SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Environmental sciences

» Geography

» Civil engineering

# Data Descriptor: A protocol to convert spatial polyline data to network formats and applications to world urban road networks

Alireza Karduni[1,2], Amirhassan Kermanshah[1] & Sybil Derrible[1]

Received: 25 January 2016
Accepted: 12 May 2016
Published: 21 June 2016

The study of geographical systems as graphs, and networks has gained significant momentum in the academic literature as these systems possess measurable and relevant network properties. Crowd-based sources of data such as OpenStreetMaps (OSM) have created a wealth of worldwide geographic information including on transportation systems (e.g., road networks). In this work, we offer a Geographic Information Systems (GIS) protocol to transfer polyline data into a workable network format in the form of; a node layer, an edge layer, and a list of nodes/edges with relevant geographic information (e.g., length). Moreover, we have developed an ArcGIS tool to perform this protocol on OSM data, which we have applied to 80 urban areas in the world and made the results freely available. The tool accounts for crossover roads such as ramps and bridges. A separate tool is also made available for planar data and can be applied to any line features in ArcGIS.

| | |
|---|---|
| **Design Type(s)** | data integration objective • observation design |
| **Measurement Type(s)** | network graph construction |
| **Technology Type(s)** | digital curation |
| **Factor Type(s)** | |
| **Sample Characteristic(s)** | city • Tokyo • Jakarta • Seoul • Delhi • Shanghai Proper • City of Manila • Karachi • New York City • Sao Paulo • Beijng Proper • Mumbai • Guangzhou City Prefecture • Moscow Federal City • City of Los Angeles • Kolkata • Dhaka • Buenos Aires • Istanbul • Rio de Janeiro • Shenzhen City Prefecture • Commune of Paris • Lima • City of Chicago • Tianjin Proper • Chennai • Bogota • Bangalore • London • Taipei City • Ho Chi Minh City • Dongguan City Prefecture • Hyderabad • Chengdu City Prefecture • Lahore • Johannesburg • Tehran • Bangkok • Wuhan • Ahmedabad • Chongqing Proper • Baghdad • Hangzhou City Prefecture • Province of Santiago de Chile • City of Fort Worth • City of San Francisco • Quanzhou City Prefecture • City of Miami • Shenyang City Prefecture • Belo Horizonte • City of Philadelphia • Nanjing City Prefecture • Madrid • City of Houston • Xianyang City Prefecture • Milan • Pune • Saint Petersburg • City of Atlanta • Surat • City of Washington • Bandung • Municipality of Surabaya • Harbin City Prefecture • City of Boston • Zhengzhou City Prefecture • Qingdao City Prefecture • Abidjan • Barcelona • Ankara • Suzhou City Prefecture • City of Phoenix • Salvador • Municipality of Porto Alegre • Roma • Recife • Province of Naples • City of Detroit • Dalian City Prefecture • Fuzhou City Prefecture • Medellin Metropolitan Area |

[1]Complex and Sustainable Urban Networks (CSUN) Lab, Department of Civil and Materials Engineering, University of Illinois at Chicago, Chicago, Illinois 60607, USA. [2]School of Architecture, University of North Carolina at Charlotte, Charlotte, North Carolina 28223, USA. Correspondence and requests for materials should be addressed to A.Karduni (email: akarduni@uncc.edu) or to A.Kermanshah (email: akerma2@uic.edu).

## Background & Summary

From rivers, roads and pipelines to electrical and telecommunication lines, many geographical systems are composed of collections of elements that are connected in space. Therefore, from studying and understanding human activities[1] to studying the relationship between different urban areas[2], conducting these studies through the lens of graphs and networks can shed light on many complex characteristics of the elements that are built upon them[3,4]. This approach is rooted in the origins of the field of Graph Theory developed in the 18th century by Euler and his Seven Bridges of Königsberg[5], and it has been applied widely ever since[6–13]. The use of graphs is further reinforced in this era of 'Big Data', where countless sources of organizational or crowd-sourced data such as locational social media (Twitter geolocations, Facebook check-ins) and geographical data (road networks and public transportation systems through OpenStreetMaps (OSM) (http://www.openstreetmap.org), and businesses location data through Yelp and Google) have become available[14].

Nevertheless, when focusing on geographical networks, and specifically on road systems, raw datasets often contain issues that make them unsuitable to conduct proper graph studies. Specifically, there are two main problems; first achieving a topologically correct dataset that represents the actual status of the street network as accurately as possible (topology problem), and second is developing a graph file format that is ready to be analyzed with available software and libraries (file format problem). In direct response to these problems, the main objective of this work is threefold; first to offer a formal protocol to convert Geographic Information Systems (GIS) data into a workable network format; second, to develop a tool to apply this protocol; and third to use the tool on a significant number of road systems and make the results available. For the purpose of this study we applied the tool to the road systems of some 80 most populated urban areas in the world (using available OSM data), and hope to supplement this database with the road networks of more cities in the future.

In the following section, we describe the existing tools and datasets that focus on network analysis of road networks followed by our methodology where, we offer a description of the protocol to effectively address the two problems mentioned earlier (topology and file format). Moreover, we offer a summary of the dataset that we have made publicly available. This is followed by a discussion of common issues that researchers might encounter when utilizing these datasets. Finally, we discuss how researchers can produce this data for any road network they might be interested in. The tool can be used with the commercial software package ArcGIS. The OSM tool takes into account roads that cross but do not intersect such as bridges and ramps when the information is present already in OSM. We have made a second version of the tool available that transforms any line feature that does not include elevation and intersection information (e.g., pipes, rivers, rails) into a spatial planar graph (e.g., a graph with no intersecting edges[15]). The latest version of the tools can be found on the Data & Tools page of the Complex and Sustainable Urban Networks (CSUN) Laboratory at http://csun.uic.edu/data.html#GISF2E (Accessed Jan 19 2016). Moreover, permanent versions of the tools along with the results for the 80 cities are permanently stored on Figshare (Data Citation 1, Data Citation 2).

There are several existing tools and software that enable researchers to conduct network and graph analyses. ArcGIS Network Analyst and QGIS Network Analysis Library are two popular toolsets, both of which create network datasets from road network files easily. However, the tools only allow users to conduct certain studies, such as shortest path calculations from a series of points to any other points, similar to origin destination matrices. Yet they do not provide a method to measure the whole system through a graph analysis and to calculate various graph metrics such as betweenness and closeness centralities[16]. Although ArcGIS Network Analyst allows some degrees of topology correction within the software's ecosystem, there is no straightforward method to convert the network datasets to a workable graph format such as an edge list (i.e., list of edges/links) or an adjacency matrix (i.e., square matrix of all nodes, containing 0 or 1 s when two nodes are connected).

DepthMapX (https://varoudis.github.io/depthmapX/) which comes in the form of standalone software, as well as a plugin for QGIS, allows the user to calculate various network metrics for road systems, but only works for a certain type of graph analysis, Space Syntax, as developed by Bill Hillier[17]. DepthMapX works with axial maps, which are a specific type of spatial graphs[18] as opposed to regular road maps, and takes the input data in many formats including AutoCAD (DWG format).

In contrast to the lack tools to convert a GIS line feature into a network, there are an abundance of libraries and software packages to conduct graph analyses, all varying in how much expertise is required to run them. Gephi[19] is a graph analysis software with a simple and intuitive graphical user interface. NodeXL[20] enables users to conduct graph analysis from Microsoft Excel. NetworkX[21] and igraph[22] are libraries for python that enable users to conduct graph analyses with minimal programming background. All of the mentioned libraries and software packages can input a series of standard graph file format such as an edge list and an adjacency matrix as described above.

In regards to the road system graph data, there are some datasets available, to name a few, the Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/#road) consists of many ready graph datasets, which include road graph files for three American states. As well, the school of computing at the University of Utah also has a series of graph files for roads networks available as edge lists[23]. However, none of these datasets can be imported back in a GIS environment, and no information could be found on how topologically correct they are.

Our toolset and dataset bridge the gap between semi-enclosed ecosystems such as ArcGIS and QGIS, and graph analysis libraries such as Gephi and igraph. This is achieved by providing both shapefiles/feature classes and network edge lists that are connected to each other with unique identifiers. Our dataset enables GIS users to easily conduct graph analyses for road systems of the 80 most populated urban areas in the world, by providing accurate data that can be easily inputted into the various graph analysis libraries listed above. The results can then be imported into a GIS environment to conduct geographic analyses and visualizations[24]. The provided toolset will enable users to create topologically correct graph edge lists from OpenStreetap (OSM), and planar graph edge lists from any road network shapefile that lacks the required information. The toolset can in fact process any line features, from roads and rail systems, to water conduits, electrical systems and even rivers.

## Methods

In a reverse engineering fashion we first note how we want our final results to be, which drives the entire procedure. Essentially, we need our network information to consist of a data set of nodes as intersections and a data set of edges/links as road segments that connect the nodes. Moreover, we need to calculate the length of edges in our dataset in a way that takes into account the curvature of the road segments. In addition we need to ensure, given the existence of required road information, any non-intersection such as highway over passes, bridges, or tunnels that overlap but that do not intersect are not counted as nodes in our graph system (Fig. 1).

To solve this problem, we establish the following five-step protocol:

1. Separate the data set into different road crossing categories based on OSM highways tags: (a) bridge and (b) tunnel
2. For each category:
   a Split the lines at their intersection
   b Create nodes at the start and end point of each split line
3. Merge the edges together
4. Merge the nodes together
5. Remove duplicate nodes

As mentioned before, our tool utilizes OSM data as the data source. The line data in OSM includes many crowd-sourced attribute tags, such as street type, name and so on. The key attribute information we have used to create our graph files are two specific OSM highway tags:

a bridge tag (0 or 1), which is defined as any road that crosses over another (See http://wiki.openstreetmap.org/wiki/Key:bridge);
b tunnel (0 or 1) which is defined as 'any underground path for a road or similar' (See http://wiki.openstreetmap.org/wiki/Key:tunnel). The required data format for our tool to perform is as follows: 'UniqueID', 'bridge' (1 for true and 0 for false), 'tunnel' (1 for true and 0 for false).

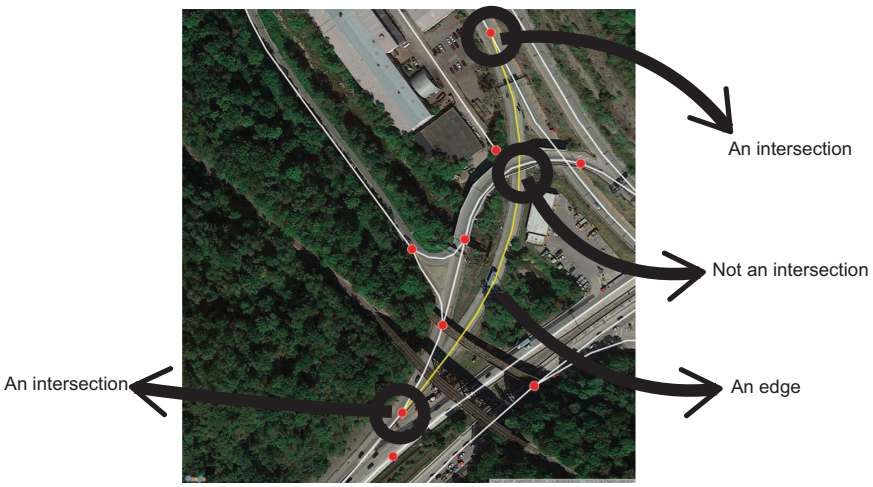


**Figure 1.** Superposing a generated graph on top of the real image of the area to define 3D aspects of the graph (i.e., define nodes and edges).

Our method solves two major topological problems with OSM line data. First it produces nodes and edges, only at locations where road intersections exist, which result in a non-planar graph file whereas if edges do not spatially intersect their line intersections are not considered as nodes (Fig. 1). This finding is significant because most line features consist of multiple segments between the intersections of two lines (i.e., many roads are artificially split in multiple segments). Most network analysis tools produce nodes at the start and end of all of these segments; inflating the actual number of nodes and edges, and reducing the length of most road segments. As illustrated in Fig. 2, our toolset, produces no extra nodes or edges, and enables an accurate calculation of edge length using ArcGIS which automatically takes into account the curvature of the road segments (e.g., in meters).

The data for the 80 most populated cities in the world are based on world atlas website (Table 1), and were collected from the OSM database during first two weeks of June 2015 (For a map of all the cities in the dataset see: https://goo.gl/DB576i). Table 1 lists the data sources used to create the data set. To build the road networks of the cities, we used boundary shapefiles to clip the road network from OSM data sets. We projected our data set using WGS 1984 UTM projected coordinate system (See http://geokov.com/education/utm.aspx) which enables our tool to calculate edge length in meters.

### Code availability

The dataset of 80 road networks from the most populated cities in the world was accessed from OpenStreetMaps and processed with our tool created in ArcGIS 10.2 Model Builder with an advanced (ArcInfo) license. The planar version of the tool was created with the same software package. Both tools, with a thorough tutorial on how they can be used, have been made publically available as an ArcGIS toolbox in CSUN's website (http://csun.uic.edu/data.html#GISF2E) and on Figshare (Data Citation 1). With the required ArcGIS licenses, both tools can be used to create new datasets. Moreover, the tools can be freely downloaded, modified and improved to fit future research needs.

## Data Records

In order to create the final datasets (Data Citation 2), we created an ArcGIS tool (Data Citation 1) and utilized it to create a dataset of 80 road network shapefiles and edge lists. Essentially, our tool creates two new GIS layers, one with all nodes and one with all edges as well as an edge list in a Comma-Separated Values (CSV) file. An edge list is a list of all edges/links in the network with start node ID, end node ID, and edge ID. These unique ids correspond to the points and lines in the generated GIS files, and can be later converted back to any GIS platform to conduct analysis or spatial visualization. More specifically, an edge list is a standard method of graph representation and can be read by many graph analysis software packages or libraries (e.g., Gephi, NodeXL, or python's igraph).

The datasets are released in the Figshare account, handled by the Complex and Sustainable Urban Networks (CSUN) Laboratory at the University of Illinois at Chicago. For each city, the data consist of two shapefiles, one for nodes and one for edges (Fig. 3), and one edge list (e.g., Table 2) for each network (e.g., Boston_Nodes.shp; Boston_Links.shp; Boston _Edgelist.csv).

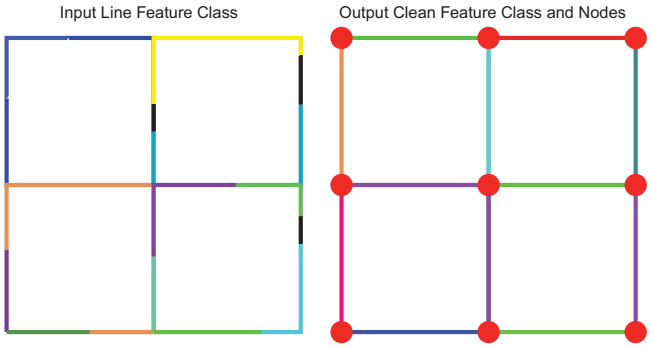


**Figure 2.** Cleaning up the OSM input data followed by creating the edges and nodes feature classes.

| Data | Websites where data were collected | Accessed Date |
|---|---|---|
| Cities information | http://www.worldatlas.com/citypops.htm | 1–14 Jun 2015 |
| Road shapefiles | http://download.geofabrik.de/ | 1–14 Jun 2015 |
| Boundaries | http://www.gadm.org/country | 1-14 Jun 2015 |

**Table 1.** Sources of raw data acquisition.

## Technical Validation

The original roads data sets, as well as the tunnel and bridge information, along with the positional accuracy of the data, are validated by the OSM community (see http://wiki.openstreetmap.org/wiki/Accuracy). The resulting crowdsourced datasets have varying levels of positional and geographic accuracy depending on the location of the data[25]. In many cases the quality of the data has been constantly improving[26]. The data generated with our tool is first cleaned, noting that all duplicate nodes and edges are handled within the tool, then the data integrity and cleanness checked by running it through the topology validation tool provided in ArcGIS. The topology tool checks for overlapping edges, nodes, or edges that are not connected to a node. Finally, the edge lists were tested by conducting simple graph analyses and by joining the data back to GIS shapefiles. In other words, the data is accurate if all of the nodes and edges present in the CSV file generated correspond to actual nodes and edges present in the two GIS layers.

## Usage Notes

The main objective of this article is to present a protocol to convert any line feature data in GIS into a workable network format, consisting of a list of edges, a node layer, and an edge layer. This protocol was created as an ArcGIS tool and applied to the 80 most populated urban areas in the world using OSM data (where overlapping lines were taken into account). Another tool was made available for planar analyses where no edges intersect, or additional intersection data is not available.

Overall enormous amounts of GIS data sets are becoming increasingly available. Simultaneously, the study of spatial systems as graphs and networks has emerged as a substantial field in the research community. The protocol introduced along with the ArcGIS tool and the data made available for the 80 urban areas further strengthen these recent advances, and can lead to the study of more spatial systems as networks, from rivers and water pipelines, to telecommunication and electrical networks.

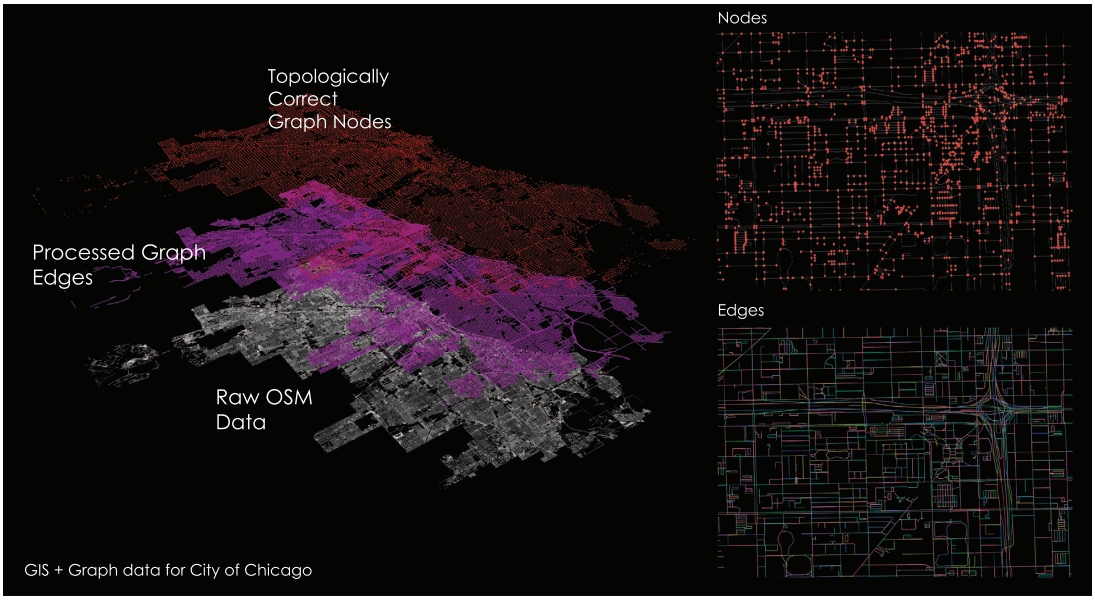


**Figure 3.** Schematic outline of the entire process to achieve nodes, edges and edge lists for a sample city like Chicago.

| XCoord | YCoord | START_NODE | END_NODE | EDGE | LENGTH |
|---|---|---|---|---|---|
| 329880.2 | 4594299 | 1 | 2 | 1 | 14.64321 |
| 329880.2 | 4594299 | 1 | 3 | 2 | 24.43203 |
| 329880.2 | 4594299 | 1 | 6 | 3 | 272.6202 |
| 329874.6 | 4594286 | 2 | 1 | 1 | 14.64321 |
| 329904.6 | 4594298 | 3 | 1 | 2 | 24.43203 |
| 329793 | 4594549 | 6 | 1 | 3 | 272.6202 |
| 329793 | 4594549 | 6 | 7 | 5 | 82.32773 |

**Table 2.** Sample edge list for small part of Boston's road network.

The GIS protocol consists of three major steps:

1. Inputting lines, cleaning up the data and creating nodes Feature Class.
2. Generating edge list data.
3. Cleaning up the results, and generating the CSV edge list output.

There is a comprehensive tutorial for using this GIS tool available at http://csun.uic.edu/data.html#GISF2E (Accessed 19 Jan 2016) and on Figshare that explains all the process in detail (Data Citation 1). The tools are available as an ArcGIS Toolbox which allows users and researchers to modify the toolbox to fit their data needs.

## References


1. Salingaros, N. A. Theory of the urban web. *J. Urban Des* **3,** 53–71 (1998).
2. Zhong, C., Arisona, S. M., Huang, X., Batty, M. & Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.* **28,** 2178–2199 (2014).
3. Kermanshah, A., Karduni, A., Peiravian, F. & Derrible, S. Impact analysis of extreme events on flows in spatial networks. in 29–34 (IEEE, 2014)doi:10.1109/BigData.2014.7004428.
4. Peiravian, F., Kermanshah, A. & Derrible, S. Spatial data analysis of complex urban systems. in 54–59 (IEEE, 2014)doi:10.1109/BigData.2014.7004432.
5. Euler, L. Solutio problematis ad geometriam situs pertinentis. *Comment. Acad. Sci. Petropolitanae* **8,** 128–140 (1741).
6. Xie, F. & Levinson, D. Measuring the structure of road networks. *Geogr. Anal.* **39,** 336–356 (2007).
7. Perret, J., Gribaudi, M. & Barthelemy, M. Roads and cities of 18th century France. *Sci. Data* **2,** 150048 (2015).
8. Barthélemy, M. Spatial networks. *Phys. Rep* **499,** 1–101 (2011).
9. Derrible, S. & Kennedy, C. Applications of graph theory and network science to transit network design. *Transp. Rev.* **31,** 495–519 (2011).
10. Derrible, S. & Kennedy, C. Network Analysis of World Subway Systems Using Updated Graph Theory. *Transp. Res. Rec. J. Transp. Res. Board* **2112,** 17–25 (2009).
11. Derrible, S. & Kennedy, C. Characterizing metro networks: state, form, and structure. *Transportation* **37,** 275–297 (2010).
12. Derrible, S. & Kennedy, C. The complexity and robustness of metro networks. *Phys. Stat. Mech. Its Appl* **389,** 3678–3691 (2010).
13. Derrible, S. Network Centrality of Metro Systems. *PLoS ONE* **7,** e40575 (2012).
14. Cottrill, C. D. & Derrible, S. Leveraging Big Data for the Development of Transport Sustainability Indicators. *J. Urban Technol.* **22,** 45–64 (2015).
15. Batty, M. *The new science of cities*. (Mit Press, 2013).
16. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Netw* **1,** 215–239 (1979).
17. Hillier, B. *Space is the machine: a configurational theory of architecture* (Space Syntax: London, 2007).
18. Desyllas, J. & Duxbury, E. Axial maps and visibility graph analysis. in *Proceedings, 3rd International Space Syntax Symposium* 27–1 (2001).
19. Bastian, M., Heymann, S. & Jacomy, M. & others. Gephi: an open source software for exploring and manipulating networks. *ICWSM* **8,** 361–362 (2009).
20. Hansen, D., Shneiderman, B. & Smith, M. A. *Analyzing social media networks with NodeXL: Insights from a connected world* (Morgan Kaufmann, 2010).
21. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. in *Proceedings of the 7th Python in Science Conference (SciPy2008)* 11–15 (2008).
22. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst* **1695,** 1–9 (2006).
23. Li, F., Cheng, D., Hadjieleftheriou, M., Kollios, G. & Teng, S.-H. in *Advances in Spatial and Temporal Databases*. 273–290 (Springer, 2005).
24. Kermanshah, A. & Derrible, S. A geographical and multi-criteria vulnerability assessment of transportation networks against extreme earthquakes. *Reliab. Eng. Syst. Saf.* **153,** 39–49 (2016).
25. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Des* **37,** 682–703 (2010).
26. Neis, P. & Zielstra, D. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet* **6,** 76–106 (2014).


## Data Citations


1. Karduni, A., Kermanshah, A. & Derrible, S. *Figshare* https://dx.doi.org/10.6084/m9.figshare.2065320.v1 (2016).
2. Karduni, A., Kermanshah, A. & Derrible, S. *Figshare* https://dx.doi.org/10.6084/m9.figshare.2061897.v1 (2016).


## Acknowledgements

The authors would like to thank our colleagues Eduardo Schaefer Sombrio and Sara Ellis for their great help in gathering the data for this article and revising and editing our manuscript.

## Author Contributions

Conceived and designed the tool: A.Kar, A.Ker and S.D. Gathering the data: A.Kar, A.Ker and S.D. Preparing the data set: A.Kar, A.Ker and S.D. Wrote the paper: A.Kar, A.Ker and S.D.

## Additional Information

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Karduni, A. *et al.* A protocol to convert spatial polyline data to network formats and applications to world urban road networks. *Sci. Data* 3:160046 doi: 10.1038/sdata.2016.46 (2016).