Title
Street Network Models and Measures for Every U.S. City, County, Urbanized Area, Census Tract, and Zillow-Defined Neighborhood

Permalink
https://escholarship.org/uc/item/9v99703k

Journal
Urban Science, 3(1)

Author
Boeing, Geoff

Publication Date
2019-03-01

DOI
10.3390/urbansci3010028

Peer reviewed
Communication

Street Network Models and Measures for Every U.S. City, County, Urbanized Area, Census Tract, and Zillow-Defined Neighborhood

Geoff Boeing

School of Public Policy and Urban Affairs, Northeastern University, 360 Huntington Ave, 360Y RP, Boston, MA 02115, USA; g.boeing@northeastern.edu

Received: 28 December 2018; Accepted: 14 February 2019; Published: 1 March 2019

Abstract: OpenStreetMap provides a valuable crowd-sourced database of raw geospatial data for constructing models of urban street networks for scientific analysis. This paper reports results from a research project that collected raw street network data from OpenStreetMap using the Python-based OSMnx software for every U.S. city and town, county, urbanized area, census tract, and Zillow-defined neighborhood. It constructed nonplanar directed multigraphs for each and analyzed their structural and morphological characteristics. The resulting data repository contains over 110,000 processed, cleaned street network graphs (which in turn comprise over 55 million nodes and over 137 million edges) at various scales—comprehensively covering the entire U.S.—archived as reusable open-source GraphML files, node/edge lists, and GIS shapefiles that can be immediately loaded and analyzed in standard tools such as ArcGIS, QGIS, NetworkX, graph-tool, igraph, or Gephi. The repository also contains measures of each network’s metric and topological characteristics common in urban design, transportation planning, civil engineering, and network science. No other such dataset exists. These data offer researchers and practitioners a new ability to quickly and easily conduct graph-theoretic circulation network analysis anywhere in the U.S. using standard, free, open-source tools.

Keywords: street networks; graphs; network science; OpenStreetMap; urban planning; transportation

1. Introduction

Urban planners and transportation engineers have examined and modeled street networks for decades to explore household travel behavior, accessibility and equity, urban form and design patterns, connectivity, and centrality [1–26]. Complex networks such as street networks have also been explored from the perspective of statistical physics to assess structure and performance [27–44]. However, large volumes of cross-sectional street network data—in a format well-suited for graph-theoretic analysis—have often been difficult to come by, especially in an open-source, scalable, automatable way.

This paper presents a new data repository to address this gap. It describes a research project that downloaded raw OpenStreetMap street network data for the entire U.S., cleaned these data, then constructed graph-theoretic models of these data at multiple scales for fast and rigorous urban analysis. It then calculated dozens of metric and topological measures of each network, nationwide. These models and measures can save researchers weeks or months of ad hoc data collection and analysis. However, they are even more useful for urban planners and policymakers who often lack the technical capacity to write their own API scripts in custom query languages or implement their own network science algorithms to understand the built form and urban circulation.

This new data repository provides four significant value-additions for urban science and analytics. First, the repository contains graph-theoretic models in common reusable formats, immediately suited
to spatial network analysis—unlike OpenStreetMap’s raw geometry data downloads—out of the box. Second, these models all have meaningful spatial extents (municipalities, counties, urbanized areas, census tracts, and neighborhoods) that correspond to administrative boundaries and socio-spatial units in useful ways for urban analysis. Third, these graphs’ topologies have all been substantially cleaned-up such that their nodes exclusively represent intersections and dead-ends, and their edges exclusively represent the street segments connecting them. Finally, this repository contains dozens of metric and topological measures calculated for each of the 110,000 study sites—no such database previously existed.

Building on recent work that developed the underlying data collection and modeling methodology [45], and presented initial empirical findings of geometric and topological trends in U.S. street network design [46], this paper documents a new undertaking of data collection, organization, and archiving to produce a newly-generated public repository of street network models and measures. This paper first briefly describes the background of street network models and data. Then it presents the methodology used to construct these street network graphs, analyze them, and produce the resulting models and indicators repositories. Finally, it concludes with a discussion of this work.

2. Background

2.1. Street Network Models

Graphs are mathematical models of networks [47–53]. A graph $G$ consists of a set of nodes $N$ connected to each other by a set of edges $E$. An edge connects one node to another, or one node to itself in the case of self-loops. In a directed graph, each edge points in a single direction from one node to another. For example, an edge $e_{uv}$ points from node $u$ to node $v$ but not reciprocally from $v$ to $u$. However, in an undirected graph, an edge points reciprocally in both directions between the nodes that it links. If a graph’s edges intersect in two-dimensions exclusively at nodes, it is planar; otherwise, it is nonplanar [54–68].

In a primal representation of a street network, nodes represent intersections and dead-ends, while edges represent the street segments that link them [69]. Dual representations invert this topology such that nodes represent streets and edges represent their intersections [70]. This project only considers primal graphs, as they account for the full spatial and geographic characteristics of the street network [71,72]. Walkable networks allow undirected flow in both directions along an edge, but many (drivable) street networks include one-way streets that necessitate directed graphs for accurate modeling. Many street networks include self-loops as well as parallel edges (i.e., when two nodes are linked by multiple edges). A multigraph is a graph that allows for such parallel edges. This project models street networks as primal nonplanar directed multigraphs with possible self-loops.

2.2. Street Network Data

Street network data around the world come from disparate sources, resulting in inconsistencies in digitization, metadata, file formats, availability, and acquisition. Traditionally, most studies of U.S. street networks used the U.S. census bureau’s TIGER/Line roads shapefiles (an official vector dataset of roads in the U.S.). While these shapefiles provide network geometry, they provide little information about topology. Street networks are embedded in geographic space, and thus are defined by both their spatial geometry and their topology: the connections and configurations of nodes and edges [73,74]. Thus, geographic shapefiles alone do not lend themselves immediately or easily to graph-theoretic network analysis. Moreover, TIGER/Line data contain substantial spatial inaccuracies and coarse-grained attribute data [75].

Online geographic information systems offer another source of data. However, “[w]hile proprietary systems such as Google Maps may be suitable for trip planning and similar applications, they cannot be used for most research and analytic purposes” [76] due to licensing terms, costs, access restrictions, and usage limitations. OpenStreetMap offers an alternative source of street network
data [77]. It is an open-source, worldwide, collaborative mapping project. OpenStreetMap provides geospatial information about streets and intersections, along with attribute data about road types, names, and (sometimes) speeds, widths, and numbers of lanes. However, its data cannot by default be automatically extracted into a graph-theoretic object for network analysis [78]. Furthermore, the network topology must be substantially cleaned to correctly represent nodes exclusively as intersections and dead-ends. Prior to this wider project, no tools or data repositories enabled the automatic, at-scale, configurable acquisition of OpenStreetMap data and construction of graph-theoretic data objects for analysis.

3. Methods

Given this background and motivation, this project originally developed a new open-source Python-based software toolkit called OSMnx [45]. OSMnx can download data from OpenStreetMap using configurable user queries, then construct a nonplanar, directed multigraph, and finally clean the topology [79]. Python was chosen to develop this tool for three reasons. First, Python is one of the most popular programming languages in the world, giving it a broad audience. Second, Python offers particularly simple and straightforward syntax, making it easy for newcomers to learn and lowering the scientific barriers to entry. Third, it has become a standard language for data science research and practice, with an extensive ecosystem of related packages for scientific, network, and geospatial analysis. Finally, other similar tools in this space include dodgr (an R tool for distance calculations on weighted directed graphs), shp2graph (an R tool for converting spatial networks into igraph objects) [80], pandana (a Python tool for network accessibility queries) [81], the Urban Network Analysis Toolbox plugin for ArcGIS and Rhino [82], and GIS2F2E (a Python tool that processes shapefiles into edge lists) [78]. However, none of these offer the end-to-end capabilities of OSMnx to download network data directly, build models, clean the topology, and conduct statistical analyses and simulations.

To construct this data repository, this project used OSMnx to download network data and construct graphs for the drivable street networks of every U.S. city/town, county, urbanized area, census tract, and Zillow-defined neighborhood (Zillow is a large online real estate database company that defines neighborhood boundaries in many cities and towns). It saved these graphs as shapefiles, GraphML files (a standard, open-source format for graph serialization), and node/edge lists. Finally, it analyzed these networks to assess the geometric, topological, and morphological characteristics of U.S. street networks and how they reflect various urban planning eras, transportation technologies, economic conditions, and design paradigms. These study sites are presented in Figure 1.

3.1. Graph Production

To produce the data in the repository, we loaded five publicly-available input datasets defining these study sites into OSMnx version 0.8.1 in May 2018:

1. U.S. census bureau 2017 places shapefiles for all 50 states. Census “places” comprise cities, towns, equivalent entities, and census-designated places. For this project, we discard the latter (small unincorporated communities) to retain and analyze every city and town ($N = 19,678$) in the U.S.
3. U.S. census bureau 2017 nationwide urban areas shapefile. Census “urban areas” comprise urbanized areas and urban clusters. For this project, we discard the latter (small agglomerations) to retain and analyze every urbanized area ($N = 497$) in the U.S. [83].
5. Zillow’s neighborhood boundaries ($N = 17,037$) shapefiles, per-state, March 2018 release [84–86]. Zillow is a prominent real estate database company, and their boundaries dataset covers large U.S. cities.
Figure 1. The dataset’s study sites’ coverage (black): 19,678 cities/towns, 17,037 Zillow-defined neighborhoods, 497 urbanized areas, 3233 counties, and 74,133 census tracts. Alaska and Hawaii not to scale.

One at a time, for each city, county, urbanized area, tract, and neighborhood boundary defined in the above shapefiles, we downloaded the drivable public street network within its boundaries from
OpenStreetMap using OSMnx. To acquire these raw data, OSMnx buffers each boundary polygon by 500 m then downloads the streets within this geometry, filtering them based on attribute data. It then constructs a nonplanar directed multigraph. In the case of one-way streets, a directed edge is added from the origin node to the destination node. However, for bidirectional streets, reciprocal directed edges are added in each direction between the two nodes.

Next, OSMnx cleans the graph’s topology to retain nodes only at intersections and dead-ends (detailed below). However, the full edge spatial geometry and length are retained in the cleaned graph. Then, it calculates node degrees and node types before truncating the graph to the original boundary polygon. This buffer/truncate workflow attenuates perimeter effects [87] and guarantees that true intersections are not incorrectly considered pseudo-nodes or dead-ends if an incident edge links to a node outside the boundary polygon. The final graph may be strongly connected, weakly connected, or neither. If it is not connected, OSMnx returns all connected components as a single graph object.

To clean the graph’s topology, OSMnx only retains nodes that represent the junction of multiple streets, as depicted in Figure 2. First, it identifies all non-intersection pseudo-nodes (i.e., all those that simply form an expansion graph). Next, it removes these pseudo-nodes while maintaining the true spatial geometry and attribute data of the street segment between the true intersection nodes. In strict mode, OSMnx considers two-way “intersections” to be topologically identical to a single street that bends around a curve. Conversely, to retain these intersections when the incident edges have different OpenStreetMap IDs, we can use OSMnx’s non-strict mode. This cleaning step is critical to this dataset, providing additional value beyond extraction of data from OpenStreetMap itself by producing models more suitable to urban design/morphology and transportation analysis by representing intersections/dead-ends as nodes and linear blocks’ sides as edges.

Figure 2. OSMnx street network topology cleaning. (Top-left) original graph (nodes in white, edges in gray). (Top-right) original graph (strict-mode pseudo-nodes highlighted in red). (Bottom-left) graph simplified in strict mode. (Bottom-right) graph simplified in non-strict mode.
Once we have constructed and cleaned the graph, we use OSMnx to save it to disk as node/edge lists (formatted as comma-separated values), as ESRI shapefiles to work with in GIS software, and as a GraphML file (an open-source, standard format for serializing graphs) to work with in common network analysis software packages such as NetworkX, Gephi, graph-tool, or igraph [88–90]. OSMnx’s saved shapefiles include separate node and edge layers. When saving shapefiles, OSMnx simplifies the network to an undirected representation but preserves one-way origin–destination directionality as edge attributes for subsequent GIS-based routing applications.

3.2. Graph Analysis

Finally, for this repository we calculate several metric and topological measures of each of these networks, common in the urban design, transportation planning, and network science disciplines. These measures include each network’s area (km$^2$), mean average neighborhood degree, mean average weighted neighborhood degree, average circuity, average clustering coefficient, average weighted clustering coefficient, average degree centrality, edge density (m/km$^2$), average edge length (m), total edge length (m), count of intersections, intersection density (per km$^2$), count of dead-ends, proportion of nodes that are dead-ends, count of three-way intersections, proportion of nodes that are three-way intersections, count of four-way intersections, proportion of nodes that are four-way intersections, average node degree $\bar{k}$, count of edges $m$, count of nodes $n$, node density (per km$^2$), maximum and minimum node PageRank (a measure of node importance based on the structure of incoming edges), proportion of edges that self-loop, street density (m/km$^2$), average street segment length (m), total street length (m), count of street segments, and average number of streets per node.

These measures are described in Table 1 and extended technical definitions and mathematical derivations can be found in [3,16,32,45,48,51,73,91–106]. Note that we distinguish between “edges” and physical “streets” in certain instances to not double-count bidirectional streets with reciprocal edges pointing in both directions. Dozens of additional graph-theoretic measures can be easily calculated by loading these GraphML files into OSMnx [45].

**Table 1.** Descriptions of network measures in the dataset.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td>land area of cities, counties, and urbanized areas; gross area of neighborhoods (km$^2$)</td>
</tr>
<tr>
<td>mean avg neighborhood degree</td>
<td>mean of all average neighborhood degrees in network</td>
</tr>
<tr>
<td>avg weighted neighborhood degree</td>
<td>mean degree of nodes in the neighborhood of each node, weighted by edge length</td>
</tr>
<tr>
<td>mean avg weighted neighborhood degree</td>
<td>mean of all weighted average neighborhood degrees in network</td>
</tr>
<tr>
<td>avg circuity</td>
<td>total edge length divided by sum of great circle distances between the nodes incident to each edge</td>
</tr>
<tr>
<td>avg clustering coefficient</td>
<td>mean of clustering coefficients (extent to which node’s neighborhood forms a complete graph) of all nodes</td>
</tr>
<tr>
<td>avg weighted clustering coefficient</td>
<td>mean of weighted clustering coefficients of all nodes</td>
</tr>
<tr>
<td>avg degree centrality</td>
<td>mean of all degree centralities in network</td>
</tr>
<tr>
<td>avg edge length</td>
<td>mean edge length in network (m)</td>
</tr>
<tr>
<td>total edge length</td>
<td>sum of edge lengths in network (m)</td>
</tr>
<tr>
<td>edge density</td>
<td>total edge length divided by area</td>
</tr>
<tr>
<td>intersection count</td>
<td>number of intersections in network</td>
</tr>
<tr>
<td>intersection density</td>
<td>intersection count divided by area</td>
</tr>
<tr>
<td>dead-end count</td>
<td>number of dead-end nodes in the network</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>dead-end proportion</td>
<td>proportion of nodes that are dead-ends</td>
</tr>
<tr>
<td>three-way intersection count</td>
<td>number of three-way intersection in the network</td>
</tr>
<tr>
<td>three-way intersection proportion</td>
<td>proportion of nodes that are three-way intersections</td>
</tr>
<tr>
<td>four-way intersection count</td>
<td>number of four-way intersection in the network</td>
</tr>
<tr>
<td>four-way intersection proportion</td>
<td>proportion of nodes that are four-way intersections</td>
</tr>
<tr>
<td>( n )</td>
<td>number of nodes in network</td>
</tr>
<tr>
<td>( m )</td>
<td>number of edges in network</td>
</tr>
<tr>
<td>avg node degree ((\bar{k}))</td>
<td>mean number of inbound and outbound edges incident to the nodes</td>
</tr>
<tr>
<td>node density</td>
<td>( n ) divided by area in square kilometers</td>
</tr>
<tr>
<td>maximum PageRank</td>
<td>highest PageRank (ranking of nodes based on structure of incoming edges) value of any node in the graph</td>
</tr>
<tr>
<td>minimum PageRank</td>
<td>lowest PageRank value of any node in the graph</td>
</tr>
<tr>
<td>self-loop proportion</td>
<td>proportion of edges that have a single incident node (i.e., edge ( e_{uv} ) where ( u = v ))</td>
</tr>
<tr>
<td>street density</td>
<td>total street length divided by area in square kilometers</td>
</tr>
<tr>
<td>avg street segment length</td>
<td>mean edge length in undirected representation of network (m)</td>
</tr>
<tr>
<td>total street length</td>
<td>sum of edge lengths in undirected representation of network (m)</td>
</tr>
<tr>
<td>count of street segments</td>
<td>number of edges in undirected representation of network</td>
</tr>
<tr>
<td>avg streets per node</td>
<td>mean number of physical streets that emanate from each node (intersections and dead-ends)</td>
</tr>
</tbody>
</table>

3.3. Data and Code Availability

These data are freely available online at the Harvard Dataverse at https://dataverse.harvard.edu/osmnx-street-networks. This repository includes the graphs of the street networks for every U.S. city, county, urbanized area, census tract, and Zillow neighborhood as a GraphML file, node/edge shapefiles, and node/edge lists. These files can be loaded into any standard GIS software or network analysis package. It also contains the analytical measures for every street network. The datasets are available in the following two repositories:


All the code to download and analyze these street networks with OSMnx (version 0.8.1) is open-source and available on GitHub at https://github.com/gboeing/dataverse-street-networks. The OSMnx software itself is also open-source and freely available for download and installation from GitHub (https://github.com/gboeing/osmnx), the PyPI package repository (https://pypi.org/project/OSMnx/), and the Anaconda package repository (https://anaconda.org/conda-forge/osmnx).

4. Discussion

This paper described a new urban science data repository constructed from raw OpenStreetMap data that provides four significant value-additions. First, the repository contains graph-theoretic models in common reusable formats, immediately more useful than typical raw geometry data downloads out of the box. Second, these models all have meaningful spatial extents (municipalities,
counties, urbanized areas, census tracts, and neighborhoods) that correspond to administrative boundaries or social units for urban analysis and simulation. Third, these graphs have substantially cleaned-up topologies such that nodes exclusively represent intersections and dead-ends and edges represent the street segments connecting them. Fourth, this repository contains dozens of metric and topological measures calculated for each graph—no such database previously existed. These measures can be used to analyze the urban fabric’s texture and walkability (via intersection density, node/edge density, and average street segment length), connectivity (via average number of streets per node and circuitry), and network resilience and importance (via centrality measures and PageRank).

In total, this data repository contains over 110,000 street networks which in turn collectively comprise over 55 million nodes and over 137 million edges. While these data fill a gap in helping researchers quickly ramp-up graph-theoretic street network analyses in the U.S. without having to spend weeks developing their own ad hoc data collection, modeling, and analysis workflows, they fill a larger gap in opening up these scientific modes of urban analysis to planners and policymakers without the individual technical or institutional capacity to do so otherwise. Yet to use these data accordingly, they must reasonably model the real world. Validation of these data can be considered from two perspectives. The first considers how faithfully the OpenStreetMap data represent the real-world street network. The second considers how faithfully the repository’s graphs in turn represent the OpenStreetMap street network.

Regarding the former perspective, various authors have explored this subject in detail [76,107–114]. OpenStreetMap’s road data quality is generally quite high—for example, Garmin consumer GPS devices can use OpenStreetMap roads data for navigation. Although data coverage varies worldwide, it is generally good when compared to corresponding estimates from the CIA World Factbook. In the US, OpenStreetMap imported the 2005 TIGER/Line roads in 2007 as a foundational data source. Since then, numerous corrections and improvements have been made. More importantly, many additions have been made beyond what TIGER/Line captures, including richer attribute data describing the characteristics of features and finer-grained codes for classifying streets. Of course, much of these data are crowd-sourced and user-generated, and errors thus occasionally exist. However, the data are validated and vetted by the OpenStreetMap community, resulting in high quality overall. Most relevant to this project’s study area, the U.S. network is essentially complete on OpenStreetMap [76].

Regarding the latter perspective, we comprehensively tested the final dataset’s quality, adapting the methodologies of [76,78] against three reference datasets: the TIGER/Line roads, Google Earth satellite imagery, and the OpenStreetMap raw source data. The first step uses QGIS to compare these data spatially against the TIGER/Line roads shapefile in a random sample of 100 cities to identify any edges appearing in one of these datasets but missing in the other. Discrepancies were then checked one at a time to ensure they correctly matched the OpenStreetMap source and secondarily against Google Earth satellite imagery for real-world verification. Finally, tests are performed on each graph to ensure they can be loaded, analyzed, and routed. The validation confirmed that this project’s algorithms reconstructed the OSM data properly with nodes exclusively at intersections and dead-ends.

Comprehensive documentation for using OSMnx is available at https://osmnx.readthedocs.org and tutorials, examples, and demonstrations are available at https://github.com/gboeing/osmnx-examples. To reuse this dataset, researchers can install OSMnx according to the installation instructions in the documentation. Then, they can load the GraphML files using OSMnx’s load_graphml function. The graphs may be similarly loaded in NetworkX, graph-tool, igraph, and other similar network analysis tools. To load these graphs with Gephi, first load the GraphML file in OSMnx, re-export it using OSMnx’s save_graphml function with argument gephi=True (to add additional customization for Gephi compatibility), then open the exported file in Gephi. The shapefiles may be loaded in standard fashion in any GIS software, such as QGIS, ArcGIS, or geopandas.
Funding: This research received no specific external funding.

Acknowledgments: The author wishes to thank Paul Waddell, Robert Cervero, David O’Sullivan, Elizabeth Macdonald, and Luis Bettencourt for their helpful comments and suggestions during the development of this project.

Conflicts of Interest: The author declares no conflict of interest.

References


69. Law, S. Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach. *Cities* 2017, 60, 166–179. [CrossRef]
78. Barrington-Leigh, C.; Millard-Ball, A. The world’s user-generated road map is more than 80% complete. *PLoS ONE* 2017, 12, e0180698. [CrossRef] [PubMed]

80. Lu, B.; Sun, H.; Harris, P.; Xu, M.; Charlton, M. Shp2graph: Tools to Convert a Spatial Network into an Igraph Graph in R. ISPRS Int. J. Geo-Inf. 2018, 7, 293. [CrossRef]


95. Jiang, B.; Claramunt, C. Integration of space syntax into GIS: New perspectives for urban morphology. Trans. GIS 2002, 6, 295–309. [CrossRef]


© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).