and suffers from high air pollution due to heavy traffic.

**Summary** Locals of Budapest, like in most European cities, traditionally values downtown areas. The relative closeness to CBD, good access to public

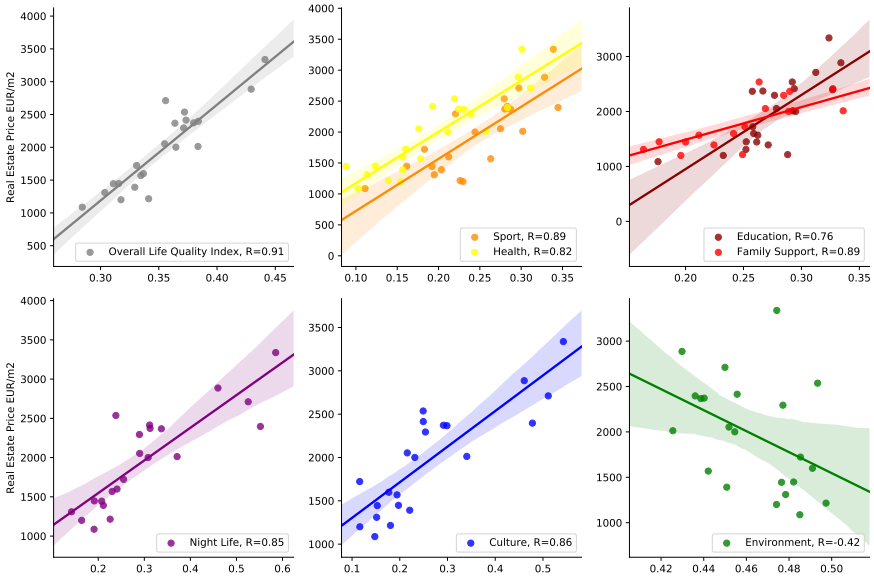



**Figure 3.3.** *Districts of Budapest Life Quality Index (LQI) and its components correlated with real estate prices ($m^2$ / EUR)*

transport, and vital city life kept it as a desirable area for living [**?**]. However, in recent years, the city is facing new challenges: due to gentrification [**?**] and over-tourism (eg.:Airbnb) real estate prices are sky-rocketing in downtown areas. In contrast with the early 2000-s when (upper) middle-class moved to the suburbs, nowadays, lower-income families and young professionals are leaving the downtown behind in hope for more affordable living.

As our findings show, Budapest is quite centralized and the quality of life highly correlates with real estate prices, which possibly lead to even more inequalities in the future. This spatial discrimination with longer traveling time, less fulfilling environment, and potential segregation reduces the chances of upward mobility and the quality of life of individuals [**?**].

## 3.5 Discussion

We have proposed a methodology to quantify life quality as a function of walkability on urban networks. We have used open data to capture inequalities between neighborhoods and districts in the city. We have shown that the real estate market reflects the life quality that our methods found.

A data-driven approach for quantifying life quality at such a granular level like our proposed method can help decision-makers to tackle social and envi-

ronmental challenges better. Designing compact, liveable neighborhoods, considering also the upcoming environmental crisis is the number one priority of many cities worldwide.

The use of open data sources and algorithmic approaches adds up towards a systematic framework for understanding urban liveability. Our current approach is not the last word in this development since it does not yet account for multiple other variables, such as the quality of services and infrastructure, and other qualitative variables. To capture the more specific indicator of liveability in different cities it would be necessary to work with more granular and city-dependant data.

We anticipate a future stream of research focused on the use of worldwide open data sets to quantify urban liveability, including longitudinal studies in multiple cities, along with algorithmic modeling, simulations, and machine learning approaches, to first quantify the liveability, propose changes and test them with the ground truth data.

CHAPTER 4

# APPENDICES

## 4.1 Life quality as walkability

### 4.1.1 Secondary data sources

- Sport associations in Budapest [**?**]
- Kindergartens, daycares, primary and secondary education [**?**]
- Art and music schools [**?**]
- Child health services [**?**]
- Social welfare system (eg.: elderly care) [**?**]
- Culture centers [**?**]
- Indoor playgrounds [**?**]
- Healthcare (hospitals, private and public clinics, specialists) [**?**]
- Fitness and training facilities [**?**]
- Outdoor fitness facilities [**?**]
- Thermal baths and spa [**?**]
- Playgrounds and parks [**?**]

### 4.1.2 Weights used in the calculations

The weights of the different $Q$ indices in the final aggregation as well as in sub-categories highly depends on the context and the nature of the problem. Here we present the values we used to generate the results of this study, that were agreed upon consulting with experts.

The weights of the sub-indices from equation (3.1) are of the following values:

- $w^{\text{services}} = 0.7$;
- $w^{\text{safety}} = 0.1$;
- $w^{\text{environment}} = 0.2$

The category weights used in equation (3.2), aggregating $Q^{services}$ are:

- $w^{\text{family}} = 0.3$;
- $w^{\text{health}} = 0.3$;
- $w^{\text{culture}} = 0.15$;
- $w^{\text{sport}} = 0.15$;
- $w^{\text{night life}} = 0.1$

## 4.2 Data-driven strategies for optimal bicycle network growth

### 4.2.1 Data

Figure 4.1 shows the connected component size distribution $P(N_{cc})$ for all considered layers and cities.

### 4.2.2 Bicycle network improvement

Here we show the improvement of the bicycle network after the implementation of the algorithms. We measure the improvement with four different metrics. Two of them implement the notion of connectedness: i) Fraction of nodes inside the largest connected component compared to the total number of nodes

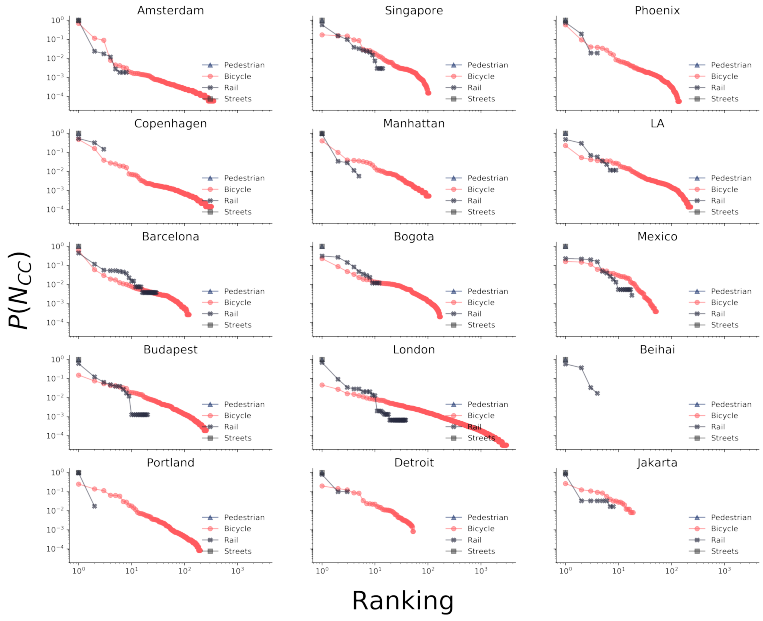



**Figure 4.1.** *The connected component size distribution [$P(N_{cc})$] for all cities and layers is well connected except in the bicycle layer. London has the most fragmented bicycle infrastructure layer, with more than 3000 components*

in the bicycle layer, and ii) the fraction of link kilometers inside the largest connected component. In Figure 4.2 and **??** we show these two measures for fourteen different cities. We also quantify iii) bicycle-to-car directness to answer the question "how direct are the average routes of bicycles compared to cars?". Finally, in order to measure the cumulative efficiency of our algorithms, we define the metric: iv) $G_{LCC}$ as the relative gain of bicycle path kilometers in the largest connected component. In Figures 4.5 and 4.6 we report these two measures for all algorithms and cities considered.

### 4.2.3 Bicycle network and 30 km/hr streets

We applied the algorithms to the bicycle infrastructure and all the bikeable streets, those with a speed limit of 30 $km/hr$ or less (see Figures 4.7, 2.4, 4.8, and 4.9).

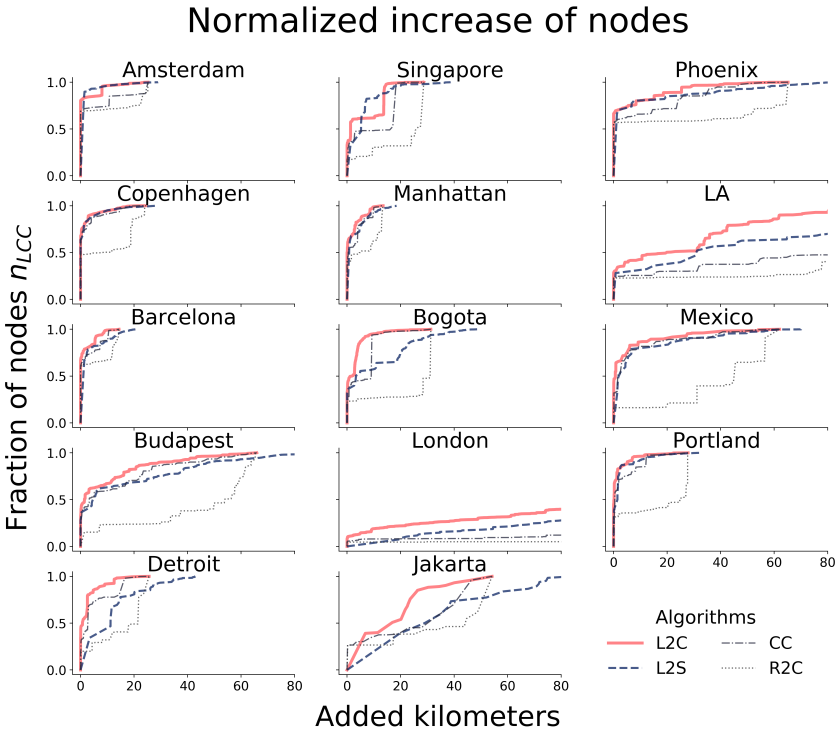



**Figure 4.2.** *Normalized increase in nodes inside the largest connected component ($n_{LCC}$).*

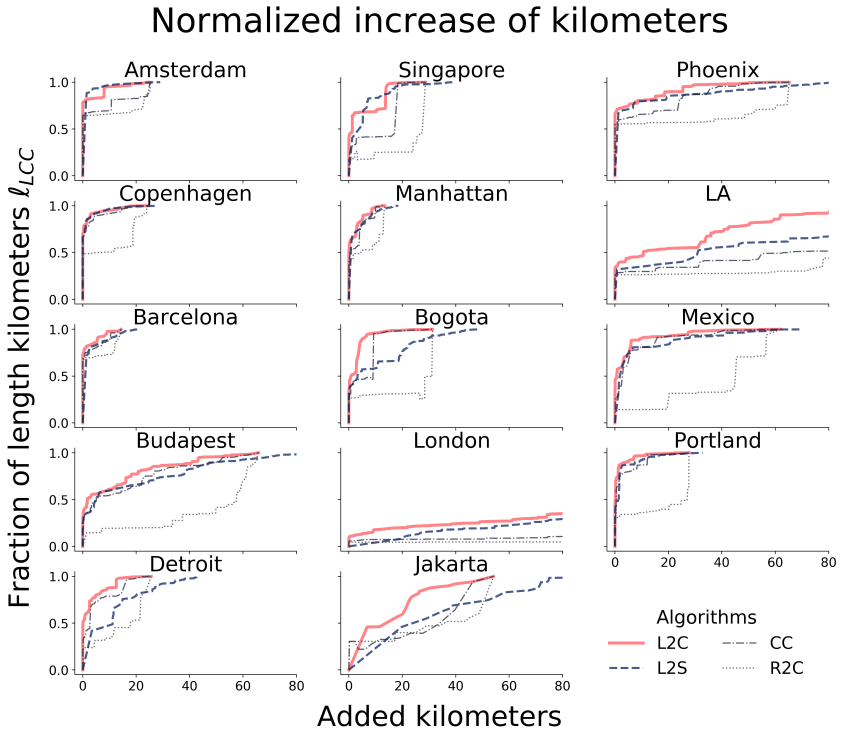



**Figure 4.3.** *Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$).*

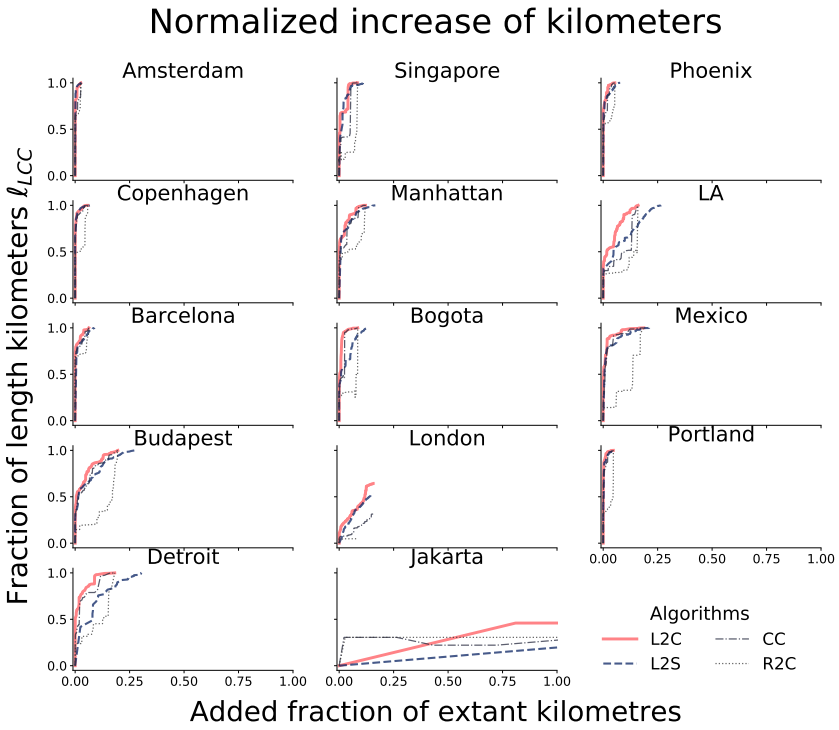



**Figure 4.4.** *Normalized increase in kilometers inside the largest connected component ($\ell_{LCC}$) versus the fraction of extant kilometers to be added.*

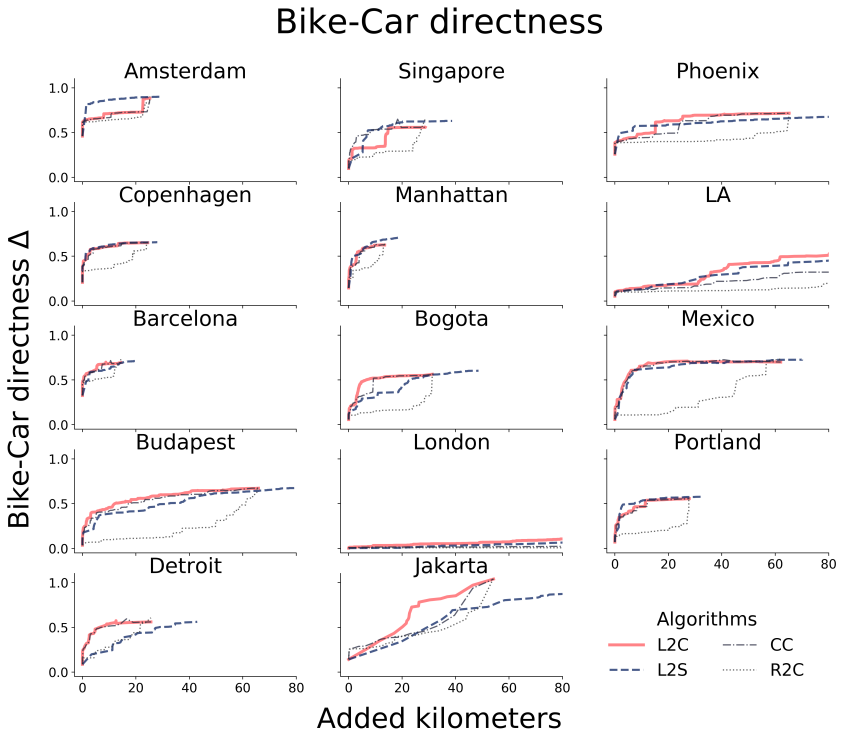



**Figure 4.5.** *Bike-car directness Δ per invested kilometers.*

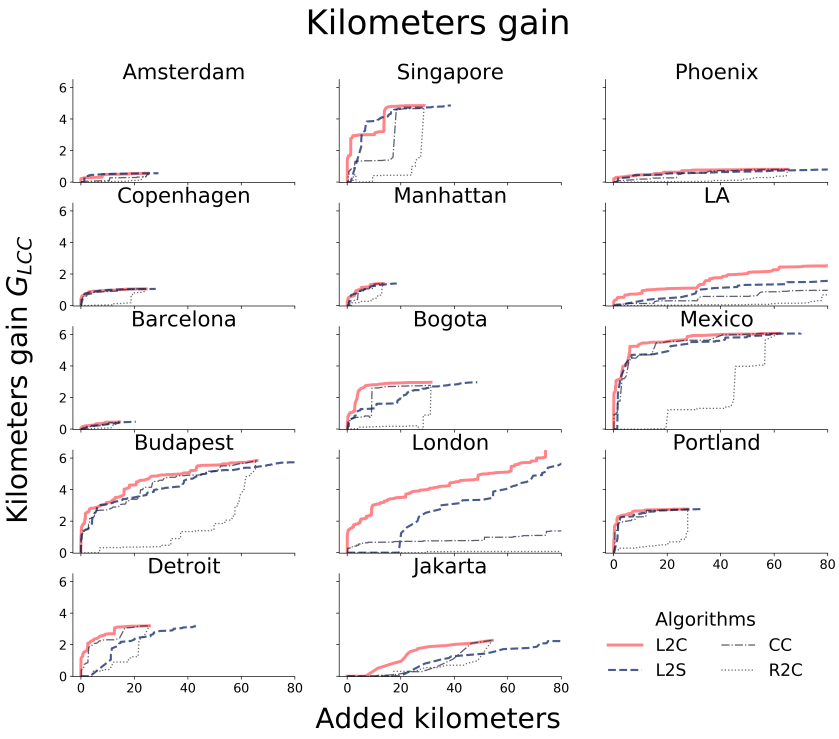



**Figure 4.6.** *Kilometers gain in the largest connected component.*

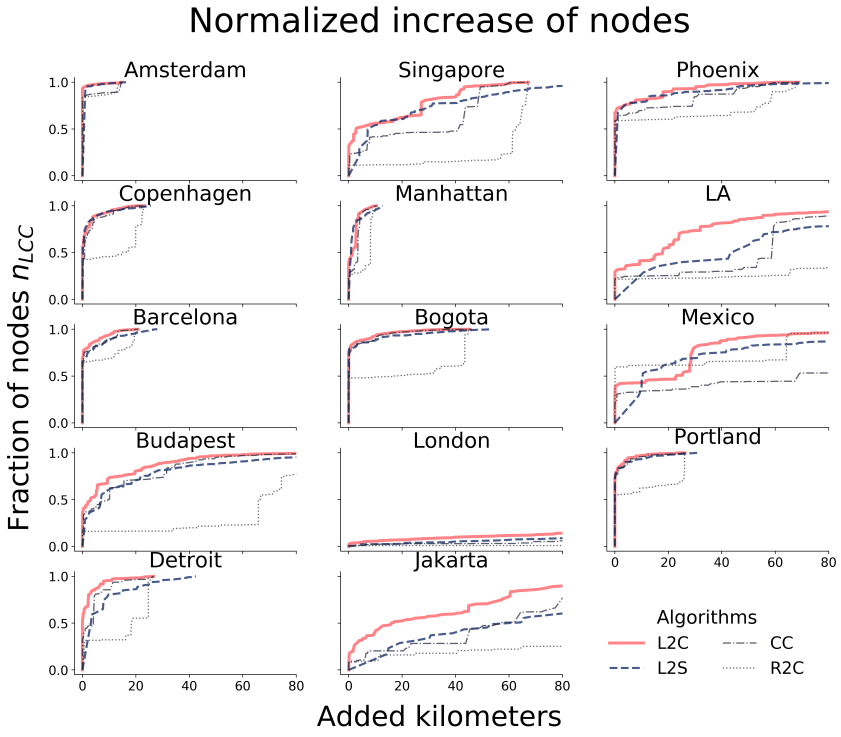



**Figure 4.7.** *Normalized increase in nodes inside the largest connected component ($n_{LCC}$).*

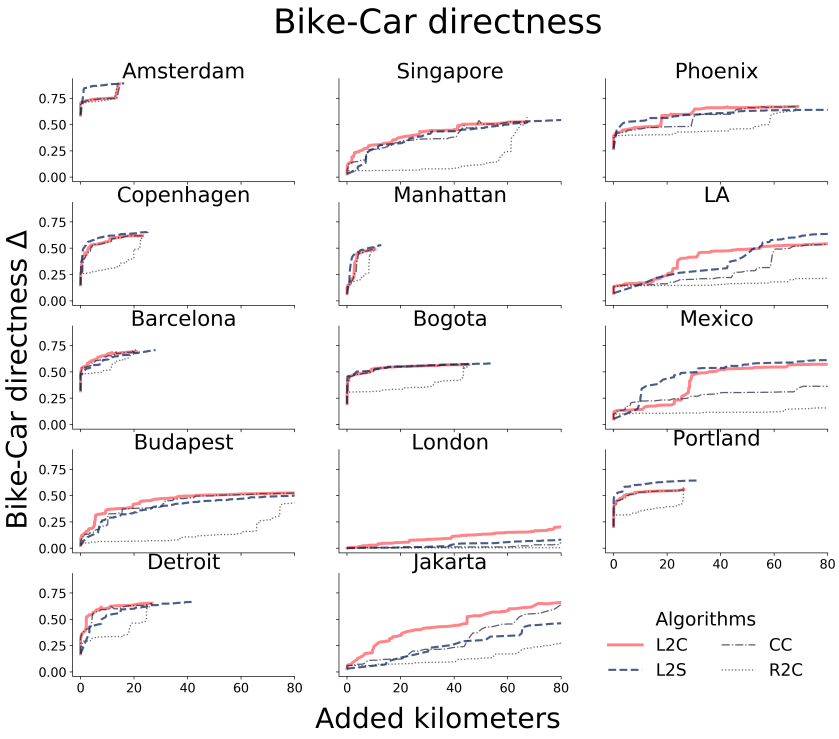



**Figure 4.8.** *Bike-car directness Δ per invested kilometers.*

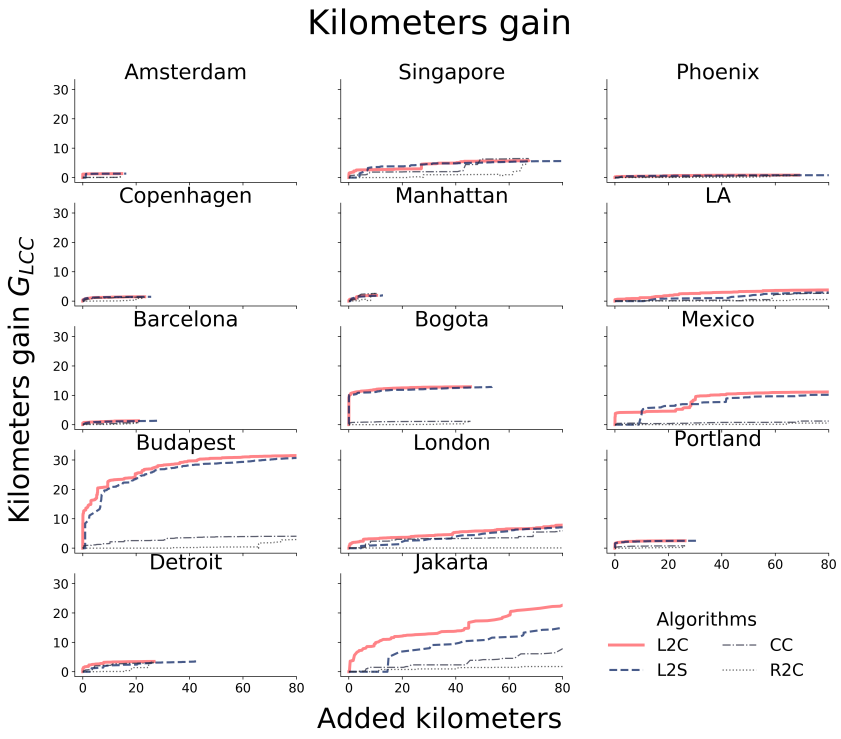



**Figure 4.9.** *Kilometers gain in the largest connected component.*