The Long and Winding Road: Combining Least Cost Paths and Network Analysis Techniques for Settlement Location Analysis and Predictive Modelling

Philip Verhagen
VU University Amsterdam, The Netherlands

Tom Brughmans
University of Southampton, United Kingdom

Laure Nuninger
CNRS, UMR6249, Besançon, France

Frédérique Bertoncello
CNRS, UMR 7264, Nice

Abstract:
In this paper, we describe an exploratory analysis of the possibilities of combining least cost path analysis and network analysis techniques. Accessibility is a potentially important site location factor. So far, the definition of accessibility has been approached through the creation of accumulated cost surfaces and least cost paths. However, these methods do not provide direct information on the foci of movement. Starting from networks created from least cost paths, network analysis and space syntax were used to obtain additional information on the structural features of the network. It is concluded that both techniques can be used with least cost path-based networks, and will provide new insights into the characteristics of the network. For most applications however the space syntax measures that take the geographical dimension into account seem to be preferable to the simple network analysis measures used here.

Keywords:
Network Analysis, Space Syntax, Least Cost Paths, Accessibility, Movement

1. Introduction

When thinking about socio-cultural factors influencing settlement location choice, the accessibility of places in the landscape is a potentially important variable to take into account. Among the possible factors determining settlement location, access to resources and ease of movement in the landscape may have been important elements. For example, settlements might be preferentially located in areas that offer good access to prime agricultural land, and that allow them to interact easily with neighbouring settlements. How to define accessibility in such a way that it might be used as a variable for site location analysis and predictive modelling is however still very much open to debate. Most published research considering landscape accessibility limits it to the ease with which humans can reach a certain location. So-called hiking equations are often used to obtain cost surfaces of accessibility, and accumulative cost surfaces are then applied to find the travel time or energy expenditure needed to reach a single destination from all points (pixels) in the area studied. By adding up these accumulated cost surfaces for each and every pixel, a map of differential accessibility of the landscape can be obtained (total path costs; Llobera 2000). This accessibility can also be analyzed for different travel times (short/medium/long distance; see also Mlekuž and Vermeulen in press).

Corresponding author: j.w.h.p.verhagen@vu.nl

However, these methods do not provide much information on the possible foci of movement in the landscape. We argue that to this purpose some additional steps are needed. We depart from the creation and addition of least cost paths (LCPs) from and to multiple locations in the landscape (see e.g. Zakšek et al. 2008; Whitley and Burns 2010; Murrieta Flores 2012; Verhagen in press). This will

***Figure 1**. Location of the study area in the Netherlands, and a general impression of the landscape in the area.*

result in structures of cumulative least cost paths that resemble networks with different weights attached to the edges, but that do not have any real 'nodes'. We can then assume that edges with a high weight will have been most attractive to travel, and thus may have been more attractive to settlement as well. However, since these 'networks' are raster-based, they do not allow us to explore their topological characteristics, as can be done through network analysis. The analysis of the connections and distances between network nodes can tell us something about their potential for interactions. We used simple exploratory network analysis measures and space syntax to examine a range of structural features of the junctions in the landscape that were identified by the multiple LCPs.

2. Case Study

As a case study, we have taken the region of Zuid-Limburg in the Netherlands (Fig. 1). This area, roughly measuring 32 by 35 kms, is characterized by undulating hills, dissected by valleys. Elevation values range between approx. 35 m a.s.l. in the Meuse Valley to more than 300 m in the southeast corner of the region (Fig. 2). Accessibility of the landscape is not greatly hampered by terrain, but the differences in elevation and slope are large enough to suppose that terrain conditions will have influenced movement preferences, as is in fact witnessed by the location of roads on historical maps. The elevation model used was created using the ArcGIS 9.3 *Topo to Raster* tool[2] on the basis of contour lines digitized from 1920s topographical maps, scale 1: 25 000[3]. The vertical accuracy of the contour lines is 2.5 meters; elevation point measurements were added where available. The horizontal resolution used is 50 x 50 m.

A cumulative cost path map was then created using the method described in Verhagen (in press). Cost surfaces were calculated for sample points 250 m apart, using Tobler's (1993) hiking equation to specify the costs of movement.[4] For each cost surface, LCPs were then created departing from 72 radially distributed points at a distance of 5 km from the sample point.[5] The LCPs for each sample point were then added to create a cumulative cost path map, that represents the attractiveness of the landscape for movement within a 5 km radius (Fig. 3). It shows a dense network of paths, with a

2 This is a discretized thin plate spline technique, for which the roughness penalty has been modified to allow the fitted DEM to follow abrupt changes in terrain, such as streams and ridges.

3 Chromotopografische Kaart des Rijks or Bonnebladen, the first detailed topographical maps of the Netherlands, made between the 1890s and 1930

4 $W = 6e^{3.5|s+0.05|}$ where
W = walking speed in km/h; e = the base of natural logarithms;
s = slope in m/m

5 These were created using the ArcGIS (anisotropic) *PathDistance* and *CostPath* modules

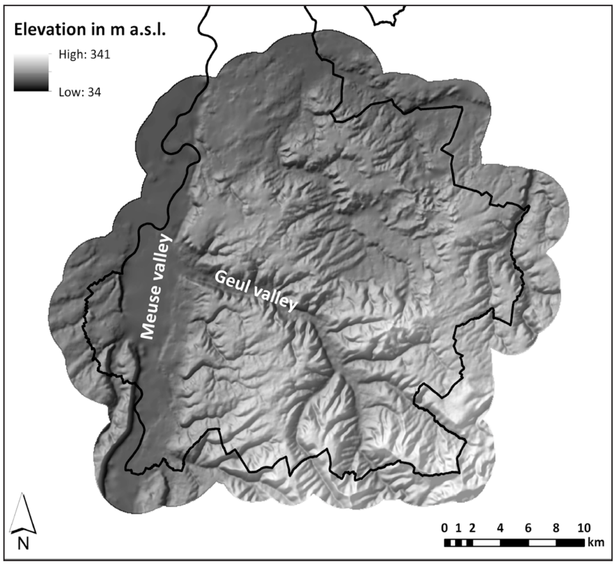

Figure 2. Digital elevation map of the study area.

concentration of movement potential in the valleys and on ridges, i.e. the areas with the gentlest slopes.

In order to investigate the potential of network analysis techniques, two approaches have been tested. The first one is based on simple node-based network analytical tools on undirected and unweighted edges (Newman 2010). The measures used include the number of nodes, connected components, clustering coefficient, density, heterogeneity, diameter, average shortest path length, average degree, closeness centrality, and betweenness centrality, they are all described in the next section. We decided to use these simple network measures because they reflect and summarise key structural features of networks (e.g the ratio of links over nodes, the distribution of paths, the level of clustering), and they are therefore most commonly used to compare different network structures (Newman 2003). This case study aims to evaluate whether these measures can also be used for exploring network structures where geographical distance between nodes plays an important role. The open source software platform Cytoscape[6] was used for this (Smoot et al. 2011), but other network analysis packages could be applied as well.[7] This node-based approach was therefore confronted with a second approach that focuses on the connections. For this we applied measures commonly used in space syntax (Hillier and Hanson 1984, Bafna 2003)

6 www.cytoscape.org

7 Alternatives include Pajek (de Nooy et al. 2005), UCINET (Borgatti et al. 2002), and Gephi (Bastian et al. 2009).

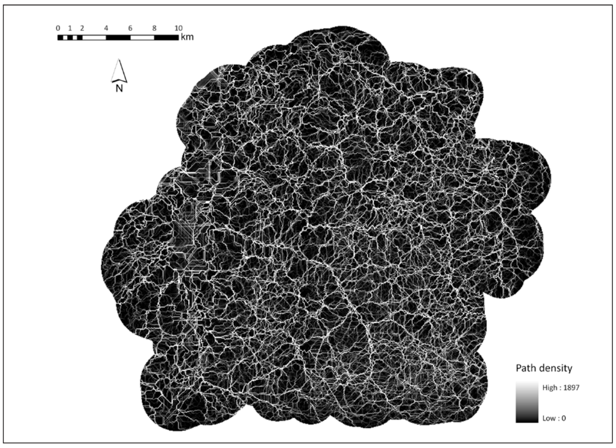

Figure 3. Cumulative cost path (or path density) map of the study region.

and implemented using Depthmap[8] (freeware; Turner 2004); no competing products are available.

In order to prepare the raster-based cumulative cost paths for analysis in the vector-based network analysis packages, some pre-processing had to be done. We assumed that only the paths that are most frequently chosen in the LCP modelling would have been capable of supporting long-distance connections. We therefore extracted the upper 10% and 20% of most frequently chosen paths from the cumulative cost paths map. These needed to be thinned and converted to polylines, before exporting the resulting edges to Cytoscape and Depthmap.

3. Cytoscape Results

Only the most common network analysis measures provided in Cytoscape were used for this case study. On the one hand, the software offers a number of global measures that hint at structural characteristics of the total network. On the other hand, local measures are available that provide information on the structural position of individual nodes in the network. Definitions of measures and terms in network analysis can differ between software packages, so the definitions given here are specific to Cytoscape.

The data included a lot of loops (i.e. edges from one node to itself). These were removed before any analysis took place. A few edges are also duplicated because they appear in both directions (e.g. from A to B and from B to A). Since the LCPs do not

8 www.vr.ucl.ac.uk/depthmap

	Network 10%	Network 10% largest connected component	Network 20%	Network 20% largest connected component
Number of nodes	1605	1417	6525	6495
Connected components	74	1	16	1
Density	0.001	0.002	0.000	0.000
Heterogeneity	0.405	0.357	0.158	0.152
Clustering coefficient	0.045	0.049	0.059	0.059
Diameter	67	67	130	130
Average shortest path length	28.787	28.793	52.840	52.840
Average degree	2.314	2.456	2.886	2.895

Table 1. Results global measures Cytoscape.

represent directed networks these multiple edges are meaningless and they were removed before performing any analysis.

3.1 Results global measures

The following definitions apply to the global measures calculated by the Network Analyzer plug-in in Cytoscape:

- A **component** is a connected subnetwork, i.e. nodes in one component are not connected to nodes in other components.
- The **average degree** is the average of the number of neighbours of all nodes.
- The **density** is a normalized version of the average degree that represents the fraction of all possible connections that are actually present (self-loops and duplicated edges are ignored).
- The **network heterogeneity** reflects the tendency of a network to contain hub nodes that are linked to many other nodes.
- The **clustering coefficient** calculates the average probability that two neighbours of a vertex are themselves neighbours, as a ratio of the number of edges between the neighbours of a given node and the maximum number of edges that could possibly exist between these neighbours (Watts and Strogatz 1998). The network clustering coefficient used here is the average of the clustering coefficients for all nodes in the network.
- The **network diameter** is the largest path length distance between two nodes. If a network is disconnected, its diameter is the maximum of all diameters of its connected components.
- The **average shortest path length** is the average of all shortest path scores between all possible pairs of vertices in the network.

The analysis results are shown in Table 1. Both the 10% and 20% networks consist of one big connected component that includes over 80% of all nodes. All other nodes are part of a large number of small components or are isolated. It was decided to only analyze the largest connected component of each network. The networks are very linear, i.e. nodes tend to follow one after the other and rarely branch off. This is reflected in the low average degree, and is a direct result of using the cumulative cost path approach. This is also the reason why the density score and clustering coefficients are low, the diameter high and the average shortest path length very high. The overall structure does not seem to be very much affected by the network size, since both the 10% and 20% networks show similar results. This network structure means that it is highly likely for a number of local measures (like node centrality) to emphasize one or two paths. Also, the geographical and linear nature of this network makes it extremely susceptible to edge-effects: the peripheral nodes will almost always be underrepresented and will contribute to the overrepresentation of the core nodes.

3.2 Results local measures

A number of local measures were used but only two of them provided interesting results. The following definitions apply to the local measures

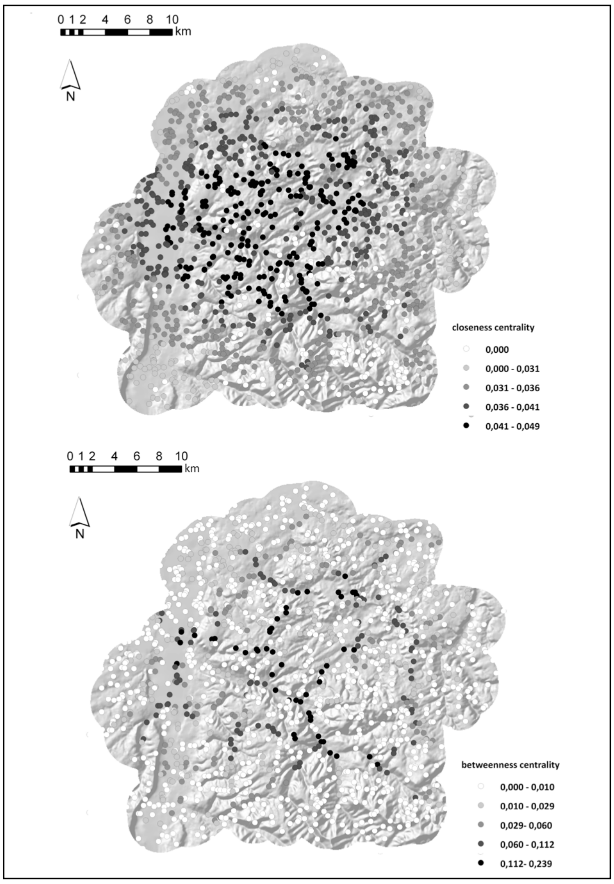

Figure 4. *Closeness and betweenness centrality for the 10% network, obtained from Cytoscape and plotted on a hillshaded relief map of the study region.*

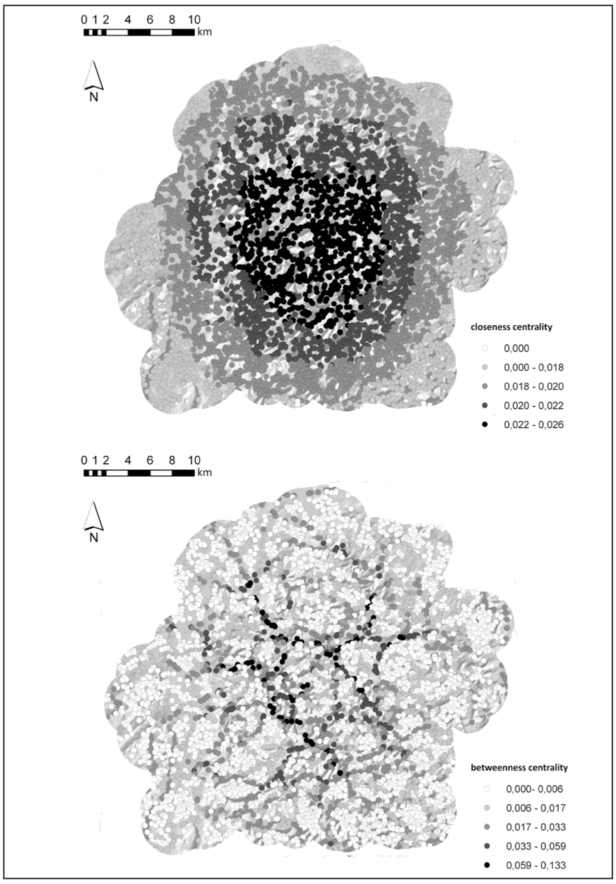

Figure 5. *Closeness and betweenness centrality for the 20% network, obtained from Cytoscape and plotted on a hillshaded relief map of the study region.*

calculated by the Network Analyzer plugin in Cytoscape:

The **closeness centrality** of a node is "the number of other vertices divided by the sum of all [topological] distances between the vertex and all others" (de Nooy et al. 2005, 127). Closeness centrality is often considered a measure of how fast information spreads from a given node to other reachable nodes in the network.

The **betweenness centrality** of a node is the proportion of all shortest (topological) paths between pairs of other vertices that include this vertex (de Nooy et al. 2005, 131). The betweenness centrality of a node is often taken to reflect the amount of control that this node exerts over the interactions of other nodes in the network. This measure favours nodes that join communities (dense subnetworks), rather than nodes that lie inside a community.

The results of the analysis are displayed in Figures 4 and 5. For the 10% network the closeness centrality scores are generally low, and their frequency follows a normal distribution. A substantial number of nodes has a higher score, these seem to lie at the centre of the region and partly in the NW corner. The lowest scores are for nodes at the periphery of the network, particularly those at the eastern and southern edges. This entire pattern is a result of the sensitivity to edge-effects of this measure.

The frequencies of the betweenness centrality scores are more extremely distributed than for the closeness scores. Many nodes have a score of 0 or an extremely low score. Some have a relatively high score, which are the nodes that are frequently traversed. These higher scores seem to follow a number of distinct paths picked up by the least cost path analysis.

For the 20% network the core area with the highest closeness scores is still situated at the centre of the region, but no longer in the NW corner. The results are very similar to those of the 10% network, again stressing the influence of edge-effects on this measure.

The betweenness centrality scores show a similar distribution of scores as for the 10% network. The analysis again seems to pick up a number of more prominent paths. Although they are more or less in the same area as those of the previous network, the paths do seem to be different. It seems then that this measure is quite sensitive to network size and most importantly to the presence or absence of possible shortcuts which cause key paths to be diverted.

4. Depthmap Results

Space syntax applies a different terminology than many network analysis approaches, although there is strong similarity in the measures used since both approaches are rooted in graph theory. The Depthmap measures used here are calculated for the edges of networks and do not provide any direct information on the characteristics of the nodes themselves. Some of the measures are very similar to those used in Cytoscape, however, but applied to edges rather than nodes. It therefore becomes an interesting exercise to compare results of a node-based network analysis in Cytoscape with an edge-based space syntax analysis in Depthmap. This will allow us to evaluate which approach using freely available user-friendly software is most suitable for exploring geographical LCP networks. We have calculated the following measures in Depthmap:

- **Step depth** is the topological distance covered between two nodes, when passing through all the nodes between them, similar to calculating a path in Cytoscape.

- **Mean depth** is the topological distance from a node to all other nodes summed and divided by the number of nodes minus one. This is the inverse of closeness centrality in Cytoscape, where the number of nodes is divided by the sum of all paths.

- **Relative/relativised asymmetry** is the normalized mean depth on a scale of 0-1.

Mean depth and relative asymmetry are used to analyze the **integration** of the network. A highly integrated network has a large number of direct connections, it is symmetric. A segregated network has few direct connections, it is asymmetric.

- **Control** is the number of nodes that are connected to a node relative to the number of nodes that these have connected to themselves.

- **Controllability** is the ratio of nodes with depth 2 and nodes with depth 1 from a node

Control is used to calculate the **distributedness** of the network, i.e. if nodes dominate the structure of the network locally. If the structure depends on a relatively large number of nodes, then the network is distributed. If few nodes dominate the network, then it is non-distributed.

- **Degree centrality or connectivity** is the number of neighbours of a node (i.e. the degree of a node).

- **Closeness centrality** is the average topological distance from one node to all other, like the same measure in Cytoscape.

- **Choice** is number of shortest paths that passes through a node (i.e. the same as betweenness centrality in Cytoscape).

Note that the definitions are given here as measures for nodes. From the documentation it is not immediately clear how the measures are calculated, i.e. if measures derived from nodes are attached to edges, or calculated directly on the edges; the results obtained however suggest that the measures are calculated on the edges. Somewhat confusingly, the term 'node count' in Depthmap refers to the number of edges.

It is not possible to analyze subnetworks separately in Depthmap without breaking up the network, and running a new analysis on the subnetwork. Since Depthmap operates on edges that are placed in a (geographical) coordinate grid, rather than on nodes that only have topological connections, it can also include node-to-node metric distances in its analysis. Topological distance only takes into account the number of steps it needs

	Network 10%	Network 20%
Node count (number of edges)	5728	17703
Average connectivity	2.47727	3.05293
Maximum connectivity	5	6
Average step depth	67.8382	96.1734
Maximum step depth	133	203
Average mean depth	82.0956	98.7524
Maximum mean depth	133.388	147.438
Average integration	0.143313	0.120811
Average control	1	1
Average controllability	0.460255	0.407049

***Table 2**. Results global measures Depthmap.*

to move from node A to node B. Depthmap can also calculate the actual metric distance it takes to move from A to B when passing through all the nodes between them, and in this way obtain metric measures as well. This leads to some differences in the analysis results. Depthmap can also perform analyses at different scale levels, by setting the radius of analysis to a maximum topological or metric distance.

4.1 Global results

The global measures obtained with Depthmap are shown in Table 2. The results are in line with those obtained through Cytoscape. The connectivity of the network is low, nodes will have few direct connections, resulting in a low average integration. In order to reach nodes in the network, a large number of other nodes has to be passed. This is the consequence of using the cumulative cost path approach.

4.2 Local results

The results for closeness centrality are comparable to those obtained with Cytoscape. More interesting are the results for the metric choice measure (Fig. 6). It shows very clear preferred pathways for the 10% network; for the 20% network patterns are less clear with a relatively even distribution of preferred pathways over the area. In both cases, the centrally located areas are given a higher weight, as a consequence of edge effects. For the 10% network, it is very clear that the preferred paths follow the valleys. A preferred route is also present on the north side of the region; this is related to the effect discussed in 3.2, with the concentration of high closeness values in the NW corner of the area. On the basis of the Depthmap results, we can conclude that this is due to the lack of a direct connection in the central valley of the river Geul, forcing routes to the north instead of following the more direct east-west connection. In the 20% network, this connection is present and therefore the high closeness and metric choice values are concentrated in the centre of the area.

The differences between the choice (not shown in the figure) and metric choice parameters are not very conspicuous when visually comparing the

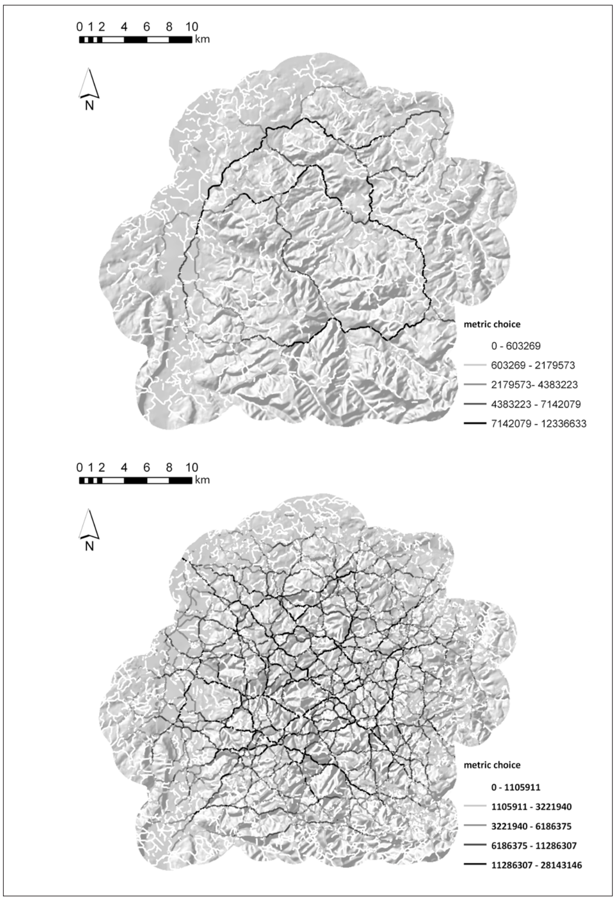

***Figure 6**. Metric choice for the 10% and 20% network, obtained from Depthmap and plotted on a hillshaded relief map of the study region.*

results, although they are more striking for the 20% than for the 10% network. This is partly because of the presence of artefacts of the LCP modelling, that creates tightly packed clusters of nodes in places where many LCPs meet. This effect is stronger for the 20% network. Passing these clusters of nodes will increase the topological distance much quicker than metric distance.

5. Discussion

There are some clear differences between the two approaches used in this case study.

Cytoscape was used to only calculate measures for nodes, Depthmap for edges. It depends on the application whether one is preferable to the other. In archaeology, the characteristics of settlements in a network might be interesting to analyze, but it can also be useful to see where the major connections are. In the first case, a network package like Cytoscape is to be preferred, in the second case it is better to use space syntax. Exporting data from GIS to Depthmap is easier since it only needs an export to MIF/MID format. For use in Cytoscape, a list of connections in the format of source nodes and target nodes has to be produced.

Visualisation in both packages is completely different. Depthmap is very much comparable to a GIS, and retains the original spatial configuration of the edges, which makes interpretation of the patterns easier (at least for the purposes of this case study). The lay-out algorithms embedded in Cytoscape can display the nodes in a wide array of configurations, none of which take the spatial configuration into account. This has the advantage of emphasizing and visually identifying topological features through different lay-outs. In this case study, however, the networks were too large and the network analysis results not striking enough to motivate a visual topological exploration. In order to visually compare the analysis results from both packages, the data has to exported back to GIS.

Both Cytoscape and Depthmap offer global and local network analysis measures. These are not fully comparable, even when strong similarities are found. We have not undertaken an exhaustive survey of all the differences and similarities between the techniques, even when this would be a useful exercise in itself. It seems that Depthmap is less well documented; measures in the manual (Turner 2004) are usually described in fairly general terms, with references to the corresponding scientific papers. The analysis of general network structure and the identification of the most important connections can equally well be done in both packages. Cytoscape is somewhat stronger on the global measures, they are easier to extract using the Network Analyzer plugin and offer a wider range of options. The calculation of the control measure however is specific to Depthmap. The clustering coefficient of nodes calculated in Cytoscape is similar in that it reflects local differences in degree, but this measure did not provide interesting results for this case study.

The strongest point of Depthmap is that it offers the opportunity to calculate metric measures alongside the topological ones. In geographical space, real distance is an important factor to take into account when calculating network characteristics. Network analysis software packages do not offer many spatial analysis techniques as a standard feature. An alternative approach would be to add geographical distance to edges, but since these distances are relative to coordinate systems and projections a GIS import tool would make more sense, or a combination of a spatial database with a network analysis extension.

Furthermore, Depthmap can limit the calculation of centrality measures to a certain distance, so centrality will then be measured within a set radius, instead of the whole region. This reduces edge effects, and emphasizes the local areas of through-movement, offering a wider range of network features to explore.

The results obtained show that network analysis techniques offer some additional information on the network structure of the cumulative cost path maps that cannot be obtained from the raster-based maps. The networks created from the cumulative cost paths show a weak integration and high distributedness; many nodes have to be passed to go from one location to the other, and junctions usually connect only three edges. This is a consequence of using the cumulative cost path approach; using a different set of connections, e.g. between settlement locations, would lead to different results.

The major connections that are obtained through the calculation of betweenness centrality (choice) follow the valleys rather than the ridges. This is because the pathways created on the ridges are not very well connected to the rest of the paths. The steep slopes on the valley sides prevent those routes of becoming well-connected to other paths, and thus through-movement is concentrated in the valleys instead. However, the centrality measures calculated for the whole region also tend to emphasize the centrally located nodes; they are very sensitive to edge effects. Furthermore, the choice of network size (10% or 20% most frequently chosen paths from the cumulative cost path map) has a strong influence on the betweenness centrality. When less frequently used paths are included, the general structure of the network becomes less clear.

The results also show, however, that network measures incorporating geographical space provide the most interesting results. The global non-geographical network analysis results helped to identify the general structure of the dataset, the local network measures that would be most suitable and the kinds of results we could expect to emerge. All of the results of local measures we derived with Cytoscape, however, found an equivalent in results derived with Depthmap. It seems then that the spatial nature of networks derived by the LCPs is not trivial and significantly affects network analysis results. In order to attain non-trivial network analysis results the networks and measures need to address this spatial nature explicitly.

Acknowledgements

Part of this study was made possible through a VENI-grant awarded to Philip Verhagen by NWO ("Introducing the human (f)actor in predictive modelling for archaeology"). Travel funds were made available to Philip Verhagen and Laure Nuninger by the Réseau Franco-Néerlandais (Van Gogh-grant) and the French Ministry of Foreign Affairs (PHC Van Gogh).

References

Bafna, S. 2003. "Space Syntax: A Brief Introduction to Its Logic and Analytical Techniques." *Environment and Behavior* 35:17-29.

Bastian, M., S. Heymann, and M. Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." Paper presented at the *International AAAI conference on weblogs and social media*.

Borgatti, S.P., M.G. Everett, and L.C. Freeman. 2002. *Ucinet 6 for Windows: Software for Social Network Analysis, Analytic Technologies*. Cambridge: Harvard University.

Hillier, B., and J. Hanson. 1984. *The Social Logic of Space*. Cambridge: Cambridge University Press.

Llobera, M. 2000. "Understanding movement: a pilot model towards the sociology of movement." In *Beyond the Map. Archaeology and Spatial Technologies*, edited by Gary Lock, 65-84. Amsterdam: IOS Press/Ohmsha.

Mlekuž, D., and F. Vermeulen. *In press*. "Time Geography, GIS and archaeology." In *Fusion of Culture. Proceedings of the XXXVIII Conference on Computer Applications and Quantitative Methods in Archaeology, 6-9 April 2010, Granada*, edited by F. Contreras, M. Farjas and F. J. Melero. Oxford: Archaeopress.

Murrieta Flores, P. 2012. "Understanding human movement through spatial technologies. The role of natural areas of transit in the Late Prehistory of South-western Iberia." *Trabajos de Prehistoria* 69:103-22.

Newman, M.E.J. 2003. "The structure and function of complex networks." *SIAM review* 45:167-256.

Newman, M.E.J. 2010. *Networks: an introduction*. Oxford: Oxford University Press.

Nooy, W. de, A. Mrvar, and V. Batagelj. 2005. *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.

Smoot, M., K. Ono, J. Ruscheinski, P-L. Wang, and T. Ideker. 2011. "Cytoscape 2.8: new features for data integration and network visualization." *Bioinformatics* 27:431-32.

Tobler, W. 1993. *Three presentations on geographical analysis and modelling. 1) Non-isotropic modelling 2) Speculations on the geometry of geography 3) Global spatial analysis*. Santa Barbara: National Center for Geographic Information and Analysis.

Turner, A. 2004. *Depthmap 4. A Researcher's Handbook*. London: Bartlett School of Graduate Studies, University College London.

Verhagen, P. In press. "On the road to nowhere? Least cost paths, accessibility and the predictive modelling perspective." In *Fusion of Culture. Proceedings of the XXXVIII Conference on Computer Applications and Quantitative Methods in Archaeology, 6-9 April 2010, Granada*, edited by F. Contreras, M. Farjas and F.J. Melero. Oxford: Archaeopress.

Watts, D.J., and S.H. Strogatz. 1998. "Collective dynamics of 'small-world' networks." *Nature* 393 (6684): 440-2. doi:10.1038/30918.

Whitley, T.G., and G. Burns. 2008. "Conditional GIS Surfaces and their Potential for Archaeological Predictive Modelling." In *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007*, edited by A. Posluschny, K. Lambers and I. Herzog, 292-98. Bonn: Dr. Rudolf Habelt.

Zakšek, K., É. Fovet, L. Nuninger, and T. Pobodnikar. 2008. "Path Modelling and Settlement Pattern." In *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007,* edited by A. Posluschny, K. Lambers and I. Herzog, 309-15. Bonn: Dr Rudolf Habelt.