

Analysis & Visualization of Technology Data in the US

Colin Murray^{#%*} Weimao Ke^{#†} Hana Milanov^{#‡} Mark Meiss^{#§} Sharavan Rajagopal^{#¶} Katy Börner^{#||}

[#]Indiana University, Bloomington, IN 47405, USA

[%]University of Sydney, NSW 2006, Australia and National ICT Australia

ABSTRACT

This IEEE InfoVis 2005 Contest submission addresses all four contest questions. Diverse data analysis and visualization techniques were applied to the *Technology Data in the US* data set to (1) understand the birth, growth, and death of companies over the 15 year time span, (2) the geospatial distribution of different industries in US, (3) the correlation of the number of employees and sales increases for companies that change their geolocations at least six times, and (4) sudden increases in the number of employees and in sales for the different industries in the complete time span. While first insights could be gained, the dataset appears to be too limited to provide definite answers to questions that business experts would ask. The complementary web page with details on the presented analysis and supplementary material is available at <http://tara.slis.indiana.edu/outgoing/infovis05/>.

1 THE INFOVIS 2005 CONTEST DATA SET

The *Technology Data in the US* data-set comprises 84,472 unique technology companies. The location (city, state, zip), sales, employment, primary industry, and product type is given for companies for the years 1989 to 2003. Unfortunately, the industry field is only available for a few companies for the years pre-1993. As to be expected for a technology company data-set, about 42.9% of all companies are in telecommunications & internet or computer software.

Companies are categorized into 18 primary industries: Factory Automation, Biotechnology, Chemicals, Computer Hardware, Energy, Environmental, Transportation, Manufacturing, Defense, Advanced Materials, Medical, Pharmaceuticals, Photonics, Computer Software, Subassemblies & Components, Telecommunications & Internet, Not Primarily High-Tech, Test & Measurement.

The geographic location of the companies can be derived from 42,193 unique zip codes. In the data-set, not all company zip codes have their latitude and longitude defined. We used the more complete latitude and longitude file from the Chizu system described in section 3. For those zip codes that were still missing we set their latitude and longitude to that of the closest zip-code. We use the zip code field to track company movements. In the data-set, 1 company moves 7 times and 5 companies move 6 times.

*e-mail: colin.murray@nicta.com.au

†e-mail: wke@indiana.edu

‡e-mail: hmilanov@indiana.edu

§e-mail: mmeiss@steinbeck.ucs.indiana.edu

¶e-mail: shrajago@indiana.edu

||e-mail: katy@indiana.edu

2 DATA ANALYSIS AND VISUALIZATION

All subsequently shown visualizations were designed in response to questions that business experts are confronted with in their jobs. All visualizations use the simplest and most readable spatial substrate, e.g., a map of US, and a simple data overlay to communicate the patterns, correlations, and trends of interest. Adding an overlay to a geographical map is a common and effective visualization technique for geographically constrained data. An example of such a visualization is zipdecode [1].

3 GEOVIS OF THE GROWTH OF COMPANIES (QUESTION 4)

Understanding of the birth, growth, and death of companies is important for practitioners as well as academics. Effects of shattering events like 9/11, deregulation policies or a bubble burst can affect companies in different ways. To show the evolution of the technology companies in the US a 15 frame animation was created representing the years 1989 to 2003. Each time frame uses the very same base map: a map of the US in which states are color coded by the total sales of all their companies. Richer states are given a darker shade. Overlaid over this map are nodes representing the companies that are in existence in this particular year. The area of nodes reflects company sales in millions of dollars. Node color represents the age of companies - younger companies are given a lighter color while well established companies appear darker. Companies that are in their first year of existence are colored yellow and companies in their last year of existence are colored red. The maps were generated by extending the Chizu system implemented by Mark Meiss, online at <http://steinbeck.ucs.indiana.edu/~mmeiss/L579/project4/final.html>. The time frame 2002 of the 15 year animation is shown in Figure 1.

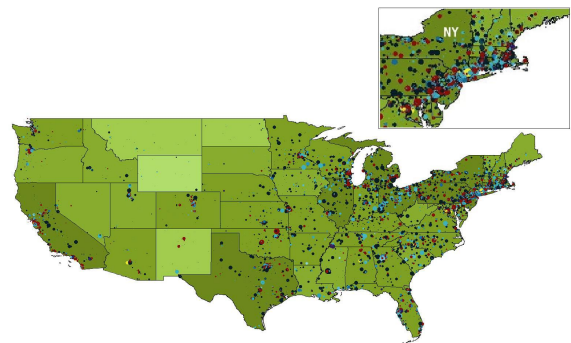


Figure 1: Geospatial locations, total annual sales, and ages of all companies in 2002. Inset shows a close-up for NY.

A closer examination of the 2002 map reveals that California and New York have the most companies with the largest sales. We also see clusters around major cities. By zooming into a local region we can access details (see inset in Figure 1). We labelled the headquarters of some major US companies as a frame of reference.

Unfortunately, the contest data provides no information on why companies emerged, thrived, or died. It would be interesting to know if a red node is a consequence of a company's bankruptcy or an acquisition or merger with another firm. One could speculate that 'smaller' red nodes were either acquired or went bankrupt due to the liability of being a small or new company. We could also speculate that two large red nodes in one year could have formed an even larger yellow node (new companies) in the subsequent year suggesting big mergers, like the one that formed Verizon in 2000. Ideally, the maps could show the merger waves characterizing many industries, reflecting the general trend of consolidation in some industries.

3.1 GeoVis of Each Industry (Question 2)

Next, our experts were interested to see the geospatial distribution of different industries in US. To ease the readability of the visualizations, the very same substrate map was used to generate density plots of all 18 industries. There is a node for each company each year placed at the companies location that year. The recent years are colored lightly and the early years are colored darker. Nodes for recent years are drawn on top of nodes from earlier years. The node size is once again mapped to total sales. All 18 maps are shown on the complementary web page.

In general, there is little difference in industries in terms of geographical distribution. However, there are major differences in the total sales and number of companies for different industries. Some industries, e.g., telecommunications & internet, experienced extensive growth in recent years.

The complementary web page shows a pair of interdependent industries on one map, biotechnology and pharmaceuticals. Old established pharmaceuticals are highly dependent on biotechs for their knowledge and innovation, whereas the biotechs are dependent on the pharmaceutical companies' capital for research testing money, as well as product development and different legal requirements (contacts in FDA, etc.). These interdependencies can be related to geographical proximity, which can be seen on the map. We can see biotechnology companies around large pharmaceutical companies.

3.2 GeoVis and Chart of Company Movement (Question 1)

Why do companies move? How often do they move? What happens to their sales and numbers of employees after they move? Figure 2 shows a data graphic that aims to communicate the movement trail of one company as well as the associated sales numbers and employee counts over a 15 year time span. The curves represent the sales and employees of the company and the bars indicate when moves have occurred and the extent of the moves. The background map shows the trajectory of the company.

Plotting trajectories of all companies over the map of US results in a rather cluttered visualization (see complementary web page). A closer examination shows that there is a small set of persistent patterns of relocation strategies. Many companies move toward large cities and there is a great deal of movement between the major business centers of the East and West coasts.

There are plenty of questions that the visualizations invite. Unfortunately, the contest data does not provide insight for the reasons of relocations. Those would naturally depend on which part of the company changed the zip code. Here we speculated it is the headquarters of the company. Given richer data on R&D centers and production facilities locations, it would be fascinating to see potential emerging patterns in relocations of companies laboratories into certain locations, potentially forming a geographically distinct cluster of knowledge (e.g., caused by collaboration with strong research university centers). The results of these analyses could then be linked to differences in innovation rates between companies that

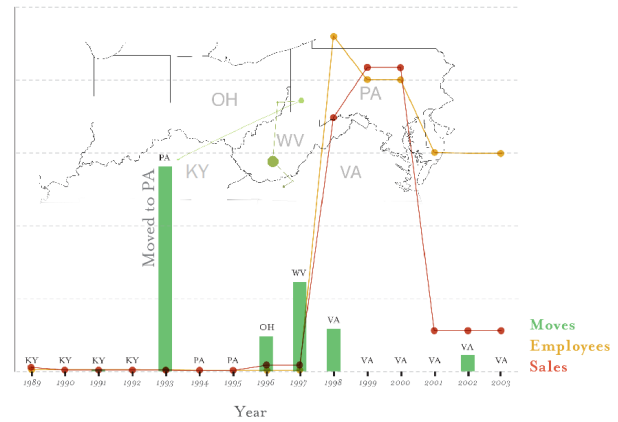


Figure 2: Geospatial trajectory, annual sales, and employee counts for one highly moving company over 15 years.

relocated R&D facilities and those that did not. Intensive R&D is expensive, and some industries (e.g., biotechnology) heavily rely on cooperation (which is facilitated by geographical proximity) seeing such patterns evolving on geospatial maps would be really interesting, especially given potential consequences for firms performance and innovation.

Direction of movement is a fruitful area for visualization, as it reveals possible patterns which then invite different questions. Reasons for companies relocations can be multiple, each of them interesting in its own way. In addition to the aforementioned formation of knowledge clusters, other reasons for relocations could be a result of specific developmental policies creating motives for relocation of companies in search for fruitful opportunities. On the other hand, some policies (e.g., placing strict ecological standards) might increase operating costs and drive companies away.

3.3 Charting Bursts of Sales and Employees (Question 3)

What industries experience a sudden increase in the number of companies? Which ones had a sudden increase in their sales or employees? Bursts in the number of employees and sales might indicate that the a company finally developed a marketable product while prior to that the company was mostly involved in research and development. Hence a burst would define the year when their product hit the market and turned first sales. For each year, we calculated the number of companies in each industry whose sales doubled from the previous year. We then plotted this information on a line chart. The same procedure was applied for the number of employees. These charts can be seen on the complementary web page. They show when the different industries experienced major growth.

4 ACKNOWLEDGEMENTS

We appreciate the effort by G. Grinstein, U. Cvek, M. Derthick, M. Trutschl in providing the *Technology Data in US* contest data set. This work is supported by a NSF CAREER grant under IIS-0238261, NSF grant IIS-0513650, and an equipment grant by UITS, Indiana University.

REFERENCES

[1] Zipcode. <http://acg.media.mit.edu/people/fry/zipcode/>.