Generation of a Weighted Network Graph based-on Hybrid Spatial Data

Markus Prossegger
Dept. of Network- and Communication Engineering
Carinthia University of Applied Sciences
Klagenfurt, Austria
m.prossegger@cuas.at

***Abstract*—State-of-the-art network simulation and optimization techniques rank among the most studied problems in the field of operations research. While the mathematical models are studied in detail and nearly each network optimization problem has its already known solution in form of an optimal or heuristic algorithm, the underlying spatial data are the one key factor with respect to the optimization results. This paper examines the generation of weighted network graphs based on heterogeneous spatial data. Based on a general format, the normalized geobasisdata, an initial graph, is constructed. This graph is then used as input into our rule-based system to select and weight the edges to be in the final graph. The successful reduction of the complexity of the generated graph is shown in the experiments.**

Keywords*—geographic information; spatial data; mathematical optimization; simulation; network construction; graph theory.

I. INTRODUCTION

Mathematical models of state-of-the-art network optimization and simulation techniques are based on network graphs consisting of vertices (i.e., points-of-interest) and pair wise joining edges between them. This work focuses on the missing link between the real world and the mathematical modeling - the weighted network graph based on spatial data. The graph is generated using spatial polygon data describing the land use of an area on the one hand, and spatial line- and point-objects, describing existing infrastructure on the other hand. Based on this spatial data originating from a number of hybrid sources, a rule-based expert system is used to construct a network graph as vital input in subsequent mathematical models. We focus on optimization models within the scope of telecommunication network construction. The models are intended to minimize network construction costs (including underground work and cable laying costs) as well as to maximize the number of customers that can be connected to the communication network infrastructure. An instance of such a model using weighted graphs is the simulation and optimization engine of fiber optic communication networks described in [1]. Here, the land use polygons are used to generate a graph originating from a cost raster, which describes the underground construction costs. While using a cost raster is a feasible way to generate a network graph representing network construction costs, a more sophisticated approach is needed to generate graphs by taking into consideration all kinds of real world information and being able to be computed in a reasonable time.

The present paper is divided into a preliminaries section, the section dedicated to the definition of normalized geobasisdata, details of our approach and experimental results followed by the conclusion.

II. PRELIMINARIES

The subsequent simulation and optimization algorithms require undirected graphs as the fundamental data structure. The graph $G = (V, E, d)$ consists of $n = |V|$ vertices and $m = |E|$ edges. The distance of an edge $e_{ij} \in E$ connecting the two vertices $i \in V$ and $j \in V$ is given as a cost function $d_{ij} : E \to \mathbb{R}$. When calculating the shortest path or the minimum Steiner Tree [2], [3] as typical routing problems, the distance d can be the Euclidean distance. The more sophisticated algorithms use travel time as the distance between two vertices. In case of scenarios considering the build-up of wired networks, the distances have to be construction costs, such as underground work or the costs for building cable poles.

The two-dimensional geographic data originate from hybrid sources, thus these data need to be prepared to serve as the basis for the construction of network graphs. There are three main sources for the geographical data:

(a) The geographic information system (GIS).
(b) The network information system (NIS).
(c) The Austrian digital cadastral map (DKM).

Each of the above listed items is needed for the construction of a consistent data source.

A. GIS

In our case, the geographic information system includes typical information used in marketing scenarios and strategic decisions. It is important to know where the potential customers are located. The data showing the population density or the number of households collected in a population census are incorporated in the GIS as well. The GIS contains statistical data aggregated from public sources together with information gathered by the prosecuting company itself. The most important information for a network construction company is the information about the location of potential customers, the expected benefit, and the classification as private or public.

 ISBN: 978-1-61208-251-6

The network operator knows the exact location and the return on investment of the current customers, but not for potential customers. The marketing division uses market surveys and other statistical data to predict the location of potential customers and the likely yearly sales.

B. NIS

The network information system contains the information regarding all hardware components of the communication network as well as all logical links between these components. Typically, the NIS contains the most important business secrets of a network operating company. The following gives a list of the typical content of a NIS:

- Current and former customers.
- Network components.
- Physical cabling plan.
- Logical cabling plan.

C. DKM

The Austrian Digital Cadastral Map is part of the official boundaries cadastre, which is the binding evidence of all parcel's boundaries. The DKM contains all public and private property and is available nationwide. Furthermore, it documents the type of land use of each parcel as well as buildings. Similar information is held in layers and together they form the DKM (a comprehensive interface description can be found in [4]):

- Boundaries.
- Parcel numbers.
- Types of land use (building land, forest, running water, standing water, etc.).
- Buildings.
- Control and Boundary points.

Formerly available only as an analog hard copy, the DKM was not only digitized but also enhanced using other official sources like Orthophotos and partition plans. Due to this reason, the quality of the digital map exceeds the quality of the analog version but it may include historical failures as well. There is a list of papers describing the aspects of spatial data quality [5], [6], [7], [8] as well as an ISO Standard regarding the quality of spatial data [9].

While the spatial accuracy is acceptable in most of mathematical simulation and optimization scenarios, the topological quality of the input data has to be ensured. There is some work proposed to identify spatial inconsistencies and incorrect object classifications using either manually defined spatial integrity constraints [10], [11], [12] or an automatic and incremental approach using decision trees proposed in [13] and improved in [14].

The Austrian DKM is used as one of the basic input into our approach and has to be normalized together with the other spatial input data.

III. NORMALIZED GEOBASISDATA

The subsequent graph generation is designed as a completely automated process without the need of any user interaction. Due to this fact, the input data are stored in a predefined digital map format and the spatial objects must meet a set of conditions. In our approach, we have decided to use the ESRI Shapefile [15] as the digital map format. This open format is widely used and supported and stores spatial geometry and attributes as elements representing points, lines, and polygons.

The Normalized Geobasisdata (NGB) format [16] is an add-on to the ESRI Shapefile specifying the minimum qualitative and logical requirements of the spatial objects. It was developed in order to allow the automated generation of weighted network graphs based on any two-dimensional spatial data that represent surface data (i.e., land uses) in form of polygons at least. If the hybrid spatial data fulfill the specified NGB format, they qualify as input into the graph generation process.

The NGB format was originally developed for a simulation model dealing with the layout planning of a fiber-optics communication network in the year 2009. Since then it has been continuously adapted to the specific requirements of individual projects.

The majority of mathematical simulation and optimization models dealing with cable infrastructure planning or routing in general rely on spatial data as the main input source to stay real world compatible. We have defined the following list of objects as the required spatial information to be covered in the NGB format:

TABLE I
NGB OBJECT TYPES AND THEIR SPATIAL REPRESENTATION.

Id	Object class	Spatial representation
a	Project area	Polygon
b	Land use	Polygon
c	Usable (own or third-party) infrastructure	Polyline
d	Infrastructure points	Point
e	Access points	Point

The polygon describing the project area (a) as well as each polygon describing the land use (b) must show the following characteristics (in addition to some mandatory attributes described below):

- Valid and closed polygon.
- No crossing edges.
- Degree-two vertices only.
- No overlapping or equality with other polygons (see Figure 1).

A polyline referred to as usable infrastructure (c) can be associated with own infrastructure (e.g., the copper cabling in the access domain) or third-party infrastructure (e.g., leased lines). The spatial object must be a valid polyline. In case of any attributive restriction in the accessibility, there must be at least two infrastructure points assigned; hence the infrastructure can only be accessed via one of the two points (an open line infrastructure can be accessed at the masts only).

The infrastructure points (d) are attributive assigned to exactly one polyline of the usable infrastructure and represent any type of infrastructure objects that can be localized on exactly one position (e.g., shafts to access pipes, masts, hardware like splitters or routers, etc.).

 ISBN: 978-1-61208-251-6

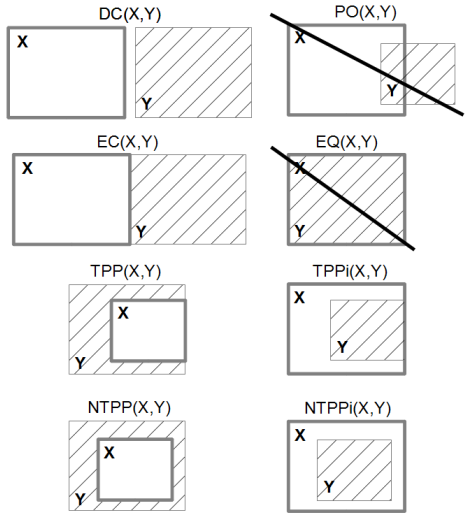
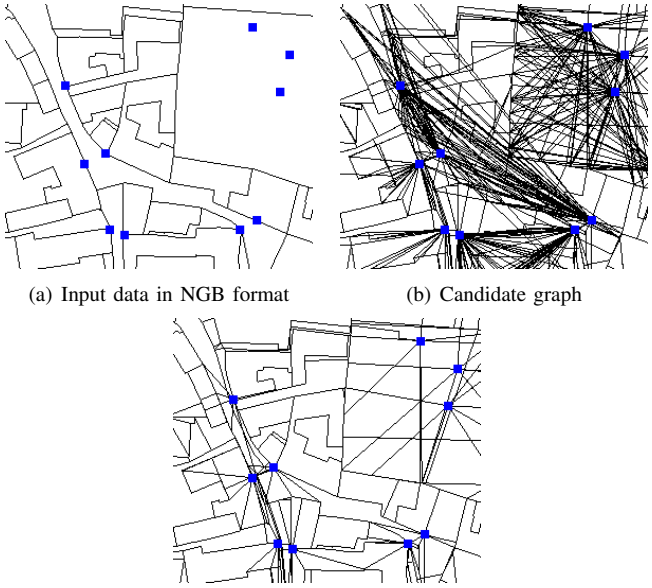


Fig. 1. All existing relations between pairwise polygons. EQ (equal) and PO (partly overlapping) relations are not allowed in NGB; Based on Region Connection Calculus (RCC-8) [17].



(a) Input data in NGB format (b) Candidate graph

(c) Generated graph

Fig. 2. (a) Preprocessed and enhanced input data, (b) the candidate graph according to Algorithm 1, and (c) the resulting generated graph . Access points are shown as squares.

The last object class is the class of access points (e). These are points representing the terminals in a following optimization. We distinguish between existing access points that are currently supplied by one or a group of connected usable infrastructure polylines, and potential access points that are not yet connected.

The presented approach of the graph generation makes use of geometric algorithms and algorithms from operations research. Thus, ambiguous relations or even holes between spatial objects cannot be allowed. This distinguishes between common geodatabase systems and spatial data in the NGB format, because there are no tolerances allowed.

Two points meant to be on the same position have to share the same coordinates. Furthermore, there are general specifications which cover the notation, the coordinate system, the locale, default attribute values, and a list of common abbreviations.

The next section describes our graph generation approach, that is based on spatial input data in NGB format.

IV. Graph Generation

The process that we follow to generate a weighted network graph consists of three consecutive stages (see Figure 2 for the individual results):

- NGB preprocessing and enhancement.
- Generation of the candidate graph.
- Running the rule-based system.

In the following section, further details to the stages will be given.

A. NGB preprocessing and enhancement

The preprocessing and enhancement of the input data are fully automated processes. As long as the input data fulfill the requirements in the NGB format, the generation of a weighted network graph will succeed. Moreover, the quality and usability of the generated graph are crucial in terms of topological errors within the spatial data. Furthermore, the succeeding process of assigning the correct weights to all edges in the graph is sensitive to the correct spatial classification.

To ensure high quality in the resulting graph, the input data are validated running the decision tree approaches described in [13] and [14]. Based on error free spatial data covering provincial, rural, suburban, and urban areas, a representative decision tree was constructed. Both approaches use this decision tree to validate the input data. The process will output warnings in case of topological errors and reclassify spatial objects according to the decision tree.

The validation is followed by the enhancement of the input data. Since underground work in crossroads areas and the subsequent obstruction in traffic should be avoided, the roadways are supplemented by polygons representing crossroads areas. Each crossing of at least two center lines of street polygons is identified and replaced by a polygon classified as crossroad area. The process runs automatically and produces suitable results (see Figure 3).

B. Generation of the candidate graph

The goal of any graph-based mathematical optimization is to connect the access points to a given or new access network. As a result of the graph generation, each spatial object will be connected with other objects in the candidate graph. Algorithm 1 describes the construction of the initial candidate graph that is used as an input into the rule-based system.

 ISBN: 978-1-61208-251-6

(a) (b)

Fig. 3. (a) Original NGB street polygons and (b) enhanced by crossroads area polygons.

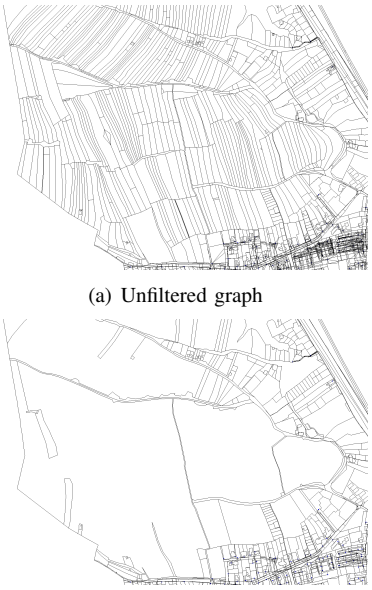
Algorithm 1 : Generation of the candidate graph

```
Import all polygons P.
Import all infrastructure polylines L.
Import all infrastructure points I.
Import all access points A.
Import specifications regarding additional crossings.
for all polygons p ∈ P do
    Create edges representing the border of p.
    if p encloses an access point a ∈ A then
        Create (orthogonal) projections from a to p.
    end if
    if p should be enhanced with crossings then
        Create a crossing all x meter.
    end if
end for
for all polylines l ∈ L do
    if l has restricted access at two or more points i_{1..n}
    then
        Create edges between the points i_{1..n} representing l.
    else
        Create edges from the polyline l.
    end if
end for
for all infrastructure points i ∈ I do
    if i hits any created edge e then
        Split e into e_1 and e_2 at location i.
    else
        Create (orthogonal) projections to connect i.
    end if
end for
```

C. Running the rule-based system

The rule-based system is based on the expert knowledge of network constructors. It is applied to the generated candidate graph to test for qualified edges. Each of the following questions will be answered with yes or no and determine the appearance of the edge in the final graph (edge is rejected, if all answered with no):

1) The edge is needed to ensure a connected graph.
2) The edge is part of the existing infrastructure.
3) The edge is not part of any land use to be filtered (compare Figure 4).
4) The edge is part of the best (least costly) possibility to connect spatial objects.



(a) Unfiltered graph

(b) Filtered graph

Fig. 4. (a) Unfiltered graph and (b) filtered by agricultural land uses.

While for the first three questions the answer is quite easy to find, the last question requires a combination of a list of rules with an algorithm from the area of combinatorial optimization to generate a valid answer.

After identifying all edges that are members in the final graph, the rule-based system is used to weight the edges. Depending on the originating land use or infrastructure, weights representing the network construction costs (or network usage costs in case of infrastructure) are assigned.

In the next section, we will describe the outcome of experiments we ran using input data from four different classification areas.

V. Experimental Results

After running a series of experiments, the results are not only promising, but are proven to be valid in case of simulation and optimization scenarios in the field of the construction of hybrid communication networks.

The three main factors determining the quality of the generated graphs are:

(a) The real-world correlation.
(b) The correct cost assignment.
(c) The usability.

The (a) correlation and (b) weighting can only be tested by a comparison of a manually planned scenario with the result of an optimization using the graph.

The (c) usability is primarily defined by the focused optimization algorithm. The number of vertices and edges in the weighted graph must be small enough to allow the application of heuristics or optimal algorithms but at the same time it also must be large enough to be non-restrictive.

 ISBN: 978-1-61208-251-6

In our experiments, we selected input data from four different classification areas:

- Urban;
- Suburban;
- Rural, and
- Provincial.

Table II provides the classification of the exemplary selected input data together with the dimensions.

TABLE II
SELECTED CLASSIFICATIONS AND DIMENSION OF ENCLOSED OBJECTS.

Area	# Polygon Objects	# Line Objects	# Point Objects
urban	20,569	59,408	33,090
suburban	18,497	30,997	16,361
rural	5,255	21,850	11,800
provincial	792	1,076	508

The results of applying our approach to the experimental data can be seen in Table III. The number of edges and vertices of the resulting graph is compared with a graph assumed to be fully connected (using all vertices of the spatial data objects). A fully connected graph represents the best possible real-world correlation (because nearly every path is present) but the least quality with respect to the applicability. Tests on small cut-outs of the experimental areas have shown that the optimization result using the generated graph equals the result using the fully connected graph.

TABLE III
DIMENSIONS OF THE GENERATED GRAPHS.

Area	Fully connected graph		Generated graph	
	# Edges	# Vertices	# Edges	# Vertices
urban	$9{,}384^6$	137^3	346^3	228^3
suburban	$6{,}850^6$	117^3	263^3	176^3
rural	870^6	41.7^3	139^3	88^3
provincial	88.4^6	13.3^3	17^3	14^3

To evaluate the performance of the graph generation process, we measured the duration of the described approach. The application was realized using programming languages from the .NET Framework. Each algorithm, the user interface, as well as the control logic was implemented in the object-oriented programming language C#. The described rule-based system was realized using the functional programming language F#. The experiments were carried out on a standard personal computer with the following setup:

- OS: Microsoft Windows Server 64bit.
- CPU: 3.5GHz Dual Core.
- RAM: 8 GByte.

TABLE IV
PERFORMANCE OF THE APPROACH.

Area	Duration in seconds
urban	3,600
suburban	2,043
rural	882
provincial	519

The performance of the graph generation process can be seen in Table IV. As expected, the duration corresponds to the number of spatial objects enclosed in the area. For the simple reason that the graph needs to be generated only once, the duration is reasonable and acceptable.

VI. CONCLUSION

In this paper, an approach for the generation and weighting of a network graph has been proposed. Introducing a normalization format called NGB to support a fully automated process of generating graphs using a wide range of hybrid spatial data as the input.

The experiments show that the approach is effective and efficient and that the weighted graph can be used as basic input into mathematical optimization algorithms. As a future investigation, we intend to explore ways to reduce the time needed for the graph generation process.

REFERENCES

[1] P. Bachhiesl, M. Prossegger, H. Stoegner, J. Werner, and G. Paulus, "Cost optimal implementation of fiber optic networks in the access net domain," in *International Conference on Computing, Communications and Control Technologies*, 2004, pp. 334–349.

[2] A. Ivanov and A. Tuzhelin, *Minimal Networks: The Steiner Problem and its Generalizations*. CRC Press, 1994.

[3] R. D. Hwang, F. and P. Winter, *The Steiner Tree Problem*. North-Holland, 1992.

[4] "Katastralmappe SHP Schnittstellenbeschreibung, Version 2.0.1." BEV - Bundesamt fuer Eich- und Vermessungswesen, 2012.

[5] R. Wang and S. D.M., "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, pp. 5–34, 1996.

[6] L. Leo Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[7] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: product and service performance," *Commun. ACM*, vol. 45, no. 4, pp. 184–192, Apr. 2002.

[8] A. Jakobsson and F. Vauglin, "Status of data quality in european national mapping agencies," in *Proceedings of the 20th International Cartographic Conference*, vol. 4, 2001, pp. 2875–2883.

[9] "19113 Geographic Information - Quality Principles," in *ISO/TC 211*. International Organization for Standardization (ISO), 2002.

[10] T. Ubeda and M. J. Egenhofer, "Topological error correcting in gis," in *Proceedings of the 5th International Symposium on Advances in Spatial Databases*, ser. SSD '97. London, UK, UK: Springer-Verlag, 1997, pp. 283–297.

[11] K. A. V. Borges, C. A. Davis, Jr., and A. H. F. Laender, "Database integrity," J. H. Doorn and L. C. Rivero, Eds. Hershey, PA, USA: IGI Publishing, 2002, ch. Integrity constraints in spatial databases, pp. 144–171.

[12] M. Mostafavi, G. Edwards, and R. Jeansoulin, "An ontology-based method for quality assessment of spatial data bases," in *Proceedings fo the Third International Symposium on Spatial Data Quality*, vol. 28, 2004, pp. 49–66.

[13] M. Prossegger and A. Bouchachia, "Incremental identification of topological errors in spatial data," in *The 17th International Conference on Geoinformatics*, Aug. 2009, pp. 1–6.

[14] M. Prossegger and A. Bouchachia, "Incremental semi-automatic correction of misclassified spatial objects," in *Proceedings of the Second international conference on Adaptive and intelligent systems*, ser. ICAIS'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 16–25.

[15] "ESRI Shapefile Technical Description," in *An ESRI White Paper*, July 1998.

[16] M. Prossegger, "Normalized Geobasisdata (NGB) - technical requirements v.3.0," FHplus Project Netquest, Carinthia University of Applied Sciences, Tech. Rep., October 2012.

[17] A. David, Z. Cui, and A. Cohn, "A spatial logic based on regions and connection," in *Proc. KR-92*, 1992, pp. 165–176.

 ISBN: 978-1-61208-251-6