Boom and Bust of Technology Companies at the Turn of the 21st Century

Heike Hofmann * Hadley Wickham Dianne Cook Junjie Sun Christian Röttger

Iowa State University

Abstract

Overall, technology companies increased in number, sales, products and employees, over 1989-2003. Most of the increase can be attributed to a few types of industries, such as telecommunications, and primarily non-technology. On an absolute scale the east (New York, Boston) and west coasts (San Francisco, Los Angeles) dominate, but on a relative scale there are hot pockets of economic activity all over the country. Many of these appear to be related to development after a natural disaster. We were surprised by many of the features we uncovered, particularly one with apparent political implications.

1 Introduction

The enormity of the data poses a huge challenge. It is large, so that loading the full data - all years, all companies - into any software is slow and tedious. It is also complex and multi-faceted so there are many, many different features to investigate. Deciding which directions to pursue, and what variables to derive, took a substantial amount of time.

Our process was to load the data into a mysql [4] database to readily subset and tabulate the data in different ways. Extracts from the database were read into R [5] for calculations, plotting, modeling, and output to other data formats. Metadata sets were created for interactive visual investigation using GGobi [1]. Subsets were also loaded into MANET [3] for interactively studying missing values and categorical variables. Visual discoveries were double-checked by further number crunching in R.

We interpreted and refined the competition tasks as we progressed into the analysis. There are some surprising results!

2 Data Description

This data [2] contains information on 84472 technology companies between 1989-2003. The companies produced 154912 unique products in this period. The time frame is notable for technology innovations, such as the rise of the internet, the dot-com bubble and crash, Y2K, the 911 tragedy and changes between democratic and republican control of government.

3 Task 1: Trends and multivariate relationships

Each of the variables are tabulated by year. The results are plotted as line plots. A robust linear regression (dashed line) is fitted where appropriate (Figure 1).

The number of companies increases each year. The increase is linear between 1989 and 1999, increases dramatically in 2000, and decreases after 2001. This most likely reflects the dot com bubble and crash. Products are similar except for the first few years of the database. Sales and employees both have an increasing trend with a flat period 1990-1993 and a rapid rise 1999-2001.

*e-mail: hofmann@iastate.edu

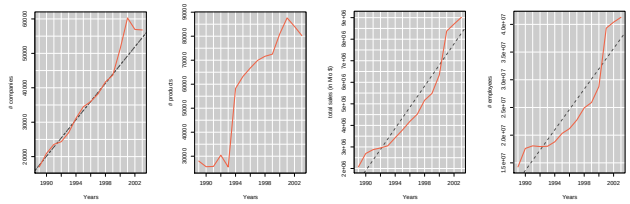

Figure 1: Overview of data. Company count, product count, sales and employees are plotted by year. Sales and employees generally increase with flat periods 1990-1993 and sharp rises 1999-2001.

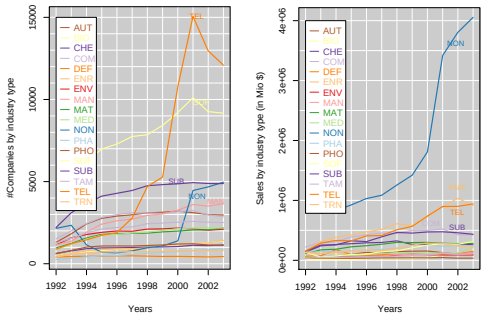

Figure 2: Company count and total sales plotted against year by industry type. Most of the increase in companies in the boom years is due to telecommunications and primarily non-technology companies, with a smaller contribution from software companies.

The number of telecommunications (TEL) companies dramatically increases in 1999 (Figure 2, left). Software (SOF) and subcomponent (SUB) companies are high throughout the period. A curiosity is the industry type NON (primarily non-technology companies), with lows between 1993 and 2000 which is a suspiciously political pattern. The dip corresponds to the Clinton years.

The dominating industry type for sales for the entire period is primarily non-technology (NON) companies (Figure 2, right). This is a surprise! There is a dramatic increase in sales for this industry type after 2000. The next two largest categories in sales are telecommunications and energy. The number of employees has a similar trend (not shown).

4 Task 2: Clusters

The density of company counts was computed with respect to geographic location, for each year. The results are displayed as colored maps and animated over time. The calculations are done on two scales, absolute and relative. Relative growth is measured as the difference in the number of companies of two consecutive years divided by the number of companies in the earlier year.

On the absolute scale (not shown) the east and west coasts dominate in the number of companies. Apart from that, Seattle, the Twin Cities, Chicago and Houston are visible.

On the relative scale different hot spots pop up each year (Figure 3). With some background research using the internet these appear to be economic activity after natural disasters.

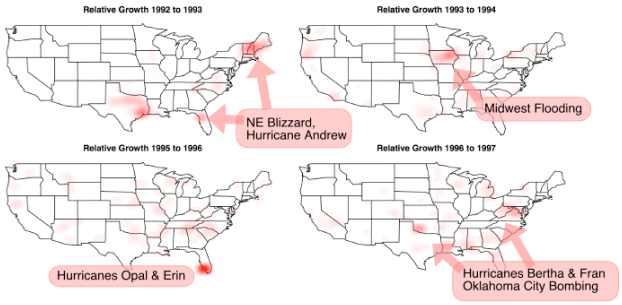

Figure 3: The density of relative company counts displayed on a map. Hot spots are related to natural disasters, located by internet search.

5 Task 3: Unusual features

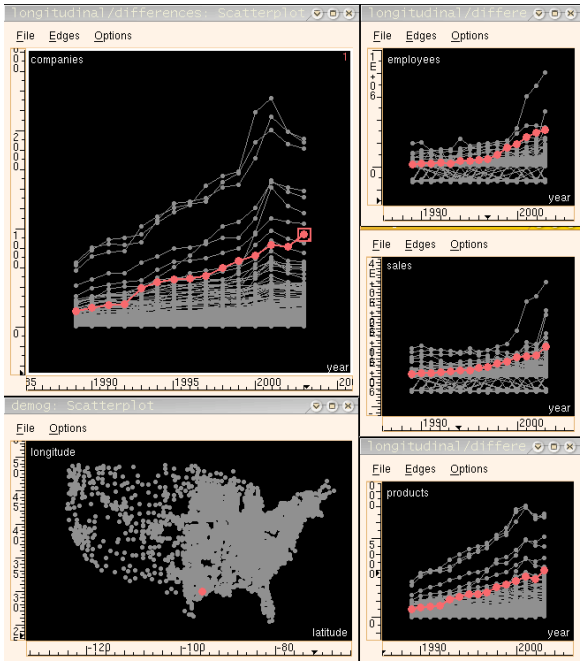

Figure 4: Harris County is the only county in the USA where the number of companies rises substantially after 2001.

The data was aggregated by county, a metadata set containing longitudinal measurements, year-to-year differences, and county demographics and statistics was created for loading into ggobi.

In the plot of the longitudinal data, most counties follow a pattern of increasing number of companies over time, and a strong drop after 2000 (Figure 5). There is one noticeable exception to this pattern: Harris County, TX. This county has a dramatic increase of 109 companies from 2001-2003, which represents a 14% increase. There next highest increase is a fifth as much, 26 companies. Is there something unique in Harris County, TX? Harris County, TX, is the home of the Johnson Space Center. It is also the county where George Herbert Walker Bush claims a homestead exemption on his residence. The increase in number of companies is explained mostly by a 50% increase in energy companies, from 117 to 172, with 26% (62 to 91) explained by primarily non-technology related companies. Sales and number of employees increase from 2001-2003 but not so much differently from other counties. The number of different products jumps, and this is noticeably different from other counties.

Other unusual patterns that we discovered include two counties in Michigan, counties near San Francisco, high company counts around Harvard and MIT in Boston, high sales in in the Mall of America in Minneapolis.

6 Task 4: Other, Data cleaning

We spent of lot of energy early in the data release finding anomalies in the data and reporting these. This resulted in numerous revisions of the competition data. Some of the problems were fixed but there still seem to be numerous problems with this data. It is a very large data set, and hence quality is difficult to control.

Founding year of a company seems to be particularly affected: after earlier issues with huge numbers of companies founded in 2000, which got revised in the data, there are still (unbelievably?) huge numbers of companies founded in 2000 - see figure 5.

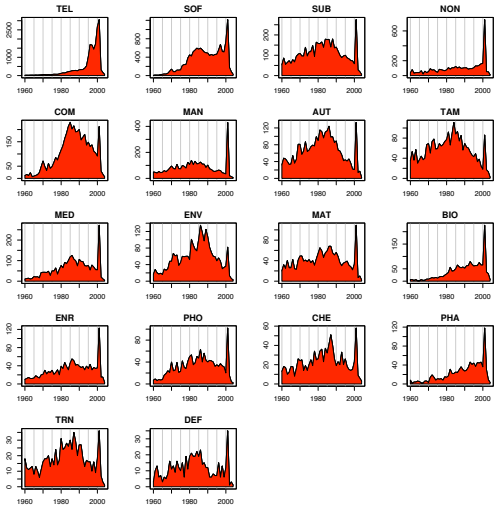

Figure 5: Founding year of all companies by industry type .

7 Conclusion

We were very surprised about two of the results: the relationship of relative economic growth with natural disasters and the pattern of Harris County, TX. When we started we expected to see the bubble pop in Silicon Valley, some economic effects in the New York region after September 11, 2001, the effects of Microsoft developing in the Seattle area. And we saw these. We also had other expectations that didn't pan out: companies that move a lot might be more likely to go bankrupt (disappear from the database), that there might be movement from away from the coasts after the bust to the mountain states and the Midwest. There is some movement of companies but these results were less interesting.

With a database this rich, there are many, many more directions we could investigate!

References

[1] GGobi. Data Visualization System. `http://www.ggobi.org/`, 2005.

[2] G. Grinstein, U. Cvek, M. Derthick, and M. Trutschl. Ieee infovis 2005 contest, technology data in the us. `http://ivpr.cs.uml.edu/infovis05`.

[3] MANET. Missings Are Now Equally Treated. `http://www1.math.uni-augsburg.de/Manet/`, 2005.

[4] Mysql. Open source database project. `http://www.mysql.org/`, 2005.

[5] R. The R Project for Statistical Computing. `http://www.r-project.org/`, 2005.