

Information Visualization Techniques for Big Data

Analytics using Heterogeneous Data in Spatiotemporal Domains

William H. Hsu

Kansas State University

This article presents challenges and recommended practices for visualizing data about phenomena that are observed or simulated across space and time. Some data may be collected for the express purpose of answering questions through quantitative analysis and simulation, especially about future occurrences or continuations of the phenomena – that is, prediction. In this case, analytical computations may serve two purposes: to prepare the data for presentation and to answer questions by producing information, especially an informative model, that can also be visualized. These purposes may have significant overlap. Thus, the focus of the chapter will be about analytical techniques for visual display of quantitative data and information that scale up to large data sets. It begins by surveying trends in educational and scientific use of visualization, and reviewing taxonomies of data to be visualized. Next, it reviews aspects of spatiotemporal data that pose challenges, such as heterogeneity and scale, along with techniques for dealing specifically with geospatial data and text. An exploration of concrete applications then follows. Finally, tenets of information visualization design, put forward by Tufte and other experts on data representation and presentation, are considered in the context of analytical applications for heterogeneous data in spatiotemporal domains.

1. Trends in Data Visualization

1.1 Learning and Analytics Tasks

This section provides a brief history of information visualization for educational and scientific applications, followed by a survey of challenges and tools encountered in visualizing data.

1.1.1 Brief History of Prediction

Information visualization is the study of (interactive) visual representations of abstract data to reinforce human cognition. ("Information visualization", Wikipedia, 2013) Abstract data include both numerical data such as geospatial locations or other physical measurements, and non-numerical data such as text. However, information visualization differs from scientific visualization: Munzner (2008) advises using the term “*infovis* (for information visualization) when the spatial representation is **chosen**”, and “*scivis* (scientific visualization) when the spatial representation is **given**”. According to Friendly (2009), scientific visualization is primarily concerned with the “visualization of three-dimensional phenomena (architectural, meteorological, medical, biological, *etc.*), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth, perhaps with a dynamic (time) component”.

Input data for visualization includes observational data, collected for the express purpose of answering questions through quantitative analysis, and simulated data, which is generated using a mathematical model. One particular type of simulated data consists of future occurrences or continuations of the phenomena – that is, *prediction*. Modeling of phenomena for the purpose of forecasting predates computational realization of the methods used, including econometrics (Frisch, 1929) and statistical hypothesis testing (Neyman & Pearson, 1933; Fisher, 1935). Some of the earliest methods for nonlinear time series prediction were extrapolation, interpolation, and smoothing methods derived by: Wiener (1949); Brown (1956), Holt (1957), and Winters (1960); and Box and Jenkins (1970). These contributions comprise fundamental representation and estimation methods that underlie spectral analysis approaches to signal identification, including autoregressive moving average (ARMA) process models.

Although specifically geared towards time series and geospatial data, the visualization approaches covered in this chapter are generally applicable to a variety of data sets and to the behavior and output of many type of machine learning algorithms. Hall *et al.* (2009) give a much more detailed catalogue of the models and algorithms implemented in the *Waikato Environment for Knowledge Analysis (WEKA)*, to which we refer the interested reader. Predictive visualization, the aspect of information visualization that especially focuses on the continuation of time series beyond historical observations, often poses questions of evaluation using previous unseen data. Watson and Wixom (2009) describe architectures for this type of analytical modeling, among others, in the domain of *business intelligence (BI)*. Business intelligence comprises theories, methodologies, and technologies that serve to transform raw data into

information for business decision making. Similar uses of prediction and visualization can be found in most fields where sequences and time series are observed as signals. This includes neuroscience, where such measurements are fundamental, giving rise to the work of scientists such as Elger and Lehnertz (1998).

1.1.2 Challenges of Heterogeneity in Big Data

The term *heterogeneous data* refers to variables that are fundamentally diverse in character, particularly their source and means of acquisition. One of the key challenges to working with heterogeneous data is that multiple dimensions and a very high volume of data may result from differences in data provenance (origin and preprocessing history). This issue gives rise to the problem of designing visual representations that can consistently support the display of such data. Heer, Kong, and Agrawala (2009) present adjustable parameters such as layering and chart sizing, and discuss the perceptual effects of introducing such degrees of freedom. Monmonier (1990) discusses methods from statistics for coping with the additional technical challenge of working with spatial data over time.

A further challenge is that of *big data*, a generic term used to refer to data of high complexity (especially intrinsic complexity), the value that can be derived from the data using various analytical methods, and longitudinal information. *Mike 2.0* (2013) notes furthermore that big data does not necessarily mean extremely large in size, if the other aspects of analytical task complexity are high. Kumar *et al.* (2005) present some specific approaches to scaling up visualizations to large data sets, such as introducing trend lines, quartile boxes, *etc.* The provenance of big data, like that of visualization data, plays an important role.

Cox and Ellsworth (1997) discuss facets of complexity, distinguishing between *big data collections*, which “typically arise in fields with acquired data, as from remote sensors and satellite imaging”, and *big data objects*, which “typically are the result of large-scale simulations in such areas as Computational Fluid Dynamics (CFD), Structural Analysis, Weather Modeling, and Astrophysics”. With respect to collections, they note that data within them often:

1. Tend to be “*distributed* among multiple sites”
2. Are stored within “collections of *heterogeneous databases* (each the repository for data acquired or processed at that site)”
3. Have “generally incompatible data interfaces and *representations*” and are “generally *not self-describing*”
4. Admit “no platform-independent definition of the data types in the underlying data, and the relationship between them” (*e.g.*, no well-defined relational *data model*)
5. Do not have “*meta-data* that facilitate discovery and use” stored with them (*e.g.*, describing where and when they were collected, what calibration was applied, what their units, are *etc.*).

Such metadata “may (and probably should) also include compressed and/or condensed representations of the underlying data” to “enable browsing of a large collection”

6. Are difficult to retrieve or *locate*: “Visualization can serve an important role in data location, in particular by compressing summary information in a format that can be visually understood quickly”
7. Admit large *storage requirements*, “requiring partitioning between disk and tape” (or between secondary and tertiary storage in general).
8. Admit “*poor locality* in the queries for the data (since, for example, requests may be to diverse variables measured at arbitrary times). Any particular request is more likely to require data from tape than from disk”, meaning that responses may need to be freshly computed or observed rather than retrieved from archives.
9. Require “raw *bandwidths* required to satisfy requests [that] may be quite large for any actively used collection. These include bandwidths from tape to disk, from disk to memory, from memory to network, and across the network.”

An independent set of characteristics for data objects is that they present challenges to:

1. *Data modeling*: multi-dimensionality, lack of a standardized representation and interfaces for structured data (as above). “As a result, visualization codes typically must handle multiple file formats and data representations.”
2. *Data model evolution*: dynamicity of representations including relational schemas and formats, e.g., alternative grid structures in Computational Fluid Dynamics.

Specific additional challenges for big data objects are:

3. *Data management*: “there is generally not a clean division between the data models (where they exist) and data management”. This requires “special handling of data sets that do not fit in main memory”. In more current terminology, there is a resultant lack of good data description languages as used in semistructured data (especially XML, which includes a standard for document type definitions).
4. “*Data too big to be memory-resident*”: “Often a single data object does not fit in main memory.” CFD applications earlier than the late 1990s produced time-varying data objects on the order of 10 gigabytes (Gb) and those of 1997 were 100-200Gb. As of 2013, this typical figure has grown by three decimal orders of magnitude to 300+ terabytes (Tb).
5. “*Data too big for local disk* (perhaps even for remote disk)”: this meant hundreds of gigabytes (more than 650, approximately 7×10^{11} bytes) in 1997 and hundreds of terabytes to petabytes (10^{14} - 10^{15} bytes) at the time of this writing.
6. “*Bandwidth and latency*”: the need, resulting in part from the preceding challenges, to find alternatives to secondary storage-based virtual memory with a high bandwidth, low latency data pathway between the data store and main memory.

Steele and Illinsky (2010) survey and catalog the above challenges in several modern domains, viewing these requirements in the context of functional and aesthetic criteria. Cuzzocrea, Song, and Davis (2011) further discuss the challenges inherent in dealing with large-scale multidimensional data analytics, especially moving towards NoSQL query languages and Map-Reduce platforms such as Hadoop. A 2011 panel at the DOLAP conference on online analytical processing that discussed visualization issues, among other aspects of analytics, advocated an emphasis on decision support context and the ability to explore multidimensional spaces.

1.1.3 Current and Emerging Technologies for Visualizing Data

Goldstein and Roth (1994) introduced a hierarchical framework for data browsing, comprising an interface mechanism they called the Aggregate Manipulator (AM). Combined with the Dynamic Query (DQ) mechanism of Ahlberg, Williamson, and Shneiderman (1992), this system has served as a widely adopted paradigm, much like the model-view-controller framework of Reenskaug (1979) among graphical user interface (GUI) developers.

The term *visual analytics* was coined to describe “an outgrowth of the fields of information visualization and scientific visualization that focuses on analytical reasoning facilitated by interactive visual interfaces”. (Kovalerchuk & Schwing, 2004; “visual analytics”, Wikipedia, 2013) Active clientele of this technology include developers of business decision support technology. In particular, the use of visualization in e-commerce has recently risen, exemplified by *customer relationship management (CRM)* and business intelligence tools that incorporate data mining-based analytics and visualization. (Cadez *et al.*, 2000) Cannataro and Talia (2003) present an architecture they call the *knowledge grid* that implements a scalable infrastructure for widespread access to online services as visualization or analytics.

Meanwhile, Keim *et al.* (2006) note in a survey article that visual analytics “combines strengths from information analytics, geospatial analytics, scientific analytics, statistical analytics, knowledge discovery, data management and knowledge representation, presentation, production and dissemination, cognition, perception, and interaction”. They also discuss the following specific open problems of visualization in detail:

1. Synthesis of heterogeneous types of data
2. Human interpretability requirements
3. Data transformations and integration constraints
4. "Problem solving, decision science, and human information discourse"
5. Semantics for future analytical tasks and decision-centered visualization (a key feature of user modeling, adaptation, and personalization)
6. Promotion of user acceptability by “addressing the challenges of new systems to be pervasive, embedded, nomadic, adaptable, powerful, intentional and eternal” *cf.* MIT Project Oxygen (2004)

7. Integration into intelligent user interfaces
8. Derivation of objective evaluation criteria

Summarizing common design tenets that for the above requirements, Keim *et al.* derive the following “visual analytics mantra”:

“Analyse First –
Show the Important –
Zoom, Filter and Analyse Further –
Details on Demand”

1.2 Taxonomies: Scientific, Data, and Information Visualization

As Munzner (2008) notes, the term *scientific visualization* is distinct from *information visualization*, in that scientific visualization aims at rendering phenomena, to a given spatial representation. That is, scientific observations entail a set of requirements and a specification – sometimes constrained by the instrumentation or use cases, such as geographic information systems (GIS) software. Information visualization typically leaves more degrees of representational freedom to the designer, having performance criteria related to usability and graphical excellence, rather than using predetermined constraints. Finally, *data visualization* gives some of the design freedom of information visualization but involves abstraction of raw observation data to some schematic form, and generally involves less preprocessing, analysis, and annotation of the source data.

2. Aspects of Spatiotemporal Data

We now discuss some aspects of data that include a spatial (“length-scale”) or temporal (“time-scale”) measure. These spatiotemporal data are most frequently presented in map and timeline form, but can be captured and represented in other ways.

2.1 Space and Time

A *spatiotemporal* sequence is a data set whose points are ordered by location and time. Spatiotemporal sequences arise in analytical applications such as time series prediction and monitoring, sensor integration, and multimodal human-computer intelligent interaction. Learning to classify time series is an important capability of intelligent systems for such applications. Many problems and types of knowledge in intelligent reasoning with time series, such as diagnostic monitoring, prediction (or *forecasting*), and

control automation can be represented as classification. The machine learning task that yields the capability of forecasting is described using the term *modeling*, while higher-level descriptive characterization of the time series is referred to as *understanding*. Spatiotemporal physical data can include scientific and engineering measurements at microscopic or sub-microscopic orders of spatial magnitude, and similarly small orders of temporal magnitude. Some domains such as genome biology and particle physics admit both information visualization and scientific visualization. Infovis may be feasible using descriptive statistics or summary attributes of an entire entity such as the genome of an entire organism, whereas scivis may use a standard visual metaphor such as a ladder depiction of the DNA double helix or the colored-ball molecular renderings produced by *RasMol* (2013).

2.1.1 Capturing and Creating Spatial Data: Georeferencing

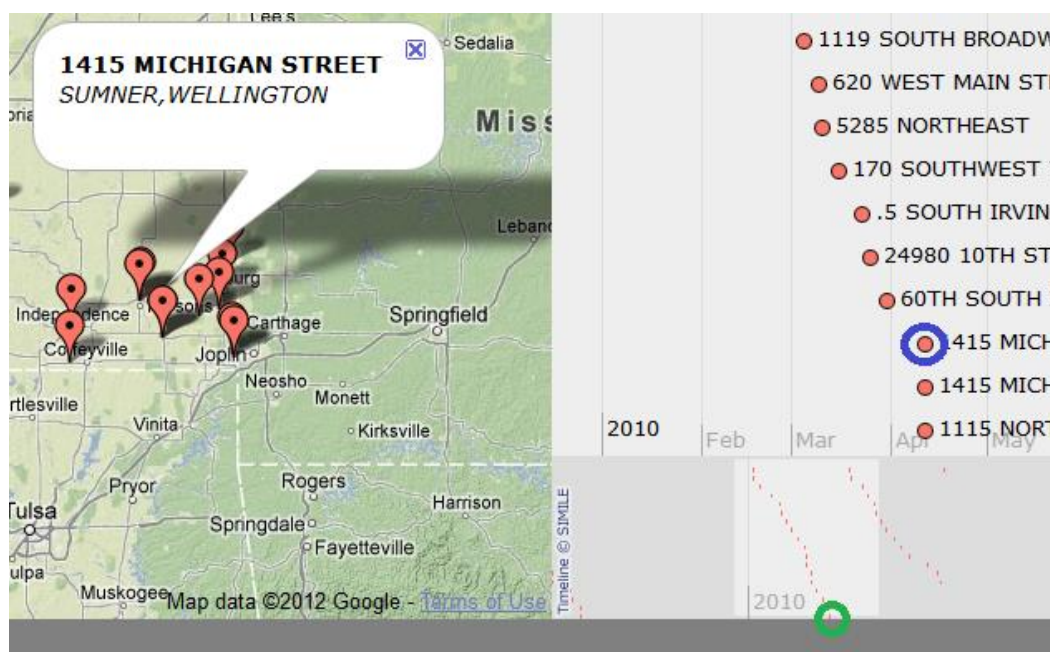


Figure 1. Map and timeline visualization of meth lab seizure events (2004-2011) using Google Maps and MIT SIMILE. Seizures from the first half of 2010 are depicted, with one event selected.

Extracting spatially-referenced events, even using structured data sources, entails a straightforward but data-intensive georeferencing task: looking up the coordinates (latitude and longitude) of locations where events are reported to have occurred. These may be identified as buildings or landmarks. The resulting coordinates are placed into a spatial database management system (SDBMS) for visualization using software libraries and services, as shown in Figure 1. Such a system uses two access layers with a unified representation and geographic information system (GIS) data model: one based on a geospatial markup language such as Google's Keyhole Markup Language (KML) and a file-based application programmer interface (API), while the second layer is based on a server-side interface to a relational database implementing the schema.

2.1.2 Capturing Temporal Data

Figure 1 also depicts the data integration between the map and timeline visualization subsystems. The seizure event in April, 2010 is represented on the map by a pop-up note, on the monthly scale timeline (upper right) by a circled dot, and on the yearly scale timeline (lower right) by a circled point.

2.1.3 Predictive Analytics: Forecasting, Modeling, and Understanding

A *spatiotemporal* sequence is a data set whose points are ordered by location and time. Spatiotemporal sequences arise in analytical applications such as time series prediction and monitoring, sensor integration, and multimodal human-computer intelligent interaction. Learning to classify time series is an important capability of intelligent systems for such applications. Many problems and types of knowledge in intelligent reasoning with time series, such as diagnostic monitoring, prediction (or *forecasting*), and control automation can be represented as classification. The machine learning task that yields the capability of forecasting is described using the term *modeling*, while higher-level descriptive characterization of the time series is referred to as *understanding*.

2.2 Big Data: Multidimensionality and the Need for Scalability

Visualization can help make both multi-dimensional data and the results of an analytical or optimization algorithm more accessible. Oftentimes, it is used to project data and objects into a lower dimension space to make the data set more manageable. It can also help the user grasp the broad characteristics of the data. The complexity challenges of big data applications underscore a need to limit the growth of the complexity of visualization. Like analytics algorithms for text and numerical data, visualization algorithms must generally have worst-case asymptotic running time linear in the problem size.

2.3 Heterogeneity

Software tools for visualization also play a key role in coping with heterogeneity, as they allow disparate aspects to be hidden. This is comparable to projecting only the shared attributes of two tables in a relational database, corresponding to the intersection of their schemas. The key challenge in visualization is accounting for relevant and interesting differences in disparate parts of a data set in order to reveal variation in the data to the user. Some text domains exhibit *semantic heterogeneity*, a disparity in the sense or meaning of terms used. Coping with this is an open research problem in general, but presents a specific set of issues for visualization, in that showing the differences (and similarities) does not necessarily confer an ability to disambiguate a model manually – *i.e.*, based on inspection.

3. Techniques for Analyzing and Learning from Big Data

3.1 Thematic Maps

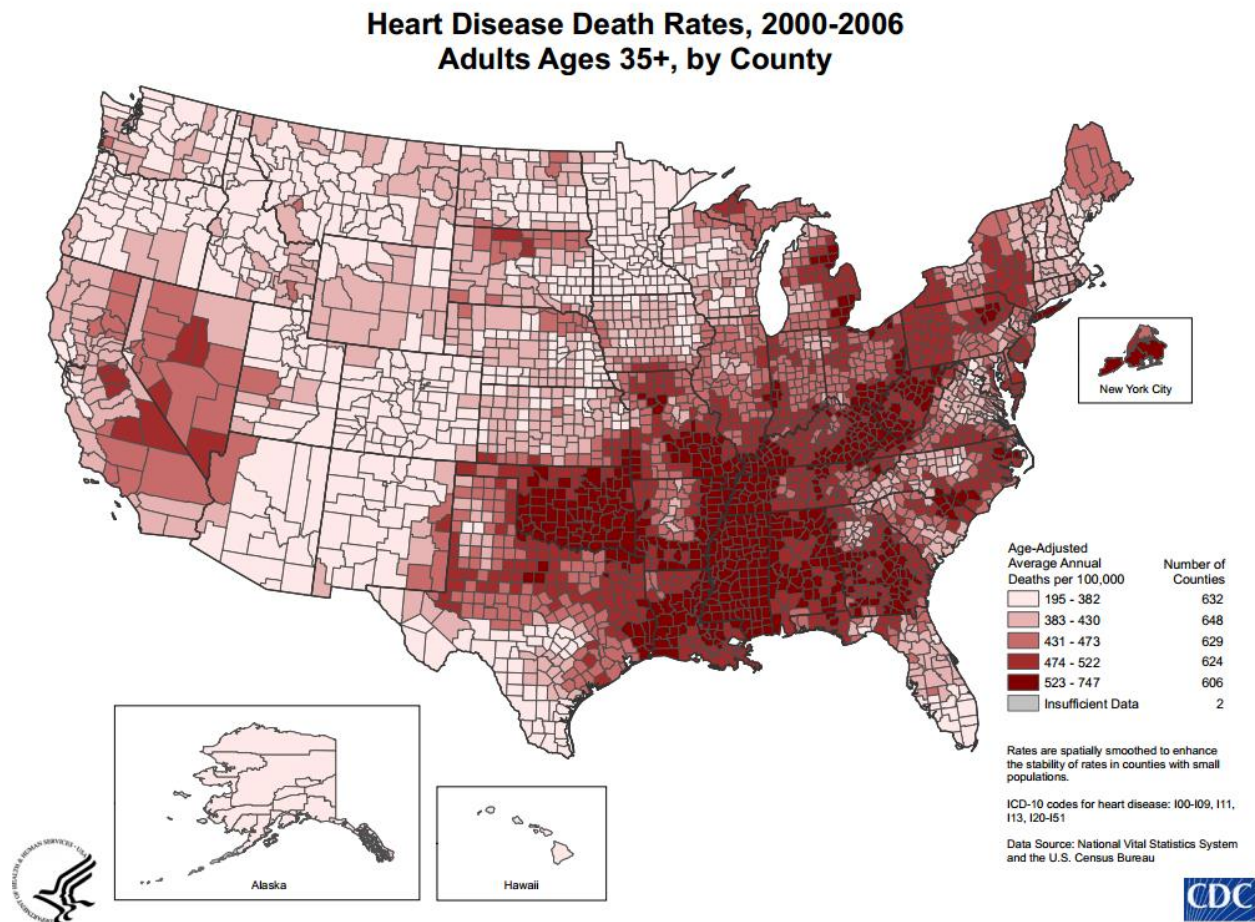


Figure 2. Choropleth map of heart disease death rates in the USA by county, 2000-2006. (CDC, 2011)

The object of thematic mapping is to depict phenomena and trends in a geospatial context. Toward this end, most frameworks allow data to be superimposed or blended with a color or grayscale map that depicts region boundaries, baseline, *etc.* This includes choropleth maps with dynamically computable color palettes.

Figure 2 is a choropleth map depicting the rate of death attributed to heart disease by county in the USA from 2000 through 2006. This map, which was published by the Centers for Disease Control, takes inter-county population distribution (but not intra-county population distribution) into account by normalizing to a death rate per 100,000 individuals. For information retrieval applications, it does not provide any drill-down interface *cf.* *HealthMap* or similar event visualization services. One of the reasons for the development of the geospatial visualization components of an event monitoring system is to facilitate information retrieval and multimodal information access using well-established visualization techniques such as thematic mapping and small multiples.

3.2 Text Analytics: Visualizing Natural Language

3.2.1 Text Annotation and Markup

Named entity recognition (NER) seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, etc. The list of categories can be extended to include disease names in the biomedical paradigm. Most NER research consists of taking an unannotated, unstructured block of text such as:

Chancellor Ridcully hired three new faculty to work at Unseen University in Ankh-Morpork.

and producing an annotated output such as:

{Chancellor Ridcully}\PERSON hired {three new faculty}\QUANTITY to work at {Unseen University}\ORGANIZATION in {Ankh-Morpork}\LOCATION.

In addition to people, organizations, and locations, quantities and dates are often included in NER systems. Quantities are particularly relevant to textual IE in the domain of epidemiology because of the many units of measurement (length, area, volume, mass and weight, monetary units, time, etc.) that are required and the variety of agents (infectious agents and host organisms) that may be named.

Research indicates that NER systems developed for one domain do not typically perform well on other domains. Early work in NER systems in the 1990s was aimed primarily at extraction from journalistic articles. Attention then turned to processing of military dispatches and reports. Since about 1998, there has been a great deal of interest in entity identification in the molecular biology, bioinformatics, and medical natural language processing communities. The most common entity of interest in that domain has been names of genes and gene products. The Stanford Named Entity Recognizer (Stanford NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. The software provides a general (arbitrary order) implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition. Included are a good 3-class (PERSON, ORGANIZATION, LOCATION) named entity recognizer for English (in versions with and without additional distributional similarity features) and another pair of models trained on the CoNLL 2003 English training data.

3.2.2 Processing Tasks: Information Extraction and Topic Modeling

Coreference resolution (abbreviated *coref*) is the problem of deciding which of multiple expressions in a sentence or document refer to the same object, quantity, or entity. For example, in the sentence "**NIH** policy was that if **a submitted proposal** was not funded, a review report would be generated for **it** by the **study section assigned to it**", the anaphoric reference "it" refers to the submitted proposal.

Coreference resolution is a significant, though not essential task in spatiotemporal event extraction; an example can be found in the statement, "informed estimates suggest that up to **17,000,000 children** suffer from hunger throughout the **USA, many of them** living in poverty". The QA task of determining how many animals this sentence says were killed, and the textual entailment task of determining whether the children live in the United States, are relatively straightforward, whereas the coreference resolution task of identifying what subset of entities were said to be impoverished is more difficult. Thus, a high-level event wherein "17 million children suffer from hunger in the USA" can be detected even without coreference resolution, whereas the more challenging problem of tagging "locales where children in poverty suffer from hunger" requires it.

3.2.3 From Tag Clouds to Opinion Mapping



Figure 3. Tag cloud of teen's descriptions of bad behaviors they've seen online. (Voice of America, 2011)

3.3 Heterogeneous Information Networks

Heterogeneous information networks (HIN) are a type of graph-based model of communication and trust that include a large proportion of all social networks. Research on heterogeneous information networks as led to a convergence of methodologies for network modeling, incorporating classification, learning and reasoning with graphical models, frequent subgraph mining, relational representation, and link annotation, among other techniques. Many intelligent systems applications to information extraction, web search, and recommendation call for inferences to be made regarding the existence, type, or attributes of links. Some tasks, such as question answering using **information networks**, may require that inferences be based upon partial link information and made under uncertainty about participating entities and relationships.

3.3.1 Visualization of Large and Interesting Graphs

Application areas that often exhibit a need for heterogeneous information network analysis include:

- **information diffusion and sharing systems:** sensor networks, social media (opinions and sentiments, meme propagation, viral content, political commentary, *etc.*)
- **behavioral modeling:** community recruitment and mass activity, large-scale patterns, traffic, spatiotemporal effects
- **content-management systems:** version control, wikification
- **social recommender systems:** communities, experts, friends, products, reviewers, providers
- **application areas:** cybersecurity (information flow, trust networks, attack graphs, mechanism design), bioinformatics and biomedicine (genomics, proteomics, metabolomics), epidemiology

3.3.2 Descriptive Analytics of Graphs

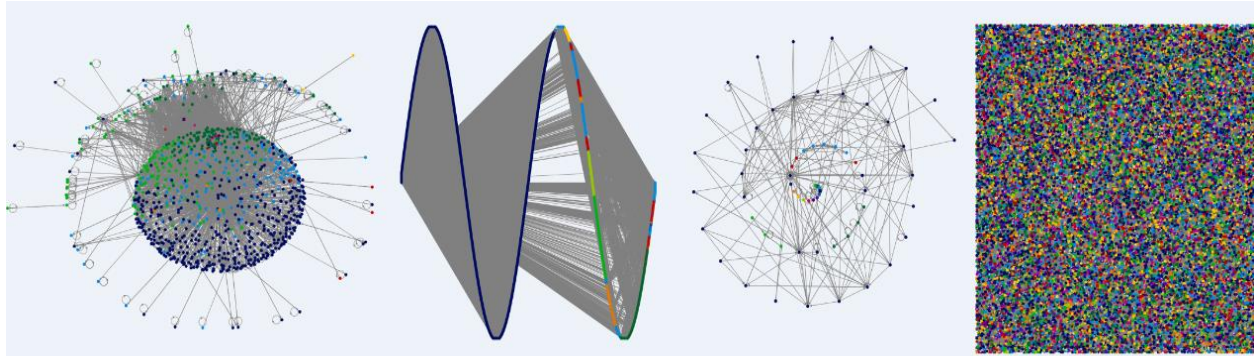


Figure 4. Different graph topologies, visualized (Part 1 of 2) – Havel-Koren fast multiscale graph, horizontal sine wave, spiral graph, uniform random graph.

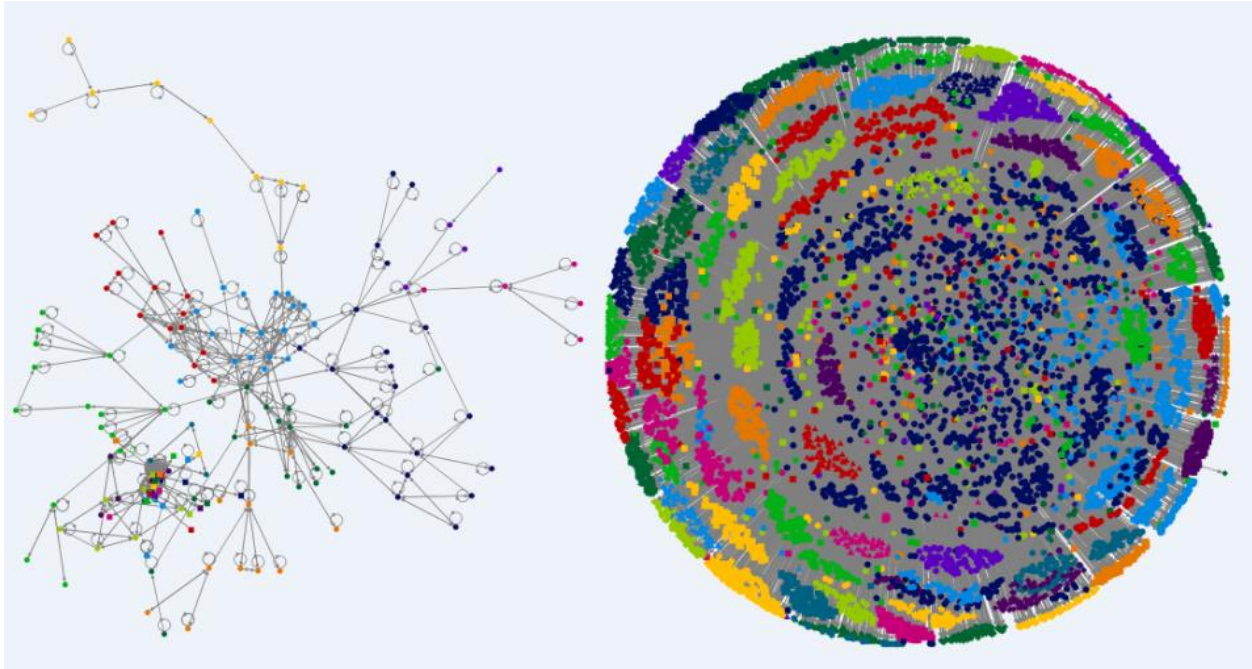


Figure 5. Different graph topologies, visualized (Part 2 of 2) – sparse (2-D) vs. dense (3-D) Havel-Koren fast multiscale graph.

As shown in Figure 4 and Figure 5, visualization can also help differentiate key features of different specified topologies, particularly when the connectivity is generated by a random or semirandom process. These include the maximum and minimum degrees, number of connected components, and centrality of the graph.

3.3.3 Emerging Techniques

Active research areas that are relevant to heterogeneous information networks include:

- community detection and formation modeling
- ranking-based clustering methods: learning to rank in information networks
- path-based similarity measures and relationship extraction
- modeling of link types and relationship strength
- applications to modeling of weblogs, social media, social networks, and the semantic web
- frequent pattern mining in graph and sequence data
- statistical relational learning
- trust networks and information sharing

4. Case Studies of Thematic Maps

4.1 Epidemiology: Historical and Predictive

Predictive epidemiology refers to the analytical study of disease dynamics to predict future outbreaks in space and time so that effective mitigation strategies can be implemented to curb the recurrence of epidemics. Since epizootic diseases such as the Foot and Mouth Disease (FMD), which periodically or sporadically break out within livestock populations, raise several political, administrative, economic and welfare issues, it is imperative to analyze the disease dynamics to facilitate adequate preventive measures, especially in countries that report recurring epidemic outbreaks instances. Figure 6 depicts a simulation described in (Roy Chowdhury, Scoglio, & Hsu, 2011) that was used to generate all of the states other than the upper-left hand corner.

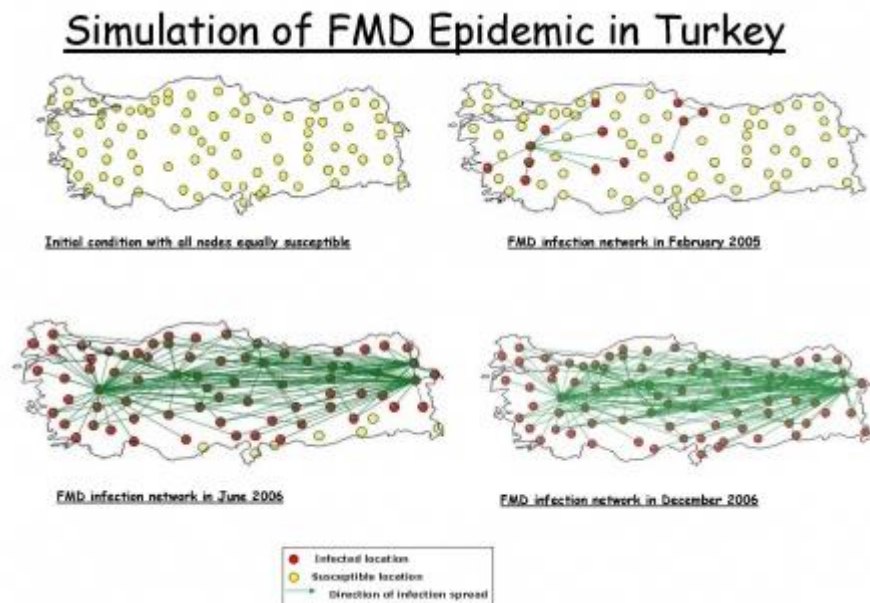


Figure 6. Simulation-based visualization of spread models for foot-and-mouth disease (Roy Chowdhury, Scoglio, & Hsu, 2011).

4.2 Text Information Retrieval and Extraction: Sentiment Analysis

The specific aims of the techniques surveyed in this section are as follows:

Aim 1. Extend known **algorithms for named entity recognition and relationship extraction**, to produce basic summaries of entities and relationships mentioned in texts. The technical objective

is to tag where basic entities and opinions are mentioned in freely available text (including both user posts and profiles), then map these tagged elements in space, time, and by topic, to acceptable levels of precision and recall.

Aim 2. Adapt basic known techniques to a particular domain of interest – specifically, extracting data from text discussions that are archived from the dark web (including blogs and forums) using web crawlers. This entails developing a means of handling entities and quantitative data that have not previously been extracted from text. Another functional requirement is some mechanism for entity reference resolution, *e.g.*, abbreviations and synonyms, for known terms. Finally, a **domain-specific ontology** of relevant aspects or attributes of the domain of discourse is proposed. This includes topics frequently discussed in blogs and forums, to better facilitate information retrieval applications such as question answering about the domain and gathering recommendations from subject matter experts.

Aim 3. Develop methods for sentiment analysis and improve existing ones, to **summarize opinions and discover patterns**. The technical objective is to relate demographic data extracted from text and profiles to qualitative data – namely, the polarity of text at the document, sentence, or aspect level, aggregated across demographic categories such as geographic region of residence. Objects of interest for sentiment analysis include recommended solutions, but can extend to aspects of the domain.

4.3 Event Tracking: Clandestine Labs

The problem of event visualization based on structured data, in the form of time-referenced and georeferenced relational tuples, and on unstructured data, in the form of free text. Information extraction systems based on named entity recognition (NER) and relationship extraction have enabled detection of events mentioned in free text and extraction of structured tuples describing the location, time, along with other attributes of an event. Identifying hotspots and trends, however, remains an open problem. One limitation is the absence of ground truth for high event activity. In some cases this is due to a lack of well-defined criteria for activity and relevance, while in some it is due to limitations in existing annotation interfaces.

We first present a basic approach to event visualization. Our general framework makes use of mapping tools such as *Google Maps*, the Google web toolkit, and timeline visualization tools such as *MIT SIMILE*. It also builds upon previous work on gazetteer-based event recognition and syntactic patterns for semantic relationship detection. Next, we show how a system developed originally for visualization of animal disease outbreaks reported in online news documents can be adapted to display reports of methamphetamine lab seizures compiled by regional law enforcement. We briefly outline the development of a domain-specific data description language for increased portability and ease of information integration. We then discuss the role of topic modeling and information retrieval approaches in filtering and ranking events.

A key technical contribution of this work is the application of topic modeling algorithms in order to compute the posterior probability of a particular spatial location, time unit, or combination given the type of event, which is treated as a topic. This allows the data to be interrogated systematically in order to display geographic regions that are more prone to events of interest. A potential application of this is to construct a time composite map of administrative divisions within a state or province, or a spatial composite time series by month or year, showing active regions. These can be visualized using a *choropleth map*: a map in which regions (geographic regions in this case) are coded by colors or grayscale intensity levels. These represent a variable of interest – in this case, event frequency. Finally, the ability to estimate marginal likelihoods over locations and times given the event type parameters can also be used to filter events, to display only those that fall within a specified frequency range. For example, the system can be configured to search for seizures of methamphetamine production labs in counties or districts where they are common or rare.

5. Effective Practices for Integrative Information Visualization

The basis tenets of information graphic design as related, critically analyzed, and demonstrated by Tufte (1990, 1997, 2001, 2003, 2006), are as follows:

1. **Show the data**
2. **Induce the reader to think about the *substance*** rather than about:
 - a) Methodology
 - b) Graphic design
 - c) Technology of graphic production
 - d) Something else
3. **Avoid distorting** what the data have to say
4. **Present many numbers in a small space**
5. **Make large data sets coherent**
6. Encourage the eye to **compare** different pieces of data
7. **Reveal** the data at different levels of detail, broad to fine
8. **Serve a clear purpose**: description, evaluation, tabulation, decoration
9. **Be closely integrated** with statistical and verbal descriptions of data

In this section, we review some realizations of this high-level vision in the specific domains of information graphics for big data analytics and spatiotemporal applications.

5.1 Information Graphic Design: Enhancing Learning and Analysis

Matlow (2000), a practitioner, instructor, and educational administrator in graphic design, describes the challenges of introducing computation into a graphic design curriculum at Harrow College, which merged in 1990 with the Polytechnic of Central London. These include the paradigm shift “from ‘page’

to ‘screen’” and the evolution of curricula. The article gives a critical analysis of contextual studies, a comparative ethnographic and cultural approach that was applied to digital media with poor initial results. It discusses how a critical theory module was successfully added to this curriculum, focusing on enabling technologies for social action and cognitive aspects of computer-mediated communication. This module went beyond incorporating technical aspects of digital arts to explore the role of language in structuring and creating meaning in cultural adaptation to technology. In particular, key topics regarding human-computer interaction and human factors, technical claims and open questions concerning computational intelligence and philosophy of mind, including the possibility of strong AI, were raised. Matlow concludes by pointing out that this kind of integrative curriculum holds positive potential for enhancing the education of digital arts practitioners, but also offers a timely warning for present-day designers of massively online open course (MOOC) material about the disparity between ease of delivery and the difficulty of ensuring the quality of content.

5.2 Tufte’s Principles Applied to Spatiotemporal Big Data

Tufte’s principles have been extensively applied to the visual display of quantitative information (Tufte, 2001), object visualization and thematic maps (Tufte, 1990), visualization of processes and storylines (Tufte, 1997), the improvement of presentation graphics (Tufte, 2003), and the critical exploration of evidence, especially for decision-making (Tufte, 2006). Of the four primary books on information graphics by Tufte (1990, 1997, 2001, 2006), *Envisioning Information* (Tufte, 1990) in particular emphasizes spatiotemporal aspects, though the other three also cover specific aspects of map visualizations and geospatial data to a degree.

Gahegan *et al.* (2001) describe the establishment of a Commission on Visualization by the International Cartographic Association, reporting on “the use of visualization within exploratory analysis, data mining and geocomputation, with the overall focus directed to the task of knowledge construction”. This paper depicts a waterfall model of knowledge discovery (here referred to as the “knowledge construction process”) proceeding from more data-driven and pattern-driven to more model-based and process-driven. In an extensive survey of exploratory visual analysis and exploratory data analysis, the Commission’s report also characterizes stages of reasoning, from more abductive in emphasis to more inductive and finally more deductive (analytical) and theory-guided. Aside from being very similar to the waterfall model of development in software engineering and knowledge engineering, this cascade process follows some computational models of cognitive processes, particularly machine learning and inference.

Bibliography

- Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000). Visualization of navigation patterns on a Web site using model-based clustering. In R. Ramakrishnan, S. J. Stolfo, R. J. Bayardo, & I. Parsa (Ed.), *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000)*, (pp. 280-284). Boston, MA, USA.
- Cuzzocrea, A., Song, I.-Y., & Davis, K. C. (2011). Analytics over large-scale multidimensional data: the big data revolution! In A. Cuzzocrea, I.-Y. Song, & K. C. Davis (Ed.), *Proceedings of the ACM 14th International Workshop on Data Warehousing and On-Line Analytical Processing (DOLAP 2011)* (pp. 101-104). Glasgow, UK: ACM Press.
- Elger, C. E., & Lehnertz, K. (1998, February). Seizure prediction by non-linear time series analysis of brain electrical activity. *European Journal of Neuroscience*, 10(2), 786–789.
- Goldstein, J., & Roth, S. F. (1994). Using aggregation and dynamic queries for exploring large data sets. In E. Dykstra-Erickson, & M. Tscheligi (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2004)* (pp. 23-29). Boston, MA, USA: ACM Press.
- Hall, M., Frank, E., Holmes, G., & Pfahringer, B. (2009, June). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), pp. 10-18.
- Heer, J., Kong, N., & Agrawala, M. (2009). Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009)* (pp. 1303-1312). Boston, MA, USA: ACM Press.
- Keim, D. A. (2006). Challenges in Visual Data Analysis. In E. Banissi, K. Börner, C. Chen, G. Clapworthy, C. Maple, A. Lobben, . . . J. Zhang (Ed.), *10th International Conference on Information Visualisation (IV 2006)* (pp. 9-16). London, UK: IEEE Press.
- Kumar, N., Keogh, E., Lonardi, S., & Ratanamahatana, C. A. (2005). Time-series bitmaps: a practical visualization tool for working with large time series databases. *Proceedings of the 5th SIAM International Conference on Data Mining (SDM 2005)*, (pp. 531-535). Newport Beach, California, USA.
- Mario, C., & Talia, D. (2003, January). The knowledge grid. *Communications of the Association for Computing Machinery*, 46(1), 89-93 .
- Monmonier, M. (1990). Strategies For The Visualization Of Geographic Time-Series Data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 27(1), 30-45.
- Steele, J., & Iliinsky, N. (Eds.). (2010). *Beautiful Visualization: Looking at Data through the Eyes of Experts*. Cambridge, MA, USA: O'Reilly Media.

Watson, H. J., & Wixom, B. H. (2007, September). The Current State of Business Intelligence. *IEEE Computer*, 40(9), 96-99 .