Robustness of Spatial Databases against Intentional Attacks and Random Errors

Finn Hedefalk and Anders Östman
Gävle GIS Institute, University of Gävle
Nobelvägen 2, 80167 Gävle, Sweden
{finhek, aon}@hig.se

ABSTRACT

Demands on the quality and reliability of volunteered geographic information have increased because of its rising popularity. Due to the less controlled data entry, there is a risk that people provide false or inaccurate information to the database. One factor that affects the effect of such updates is the structure of the database schema, which in this paper is described by network models. By analyzing GIS data models, we have found that their class diagrams have small-world properties and long-tailed distributions. Moreover, an analysis of the error and attack tolerance showed that the data models were robust against random errors but very fragile against attacks. In a network structure perspective, these results indicate that false updates on random tables of a database should usually do little harm, but falsely updating the most central cells or tables might cause big damage. Consequently, it may be necessary to monitor and constrain sensitive cells and tables in order to protect them from attacks.

INTRODUCTION

Geographic information (GI) is copyrighted in many parts of the world, and this restricts the GIS users who need cheap, unlicensed spatial data in their projects. GI may also be quickly out of date, which raises the need for online up-to-date maps (Haklay and Weber, 2008). The idea of involving users in the collection and maintenance of GI is therefore becoming more popular. For example, many open map solutions have been created, and there have been discussions about updating national databases using crowdsourcing techniques (EuroSDR, 2009). The level of public participation in the update and maintenance of GI may vary between solutions. For instance, update of sensitive and complex data may be restricted to a smaller group of experts compared to the large groups of users that are allowed to update some common open maps. Despite the differences of restrictions, allowing more, and sometimes unknown, users to update the map (i.e. the database) will increase the risk of having inaccurate data. Reasons may be different types of measurement instruments (Morris et al., 2004; Sayda, 2005), conflicts with semantics (Matyas, 2007) or that false updates occur (Flanagin and Metzger, 2008; Sayda, 2005).The latter may occur because of people's biased way of describing the world (Flanagin and Metzger, 2008). The inaccurate data causes problems, especially if the service has many users or is important in other ways. Current solutions for quality assurance are mainly based on models for measuring trust (Bishr and Mantelas, 2008; Sayda, 2005). However, false data may still pass through these filters.

The objective of this paper is to study what impact false updates, or similar harmful occurrences, will have on the information in the database. When a user updates a database, one or many tables are affected. One factor that then influences the impact is the database schema structure. For instance, falsely updating a table with many relationships may do more harm than doing the same on one with fewer relationships. A schema specifies among other things the tables, the relationships between them and their attributes (columns). In this paper, we describe the schema structure as a network, where the tables or attributes are vertices and their relationships are connections between the vertices. Our hypothesis is that the network structure may then reveal the database's robustness against intentional attacks and random errors. Intentional attacks apply to the cases where users maliciously update the

essential parts of the network. For instance, users who are familiar with the schema's network structure may attack the table with the most connections in order to do more harm. On the other hand, random errors apply to false updates on the nonessential part of the network. Such errors may be caused by unintentional acts such as operational mistakes or use of inaccurate measurement instruments, or intentional damages by users who are not familiar with the network structure. The latter case, although done on purpose, is random with respect to the network structure; thus it is classified as random errors. By applying general network descriptors, we can identify, monitor and constrain sensitive tables, attributes and cell values.

Many real-world networks have small-world and scale-free properties. Small-world networks are highly clustered (a vertex neighbors are also neighbors of each other), with few steps between any two vertices in the network. That is, they have a high *clustering coefficient C* and a short *average path length L*. The clustering coefficient for vertex v_i can be defined as

$$C(v_i) = \frac{2E_i}{d_i(d_i - 1)}, \tag{1.1}$$

where d_i is the *degree* (number of connections) of vertex v_i. E_i is the number of *edges*, i.e. undirected connections, among the vertices in a 1-neighborhood of v_i (Watts and Strogatz, 1998). The average clustering coefficient is then defined as

$$C = \frac{\sum C(v_i)}{n}, \tag{1.2}$$

where n is the number of vertices in the graph. Furthermore, the average path length can be defined as

$$L = \frac{1}{n(n-1)} \sum_{i,j} d(v_i, v_j), \tag{1.3}$$

in which $d(v_i, v_j)$ is the shortest distance between v_i and v_j (Watts and Strogatz, 1998). In scale-free networks, the degree is power-law distributed (Barabasi and Albert, 1999). This means that the frequency function p(x) can be written as

$$p(x) = P(d \in [x, x + \Delta x] = \gamma \cdot x^{\uparrow}(-\alpha), \tag{1.4}$$

where the power-law exponent $\alpha > 1$. As a result, some few vertices will have an extremely high degree, whereas most get a very small degree.

Several UML class diagrams for Object-oriented (OO) solutions have small-world and scale-free properties (Concas et al., 2007; Myers, 2003; Valverde and Solé, 2007). Many spatial data models are stored in UML class diagrams, and therefore they may have similar network properties with the ones for OO softwares. Scale-free networks are considered to be robust against random vertex removals but extremely sensitive to removals of their most central vertices (Albert et al., 2000). The information spread is also fast, especially if the spreading starts from highly connected vertices (Dezso and Barabási, 2002). Similar results have been found for small-world networks, where the information spread is fast because of the short paths between nodes (Watts and Strogatz, 1998). However, most studies regarding the characteristics of scale-free and small-world networks have been conducted on artificial networks. Robustness tests on the database models are therefore required.

The main objective of this study is to analyze spatial data models with respect to scale-free and small-world properties. Moreover, the robustness, in this case the error and attack tolerance, is to be analyzed for selected data models. To achieve the objectives, following tasks will be done

- Graph specification and transformation of data models
- Small-world and scale-free analysis of the graphs
- Error and attack tolerance of selected graphs
- Evaluation of consequences for spatial databases.

THE USE OF NETWORK ANALYSIS ON DATABASE MODELS

As mentioned, one factor that affects the impact of a false update is the database schema structure. Below, a false update of a protected area is described. Since protected areas in reality are sensitive information, it would usually not be possible for any user to have this kind of access. Nevertheless, if a database with such information was editable for some few users, inaccurate data could still occur. Moreover, in cases when the database is online, although closed for edits, the risk is higher for malicious users to illicitly access the database and provide false information (Litchfield et al., 2005). In figure 1a, the *protected area* "Gysinge" is managed by *agency* "Lst GB". It has the *activity* "Fire" specified and the *specie* "Lynx" lives within it. If then the protected area tuple is removed, the agency, activity and specie tuples become isolated (figure 1b). They lose their connections to the protected area as well as to each other. However, if instead the *activity* tuple is removed, the other tables would still be related to each other, and only one direct connection would be broken. Figure 1c-d illustrates the same operations but using graphs instead. Here, the cell values are defined as vertices and their relationships as edges.

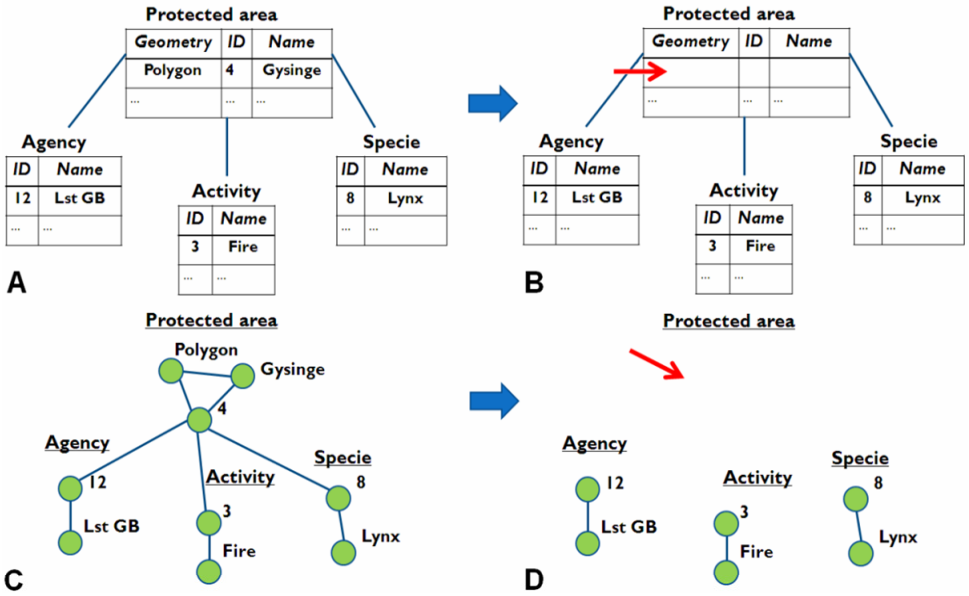

Figure 1: False updates on a database. Primary key (PK) and Foreign key (FK) relationships are omitted for simplicity reasons. A) A protected area with a related agency, activity and specie, b) the same tables being falsely updated, c) graph of the same schema, d) the graph after false updates.

MATERIALS AND METHODS

In this study, we have analyzed fourteen class diagrams stored in eXtensible Markup Language (XML) Metadata Interchange (XMI). They were downloaded from the data model directory of the ESRI support centre (ESRI, 2009). We have also studied six diagrams stored as gif and pdf files from the Swedish land survey and the Annex 1 application schemas from the INSPIRE Consolidated UML Model (INSPIRE, 2009).

Specification of graphs

The UML class diagrams used consists of spatial and non-spatial feature classes, abstract classes, relationships, attributes, lists of domain constraints and geometry classes. All of these items were included in the specifications except the domain constraints. One reason was problems in automation of the translations of the diagrams stored in XMI files. Three graphs were specified in order to estimate the distribution of classes' relationships, classes' attributes and attributes' relationships. For the first graph, the concepts of *class graph* from Valverde and Sole (2007) were used. Here, the feature classes are defined as vertices and their relationships as edges (figure 2a). Thus, there are no differences between types of relationships. For the second graph, *attribute graph*, the classes and attributes are specified as vertices. Directed connections, i.e. *arcs*, are established between each class and its attributes, so the number of every class' attributes can be measured (figure 2b).

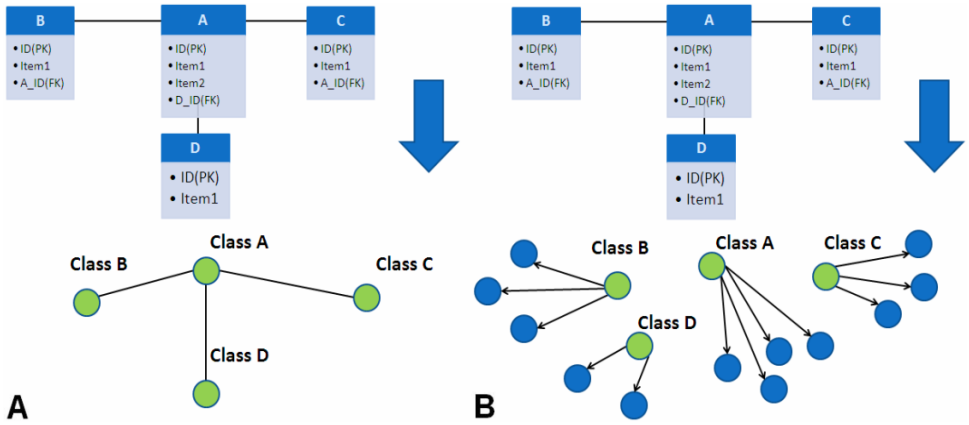

Figure 2: Graph specification. A) Class diagram transformed to a class graph, b) class diagram transformed to an attribute graph

The third graph, *schema graph,* tries to better model the physical implementation (figure 3). Hence, the vertices consist of attributes, and edges illustrate the Primary Key (PK)-Foreign Key (FK) relationships. Attributes in same tables are considered as connected. Moreover, due to the database normalization rules, vertices and edges representing relation tables are created between two classes that have the cardinality many-to-many. Many data models, however, do not describe the PK-FK relationships, which complicate the transformation process. An alternative way is then to specify the classes as artificial PKs and FKs. Then, the classes' relationships can illustrate the relationships between the PKs and FKs. A table with many FKs, however, may with this specification get an inaccurate degree.

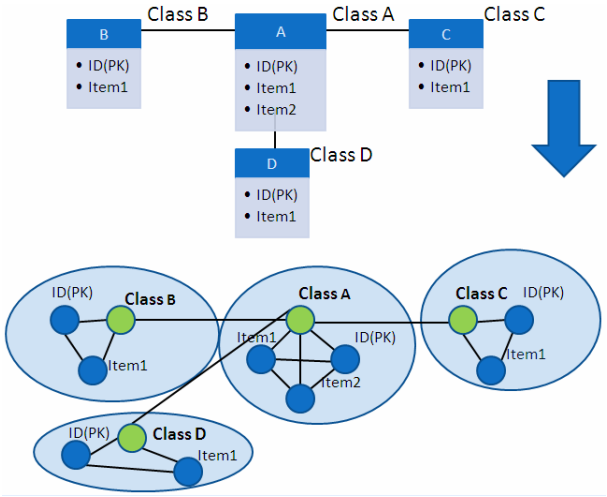

Figure 3: Class diagram specified as a graph with classes and attributes as vertices.

When analyzing the schema graphs for scale-free properties, the *betweenness* distribution is studied. De Nooy et al. (2005) define the betweenness for vertex v_i as

$$B_i = \frac{\#\ of\ shortest\ paths\ between\ v_j\ and\ v_w\ passing\ through\ v_i}{\#\ of\ shortest\ paths\ between\ v_j\ and\ v_w} \quad (3.1)$$

This parameter plays a more important role in the schema graph since many vertices will have similar degree due to the high clustering. Moreover, small-world analysis was considered relevant for the schema graphs since they intend to represent the actual database structure of the implementation.

Transformation and network analysis

The diagrams stored in XMI were transformed with eXtensible Stylesheet Language Transformations (XSLT) to graphs, in this case the ASCII-based Pajek .net format (figure 4). For the six class diagrams stored in graphic formats, the degree, related to class graphs and attribute graphs, were manually counted. No schema graphs were created for the diagrams in graphic formats since doing it manually would be too time-consuming.

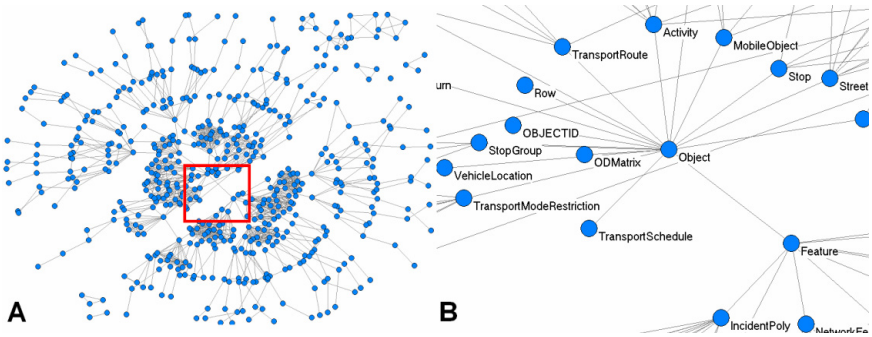

Figure 4: Example of a class diagram transformed to a schema graph. A) Schema graph of an ESRI data model for transportation viewed in Pajek, b) zoomed in view of the graph.

Compared to random networks, small-worlds are much more clustered and have similar or slightly larger separation between the vertices, i.e. $C_{small-world} \gg C_{random}$ and $L_{small-world} \gtrsim L_{random}$. This is a small-world indication (Watts and Strogatz ,1998), and therefore L and C were measured for each schema graph and then compared with L and C for equal sized random Erdős–Rényi (ER) graphs with the same average degree. During the scale-free analysis, the degree distribution for class graphs and attribute graphs, and the betweenness distribution for the schema graphs were estimated. For this, the estimation methods from Clauset et al. (2009) were used. These methods use ML estimators, K-S Goodness-of-Fit Tests, and Likelihood-ratio tests for comparing the fit of different models. The networks' error and attack tolerance were also analyzed. In this analysis, an *error* is defined as the removal of randomly chosen vertices, whereas an *attack* is the removal of the most central ones (Albert et al., 2000). In our study, an attack may illustrate a false update (for example, deletions) on the most central tables or cell values, whereas an error would be a randomly performed update. During the analysis, the vertices were removed one by one and as an indicator of the robustness, the drop in the relative size S of the largest connected network was observed. Class and schema graphs were analyzed, and their results were compared with the error and attack tolerance for ER graphs with the same size and average degree.

RESULTS

Fourteen schema graphs were analyzed for small-world properties. The results show that for around half of the schema graphs, $L_{schema} \gtrsim L_{random}$, and $C_{schema} \gg C_{random}$, which is a clear indication of small-world properties. However, for the other half, L_{schema} was around twice as large as L_{random}. Many schema graphs had small-world properties, but because all attributes in the same classes were defined as clustered vertices, most graphs would naturally be highly clustered. Furthermore, the results from the scale-free analysis show that the majority of the 20 analyzed graphs have power-law distributions, while the others are more likely to be at least heavy-tailed distributed instead of Poisson or exponentials (table 1). The latter two are common distributions for random networks.

Table 1: Most probable distributions for the graphs.

Graph	n[a]	PL[b]	PL + cut-off[c]	Heavy-tails[d]	Poisson/Exp.	None
Class graph	40-370	8	0	8	0	1
Attribute graph	40-370	11	0	5	0	0
Schema graph	170-1700	9	5	0	0	0

[a] Number of vertices
[b] Power-law
[c] Power-law with exponential cut-off
[d] Lognormal or Discrete Weibull

Although more than half of the distributions for the different graphs got support for power-laws, the populations in the tests were small, which made the estimations less accurate. In addition, some of the distributions got power-law support only for a few percentages of the population. Nevertheless, around 20 – 30% of the vertices in the class graphs and attribute graphs, and 10 – 25% of the vertices in the schema graphs, had a degree or betweenness above the mean value. This is an indication that most connections belong to some few vertices, and therefore the distributions are considered long-tailed but not strict power-laws (see figure 5 for examples).

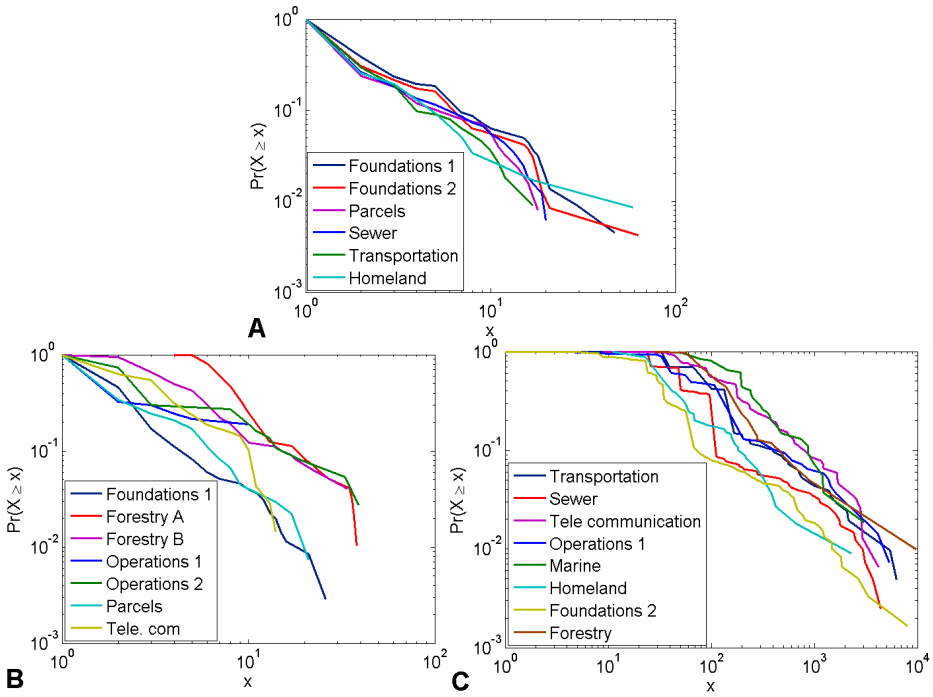

Figure 5: Degree and betweenness distributions plotted on double logarithmic scales. A) Class graphs, b) attribute graphs, c) schema graphs. Straight, diagonal lines indicate power-law distributions.

For the error and attack tolerance, the analyzed class graphs and schema graphs broke down much faster than the corresponding ER graphs during betweenness and degree attack (figure 6). Betweenness attack was most efficient for all class diagrams. For error, there were no differences between the robustness of the diagrams and their corresponding ER graphs. The results indicate that the class graphs and schema graphs are stable against errors but very fragile against attack, especially betweenness attack.

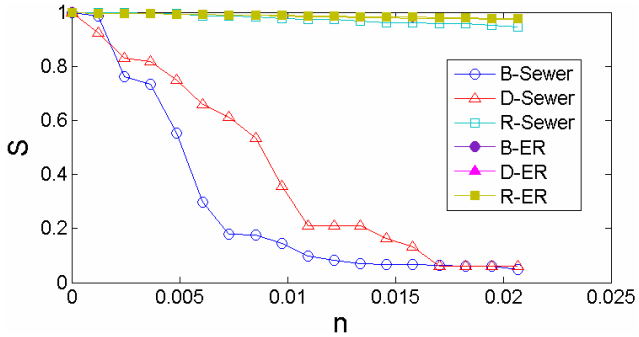

Figure 6: Example of a class graph's error and attack tolerance, in this case the ESRI data model of a Sewer system. Every point represents one vertex removal. S = relative size, n = relative number of removed vertices, B = betweenness attack, D = degree attack, R = error, ER = Erdős–Rényi random graph. The symbols for D-ER and B-ER are hidden under the ones for R-ER and R-Sewer.

CONCLUSIONS AND FUTURE PROSPECTS

The small-world analysis show highly clustered schema graphs and short average paths between the vertices. From the scale-free analysis we can see that many graphs got support for power-laws on parts of their populations, but there were no strict power-laws for all vertices. Nonetheless, there were long-tailed distributions of classes' relationships, attributes and attributes' betweenness. Moreover, selected class graphs and schema graphs were robust against errors but fragile against attack, in which betweenness attack was most efficient.

The above characteristics explain the robustness of the GIS data models, and thus their physical implementation. For example, attributes that connect to other clusters of attributes are sensitive against attacks, and short average paths in the graphs may indicate large spread of false information. Furthermore, when a false update occurs in a randomly chosen table, the information in the whole database will be slightly affected. However, it may be very large impacts due to false updates in one of the few sensitive tables with extremely many attributes or relationships. Moreover, removals of sensitive tables or cell values will quickly disconnect the network. It is important, however, to mention that this study only deals with the linkage of attributes and tables; the spatial and topological dimensions are not studied. For example, a road segment may be considered unimportant with respect to its network position in the database schema; nevertheless, for many users the correctness of this information may be crucial.

It may be necessary to pay attention to sensitive attributes and tables when modeling and managing spatial databases, especially when the databases are open to the public. Examples of actions are additional monitoring, and topological or non-topological constraints on sensitive features. Such strategies may be a complement to trust systems such as the ones from Sayda (2005) and Bishr and Mantelas (2008). To better protect the databases, topological constraints can be added. For instance, when a road touches the boundary of a land parcel, the corresponding nodes are connected to each other in the graph. Such a procedure improves the protection against false alterations or inserts.

BIBLIOGRAPHY

Albert R., Jeong H., and Barabási A., 2000 Error and attack tolerance of complex networks, Nature, vol. 406, no. 6794, pp. 378-82.

Barabási A.-L. and Albert R., 1999 Emergence of scaling in random networks, Science, vol. 286, no. 5439, pp. 509-12.

Bishr M. and Mantelas L., 2008 A trust and reputation model for filtering and classifying knowledge about urban growth, GeoJournal, vol. 72, no. 3, pp. 229-37.

Clauset A., Shalizi C., and Newman M. E. J., 2009 Power-law distributions in empirical data, SIAM Review, vol. 51(4), pp. 661-703.

Concas G., Marchesi M., Pinna S., and Serra N., 2007 Power-laws in a large object-oriented software system, IEEE Transactions on Software Engineering, vol. 33, no. 10, pp. 687-708.

Dezso Z. and Barabási A.-L., 2002 Halting viruses in scale-free networks, Physical Review E, vol. 65, pp. 055103.1-5.

De Nooy W., Mrvar A., and Batagelj V., 2005 Exploratory social network analysis with Pajek, Cambridge University Press New York, USA. ISBN 0521602629.

ESRI., 2009 Data Models – ESRI Support, viewed 1/08 2009, <http://support.esri.com/index.cfm?fa=downloads.datamodels.gateway>.

EuroSDR, 2009 Workshop report, 1th EuroSDR Workshop on Crowd Sourcing for Updating National Databases, <http://www.eurosdr.net/workshops/crowdsourcing_2009/eurosdr_crowdsourcing_2009_report.pdf>

Flanagin A. and Metzger M., 2008 The credibility of volunteered geographic information, GeoJournal, vol. 72, no. 3, pp. 137-48.

Haklay M. and Weber P., 2008 OpenStreetMap: User-Generated Street Maps, IEEE Pervasive Computing, vol. 7, no. 4, pp. 12-8.

INSPIRE., 2009 Consolidated UML Model, viewed 1/08 2009, <https://inspire-twg.jrc.it/inspire-model/index.html?>.

Litchfield, D., Anley, C., Heasman, J. and Grindlay, B., 2005 The Database Hacker's Handbook: Defending Database Servers, John Wiley & Sons. ISBN: 978-0-7645-7801-4

Matyas S., 2007 Collaborative Spatial Data Acquisition–A Spatial and Semantic Data Aggregation Approach, in 10th AGILE International Conference on Geographic Information Science, Aalborg University, Aalborg, Denmark.

Morris S., Morris A., and Barnard K., 2004 Digital trail libraries, in Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, Tuscon, AZ, USA pp. 63-71.

Myers C. R., 2003 Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs, Physical Review E, vol. 68, no. 4, pp. 046116.1-15.

Sayda F., 2005 Involving LBS users in data acquisition and update, in 8th AGILE International Conference on Geographic Information Science, Estoril Congress Center, Estoril, Portugal.

Valverde S. and Solé R., 2007 Hierarchical small worlds in software architecture, Dynamics of Continuous Discrete and Impulsive Systems: Series B; Applications and Algorithms, vol. 14, pp 1-11.

Watts D. and Strogatz S., 1998 Collective dynamics of 'small-world' networks, Nature, vol. 393, pp. 440-2.