Finding Spatial Patterns in Network Data*

Roland Heilmann, Daniel A. Keim **, Christian Panse, Jörn Schneidewind, and Mike Sips

University of Konstanz, Germany
{heilmann,keim,panse,schneidewind,sips}@dbvis.inf.uni-konstanz.de

Abstract. Data on modern networks are massive and are applied in the area of monitoring and analyzing activities at the network element, network-wide, and customer and service levels for a heavily increasing number of networks since network technology is used in almost every personal computer. This results in very large log-files containing important data about the network behavior such as http accesses, e-mail headers, routing information of backbones, firewall alarms, or messages.
Finding interesting patterns in network data is an important task for network analysts and managers to recognize and respond to changing conditions quickly; within minutes when possible. This situation creates new challenges in coping with scale. Firstly, the analysis of the huge amounts (usually tera-bytes) of the ever-growing network data in detail and the extraction of interesting knowledge or general characteristics about the network behavior is a very difficult task. Secondly, in practice, network data with geographic attributes are involved, and it is often important to find network patterns involving geo-spatial locations.
In this paper we address the problem of finding interesting spatial patterns in network data. Sharing ideas and techniques from the pattern visualization and geo–spatial visualization areas can help to solve this problem. We provide some examples for effective visualizations of network data in a important area of application: the analysis of e-mail traffic.

Keywords: Pattern Visualization, Visualization of Geo-Spatial Data, Visualization and Cartography, Spatial Data Mining

1 Analyzing and Visualizing Massive Geo-Spatial Network Data

Nowadays, we have to deal with heavily growing networks since network technology is used in almost every personal computer, cellular phone etc. Data communication networks such as the internet connect millions of computers, cellular

* This work was partially funded by the Information Society Technologies programme of the European Commission, Future and Emerging Technologies under the IST-2001-33058 PANDA project (2001-2004).

** **Correspondence:** Prof. Dr. Daniel A. Keim, Computer Science Institute, Universität Konstanz Fach D78, Universitätsstr. 10, D-78457 Konstanz, Germany, Phone: (+49) 7531 88 3161, Fax: (+49) 7531 88 3062

phones are used in almost every household, and personal communication networks are in commonplace.

Consider, for example, global telecommunication networks and its services. A voice network handles more than 250 million calls per day. Each call can be described by at least one event, yielding a total of tens of gigabytes of data daily. This means, however, that more network data than ever before are available today. An understanding of these data at full scale is of crucial importance for the analysis of the behavior of networks, for managing networks, and improving their performance and reliability from a customers viewpoint. There are many ways to analyze these data, including statistical models and techniques in the area of graph drawing. Unfortunately, these approaches have not kept step with the data volumes and often fall short of providing completely satisfactory results. Often, they are not better than simple visualizations of the structure of the network itself.

Finding interesting patterns in network data is an important task for network analysts and managers to recognize and respond to changing conditions quickly; within minutes when possible. This situation creates new challenges in coping with scale. Pattern visualization techniques have become increasingly important for achieving this goal. Presenting data in an interactive, graphical form often fosters new insights, thereby encouraging the formation and validation of new hypotheses, which lead to better problem-solving and gaining deeper knowledge of the domain.

In practice, network data with geographic attributes are of particular relevance to a large number of applications such as the investigation of the load of a large number of internet nodes at different locations and of the usage of connecting nodes in telephone networks, the detection of geo-related bottlenecks in global telecommunication networks and of geo-specific fraud in networks. Automated data mining algorithms are indispensable for analyzing large geo-spatial data sets, but often fail to provide completely satisfactory results. Usually, interactive data mining based on a synthesis of automatic and visual data mining techniques does not only yield better results, but also a higher degree of user satisfaction and confidence w.r.t. the findings [4, 5]. Although automatic approaches have been developed for mining geo-spatial data, they often do not outperform simple visualizations of the geo-spatial data on a map.

1.1 Network Data

A network consist of nodes, links, and geo-spatial information. Statistics, which may be the raw data or data summaries and may vary over time, are associated with the nodes and the links. Today, the internet is a network of networks. It comprises ten thousands of interconnected networks spanning the globe. The computers which form the internet range from huge mainframes in research establishments to modern PCs in people's home.

All these facts result in massive log-files containing important data about the network behavior. Studying log-files is an important but difficult task. Formally,

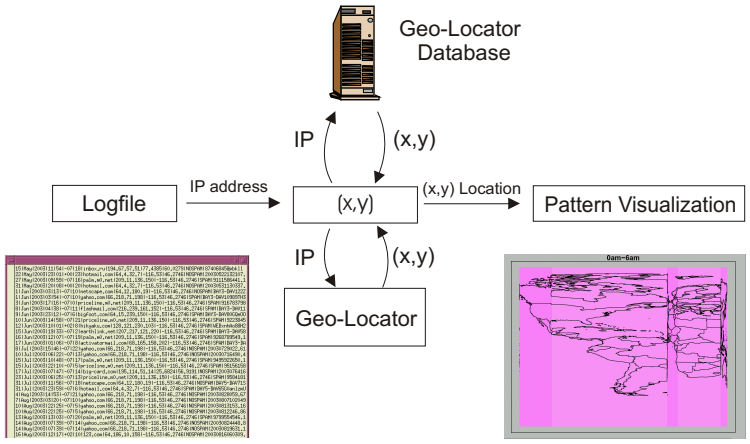

Fig. 1. Basic Idea of the Geo-Locator – From Log-Files to Visual Pattern Analysis

a log-file can be seen as a massive database with network data and possibly geo-spatial information $DB = \{\{\lambda_1, \varphi_1, t_1, a_{1,1}, ..., a_{1,m}\}, ..., \{\lambda_n, \varphi_n, t_n, a_{n,1}, ..., a_{n,m}\}\}$ where $\lambda_i \in [-\pi; \pi]$ represents longitude, $\varphi_i \in [-\pi/2; \pi/2]$ latitude, t_i the time and $a_i, \in \mathbf{R}^m$ the attribute of the i-th data item.

1.2 Geo-Locator: From Log-Files to Visual Pattern Analysis

Nowadays, real network topologies are not publicly available because ISPs generally regard their router-level topologies as confidential. Some ISPs publish simplified topologies on the Web, but these lack router level connectivity and POP structure and may be optimistic or out of date.

The basic idea of our *GeoLocator* is to scan a log-file and convert the monitored ISP network information to geo-spatial coordinates (longitude and latitude). The current main component of our *GeoLocator* is a collection of public available IP databases. These databases have a collection of IP address with their associated geo-spatial location. Figure 1 shows the basic idea of our GeoLocator.

Sadly, the quality and the accuracy of this information is sometimes doubtful and do not scale with the growth of the internet. These problems lead to the development of real ISP mapping method. The basic idea is to measure some network parameter such as the network delay using traceroute and ping method. An example of such a method is rocketfuel [18]. The goal of our future research is to get stable and accurate methods to determine the geo-spatial location of every host.

1.3 Visualization Challenges

The visualization strategy for geo-spatial data is straightforward. The geo-spatial data points described by longitude and latitude are displayed on the 2D euclidian plain using a 2D projection. The two euclidian plain dimensions x and y are directly mapped to the two physical screen dimensions. The resulting visualization depends on the spatial **dimension** or **extent** of the described phenomena and objects. A nice overview can be found in [13].

The task of geo visualization is to view geo-spatial data and to give insight into geo-processes and geo-phenomena. One of the challenges today is to find out how to deploy efficient visualization strategies for the representation of geo-spatial data. The major challenges to efficiently visualize geo-spatial data are

Highly non-uniform distribution

Since the geo-spatial locations of the data are highly non-uniformly distributed in a plane, however, the display will usually be sparsely populated in some regions while, in other regions of the display, a high degree of overplotting occurs. There are several approaches to cope with dense geo-spatial data [7, 11, 12].

Overlap/overplot problem

Nowadays, **Network Maps** are widely used. Some approaches only display the structure of networks (usually modeled as graphs) to interpret and understand the general behavior and structure of networks. The goal is to find a good geometric representation of the network on a map. However, the visualization of large networks on maps leads to the problem of overlapping or overplotting w.r.t. line segments in dense areas.

Highlighting patterns of high importance

Two types of maps, called **Thematic Map** and **Choropleth Map**, are used in cartography and GIS-systems. Thematic Maps are used to emphasize the spatial distribution of one or more geographic attributes. Popular thematic maps are the Choropleth Map (Greek: choro = area, pleth = value), in which enumeration units or data collection units are shaded to represent different magnitudes of a variable. Often, the statistical values are encoded by colored regions on the map. For both types of maps, high values are often concentrated in densely populated areas, and low statistical values are spread over sparsely populated areas. Therefore, these maps tend to highlight patterns in large areas, which, however, may be of low importance.

High dimensionality

Usually network elements contain a high number of working parameters e.g. the load of the system, the number of network packages and of faulty accesses,

or, for the higher OSI-layer, the number of e-mails arrived within a certain time slot. In this paper, we are going to visualize the score for each sender of an e-mail determined by a spam-filter [15] as SPAM classified e-mail. Of course, these data are characterized by a high number of dimensions and hence are difficult to analyze.

1.4 Connecting Pattern and Geo-Spatial Visualizations

The exploration of large network data sets in order to find interesting spatial patterns is an important but difficult problem. Our observation is, that single techniques from the pattern or geo-spatial visualization do not solve the problem. To find spatial patterns, our basic idea is to tightly integrate pattern visualization techniques with traditional information visualization techniques (such as parallel coordinates) and advanced geo-spatial visualization techniques.

1.5 Visual Data Exploration

For the effectiveness of the analysis of large geo-spatial network data sets and for the extraction of interesting patterns, it is important to include humans in the data exploration process, combining their flexibility, creativity, and domain knowledge with the storage capacity and computational power of current computer systems. Visual data exploration often follows a three step process: *Overview, zoom and filter, and details-on-demand* which has been called the Information Seeking Mantra [17]. In other words, in the exploratory data analysis (EDA) of a data set, an analyst firstly tries to get an overview. This may reveal potentially interesting patterns or certain subsets of the data that deserve further investigation. The analyst then focuses on one or more of them by inspecting the data in more detail.

2 The Scope of Cartogram Techniques for Finding Network Patterns

A cartogram can be seen as a generalization of a familiar land-covering choropleth map. Here, an arbitrary parameter vector contains the intended sizes of the cartogram's regions Hence, a familiar land-covering choropleth map is simply a cartogram with sizes proportional to land area. In addition to the classical applications mentioned above, a key motivation for cartograms as a general information visualization technique is to have a method for trading off shape and area adjustments. For example, in a conventional choropleth map, high values are often concentrated in highly populated areas, and low values may be spread across sparsely populated areas. Therefore, such maps tend to highlight patterns in less dense areas where few people live. In contrast, cartograms display areas in relation to an additional parameter, such as population. Patterns may then be displayed in proportion to this parameter (e.g. the number of people involved) instead of the raw size of the area involved. Applications include population

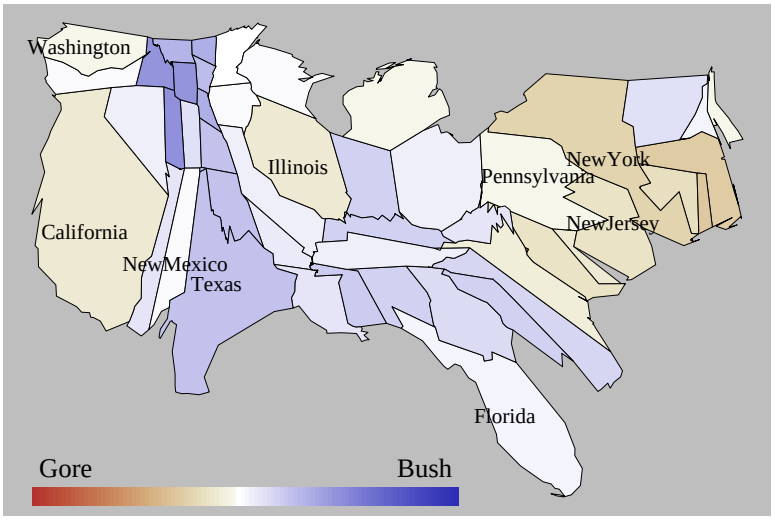

Fig. 2. The Figure displays the U.S. state population cartogram with the presidential election result of 2000. The area of the states in the cartograms correspond to the electoral voters and the color corresponds to the percentage of the votes. A bipolar colormap depicts which candidate has won each state.

demographics [8, 10, 16, 20], election results [14], and epidemiology [3]. Because cartograms are difficult to make by hand, the study of automated methods is of interest [2].

The basic idea is to compute a distortion of a map w.r.t. to a statistical value associated with the network, e.g. the idle time, overload time, or the number of hosts and other geo-related information such as census demographics (population, household income). The distortion is coping with dense regions and virtual empty regions on maps and enables us to show more details in potentially interesting regions. For example, we can illustrate the connectivity between different provider servers in the big cities of the United States.

3 The Scope of Traditional Information Visualization Techniques for Finding Network Patterns

3.1 Parallel Coordinates

Almost all network data sets consist of more than three attributes and therefore do not allow a simple visualization by 2-dimensional or 3-dimensional plots to find interesting patterns. An example for a technique which allows the visualization of multidimensional data is the Parallel Coordinates Technique [6] (see Figure 3). Parallel Coordinates display each multi-dimensional data item as a set of line segments that intersect each of the parallel axes at the position corresponding to the data value for the respective dimension.

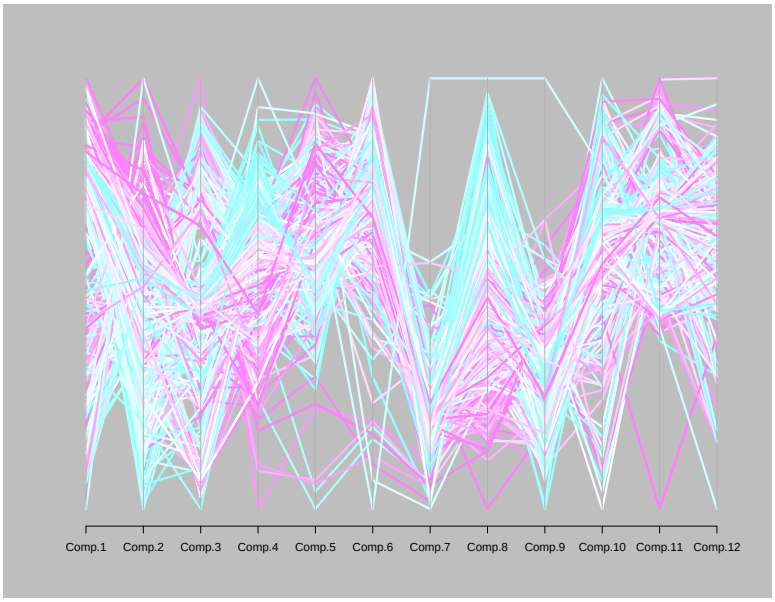

Fig. 3. The figure displays a sample of 500 SPAM e-mails arrived in 2003. The bipolar colormap encodes the Greenwich Mean Time (GMT) and the axis of the parallel coordinate plot shows 12 attributes of SPAM data (x,y,time zone,Hour,Attr1,Attr2,...)

The parallel coordinate techniques can be used to emphasize network data in such way, that the axis represent longitude, latitude, and other attributes. In our e-mail traffic analysis example, the parallel coordinates show 12 attributes of SPAM data (x,y,time zone,Hour,Attr1,Attr2,...)

3.2 Level-plot Technique

A level plot (see [1]) is a visualization technique for the representation of 3-dimensional data tuples $t = (t_1, t_2, t_3)$ by plotting constant z slices, called contours, on a 2-dimensional format. That is, given a value of z, lines are drawn for connecting the (x,y) coordinates where that z value occurs.

The level plot technique can be used to emphasize network data by associating the latitude attribute with the vertical axis, the longitude attribute with the horizontal axis, and a network variable with the iso-response values. The two geo-spatial attributes are usually restricted to a regular grid. In our e-mail traffic analysis, the level plots illustrate the time zone (represented by the horizontal axis), the 24 hours of a day (represented by the vertical axis), and the number of e-mails arrived.

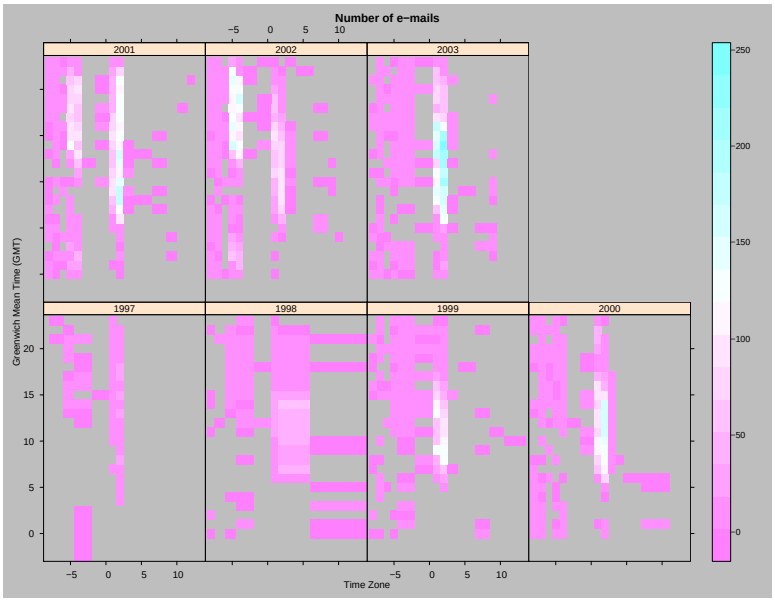

Fig. 4. 2D Level Plot of the e-mail data – The figure shows the increasing amount of the e-mail traffic as well as the spread of the e-mail communication over the years.

4 E-Mail Traffic Analysis

One of the first applications of the internet was the electronic mailing (e-mail)[19]. Messages are sent between users of computer systems to different places all over the world and the computer systems are used to hold and transport the messages. There are several advantages of electronic mailing. In particular, it is a fast, cheap, and comfortable communication method. The number of internet users has increased exponentially and therefore more and more people are able to send and receive e-mails.

4.1 Interesting Patterns of E-Mail Communication

The search for patterns of e-mail communication is lead by the following questions:

- Is there an increase of traffic?
- Where do the e-mails come from? Are there dead areas?
- When and from where do e-mails arrive? (This is important for the prevention of bottlenecks, e.g. a spam-filter needs some computational resources.)
- Are there any patterns classifying SPAM and NO-SPAM e-mails?
- Is there a spread of e-mail communication?
- Is there a certain dynamics w.r.t. the locations of the senders?

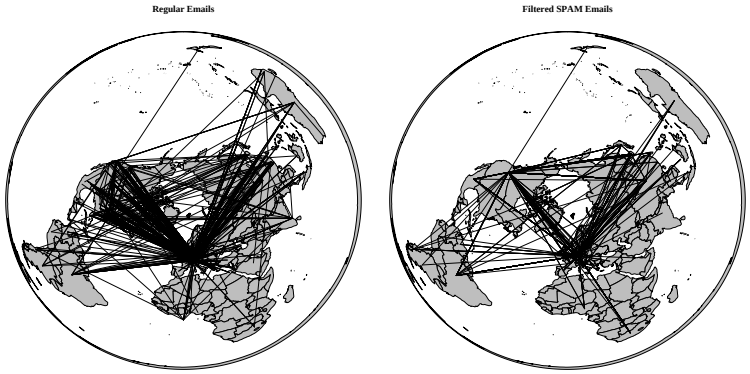

Fig. 5. The figures display the worldwide e-mail routes of one of our *IMAP* users. The *IMAP* server is located in Konstanz, Germany (*37 41.0N / 09 08.3E*). In our department, SPAM hits one fourth of our e-mail traffic.

4.2 Pattern 1: E-Mail Paths

An interesting approach is to visualize the path of SPAM e-mails to see interesting patterns and behavior. The path information can be derived from the e-mail headers. Figure 5 shows the regular and SPAM e-mails path of one of the authors. The e-mail paths displayed in the plot have been stored since 2000. Each spatial location corresponds to a computer system from which the e-mails were sent. Each line segment represents the path of an e-mail message between two computer systems. The picture on the right displays only SPAM e-mails. Visualizing e-mail paths may help to find important patterns of the e-mail traffic.

4.3 Pattern 2: Spread of E-Mail Communication

Figure 7 the spread of e-mail communication of one of the authors. Once can clearly see, that the communication of this author changes. The explanation of this phenomenon is easy. The author had a research year in the United States.

Since the e-mail header contains the time zone, we can easily distort a familiar land-covering map in such a way that the area of the map region is proportional to the number of e-mails being sent.

A cartogram algorithm can be run using the spatial information and the number of e-mails. One way of doing this has been introduced by an american geographer [21]. The distortion of the mesh can be computed by solving the following integral $x = \int_{-\pi}^{\lambda} d(\lambda) d\lambda$. A more efficient algorithm, which is based on histograms, has been introduced by [9].

Figure 6 shows clearly the different office times of our partner organization around the world. A regular pattern reflects clearly the different time zones.

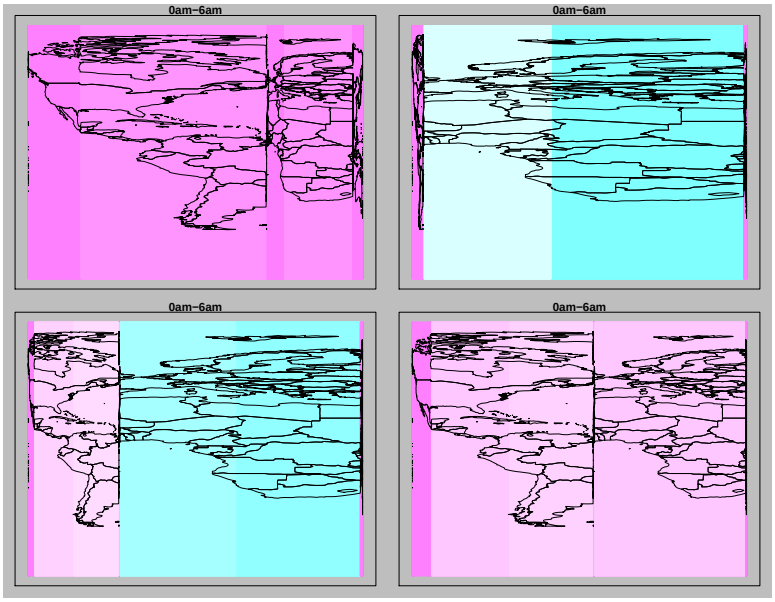

Fig. 6. HistoScale cartogram in longitude direction – The picture shows e-mail volume patterns at four different points in time (0am–6am, 6am–12am, 12am–6pm, 6pm–0am GMT) during one day. The area of a region represents the relative number of e-mails and the color indicates the absolute number of e-mails arrived.

4.4 Patterns Example 3: Spam vs. No-Spam E-Mails

In the meantime, corporate and university networks have become increasingly clogged by e-mail pitches for pornography, money-making schemes, and products. In our department, about one fourth of our e-mail traffic are SPAM's. In 2002, we had one SPAM for every 20 legitimate e-mail messages; today the ratio is closer to one in four.

Using Anti-SPAM software on specialized servers can discern SPAM from legitimate e-mails. The software can also upload potentially new forms of SPAM for analysis, and develop recognition algorithms to identify and filter new types of SPAM e-mails. The ultimate goal is to bring the power of visualization technology to every desktop to allow a better, faster, and more intuitive exploration of e-mail resources, to give the user insight into the filter step mechanism and the public a powerful tool to optimize the anti-spam software. This will not only be valuable in an economic sense but will also stimulate and delight the user.

5 Conclusion and Future Work

One of the challenges today is to find out how to deploy efficient visualization strategies to represent geo-spatial data. Among the efficient strategies to represent geo-spatial data and to interact with that data, linked combination of maps

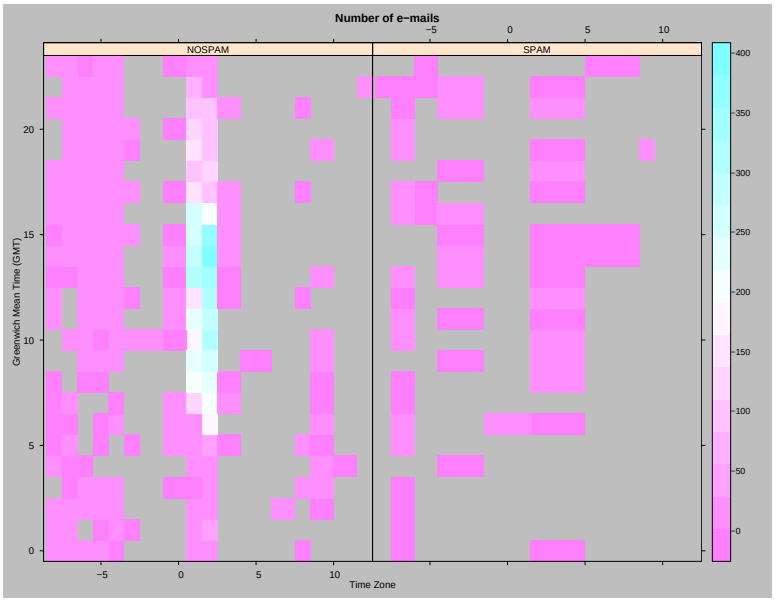

Fig. 7. 2D Level Plot of the e-mail data – The figure shows the pattern of e-mails which has been classified as SPAM and NO-SPAM over the years.

with information visualization techniques are avenues for future work. The ultimate goal is to bring the power of visualization technology to every desktop to allow a better, faster and more intuitive exploration of very large data resources. In future work, we expect to investigate:

- related approaches for visualizing large geographical data sets
- user studies to identify strengths and weaknesses of these approaches
- developing new mapping methods in our *GeoLocator* to improve the quality of the spatial location mapping
- visualization interface for spam-filter tuning

Exploring and analyzing network data is an important research area. In practice, almost every network data sources provide spatial attributes. In this article, we describe an overview of methods for visualizing massive spatial network data to find interesting spatial network patterns. Sharing ideas and techniques from the pattern visualization and geo–spatial visualization areas can help to solve this problem. We provided some examples for effective visualizations of network data in a important area of application: the analysis of e-mail traffic.

References

1. W. S. Cleveland. *Visualizing Data*. Hobart Press, Summit, New Jersey, U.S.A, 1st edition, 1993. http://cm.bell-labs.com/cm/ms/departments/sia/wsc/.

2. B. D. Dent. *Cartography: Thematic Map Design, 4th Ed., Chapter 10.* William C. Brown, Dubuque, IA, 1996.
3. S. Gusein-Zade and V. Tikunov. Map transformations. *Geography Review*, 9(1):19–23, 1995.
4. J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2001.
5. D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* MIT Press, 2001.
6. A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proc. Visualization 90, San Francisco, CA*, pages 361–370, 1990.
7. D. A. Keim and A. Herrmann. The gridfit algorithm: An efficient and effective approach to visualizing large amounts of spatial data. *Proc. IEEE Visualization, Research Triangle Park, NC*, pages 181–188, 1998.
8. D. A. Keim, S. C. North, and C. Panse. CartoDraw: A fast algorithm for generating contiguous cartograms. *Transactions on Visualization and Computer Graphics*, 10(1):95–110, January/February 2004.
9. D. A. Keim, S. C. North, C. Panse, M. Schäfer, and M. Sips. HistoScale: An efficient approach for computing pseudo-cartograms. In *IEEE Visualization 2003 DVD-ROM, Seattle, Washington, USA*, pages 28–29, October 2003. IEEE Catalog Number 03CG37496D, ISBN 0-7803-8121-1.
10. D. A. Keim, S. C. North, C. Panse, and J. Schneidewind. Visualizing geographic information: VisualPoints vs CartoDraw. *Palgrave Macmillan – Information Visualization*, 2(1):58–67, March 2003.
11. D. A. Keim, S. C. North, C. Panse, and M. Sips. PixelMaps: A new visual data mining approach for analyzing large spatial data sets. In *The Third IEEE International Conference on Data Mining (ICDM03), Melbourne, Florida, USA*, November 2003.
12. D. A. Keim, C. Panse, J. Schneidewind, and M. Sips. Geo-spatial data viewer: From familiar land-covering to arbitrary distorted geo-spatial quadtree maps. In *WSCG 2004, The 12-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, February 2004.
13. D. A. Keim, C. Panse, and M. Sips. Visual data mining of large spatial data sets. In *Databases in Networked Information Systems – Third International Workshop, DNIS 2003, Aizu, Japan*, pages 33–36, September 2003. ISBN: 3-540-20111-4.
14. C. J. Kocmoud and D. H. House. Continuous cartogram construction. In *IEEE Visualization, Research Triangle Park, NC*, pages 197–204, 1998.
15. J. Mason. Spamassassin, Dec. 2003. http://bugzilla.spamassassin.org/.
16. E. Raisz. *Principles of Cartography.* McGraw-Hill, New York, 1962.
17. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Visual Languages*, pages 336–343, 1996.
18. N. Spring, R. Mahajan, and D. Wetherall. Measuring isp topologies with rocketfuel. In *Proc. SIGCOMM*, 2002. http://www.acm.org/sigcomm/sigcomm2002/papers/rocketfuel.pdf.
19. W. R. Stevens. *TCP/IP Illustrated Volum I. The Protocols.* Addison Wesley Longman, 1994.
20. W. Tobler. Cartograms and cartosplines. *Proceedings of the 1976 Workshop on Automated Cartography and Epidemiology*, pages 53–58, 1976.
21. W. Tobler. Pseudo-cartograms. *The American Cartographer*, 13(1):43–40, 1986.