UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

UN ESTUDIO DE LA ESTRUCTURA Y DINÁMICA DE LA RED DE ACCIONISTAS CHILENOS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MAURICIO NIVALDO ANDRÉS MONSALVE MORENO

PROFESOR GUÍA:
CLAUDIO GUTIÉRREZ GALLARDO

MIEMBROS DE LA COMISIÓN:
PABLO BARCELÓ BAEZA
GONZALO NAVARRO BADINO
MARCELO ARENAS SAAVEDRA

Este trabajo contó con la cooperación del proyecto FONDECYT 1070348.

SANTIAGO DE CHILE
JULIO 2009

Resumen

En el marco del análisis de redes sociales, estudiamos la red chilena de accionistas, una red dinámica cuyos actores son empresas y accionistas, y cuya relación es la relación de propiedad. En particular, damos especial énfasis a una subred de ésta, la red de inversiones entre empresas, cuyos actores son empresas solamente.

Para modelar la red de accionistas, recuperamos y procesamos la información disponible en el sitio web de la Superintendencia de Valores y Seguros. Usando la proporción de propiedad de cada accionista en cada empresa, y los estados financieros de las últimas, modelamos esta red dirigida, dinámica y multivariable desde Diciembre de 2003 hasta Junio de 2007.

Para estudiar redes dirigidas multivariadas, diseñamos métodos empíricos y análiticos. Los métodos empíricos consisten en dos visualizaciones: perfiles de correlación de arcos y scatter-plots de redes. El método analítico consiste en reducir la topología del grafo a distribuciones de probabilidad (una para las relaciones y otra para los actores). Incluimos estas técnicas de análisis en una aplicación de software.

Finalmente, estudiamos la red con las metodologías desarrolladas. Primero, buscamos propiedades generales de la red (como el hecho que *las empresas más grandes tienden a poseer o participar en las empresas más pequeñas*). Luego, nos enfocamos en simular la red de inversión entre empresas, usando las distribuciones de probabilidad obtenidas de la red original. Concluimos que, aunque el modelamiento a nivel relacional, como la reducción de la topología a distribuciones de probabilidad, funciona relativamente bien al reproducir la red, no lo hace completamente ya que la red muestra propiedades emergentes, más allá del ámbito relacional.

Universidad de Chile
Faculty of Physical and Mathematical Sciences
Graduate School

A study of the structure and dynamics of the Chilean shareholding network

by

Mauricio Monsalve

Submitted to the University of Chile in fulfillment
of the thesis requirement to obtain the degree of
Master of Science in Computer Science

Advisor	:	**Claudio Gutiérrez Gallardo**
Committee	:	Pablo Barceló Baeza
		Gonzalo Navarro Badino
		Marcelo Arenas Saavedra

We would like to thank project FONDECYT 1070348 (Chile) for their support.

Departament of Computer Science - University of Chile
Santiago - Chile
JULY 2009

Abstract

Within the subject of social network analysis, we study the Chilean shareholding network, a dynamic network which actors are firms and shareholders and which relation is the shareholding or ownership relation. However, we place special emphasis to a subset of this network, the network of investment between firms, which actors are firms.

To model the Chilean shareholding network, we retrieved and processed the data available at the *Superintendencia de Valores y Seguros* website. With this data, we modeled this dynamic, weighted, multivariate and directed network from December 2003 to June 2007, by retrieving the proportion of shares (weights) along with the financial statements of the firms (multivariate data).

To study this network, we designed empirical and analytical techniques and developed a software application based on them. The two empirical techniques are correlation profiles and network plot graphs, which are visualization techniques for multivariate directed graphs. The analytical technique consists in reducing the topology of the graph to probability distributions.

Finally, we study the network using the above techniques, first by finding properties (such as *rich firms tend to own or participate in poorer firms*) and then by simulating it using the estimated probability distributions. We conclude that, even tough relational-level modeling techniques, such as probability distributions, work well in approximating the topology, do not completely model the network since it shows emergent, non relational behavior.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are concerned with the study of the Chilean shareholding network, a network where shareholders invest in firms. The scholarly value of this network data comes from its nature:

- As network, is directed and weighted. But more importantly, is a dynamic network.
- It is a multivariate network because we know the financial statements of the firms.
- It is a social network but also an economic network. In particular, it is a *cross shareholding* network, which models ownership.

Our study will be focused in the multivariate and network nature of the data. In spite of the interdisciplinary value of the dataset, we will be concerned with the topology and its relation with the additional information of the vertices (shareholders and firms). In particular, we will be concerned with firms since we know their *financial statements*.

1.1 Objectives

Our objective is to study the structure and dynamics of the Chilean shareholding network as a social network. This somewhat broad objective can be decomposed into several precise, particular goals:

Develop analytical methodologies to study complex networks. In what concerns to data processing and analysis, which should be of interest to computer scientists, we wish to develop methodologies for analyzing multivariate network data. As we should not interpret data, we should limit ourselves to support researchers by developing visualizations, measures, and obtaining analytical functions from the data.

Develop a software for analyzing directed and multivariate networks. Next to methodological developments, the objective of creating a software for network analysis is concerned with automatizing the data analysis process (which will help us study the network) and leaving a legacy to the research community.

Obtain the Chilean shareholding network data. This concerns to the data retrieval and preprocessing of this work. Without this data, this work would lack its empirical validation.

Discover empirical properties of the Chilean shareholding network. This objective concerns to the data analysis part of our inquiry. By discovering properties of the studied network, we will prove that the developed methods work. (Whether they are useful depends on the data and the researcher.)

Model the dynamics of the Chilean shareholding network. If the methods developed are of use, we should be able to use the properties found to develop a network evolution model.

1.2 Motivation

Our motivation to study this network comes from its multidisciplinary value: we will be studying a social, economic network. On top of that, the subject of *shareholding*, which will be studied in chapter 3, is of particular concern to finance, industrial organization, corporate governance, even to law.

In what regards to social networks, we have dynamic network data. Dynamic social network analysis is a challenge, since most analytical techniques were developed for static networks. *Centralities* are typical examples of this, since they are solely based on the topology of them (see chapter 2).

Also, our network is *multivariate*: vertices and arcs have many attributes (such as financial statements). And analyzing such rich network data is the main challenge of this thesis: how can we explain the topology in terms of this additional information? (Or viceversa.)

The novelty of our work is also a motivation. There are not closely related works, which study dynamic multivariate directed graphs. The most related works are those that study shareholding networks while somehow relating the topology of the network to financial information of firms (see chapter 3 and 4). However, we develop specific analytical techniques for relating topological and non-topological data. In particular, we are able to reproduce a subset of the Chilean shareholding network, the cross-shareholding between firms, using non-topological data (see chapter 11).

1.3 Outline - how to read

We structured this document to be complete yet fast to read. Chapters are very specific, so they can be skipped if are of no interest. For example, if the reader is not interested in the description of *Network Observer*, a program we developed, he/she can skip chapter 9.

Additionally, certain paragraphs of lesser importance can be skipped too. These portions of text have smaller fonts than the rest, and were included to reinforce ideas and add small details, for completeness.

1.3.1 Part I

The first part of this thesis, which consists in chapters two and three, is a theoretical introduction to the subject.

Chapter 2 is an introduction to social network analysis. If the reader is well informed about this subject, he/she can skip the chapter.

Chapter 3 is an introduction to shareholding networks. Section 3.1 states the importance of the research topic. Section 3.2 is an introduction to corporate finance, and introduces the reader to accountability, valuation, portfolio management, etc. Section 3.3 is a literature review on similar works (i.e. studying shareholding networks as networks).

Chapter 4 summarizes some techniques for analyzing networks in economics. It also includes a section about our methodology.

1.3.2 Part II

The second part is about the preparations for studying the network.

Chapter 5 explains how the data was obtained. It first presents the source (the website of the SVS). Then, it explains how the data was downloaded, preprocessed and stored. This chapter is not theoretically rich but rather technical.

Chapter 6 presents two visual methods to explore multivariate directed graphs: *arc correlation profiles* and *network plot graphs*. Both methods are designed for the exploratory analysis of multivariate network data.

Chapter 7 presents a theoretical development to study multivariate graphs. This method consists in reducing the topology of the network to probability distributions.

Chapter 8 is concerned with the problem of estimating joint probability distributions. It presents a methodology to estimate them, and how it works (experiments).

Chapter 9 presents *Network Observer*, a software which is packed with the previous developments.

1.3.3 Part III

The actual study of the Chilean shareholding network is presented in this section.

Chapter 10 presents a general characterization of the dynamics of the Chilean shareholding network. It shows the dynamics of the topology and how the firms related by an arc are also related through their financial assets. To do this, we studied some time series (using Gnumeric) and the topology using Network Observer.

Chapter 11 is about the simulations performed to explain the topology of the network. In this chapter we test if what we learned from the network explains its evolution. To do this, we reduced the topology of the network to probability distributions, and used them in the simulations, closely reproducing the topology.

Chapter 12 are the conclusions, the lessons learned from this work.

Part I

Theoretical preamble

Chapter 2

A brief of social network analysis

In this chapter, we review the concept of social network and its related research field: social network analysis. We cover some basic concepts as well as theoretical and practical developments.

2.1 Overview of the subject

2.1.1 What social network analysis is about

Social network analysis is concerned with the study of social networks in a quantitative or precise way. To do that, it studies social networks as graphs, and use several measures and other methods to analyze them. Of course, improving these methods is also of concern to social network analysis. But it is important to keep in mind that social network analysis is part of sociometrics: quantitative sociology. As such, social network analysis is also concerned with the sociological meaning of their results.

Social networks are networks of people or groups. They are about *actors* and the *relationships* among them. Relationships can be of any kind; they can be friendship, communication, bloodline, common interests, colleagues, employer-employee, etc. Actors can be people, communities, countries, firms, etc. And social networks may be studied within different contexts too: classrooms, cities, firms, countries, etc. It is noticeable how general social networks are, and social network analysis is.

In the last decade, social network analysis has been extended to analyze even technological and biological networks. This has been possible due to the generality of the analytical tools developed within the subject.

Another consequence of the broadness of the subject is the growing interest among different disciplines. People from mathematics, physics, ecology, management, economics, computer science, etc. started doing research in social networks.

For more on how social network analysis has been extended to other fields, see

Boccaletti et al. (2006). For more on the interest in computer science, see Hogan (2007), Getoor and Diehl (2006), and Alt et al. (2006).

2.1.2 Its basic tools

Social network analysis uses graph analysis and other tools to analyze network data.

To analyze *position* in a social network (like privileged power positions), *centrality* measures are used to estimate the position of each actor in the network. Some centralities are designed to measure fame or reputation, while others measure sensitivity to spread (like gossips and epidemics), and others measure how much the network is disconnected if the actor is removed. To computer science, this is related to discrete mathematics and algorithms.

To detect *communities* within social networks, several methodologies are used, such as k-means, simulated annealing, spectral analysis, etc. To computer science, this is another form of the problem of clustering.

Visualization is concerned with the problem of showing complex network data in an aesthetic and meaningful way. To computer science, this is both a graphic and user interfaces problem.

Also, statistics has been widely used in social network analysis, specially when network topologies are analyzed and classified. Similarly, data mining has been concerned with the *link prediction problem*, which is about knowing the likelihood of the presence of a relation (arc or edge) between two known actors.

2.1.3 Basic graph terminology

Graphs are the mathematical models of networks. Different types of graphs model different types of networks. We will start by studying the simplest type of graphs, which models symmetrical relations.

A *graph* $G(V, E)$ is a tuple composed of a set of vertices V and a set of edges $E \subseteq \{\{a, b\}, a, b \in V\}$. Edges model the symmetrical relationship between two vertices; if vertices $a, b \in V$ are related, then $\{a, b\} \in E$.

In graphs, we define *walks* as successions of adjacent edges. *Cycles* are walks which start and end in the same vertex. *Paths* are walks that do not contain cycles. The *length* of a walk or cycle is the number of edges involved in the walk or cycle. And the *distance* between two vertices is the length of the shortest path between them (if exists).

Adjacency matrices are matrices that represent graphs. An adjacency matrix A is such that its element $a_{i,j}$ equals 1 if $\{v_i, v_j\} \in E$, and 0 otherwise. (Obviously, adjacency matrices are symmetrical.) A nice property of adjacency matrices is that, if powered, they count the number of walks between two vertices: $B = A^n$ is such that $b_{i,j}$ is the number of walks of length n between v_i and v_j. This useful property is the foundation of several SNA measures.

But not all graphs have edges. Others have *arcs*, that model asymmetrical relationships. These graphs are called *directed graphs* or simply *digraphs*, and are defined as $G(V, A)$ where V are vertices and $A \subseteq V \times V$ are arcs. Walks are redefined in terms of arcs, but remain essentially the same (which applies to cycles and paths). The adjacency matrices of digraphs (J) are such that $j_{i,j} = 1$ if $(v_i, v_j) \in A$, and 0 otherwise. (Note that J is not necessarily symmetrical.) Fortunately, J^n also counts the number of walks of length n between vertices.

Another types of graphs are *weighted graphs*. These have weights (numerical values) associated to their arcs or edges. In addition to adjacency matrices, they have weighted adjacency matrices which have the weights of arcs or lengths.

Additionaly, there are graphs which allow loops (arcs or edges that start and end in the same vertices), these that allow several arcs or edges between the same vertices (*multigraphs*), and these that allow arcs or edges between three or more vertices (*hypergraphs*). (The latter two types of graph are less commonly used.)

For more on graph theory, see Diestel (2000).

2.2 Measures

Quantifying topological properties of social networks is an essential part of social network analysis. Using these measures, it is possibly to somewhat justify sociological hypothesis concerning the social structure.

One of the basic ideas behind social networks is that actors have positions within the social structure, which enable them to be famous, spread information, relevance, etc.

2.2.1 Basic graph measures

These measures are for reference, as they do not enable researchers to compare different networks (field knowledge). Among these, we find:

Order. Number of vertices: $|V|$.

Density. The density of non directed graphs is defined as:

$$\rho = \frac{2|E|}{|V|(|V|-1)},$$

and the density of directed graphs is defined as:

$$\rho = \frac{|A|}{|V|(|V|-1)}.$$

Average distance. It is the average distance between any two vertices.

Diameter. The diameter is the maximum distance between any two vertices.

Radius. Greatest distante from the *central* vertex to any other vertex. The central vertex is the one with the smallest greatest distance to any other vertex.

For more on graph measures see Diestel (2000, ch. 1).

2.2.2 Degree centrality

One measure of centrality is degree centrality. Its basic idea is that one actor is famous when he/she/it is known by many (neighborhood).

2.2.3 Bonacich's $c(\beta)$ / Katz's centrality

Another property to measure is the spreading power of an actor. For example, in a social network of gossips, a person with good position spreads communications fast. He/she reaches lots of actors within a few steps. This is the idea behind this centrality.

In terms of transitions, A^k should be higher for well positioned actors when k is small (fast spreading). So, to measure good positions for spreading (information, objects, epidemics, etc.), we could sum the powers of the adjacency matrix A using a discount rate $\beta \leq 1$. And, by summing the number of walks from an actor to each other one, we define:

$$c(\beta) = \sum_{k \geq 0} \beta^k A^{k+1} \vec{1}.$$

Note that the above formula may not converge. This is why it is important to use a suitable $\beta \in (0, 1]$. If it converges, then:

$$c(\beta) = A(I - \beta A)^{-1} \vec{1}.$$

2.2.4 Eigenvector centrality

The insight behind this centrality is described as follows: An actor's centrality is proportional to its neighboors' centralities. As an equation, it has the form:

$$c_i = \alpha \sum_{j \neq i} c_j a_{i,j},$$

where α is a constant of proportionality.

The previous equation can be written, if there are no loops, as:

$$\vec{c} = \alpha A \vec{c} \Rightarrow (I - \alpha A)\vec{c} = 0,$$

which defines α as an **eigenvalue**. And of course, $\vec{c}$ becomes an **eigenvector**. Now the problem are the multiple solutions to the eigenvalues/eigenvector problem. Well, the academic

habit is using the eigenvector related to the greatest eigenvalue. Fortunately, this eigenvector can be approximated by:

$$\vec{c} \approx \frac{A^k \vec{1}}{||A^k \vec{1}||},$$

when $k \to +\infty$. Using a big value for k, for example 40, we can approximate this centrality without further complications.

2.2.5 Closeness centrality

Another insight which can serve to define a centrality is *closeness*, or how close an actor is to the rest of the network. The proposed measure for this:

$$c_i = \frac{1}{\sum_j d_{i,j}},$$

where $d_{i,j}$ is the distance between vertices v_i and v_j.

Note that the value $\sum_j d_{i,j}$ is somewhat greater when the actor v_i is closer to the rest of the network than when is further from it. Naturally, its multiplicative inverse somewhat measures closeness. (Closeness is a vague yet intuitive concept.)

2.2.6 Betweeness centrality

Betweeness centrality is based on the idea that there are *key players* in networks, whom, if removed, largely reduce the connectivity of their networks. In the worst case, they may split the network into two.

The insight behind betweeness centrality is that an actor is more relevant (more of a key player) the more shortest path cross through him/her/it. Thus, we define betweeness centrality of an actor as *the number of shortest paths in which the actor participates*.

2.2.7 Clustering

The clustering of a vertex measures if an actor is part of a cohesive group (most likely a community). The insight behind this measure is that the vertex belongs to a group if its neighbors also belong to it, and this should be reflected as arcs or edges between them. Thus, clustering is defined as:

$$c_i = \frac{2e_i}{k_i(k_i - 1)},$$

where e_i is the number of arcs or edges between the neighbors of vertex i, and k_i is its degree. (This is a measure of density of the neighborhood of vertex i.)

The previous measure is also extended to the whole graph; in that case, the clustering is defined as the average of the clustering of its vertices:

$$c = \frac{1}{|V|} \sum_i c_i.$$

2.2.8 More measures

In the social networks analysis literature, there are several works regarding centrality definition, comparison and validation. The same applies to other types of measures, but centralities have received most attention.

Also, centrality measures are sometimes addressed as *prestige*, *status*, and *prominence*. These terms convey sociological concepts, and depend on the situation modeled.

To know more about measures in social networks, see Marsden (2002), Poulin et al. (2000), Borgatti (2005), Bonancich and Lloyd (2001), Zemljic and Hlebec (2005), and Bonacich (1987). These works make use of several modeling techniques and insights, as well as validations against evindence.

2.3 Community detection

A relevant problem in social network analysis is how to detect communities using only graph data.

Block analysis is about rearranging rows and columns of adjacency matrices, and finding communities by creating blocks around the diagonal. These blocks represent communities.

Spectral analysis is about using eigenvectors of the adjacency matrix to detect communities. By looking the definition of eigenvector centrality, it can be seen that friends should have similar centralities, regardless of the eigenvector used. So, the more eigenvectors we use, the more information we have for distinguishing communities.

K-plex and k-clique methods are based on the idea that communities are more or less complete subgraphs. Then, the problem is reduced to find complete or almost complete subgraphs.

Hierarchical clustering is about defining several clusters, and adding actors to these clusters. This is the *agglomerative* algorithm, as it merges similar clusters. One result of this algorithm is the dendogram or hierarchical tree, which is a representation of the number of clusters the algorithm had.

Inside communities, edges have low betweenness as there are several alternative paths between actors (communities are cohesive). This is the principle behind the algorithm of Girvan and Newman, which removes the edge with the highest betweenness score in each iteration, until the graph is converted into several isolated and cohesive subgraphs: the communities.

There are many more algorithms, which are based on several insights. For example, there are a few which make use of circuit analogies to detect communities (current goes through

shortest paths, which is a form of betweenness), attraction of forces, spins, random walks (they also estimate betweenness is a way), etc.

For more on community detection, see Freeman (2003) and Boccaletti et al. (2006).

2.4 Common topological properties

Some topological properties have been very popular in past years. In fact, the growth models that reproduce these properties boosted research on networks since the end of the XX century. One of them is about how close we are in the world, and the other is about inequality in social networks. This section is based in Amaral et al. (2000), Boccaletti et al. (2006), and Borner et al. (2007).

2.4.1 Small world networks

Small world networks are networks which diameter increse logarithmically with the number of nodes.

The small world property was first discovered by Stanley Milgram in the 1960s. He found that people in Nebraska found an unknown yet defined target in Boston in just six steps (path length). This amazing property was named small world (by others) as it showed that people is very close in the world.

The small world models appears everywhere. It is specially common in the Internet, from email networks to chat contacts. It has been found in several online social networks too.

There are many growth models that reproduce the small world networks. Some of them use edge rewiring, i.e. the edge $\{a, b\}$ is now $\{a, c\}$; starting from a lattice, it is possible to obtain a small world network by rewiring its edges randomly. And other models use edge addition, typically starting from a set of isolated vertices.

2.4.2 Scale free networks

Scale free networks are networks which degrees follow a power law distribution. They are also a class of small world networks.

Scale free distributions are related to inequality. For example, the distribution of wealth follows a power law. Note that this type of distribution is also called the *Mattew effect*, as it denotes inequality: the rich become richer while the poor become poorer. In fact, the basic model for generating scale free networks is the *preferential attachment model* by Barabasi and Albert, where vertices with greater degrees are more likely to obtain new edges and increase their degrees.

Scale free networks are very common in social contexts as well as in nature. The Internet, the World Wide Web, actor networks, scientific collaboration, protein networks, etc. follow scale free distributions.

Note that **scale free distributions are power laws**. Scale free is invariance to scale, which is written as:

$$f(xy) = g(x)f(y),$$

where f is the distribution and g is a *scaling* function. Note that $f(\alpha x)$ is proportional to $f(x)$, i.e. scaling the input only scales the output, therefore the label *scale free*. Now, note that:

$$f(x) = g(1)f(x) = g(x)f(1) \Rightarrow f(x) = g(x),$$

which takes us to:

$$h(x) \equiv \log f(e^x) \Rightarrow h(x+y) = h(x) + h(y) \text{ (linearity)}$$
$$\Rightarrow h(x) = \alpha x \Rightarrow \log f(e^x) = \alpha x$$
$$\Rightarrow \log f(x) = \alpha \log x \Rightarrow f(x) = x^\alpha.$$

In other words, scale free distributions are power laws.

If f were to be plotted in a log/log scatter plot, it would look like a line. And the inclination of the line would be its exponent.

2.5 Visualization

Network visualization has been widely used as a research tool rather than an aesthetic object. However, after the social networks boom (early XXI century), several works focused in the later. Fortunately, new works are focusing in the dynamics of networks, going back to its scholarly origin.

This section is based on Freeman (2000), Huisman and van Duijin (2005), and Bender-deMoll and McFarland (2006).

2.5.1 Statics

Illustrating networks require that the layout of edges and vertices do not obscure the topology of the graph. For example, node overlapping, line crossings, bends, etc. tend to indicate bad layouts. Also, excess of colors or iconography may cause layout clutter, which is also bad.

Researchers have proposed several layout algorithms using various insights. Layouts like Kamada-Kawai use force fields to arrange vertices. In particular, Kamada-Kawai relates vertices through edges, which work like springs between them. In a similar way, Fruchterman-Reingold use electrostatic repulsion for vertices, while attraction is set by edges.

Peer influence algorithms arrange vertices so that their neighbors are influenced by their position (for example, the average of their coordinates). The final layout is achieved after several iterations.

Multidimensional scaling algorithms asign vertices positions in a high dimensional space, then reduce the dimensions to two (or three) by means of several algorithms. Some of these

are strict projections whereas others are *low stress* projections (which have few line crossings, vertex overlappings, etc.).

Optimization algorithms arrange vertices so they solve an optimization problem. This problem consists of an objective function and several quality constraints. Contraint optimization techniques are used to achieve the desired solution (layout).

Other ideas are related to showing different information. For example, *centrality maps* plot vertices according to their centrality; the higher the centrality, the closer the vertex is to the center of the graph.

Different software packages use different visualizations, but most of them use the traditional layouts mentioned above.

2.5.2 Dynamics

Visualization of network dynamics is a recent issue, which was enabled by innovations in data collection, data from new fields, and the use of simulation to create [artificial] network data.

Dynamic network visualization offers challenges to theorists, specially taking into account that social networks are conceptual, while other networks exist in space, like power grids. In social networks, relations like power, friendship, common bloodlines, etc. are not visible, making their visualization difficult.

One solution to dynamic visualization is to animate the network using several static visualizations over time. This solution is quite intuitive, but revealed that different time steps show data in a different way. Small steps may show chaotic networks while big steps show them static.

Other solution is to add *movement*. Basically, the social network is transformed into a virtual space, and network dynamics is shown as movement. Note that these movements may be drawn, as in a static visualization, and vertices can move from one part to another in the virtual space.

2.6 Comments

We reviewed the basics of social network analysis, focusing in its research concerns. And now that we explained it, we move on to the next topic: shareholding networks.

References

Alt, C., Astrachan, O., Forbes, J., Lucic, R., and Rodger, S. (2006). Social networks generate interest in computer science. In *SIGCSE '06*.

Amaral, L. A. N., Scala, A., Barthelemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *PNAS*, 97(21):11149–11152.

Bender-deMoll, S. and McFarland, D. S. (2006). The art and science of dynamic network visualization. *Journal of Social Structure*, 7.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, (424):175–308.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Bonancich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, (23):191–201.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, (27):55–71.

Borner, K., Sanyal, S., and Vespigani, A. (2007). Network science. *Annual Review of Information Science and Technology*, 41.

Diestel, R. (2000). *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, 2 edition.

Freeman, L. (2003). Finding social groups: A meta-analysis of the southern women data. In R. Breiger, C. Carley, P. P., editor, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*.

Freeman, L. C. (2000). Visualizing social networks. *Journal of Social Structure*, 1(1).

Getoor, L. and Diehl, C. P. (2006). Link mining: A survey. *SIGKDD Explorations*, 7(2).

Hogan, B. (2007). Using information networks to study social behavior: An appraisal. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Huisman, M. and van Duijin, M. A. (2005). Software for social network analysis. In Carrington, P., Scott, J., and Wasserman, S., editors, *Models and methods in social network analysis*, pages 270–316. Cambridge University Press.

Marsden, P. V. (2002). Egocentric and sociocentric measures of network centrality. *Social Networks*, (24):407–422.

Poulin, R., Boily, M.-C., and Masse, B. R. (2000). Dynamical systems to define centrality in social networks. *Social Networks*, (22):187–220.

Zemljic, B. and Hlebec, V. (2005). Reliability of measures of centrality and prominence. *Social Networks*, (27):73–88.

Chapter 3

A brief of shareholding networks

Now we cover the subject of shareholding networks along with the necessary economic and financial concepts concerning them, since we work with economic data in our research.

3.1 What are shareholding networks?

We call shareholding networks to those networks which actors are firms and shareholders, and which relation is the shareholding or ownership relation. If an actor A has ownership on B, then A is a shareholder and B is a firm. Naturally, firms can be shareholders. However, people can be shareholders too.

The relation between shareholders and firms is studied in corporate finance. However, the study is usually limited to one firm, or a group of related firms (a *holding*). Dietzenbacher and Temurshoev (2008) argues that studying shareholding networks (or *cross shareholding*) is important because:

1. Indirect relations might uncover a different structure of ownership than what could be seen by simply taking direct relations into account.

2. The effect of cross shareholding has not shed light on the relation between ownership and corporate performance. Current evidence is ambiguous: there is work left to be done.

3. Dietzenbacher et al found that their measures of network complexity are related to the degree of separation between control rights and dividends (Dietzenbacher and Temurshoev, 2008). In other words, it is possible to relate the network structure to the performance of firms.

Shareholding networks seem important to other research fields, which adds some value to our research.

3.2 Basics of corporate finance and governance

3.2.1 Accountability

Accountability is about keeping a registry of all the movements of money or values. Money lent, new machinery, wages, etc. are recorded. The result are accounts, and we will talk about *assets*, *debt* and *equity*, the three basic accounts within firms.

Assets represent the market value of everything that belongs to a firm, e.g. furniture, computers, installations, etc. Assets somehow represent the size of the firm, and are increased with acquisitions (more infrastructure, etc.).

Obviously, the money used to buy the assets come from a different source: *debt* or *equity*. *Debt* increases cash while requires frequent payments to lenders. *Equity* is the money invested by the investors or the money accumulated from *earnings*.

Since the money for assets came from debt or equity, we can write the equation which rules accountability:

$$\text{Assets} = \underbrace{\text{Debt} + \text{Equity}}_{\text{Liabilities}}$$

Or simply:

$$A = D + E$$

For more on accountability and firm structure, see Samuelson and Nordhaus (2005, ch. 7).

3.2.2 Valuation

Valuation is about estimating the market value of something. It is an important part of economic evaluation, which is also concerned with market and financial design and viability.

One basic concept behind valuation is that money changes in percents. Suppose we have $\$100$ now, and that we can work it so the next year we have $\$110$. If we had $\$200$, we will not have $\$210$ the next year; having $\$200$ is the same as having $\$100$ twice, so the next year we will have $\$110$ twice, or equivalently, $\$220$. But if we had $\$50$, we cannot split it in terms of $\$100$. However, we can do half of the work to increase it (if possible) or ask other people to lend us money, and we could get $\$110$ again, but after returning $\$55$, we end up having $\$55$. In other words, we increased our money in 10%, regardless of its amount. (In reality, the more money we have, the easier is to work it this way. But we should not earn proportionally more by having less, because money can be split into several smaller sums of it.)

The next basic concept is inter-temporal comparison. Let us suppose we have $\$100$ and the next year $\$110$. If we do not have anything now, but $\$110$ the next year, it should be same for us. (Let us also suppose that we do not have necessities, otherwise this will not work.) As long as we work money in term of percents, which are multiplicative terms, we can

compare money from different moments of time. $\$100$ today and $\$110$ the next year are the same because we increase our money in 10% per year. So, having $\$100$ today is the same as having $\$110 = \$100(100\% + 10\%)$ the next year. And going backwards, having $\$110$ the next year is the same as having $\$100 = \$110/(100\% + 10\%)$ now. If we work money at a rate r, which is 10% in our example, we estimate its equivalent in the next period by multiplying our accounts by $1 + r$. And to take it to the previous period, we only have to divide it by $1 + r$. Of course, multiplying by $(1 + r)^n$ advances n periods, and if $n < 0$, goes back in periods.

To valuate an investment which periodically gives us money, we use a measurement called *Net Present Value* or NPV. NPV measures all the flows of money by taking them to the present, and summing them up. NPV is computed as:

$$NPV = -I + \sum_{k \geq 1} \frac{F_k}{(1 + r)^k},$$

where I is the investment, and F_k are the future flows of money. If a NPV is less than zero, the flows do not account for the invested money; in other words, $NPV < 0$ is a bad investment. However, the greater the NPV, the better the investment. (This criterion does not take into account that people cannot wait too much for recovering their money. There exists another measurement for that.)

Also, there are other measurements which are useful to valuators. For example, the *Internal Return Rate* or IRR, which measures the internal rate of return of a project, is also a widely used measurement. The IRR is defined as the rate at which $NPV = 0$. However, there might be multiple rates that solve $NPV = 0$ since it is a polynomial root problem. (IRR is used very often; in general, it is applied to financial instruments.)

Of course, there is much more to valuation, but we will not cover more. For more on valuation, see Merton and Bodie (2000, ch. 4, 6 and 7) and Samuelson and Nordhaus (2005, ch. 25).

3.2.3 Investments, returns and risk

Investing all of our money in a single financial asset is generally a bad idea. The risk of losing all of our money is high. However, we can reduce that risk by having a *portfolio* of investments.

If we invest in similar yet independent projects, we can reduce their joint standard deviation. Let r_1 and r_2 be random variables that follow the same probability density function (pdf), and that represent the returns of investments I_1 and I_2. If we invest everything in I_1 or I_2, we get the same expected return $\mathbb{E}(r_1) = \mathbb{E}(r_2)$. Of course, if we invest half of our money in I_1 and half in I_2, we get $\mathbb{E}(0.5r_1 + 0.5r_2) = \mathbb{E}(r_1) = \mathbb{E}(r_2)$, which is the same expected return. However, risks are different. We know that $Var(r_1) = Var(r_2) = \sigma_r^2$, but $Var(0.5r_1 + 0.5r_2) = (0.5r_1 + 0.5r_2) \circ (0.5r_1 + 0.5r_2) = 0.25\sigma_r^2 + 2 \cdot 0.25 \cdot r_1 \circ r_2 + 0.25\sigma_r^2 = 0.5\sigma_r^2$. By diversifying our investments, we reduced their joint risk. (Note: $A \circ B = Cov(A, B)$, the covariance.)

However, we can make better decisions. Returns are not completely independent; in economies, all firms are related, directly or not. And it is possible to invest in firms which are

negatively correlated. For example, firms that import and firms that export. If $\rho_{r_1,r_2} = -1$, then $Var(0.5r_1 + 0.5r_2) = (0.5r_1 + 0.5r_2) \circ (0.5r_1 + 0.5r_2) = 0.25\sigma_r^2 + 0.25\sigma_r^2 - 0.5\sigma_r^2 = 0$. In this extreme case, negative correlation virtually nullified the risk. Of course, situations like this do not exist, but are useful to show the power of negative correlations between assets.

For more on *portfolio theory*, see Merton and Bodie (2000, ch. 12 (sec. 12.3)).

3.2.4 Investing in firms

When a person invests in a firm, he/she expects that the firm returns the money and some additional profits. So, in terms of money, investing in firms is rather similar to investing in financial assets.

There are special types of firms which allow people to invest in them almost freely, while giving their owners rights on the firm. These are the *incorporated* (Inc) firms, which are also called *sociedad anónima* (SA), *limited company* (LTD), *societé anonime* (SA), *sozieta per azioni* (SpA), etc. depending on the country (Merton and Bodie, 2000, ch. 1 (box 1.3)).

Incorporated companies issue *shares* or *stocks* which are bought by people, groups or firms, who become *shareholders*. Shareholder have rights on their firms, which somewhat become their ownership.

Shareholders receive *dividends* (money) from their firms. The amount of money given to the shareholders is decided by the managers (the board). However, shareholders can fire managers through voting systems (this depends on the firm). This creates a conflict of power within the firm. However, this problem is minimized when stock markets are open.

The conflict of power can be greater. Sometimes, managers run away with the shareholders' money. Within the contract model of the firm, the separation of management and finance (shareholders) separates control and ownership, creating incentives for managerial misuse of money (Shleifer and Vishny, 1996). Even if not common, managers sometimes invest shareholders' money in their own business (for example, selling equipment below market price to his/her firms), spend their money in travels, pursuing pet projects, and even resist being fired even when they are no longer competent. Shareholders use their power mostly through contract design and votes. (This problem is called the *Agency problem*, and is widely studied in economics and finance.)

For more on this see Rappaport (1998, ch. 1 and 7) and Merton and Bodie (2000, ch. 1, 2 (sec. 2.3), 9, and 16 (sec. 16.7)). For more on agency problems, see Shleifer and Vishny (1996).

3.2.5 Measuring firm performance

It is very important for shareholders to understand the performance of firms, to know whether they should invest or retire their money from them. However, their knowl-

edge is limited because control belongs to managers. A good way to solve this problem is the use of measurements.

Several measurements have been proposed to demonstrate a firm's performance. The most widely used are *financial ratios*, because they have direct impact in valuation.

For example, let us review how firms can increase the performance of equity, which may be of interest to shareholders. Suppose a firm works its assets at a rate r_A. So, each period that firm will produce Ar_A money. But part of that money came from debt: D. And the interest on that debt is $r_D < r_A$ (otherwise that firm may not be able to pay that debt). So, the earnings after debt are $Ar_A - Dr_D$, which in turn are the performance of equity, namely Er_E. (I.e. $Er_E = Ar_A - Dr_D$.) Shareholders should seek benefit from r_E because it is the rate at which their money increases. So, by increasing debt, managers can increase their firms' value to shareholders, but at the risk of bankrupcy (which is hard to measure). Note that the ratio D/A should somewhat measure the risk of bankrupcy and explain the amount r_E in terms of r_A. These are basic financial measures.

Financial ratios are the most popular measurements of firm performance, as their economic meaning is often clear (as in the above example). Common measurements are the RoA (return over assets), RoE (return over equity), RoI (return over investment), RoS (return over sales), D/E (debt-equity ratio), etc. which often are methods to analyze financial statements (Merton and Bodie, 2000, ch. 3 (table 3.5)).

For more on financial ratios and measurements of firm performance, see Rappaport (1998, ch. 3) and Merton and Bodie (2000, ch. 3).

We are specially concerned with financial ratios because they may guide shareholders to invest in firms or not. Also, they can be easily computed from financial statements, which is public information.

3.3 Research on shareholding networks

3.3.1 Equity and investment networks

Before talking about *cross shareholding*, we review some literature regarding equity and investment networks. The following works are concerned with the topology of market networks.

Kim et al. (2002) studies the scale free topologies of minimum spanning trees (MST) built from the correlations of financial instruments. It focuses in studying the power laws governing the degrees of the financial instruments.

Bonanno et al. (2004) studies a "network of equities" in the financial market, by computing the correlations among stocks, computing a distance measure between them[1], and then building MSTs. Then, it analyzes the topological properties of MSTs

[1] The distance is $d_{i,j} = \sqrt{2(1-\rho_{i,j})}$, which is actually a metric.

built from different data (local and global levels, and different time horizons) using statistics.

Boginski et al. (2005) discusses the relation of graph analysis (social network analysis) and data mining, and studies a "market graph" built from financial data. In its graph, two financial instruments are linked when their correlation is above 0.5. Its analysis consists in studying the evolution of statistical properties of the graph (number of nodes, degrees, etc.).

Chmiel et al. (2007) analyzes a network of firms related by targeting common niches. What it is different from other works is that it models competition among firms. Again, it studies the statistical properties of that network (degrees, weights, clustering, etc.).

Cajueiro and Tabak (2007) (Banco Central do Brasil) studies the graph of brazilian banks related by interbank lending. It builds the MST and studies its topological properties (degrees, centralities).

Eom et al. (2007) studies the MSTs built from correlated financial instruments. It analyzes the statistical properties of the graphs, specially their degree distributions (with focus on the power laws).

Song et al. (2008) studies the "world investment networks" which model trade between countries. It studies different graphs that can be built using this data: undirected, directed, and weighted graphs. Then it studies the statistics of the topologies built, and the power laws governing them (which they call *allometric laws*).

The cited works do not relate the topology of the graphs to non topological data (but Bonanno et al compared the topology to an ideal model, the one factor model (Bonanno et al., 2004), which is similar to the CAPM model (Merton and Bodie, 2000, ch. 13)). However, two works, both by Souma et al, are concerned with the relation between the topology of japanese **shareholding networks** and **non-topological data**, such as assets and profits (Souma et al., 2004, 2005). They model the japanese shareholding networks (fig. 3.1), and study the relations between topological and non-topological data (fig. 3.2). These works are the closest to ours, since we are working with shareholding networks and the relation between topological and non topological data.

3.3.2 Cross shareholding

In the industrial organization literature, *cross shareholding* is the term used to talk about direct and indirect shareholding relations among firms, and includes the different structures studied in finance, such as one-sided shareholdings, mutual shareholdings (which include ring-like structures), pyramiding structures, etc. (Dietzenbacher and Temurshoev, 2008). Note that the word "shareholding" might be replaced by "stockholding" or simply "ownership".

Studying cross shareholding seems very important: the structure of sharehold-

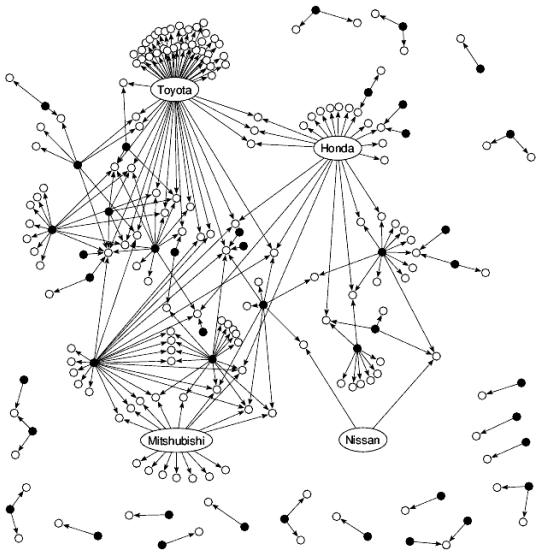


Figure 3.1: The shareholding network of the japanese automobile industry sector. Taken from Souma et al. (2004).

ing networks may predict massive financial shutdowns (Eisenberg, 1994); firm birth might be boosted with given structures (Almeida and Wolfenzon, 2006); tacit collusion of firms might be related to common shareholders between firms (Gilo et al., 2006); and even relate ownership to corporate performance (Dietzenbacher and Temurshoev, 2008).

Most works use input-output analysis (that uses Leontief matrices, which element $a_{i,j}$ says how much shareholder i owns from firm j) to study the problem of cross shareholding (Chapelle, 2005). In a Leontief matrix L, total dependence or ownership is measured as $\vec{x} = (L + L^2 + L^3 + ...)\vec{1}$, which is similar to Bonacich's $c(\beta)$ centrality, using $\beta = 1$. Chapelle (2005) studies the Belgium shareholding network using the

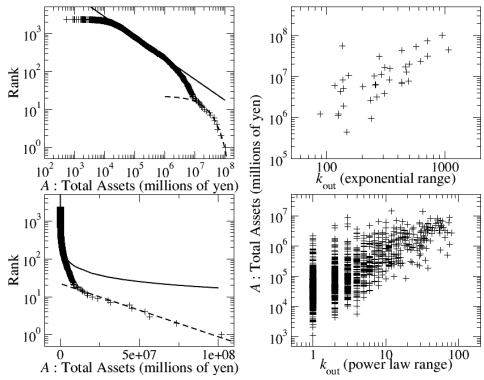


Figure 3.2: Relations between topological and non topological data in the japanese shareholding network. Taken from Souma et al. (2005).

measure of total dependence, and a special matrix for direct ownership: $a_{i,j} = 1$ if and only if i owns more than 50% of j, otherwise $a_{i,j} = 0$. Gilo et al. (2006) proves several incentive theorems using Leontief matrices. Dietzenbacher and Temurshoev (2008) defines a measure of network complexity based on this matrix (a measure similar to $\vec{M} = \sum_{k\geq 1} kA^k\vec{1}$) and compared it to firm performance.

Note that we are specially concerned with input-output analysis, since it uses *adjacency matrices* and a centrality measure as tools for analyzing the economic structure. This links social network analysis to input-output analysis, which is a method used mainly in macroeconomics.

But not all works use input output analysis. For example, Frieder et al study the social network of governing boards and its relation to firm performance, using just two firms instead of a whole network in their analysis (Frieder and Subrahmanyam, 2008). For other example, Almeida et al prove that the pyramidal ownership structure enables families to create new firms (which is harder with horizontal structures) by using proportions of ownership (Almeida and Wolfenzon, 2006).

However, these works are mostly topological and do not relate firms' financial attributes to the structure of shareholding networks.

3.3.3 Implications

Contributing to this body of literature might have effects in social network analysis to industry regulation.

By searching scholarly databases, we found that shareholding networks are of importance to several research areas: finance, macroeconomics, industrial organization and regulation, corporate governance, and law. For example, Bebchuk and Roe (1999) analyzes how ownership structures evolve given different initial conditions (which may shed light on how economies evolve), and Jones et al. (2003) studies the ownership structures that appeared soon after the privatization of public enterprises in Estonia. Also, Shleifer and Vishny (1996) shows the relation between corporate governance and firm regulation (law).

3.4 Shareholding networks in popular culture

Shareholding or ownership networks often appear in mainstream media, specially in business oriented media. However, they also appear often in activist media.

Figure 3.3 shows part of an infographic that shows who owns who within the car industry of the United States. It was taken from an activist website concerning the excess of cars issue (AFT, 2008). This infographic is part of their campaign to reduce the number of cars in the streets.

Figure 3.4 shows part of an infographic which shows who owns who within the media

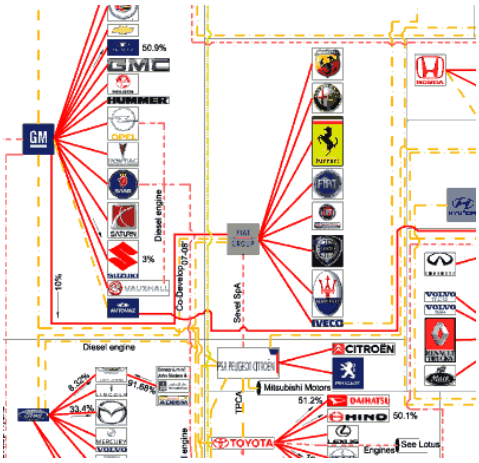


Figure 3.3: Automotive Family Tree. From AFT (2008)

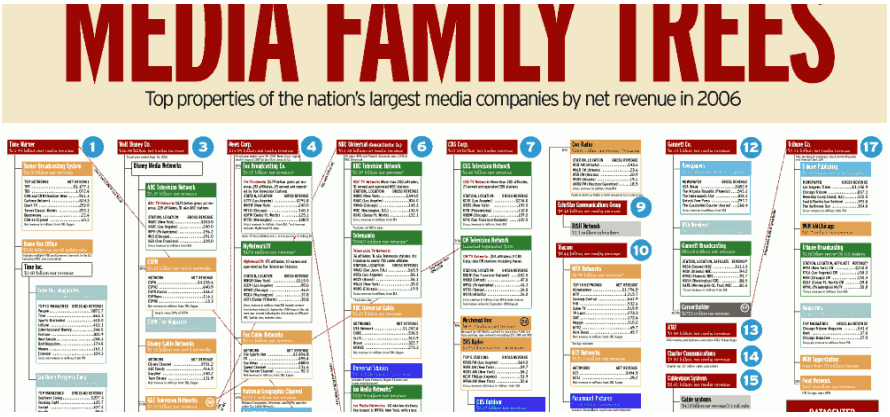


Figure 3.4: Media Family Trees. From MFT (2006)

companies. It was taken from a website which targets marketing professionals (MFT, 2006).

3.5 Comments

We studied the basics of corporate finance and governance, and then studied the literature on shareholding networks.

We found that most works do not relate topological and non topological network data. This opens a research opportunity for people who work with data analysis: analyzing complex data and developing new analytical methods and tools.

References

(2006). Media family trees. Appeared in Advertising Age.

(2008). Automotive family tree. Appeared in Too Many Cars.

Almeida, H. and Wolfenzon, D. (2006). A theory of pyramidal ownership and family business groups. *The Journal of Finance*, 61(6):2637–2680.

Bebchuk, L. A. and Roe, M. J. (1999). A theory of path dependence in corporate ownership and governance. *Stanford Law Review*, 52:127–170.

Boginski, V., Butenko, S., and Pardalos, P. M. (2005). Mining market data: A network approach. *Computers and Operations Research*.

Bonanno, G., Caldarelli, G., Lillo, F., Micciche, S., Vandewalle, N., and Mantegna, R. N. (2004). Networks of equities in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):363–371.

Cajueiro, D. O. and Tabak, B. M. (2007). The role of banks in the brazilian interbank market: Does bank type matter? Technical Report 130, Banco Central Do Brasil.

Chapelle, A. (2005). Separation between ownership and control: Where do we stand? *Corporate Ownership and Control*, 2(2):91–101.

Chmiel, A. M., Sienkiewicz, J., Suchecki, K., and Holyst, J. A. (2007). Networks of companies and branches in poland. *Physica A: Statistical Mechanics and its Applications*, 383(1):134–138.

Dietzenbacher, E. and Temurshoev, U. (2008). Ownership relations in the presence of cross-shareholding. *Journal of Economics*.

Eisenberg, L. K. (1994). Connectivity and financial network shutdown. Working Papers 95-04-041, Santa Fe Institute.

Eom, C., Oh, G., and Kim, S. (2007). Topological properties of the minimal spanning tree in the korean and american stock markets.

Frieder, L. L. and Subrahmanyam, A. (2008). Social networks and corporate governance. *European Financial Management*, 14(2):633–662.

Gilo, D., Moshe, Y., and Spiegel, Y. (2006). Partial cross ownership and tacit collusion. *RAND Journal of Economics*, 37(1):81–99.

Jones, D. C., Kalmi, P., and Mygind, N. (2003). Choice of ownership structure and firm performance: Evidence from estonia. Discussion Paper 7/2003, Bank of Finland Institute for Economies in Transition (BOFIT).

Kim, H.-J., Kim, I.-M., Lee, Y., and Kahng, B. (2002). Scale-free network in stock markets. *Journal of the Korean Physical Society*, 40(6):1105–1108.

Merton, R. and Bodie, Z. (2000). *Finance*. Pearson Education Inc., Prentice Hall Inc.

Rappaport, A. (1998). *Creating shareholder value: a guide for managers and investors*. The Free Press, 2 edition.

Samuelson, P. A. and Nordhaus, W. D. (2005). *Economics*. The McGraw-Hill Companies Inc., 18 edition.

Shleifer, A. and Vishny, R. W. (1996). A survey of corporate governance. Technical Report 5554, NBER.

Song, D.-M., Jiang, Z.-Q., and Zhou, W.-X. (2008). Statistical properties of world investment networks.

Souma, W., Fujiwara, Y., and Aoyama, H. (2004). Heterogeneous economic networks. In *9th Workshop on Economics and Heterogeneous Interacting Agents (WEHIA2004)*, pages 27–29.

Souma, W., Fujiwara, Y., and Aoyama, H. (2005). Shareholding networks in japan. In *Science of Complex Networks: From Biology to the Internet and WWW (CNET 2004)*.

Chapter 4

Analytical methods used in previous works

In this brief chapter, we cover the techniques used to study networks in economics and econophysics, specially these related to shareholding networks.

4.1 Input-Output Analysis

Input-Output (I-O) analysis is a method for studying how the Gross Domestic Product (GDP) is produced within an economy. Its creator, Wassily Leontief, was awarded the Nobel Prize of Economics for it.

I-O analysis is based on I-O tables, which summarize the fact that the input of a firm is the output of another firm. Rows and cols of I-O tables represent industries within the economy, such as figure 4.1.

	Forestry	Clothing	Construction
Forestry	0.1	0.2	0.4
Clothing	0.5	0.6	0.3
Construction	0.3	0.1	0.2

Table 4.1: Sample I-O table

Each element of an I-O table represents the incidence of the input on the output. For example, element (i, j) might mean how many products are created by j using one product of i. Another example: element (i, j) may mean how much money is earned by j per \$1 of the products supplied by i.

Let A be the I-O matrix with the data of the I-O table, and let us assume that a_{ij} means how many products from i are needed to create one product by j. Then, let us

relate the production between the industries:

$$p_i = e_i + \sum_j a_{ij} p_j,$$

where p_i is the production of industry i and e_i is the external demand, which is not explained by inter-industry relations. In terms of vectors and matrices:

$$\vec{p} = \vec{e} + A\vec{p} \Rightarrow (I - A)\vec{p} = \vec{e}$$

$$\Rightarrow \vec{p} = (I - A)^{-1}\vec{e},$$

i.e. we can explain industrial production (in detail) as a function of the external, non-industrial demand (consumers, other nations, etc.).

To see this as an application of graph theory, let us note that I-O tables are weighted adjacency matrices. Let us define a graph which vertices are industries and which arcs are the supply relation between industries: $X \rightarrow Y \Rightarrow X$ supplies Y. And let us assign weights to these arcs: how many products Y creates using one unit of X. Naturally, the weighted adjacency matrix of this graph is A, the same of the I-O table.

Now, how much produces each industry? Let us use the already defined $\vec{e}$ and see how much production induces an industry into another, in particular, i over j: directly, i induces $e_i a_{ij}$ on j; in two steps, i induces $e_i a_{ik}$ on k, which induces a_{kj} per unit on j, totaling $e_i a_{ik} a_{kj}$ and, as k is anyone, i induces $e_i \sum_k a_{ik} a_{kj}$ on j; and so on (adding more intermediaries). Using A, this is:

$$A\vec{e} + A^2\vec{e} + A^3\vec{e} + ... = \sum_{n \geq 1} A^n \vec{e} = (I - A)^{-1} A\vec{e},$$

and since that is the production/demand explained by the industries of the model:

$$\vec{p} - \vec{e} = (I - A)^{-1} A\vec{e},$$

which leads us to:

$$(I - A)\vec{p} - (I - A)\vec{e} = A\vec{e} \Rightarrow (I - A)\vec{p} = \vec{e}$$

$$\Rightarrow \vec{p} = (I - A)^{-1}\vec{e},$$

the formula that relates total production to the external demand.

As stated in the previous chapter, I-O matrices can be applied to other contexts, such as ownership; we already cited that Gilo et al. (2006) and Dietzenbacher and Temurshoev (2008) used this approach to study cross-shareholding.

I-O analysis is still being actively applied in macroeconomics, but at government level since the data is hard to retrieve. In particular, is being applied in Chile (Frigolett, 2005).

4.2 Spanning Trees

As Bonanno et al. (2004) say, spanning trees are convenient because connect the vertices of a graph without forming cycles. Minimum spanning trees (MSTs), in particular, are built using the set of edges to achieve the shortest total length.

Graphs can be directed or non-directed, but mostly are non-directed since correlations between stocks are the most studied networks using MSTs.

For example, Cajueiro and Tabak (2007) used the following procedure to obtain the Brazilian interbank network:

1. Get the matrix of bilateral exposures. It is a directed adjacency matrix A, where $a_{ij} > 0$ if bank i is creditor of bank j.
2. Reduce the matrix to a symmetrical one, S.
3. Get $max(S)$, the maximal weight of matrix S.
4. Distance d_{ij} between i and j is:

$$d_{ij} = 2 - \frac{s_{ij}}{max(S)}.$$

5. Find the MST of S.
6. Recover the directions of the MST from step 1.

Distances can vary. For example, Bonanno et al. (2004) and Eom et al. (2007) use $d_{ij} = \sqrt{2(1 - \rho_{ij})}$, where ρ_{ij} is the correlation of i and j, in the network of stocks.

4.3 Network effects

In traditional economics, *network effects* are used to explain some *externalities*[1] of networked systems. As Weitzel et al. (2003) say, network effects have been used to explain the discrepancy between private and collective gains, which may lead to suboptimal (non-paretian) equilibrium.

A classic example of network effects are streets. The more cars are in the streets, the slower (and riskier) becomes driving. Also, if an accident happens somewhere in the city, vehicular flow will decrease around the area.

[1] A externality is a cost unrelated to economic transactions.

4.4 Social network analysis

Social network analysis has not been a concern to economics until recently. Granovetter's sound criticism of Williamson's "new institutional economics" started the concern. Granovetter argued that transaction costs might not be a barrier to people in given social contexts, as in buyer-seller networks (Granovetter, 1985), while Williamson argued that, in some cases, vertical integration would be better. In a similar line, Coleman introduced the term social capital (Coleman, 1988).

Several works have studied market networks using social network analysis, but few in what regards to equity networks (as networks of stock). One of these works is Cajueiro and Tabak (2007), which studies the Brazilian interbank network using social network analysis. Cajueiro and Tabak classify banks according to their roles in the network, which determined using centralities.

4.5 Correlations

Souma et al. (2004) and Souma et al. (2005) are the closest works to us, in that they study the topology of shareholding networks and relate it to the financial statements of firms. They drew scatter plots and computed the correlations between degrees and financial data such as assets and profits, to find the relations between these variables.

4.6 Comments

We studied some basic techniques for handling graphs in economics, at mathematical and conceptual levels.

Our methodology Our approach to the study of the Chilean shareholding network is very different to previous works. To study this network, we developed visualizations and analytical techniques, to study the relations of many variables along with the topology.

To study the statics and dynamics of our network, we:

1. Compute some statistics of the network, such as number of firms per period, firm birth per period, firm death per period, etc.

2. Then, study the rather static *behavior* of the network, or why a vertex links to another one based on their financial statements.

3. Finally, reproduce the original network using simulation and the distributions of vertices and arcs.

Our work is rather different from previous works. This makes it hard to compare it to them. But one thing should be clear: our main goal is to relate network topology to financial statements.

References

Bonanno, G., Caldarelli, G., Lillo, F., Micciche, S., Vandewalle, N., and Mantegna, R. N. (2004). Networks of equities in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):363–371.

Cajueiro, D. O. and Tabak, B. M. (2007). The role of banks in the brazilian interbank market: Does bank type matter? Technical Report 130, Banco Central Do Brasil.

Coleman, J. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94(Suppl):s94–s120.

Dietzenbacher, E. and Temurshoev, U. (2008). Ownership relations in the presence of cross-shareholding. *Journal of Economics*.

Eom, C., Oh, G., and Kim, S. (2007). Topological properties of the minimal spanning tree in the korean and american stock markets.

Frigolett, H. (2005). Documentos de Proyectos. División de Estadística y Proyecciones Económicas, CEPAL, Naciones Unidas.

Gilo, D., Moshe, Y., and Spiegel, Y. (2006). Partial cross ownership and tacit collusion. *RAND Journal of Economics*, 37(1):81–99.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510.

Souma, W., Fujiwara, Y., and Aoyama, H. (2004). Heterogeneous economic networks. In *9th Workshop on Economics and Heterogeneous Interacting Agents (WEHIA2004)*, pages 27–29.

Souma, W., Fujiwara, Y., and Aoyama, H. (2005). Shareholding networks in japan. In *Science of Complex Networks: From Biology to the Internet and WWW (CNET 2004)*.

Weitzel, T., Wendt, O., von Westarp, F. G., and Konig, W. (2003). Network effects and diffusion theory: Network analysis in economics. *International Journal of IT Standards and Standarization Research*, 1(2).

Part II

Methodological developments

Chapter 5

The data

We now address how we retrieved and processed the data before performing any analysis.

5.1 Context

5.1.1 Source and available data

The data was retrieved from the website of *La Superintendencia de Valores y Seguros*, a public institution that watches over the *securities* exchange market. *Securities* are financial instruments such as insurances and stocks.

Every firm that trades stocks, insurances and other *securities* is registered in *El Registro de Valores de la Superintendencia de Valores y Seguros*, a law-based registry. These firms are required to submit updated information every three months, including financial statements, shareholders, directory, *OPAs* (*Oferta Pública de Acciones* or public stocks issue), acquisitions, essential facts (important facts that happened in the period, e.g meetings and directory changes), etc.

The data is publicly available at the website of *La Superintendencia de Valores y Seguros*, as required by law. The URL of this website is: http://www.svs.cl

5.1.2 Chosen data

This research is based on a particular subset of the available data:

1. The registry information of each firm.
2. The available financial accounts, for each period and firm.
3. Each firms' shareholders, including their share (proportion of stocks acquired), for each period and firm.

The relations between the above datasets are represented in figure 5.1. This entity-relationship diagram also models how the information is shown in the website.

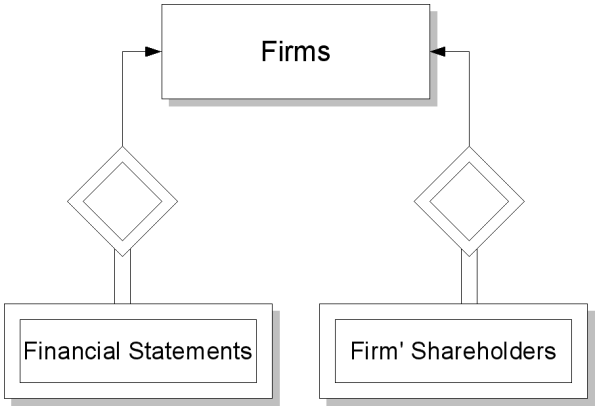


Figure 5.1: ER diagram of the available data.

Why choosing this data? As the purpose of this work is to study the investment relations between shareholders as networks, knowing the shareholders by firm is enough to build these graphs. (By *these graphs* we mean the network of investments between firms, the network of firms with common shareholders, and many other possible networks.) But it is still necessary to know the name of each firm: only then it is possible to know if a shareholder is a firm (shareholders are only known by name). Financial statements are relevant as they characterize firms as economic structures. These accounts may play a major role in the structure and dynamics of the networks. Not only that, financial statements are a sample of the behavior of the Chilean economy.

5.1.3 Possible networks

There are many networks that might be studied from the chosen data for this research.

The network of investment between firms. This network is represented by the directed graph built from the relation "A invests in B" or $A \rightarrow B$, where $A, B \in \{\text{Firms}\}$.

The network of firms related by a common shareholder. This network is represented by a non-directed graph $G(V, E)$, where $V = \{\text{Firms}\}$ and $E \subseteq V \times V$. The following condition explain the edges: $A, B \in firms, \exists C : C \rightarrow A \wedge C \rightarrow B \Rightarrow (A, B) \in E$. It is easy to see why the graph is non-directed: $(A, B) \in E \Rightarrow (B, A) \in E$ (symmetry).

The network of shareholders having shares in the same firm. Again, this network is represented by a non-directed graph $G(V, E)$, where $V = \{\text{Shareholders}\}$ and

$E \subseteq V \times V$. Note that $A, B \in V, \exists C : A \to C \wedge B \to C \Rightarrow (A, B) \in E$. Again, $(A, B) \in E \Rightarrow (B, A) \in E$ (symmetry).

5.2 Retrieval

Before performing any analysis, the data was retrieved and processed so that it could be stored in a relational database and graphs could be built from it.

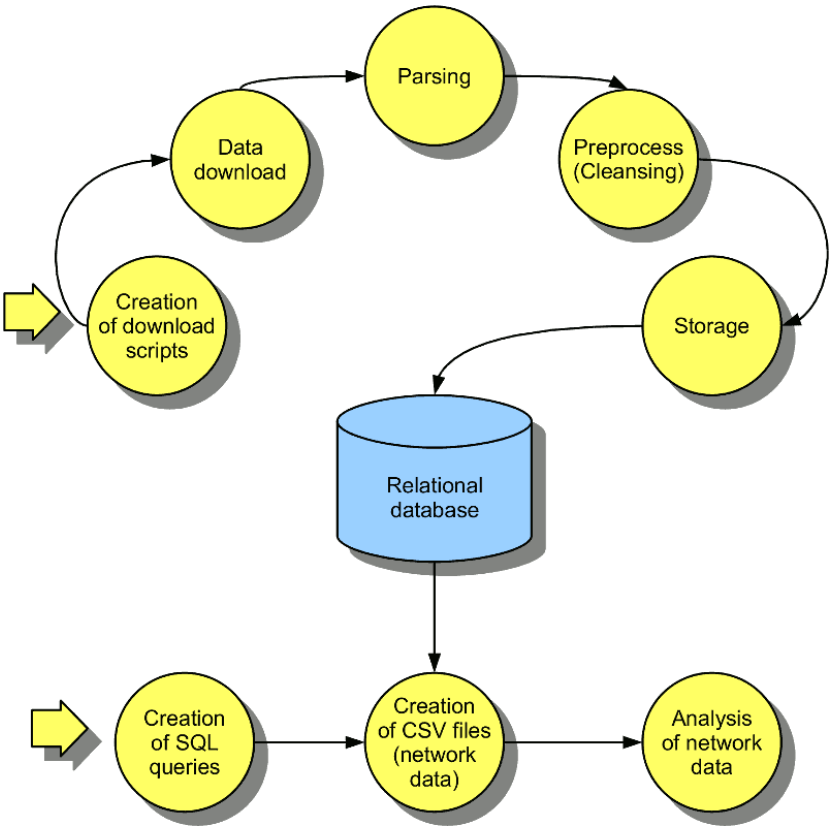


Figure 5.2: Retrieval and processing phases.

The objective of the retrieval was to download the right number of files from the website. We used the following strategy:

- Download the list of firms in the registry.

  For each firm, retrieve (download) the registry information.

- According to initial and ending terms, build the scripts for downloading the financial statements and shareholders by period.

- Run the scripts in order to download the data.

This phase took several hours (each time), from 5 to 14 hours. (This is time consuming: weeks are easily lost in script correction and extension, specially if problems are hard to detect and new data becomes available due to website updates.)

The outcome of this phase was a collection of files (html documents) containing the desired information, which required about 700 mb of disk space.

5.3 Processing

First, data was parsed from the retrieved webpages, and was stored in CSV (*comma separated values*) files. This was easy to do because most data was stored within HTML tables. In the particular case of financial statements, item names often changed, but their codes were static (numerical codes).

However, the data was filled with spelling mistakes, abbreviations, and even intentional errors (like adding zeroes or asterisks to shareholders' names). To solve these problems, two cleansing scripts were created: one for removing undesirable characters, and other for replacing words. The former was simple while the latter was difficult.

To replace words, which solves the problem of spelling mistakes and abbreviations, a list of words and their frequencies was made. By looking for short words (even single letters!), we found the abbreviations. We added these words and their corrected version in a separate table.

To identify spelling mistakes, we had to look for rare words. Spelling mistakes should not appear often (unless they are intentional). So, after identifying the spelling mistakes, we added them to a separate table with their corrected version.

After cleansing the shareholders' names, we found ourselves with yet another problem: similar names did not match. Sometimes, shareholders' names were written in detail, other times they appeared in an informal way.

To overcome the above problem, we assumed that words that are too common do not identify names. In other words, we assumed names were unique because of their keywords. Using this logic, we took the two least frequent words from each name and used it to identify them.

After we identified the shareholders, we replaced their identifiers (the two words) by numerical codes, which are easier and faster to handle. (By doing this, we hid the overall data processing phase from storage and analysis. Besides, we are only interested in numerical data; we have no use for keywords.)

We corrected some minor accuracy problems afterward (by hand), and we can say that our data is mostly correct (at least, 90% to 95% of our data is correct). We could do better, but the original data was blurry, and the whole process took several weeks because it involved lots of non automatized tasks, such as the creation of the replacement table and the visual inspection to assess the quality of the data. (We

first performed a general analysis and then we randomly took several rows from the cleansed data and verified them with accuracy.)

5.4 Integrity problems

Once the data was processed, several new problems appeared, which were caused by: non available data and publication-related problems.

Regarding non available data, we found several firms which acted as shareholders yet their financial statements were not available. These firms appear as inactive. However, they appeared active as shareholders, participating in several shareholding relations. These firms are listed in table 5.1. Since we do not have financial information regarding them, we cannot study why these firms invest in other firms.

Similarly, we found two active firms without financial statements which were not shareholders. They are shown in table 5.2.

Also, when retrieving the data, several firms appeared **without assets**. No firm should have no assets. However, studying these exceptions, we found three types of situations:

- Some mutual fund companies published their financial statements in `.pdf` format since their accountability was not standard. These firms had rather large assets (above CLP 10^{10}) and **participated in a significative number of shareholding relations**. Since they were few, and behaved differently than the rest of the firms, we left them with no assets (mostly to identify them).
- Some emerging firms had no assets. They were registered in the Registry of Values before being active. These firms had no assets for one or two periods. After that, their wealth started increasing.
- Some firms had consistently no assets. This seems like a data integrity problem from the source.

All of the firms which had no assets at some point of time are listed in table 5.3.

5.5 Storage

5.5.1 Advantages of relational databases

Looking at the networks that might be built from this data, it is possible to see that the graphs are based on a very basic relationship: "A invests in B" or $A \rightarrow B$. Other relations are based on conditions over this relationship. For example, consider the common shareholder or $(A, B) \in E \Leftrightarrow \exists C : C \rightarrow A \wedge C \rightarrow B$. The problem of building this relation is equivalent to

ID	RUT	Name
66	91333000	CIA DE PRODUCTOS DE ACERO COMPAC SA
75	92373000	INMOB E INVERSIONES ACONCAGUA SA
94	99022000	ADMINISTRACIONES E INVERSIONES SA
120	90323000	THE CENTRAL AGENCY LTD
133	94088000	AGROFORESTAL E INVERSIONES MAIHUE SA
151	90512000	CIA DE RENTAS E INVS SAN AGUSTIN SA
396	86247400	EMPRESAS AQUACHILE SA
406	91464000	INMOB ARAUCANIA SA
409	96573310	FORESTAL ARAUCO SA
781	94018000	INMOB SANTA BLANCA SA
905	93944000	INVERSIONES CABURGA SA
943	93711000	CIA PESQUERA CAMANCHACA SA
1232	93955000	CIA INMOB Y DE INVS RIO CLARO SA
1245	90743000	PROMOTORA CMR FALABELLA SA
1254	61704000	CORPORACION NACIONAL DEL COBRE DE CHILE
1449	93823000	INVERSIONES Y RENTAS COPERNICO SA
1454	93719000	SOC INMOB COPIHUE SA (EN LIQ)
1518	93388000	INMOBILIARIA CRAIGHOUSE SA
1519	93699000	CRAV ALIMENTOS SACI (EN LIQ)
1655	92247000	EMPRESA CONSTRUCTORA DELTA SA
1693	93828000	INVERSIONES Y DESARROLLOS S A
1844	92288000	DROGUERIA HOFMANN SAC
2099	90706000	ESSO CHILE SA PETROLERA
2119	61216000	EMPRESA DE LOS FERROCARRILES DEL ESTADO
2192	59056150	BANCO EXTERIOR CHILE
2201	90679000	CIA FABRICA DE PANOS BIO BIO SA
2397	93802000	SOC DE INVERSIONES SAN FRANCISCO SA
2518	90392000	CIA DE GAS DE CONCEPCION SA
2551	93832000	INMOB GENERAL SA
2622	90494000	GRACE Y CIA CHILE SA
2798	95412000	CIA INMOBILIARIA LA HISPANO CHILENA SA
2903	93755000	INVERSIONES SAN IGNACIO SA
3063	96578390	RENTAS INMOBILIARIAS SA
3146	84671700	SANTA ISABEL SA
3240	94200000	INVERSIONES JUNCAL SA
3628	93877000	SOC DE INVERS EL MAITEN SA
3659	96530650	INGENIERIA E INMOBILIARIA MANSO DE VELASCO SA
3730	61214000	EMPRESA MARITIMA SA
3748	91226000	SOC INMOB SAN MARTIN SA (EN LIQ)
3934	91199000	CIA INVERS MOB E INMOB MAR DEL PLATA S A
3954	90503000	SOC MOLINERA DE OSORNO SA
4152	87041000	INVERSIONES LAS NIEVES SA
4297	96696280	PACIFIC TRUST SA
4298	91740000	CAPITALIZACION Y RENTAS DEL PACIFICO SA
4431	92593000	EMPRESAS PENTA SA
4472	92604000	EMPRESA NACIONAL DEL PETROLEO
4476	92933000	PETROQUIMICA DOW SA
4732	92410000	PRODUCTOS AGRICOLAS PUCALAN SA
4802	90784000	INDUSTRIA DE RADIO Y TELEVISION SA
4972	92030000	COMERCIAL Y DE RENTAS RIVAS ROCES SA
5116	93601000	SOC DE INVERSIONES SAN JOSE SA
5226	95721000	SEGURAVITA SA
5576	81981500	TERCIADOS Y ELABORACION DE MADERAS SA
5591	96533620	TERRANOVA SA (DISUELTA JUNTA 011297 FUSION FORESTAL)
5611	93049000	TEXTILES ZAHR SACI
5641	96782840	SOC DE INVERSIONES TOCOPILLA SA
5701	91888000	CIA DE INVERSIONES TRANSOCEANICA SA
5815	80492300	INMOBILIARIA PEDRO DE VALDIVIA SA
5815	93478000	INVERSIONES PEDRO DE VALDIVIA SA
6003	96648500	VITAL SA

Table 5.1: Shareholder firms without financial statements.

the problem of finding a C that invests in A and B. And that problem is solved by a simple search.

As a simple exercise, let us suppose that we have the relation $R(X, Y) \Leftrightarrow (X \rightarrow Y)$. If we want to build $E(A, B)$ from $R(X, Y)$, we have to recognize that $r, s \in R$, $r[X] = s[X] \Rightarrow (r[Y], s[Y]) \in E$. (This is tuple notation. $r[X]$ means the value of the attribute X in the tuple r, pretty similar to the meaning of $r.X$ if r is considered an instance of the data structure R with fields X and Y.) So, we can build E by using relational algebra: let be $S = R$ and $T = R$, then $E = \pi_{\{S.Y,T.Y\}}(\sigma_{S.X=T.X} S \times T)$. The previous query is easily translated to SQL:

```
SELECT S.Y, T.Y FROM R AS S, R AS T WHERE S.X=T.X ;
```

The strong relation between relational handling (relational algebra) and SQL makes re-

ID	RUT	Name
121	96623460	SANTANDER SA AGENTE DE VALORES
981	91168000	FORESTAL CARAMPANGUE SA

Table 5.2: Non shareholder firms without financial statements.

ID	RUT	Name
642	98000600	AFP BANSANDER
754	99545440	BGA CHILE COMUNICACIONES SA
863	59500006	BSSF CHILE SA
865	59500007	BSSFP CHILE SA
1367	96965220	CONECT SA
1587	98001000	AFP CUPRUM
2279	76655650	INVERSIONES SANTA FE SA
2685	98000100	AFP HABITAT
3714	98000000	AFP SANTA MARIA
4544	98000500	AFP PLANVITAL
4544	98000900	AFP PLANVITAL
4544	98001200	AFP PLANVITAL
4712	98000400	AFP PROVIDA
5176	59500005	SCF CHILE SA
5237	99582620	SEMBRADOR CAPITAL DE RIESGO SA
5559	99592120	TELETUBES SA

Table 5.3: Firms which assets have been zero.

lational databases suitable for handling the retrieved data for this research. Moreover, SQL might come in handy whenever relational manipulation of this kind is required. (The restrictions are set by the expressivity of relational algebra. Nevertheless, it is possible to use recursive SQL, but this is outside the scope of the context of this research.)

Note that similar remarks were made by Adar et al (Adar and Re, 2007).

5.5.2 Some queries

We used `sqlite3` to handle our scripts. By using it, we created the files to be analyzed, using SQL.

For example, to create a file with the arcs of shareholders who invest in the same firm, we executed queries like:

```
select distinct a.id2,b.id2
from zshareholders a, zshareholders b
where a.year=2003 and b.year=2003 and a.month=12
  and b.month=12 and a.id1=b.id1
and a.id2>b.id2;}
```

To create a profile with each shareholder's attributes, we executed queries like:

```
select id1, avg(share) as AvgShare,
  count(distinct id2) as NumInvestmnt,
  max(share) as MaxShare, min(share) as MinShare,
  avg(share*share) as AvgShare2, sum(share) as TotalShare
```

```
from zshareholders
where year=2004 and month=3
group by id1;
```

To create a profile with each firms's attributes, we executed queries like:

```
select DISTINCT A.id as ID, A.assets as ASSETS,
  A.equity as EQUITY,A.debt as LT_DEBT,
  A.profits as PROFIT,A.dividends as DIVIDE,
  A.operations as OPERATION,A.investment as INVESTMENT,
  max(1.0*A.profits*A.assets/(1+A.assets)/(1+A.assets)) as RoA,
  max(1.0*A.debt/(1+A.equity)*A.equity/(1+A.equity)) as DoE,
  max(1.0*(A.debt+A.equity)/(1+A.assets)*A.assets/(1+A.assets)) as DEoA,
  max(1.0*A.investment/(1+A.assets)*A.assets/(1+A.assets)) as IoA,
  2005*4 + 3/3 - min( B.year*4 + B.month/3 ) as history
from zaccounts A, zaccounts B
where A.year=2005 and A.month=3 and B.id=A.id and A.equity>=0
group by A.id,A.assets,A.equity,A.debt,
  A.profits,A.dividends,A.operations,A.investment;
```

We created several files by changing parameters like time, amount of share, constraints on attributes (like not allowing firm with no equity), etc. This is not time consuming because these queries are created from scripts; basically we created a Perl script to create a large SQL script which, in turn, created the files to be analyzed later.

Also, by using a database with indexes, the execution of the SQL script is fairly fast. In a few minutes, all of the files with the network data are created.

Note that we used several approximations to solve issues like divisions by zero. For example, `A.equity/(1+A.equity)` is zero if `equity=0`, and as `equity` is large in general (mostly between 10^3 and 10^9), it works like a good approximation.

Also, we used the `max` aggregate operation in the SQL query because it had a `group by` operator.

5.6 Comments

We discussed how we retrieved and processed the data before performing any kind of analysis. Basically, we downloaded several webpages, parsed them to retrieve the desired information, and then processed that information so it could be used afterwards. Finally, we stored the cleansed data in a relational database, which we used to generate several views of the retrieved network.

Data processing can be very difficult. Sometimes data is not well formatted, or is blurry, like in our case. The strategy used to match shareholders' names might be of use for another purposes, as it is simple and extensible. Probably, we should get rid of the "two words" restriction. For example, to determine whether two names are the

same, we could compare the frequency of the matching words against the rest, and accept the equivalency when the similarity is above a certain threshold.

Now that the data is clean, we should work on methods for analyzing it. This is the topic of the next chapter.

References

Adar, E. and Re, C. (2007.). Managing uncertainty in social networks. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Chapter 6

Empirical methods for studying networks

Once the data was retrieved and cleansed, we found ourselves trying to analyze complex data. However, current methods were not designed for analyzing networks rich in additional information.

In the following we discuss the two empirical methods chosen for analyzing non topological network data: an existing one and a novel one.

6.1 Exploring networks

Exploratory Data Analysis (EDA) is about studying data without making prior assumptions about it. Moreover, EDA is about discovering properties from data by using tools to *explore* it. To do this, descriptive statistics is essential.

In our case, we needed to explore complex graph data. However, we could not do that with current software applications because they are intended for simpler data. In our networks, nodes do not have labels, instead have several numerical attributes, and links are weighted (we can work around this by making several views which allow certain weights). So, we needed to develop our own exploratory methods for our networks.

We decided to develop a software application to automatize the application of the exploratory methods, and support EDA for networks. The exploratory methods should be visualizations, like pie charts, scatter plots, and histograms, but oriented to graphs. Of course, our methods should be designed to handle complex data, so they do not need to be as simple as charts, but they should be clear and/or meaningful anyways.

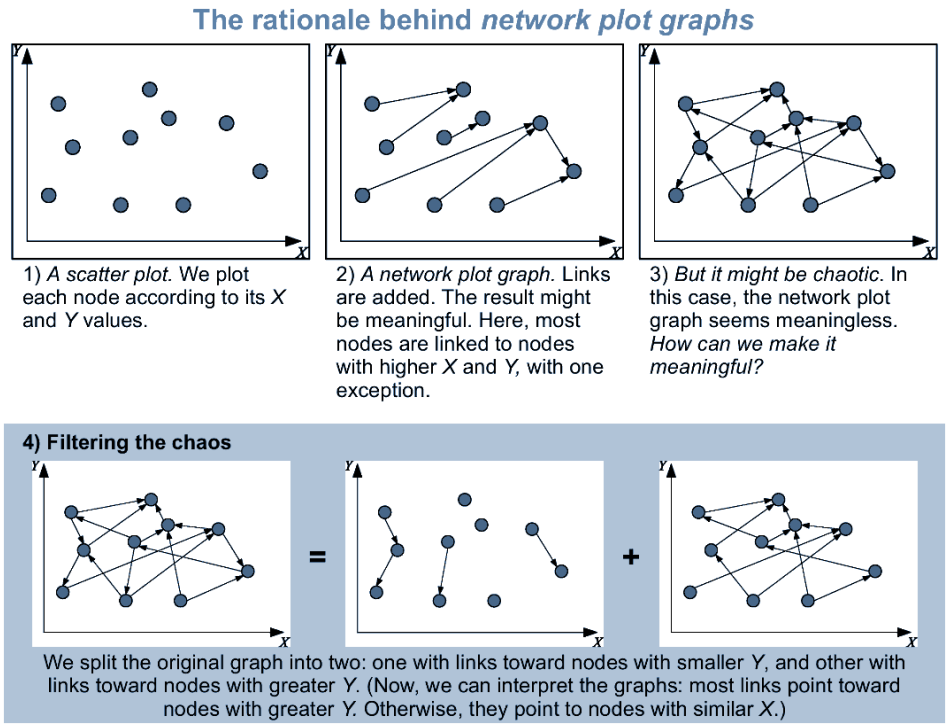


Figure 6.1: The rationale behind network plot graphs.

6.2 Visualizing graphs

6.2.1 Traditional visualizations

We already reviewed several techniques for displaying networks. However, most of these visualizations, if illustrative, do not express additional information about the data, except in the topological part.

In most visualizations, it is possible to identify communities without difficulties. This visually confirms the results of clustering algorithms.

Also, centrality maps show the relation of network structure and the centrality of the vertices. However, they do not show additional information to the researcher.

6.2.2 Network plot graphs

The objective of most network visualizations is to display a clear, aesthetic figure which aids the identification of topological properties such as clusters, hubs, patterns, etc. However, can we show different information through vertex and arc/edge arrangement? The answer is yes, by using *network plot graphs*.

In network plot graphs, vertices are arranged according to two quantitative attributes, like a plot graph. Then, after all vertices have a defined position, arcs are

drawn between them. See figure 6.1 for several examples. In particular, steps 1 and 2 show the process of drawing a network plot graph.

Additional information is always good, specially when we want to know more about something. In network plot graphs, we focus in arcs: where they begin and end. By taking this information into account, we may discover relational and non relational topological properties and how they relate to the topology. However, relational information is still visible in this visualization technique. For example, in figure 6.1, inset 2, we can observe that most vertices are linked to vertices with greater X and Y.

Networks are chaotic sometimes, at least in appearance. When this is the case, we have to resort to special strategies to simplify things up. Our solution consists in filtering arcs of certain types. This way, we may reduce the complexity of our visualization. An example of this in shown in figure 6.1, insets 3 and 4. Now, we thought of filtering arcs according to their direction: if they point toward vertices with greater X or not, and/or if they point toward vertices with greater Y or not.

But we should not stop studying relational properties. If we have the chance, we should study *paths* and small structures, like circular structures, sociometric stars, trees, etc. However, doing this could be difficult when we have too many vertices.

Network plot graphs may prove useful in some situations, however, they will be useless when working with large networks. This is also true for most graph visualizations. In these situations, choosing a different kind of visualization may be a good idea.

Another problem of network plot graphs is that they require quantitative attributes. Categorical attributes, cannot be plotted in plot graphs, unless they are converted into ordinal numbers or transformed to scores (using rubrics, for example).

6.3 Correlation profiles

6.3.1 Profiles and matrices

Correlation profiles are two dimensional histograms used to study the relation between two variables, just like scatter plots. Yet far different from the latter, correlation profiles approximate the shape of the underlying joint distributions by counting frequencies in groups or bins.

Correlation profiles can be powerful tools for the analysis of relational data. However, mainstream network analysis methods are much more related to traditional matrix arrangement or *block analysis*. Examples of block analysis and visualization are shown in figure 6.2. As it can be seen, block analysis is a useful tool for discovering patterns in adjacency matrices.

Some works resort to correlation profiles or correlation matrices instead of adjacency matrices. In these works, data is arranged in bins and shown in various ways, generally borrowed from the physics. Two examples can be seen in figure 6.3.

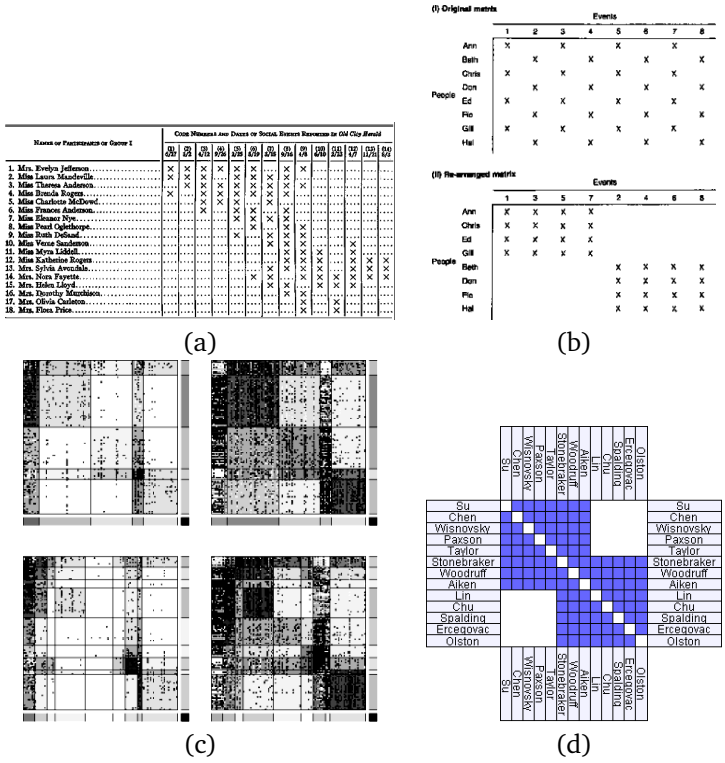


Figure 6.2: Different forms of matrix analysis. Images: (a) Southern Women data (Freeman, 2003), (b) matrix re-arrangement (Scott, 2000) , (c) products (Hidalgo et al., 2007), (d) and NodeTrix visualization (Henry et al., 2007).

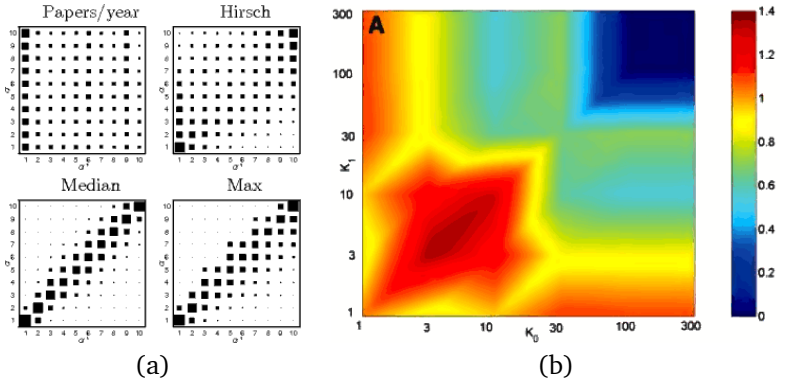


Figure 6.3: Correlation profiles in literature. Images: (a) topological measures of scientific publications (Lehmann et al., 2007) and (b) heat graph of protein-protein interaction (Maslov and Sneppen, 2002).

What are the consequences of analyzing correlation profiles instead of plain matrices? Matrices are powerful, however simple inspection is better with correlation profiles than with matrices. Correlation profiles approximate bivariate joint distributions instead of showing each element in an array.

6.3.2 Arc correlation profiles

As we discussed before, correlation profiles approximate bivariate joint distributions. How can this be useful to our analysis? **If we study why nodes are linked through their attributes, we may be able to understand the rationale behind the formation of the network.**

For example, let us imagine a social network where people become friends with other people. If we consider personal wealth, we should see that lots of people become friends with richer people (interest). However, we cannot expect the poor to be friends with the rich; often, people become friends only with people of similar wealth. If we ranked wealth as numbers from 1 to 10, and drew a correlation profile, we might see that people of rank i link to people of ranks $i + 1$ and $i + 2$, reflecting this situation. Of course, this is a hypothetical situation, but explains why correlation profiles might be useful to discover these kinds of relations.

In our particular case, we are somewhat concerned with economic rationality, so we expect to see shareholders investing in firms with particular financial statements. If so, this should be reflected in the respective correlations profiles.

However, if we use correlation profiles to study the distribution of links, we should be careful. First of all, we should split node attributes into intervals or bins. Then, we should analyze not one but several frequencies.

Let us elaborate on the friends and wealth example. We should know that the poor are much more than the rich. People in rank 1 should be much more than people in rank 10, let us say 100 in 1 and 10 in 10. This implies that if we see 10 links from people of rank 1 to people of rank 1, we should know that there are only 10 links out of the $100 \times 100 = 10000$ possible links! (a complete clique). But, if there are 2 links from people of rank 10 to people of rank 10, there are 2 links out of $10 \times 10 = 100$ possible links. If terms of probability, we are saying that two people of rank 1 are friends with probability $10/10000 = 0.001$, and that two people of rank 10 are friends with probability $2/100 = 0.02$. So, bin size IS important!

Together, we talked about three frequencies for correlation profiles: number of arcs, bin size, and arc likelihood (a probability). Probably, the last one is the most important; it shows the probability that two nodes are linked based on their attributes, without worrying about the distribution of nodes (which makes up for bin sizes). However, bin sizes might be too small sometimes, and having 1 arc out of 20 possibilities might be misleading.

For example, let us suppose we have 20 links in one bin (from rank 1 to rank 1) and just one link outside of it, in rank 5 to rank 7, which bin size is 15. Sadly, **one observation does not make up for a probability.** It is statistically meaningless to justify this isolated link; we could say that a person of rank 5 is friends with a person of rank 7 with probability $1/15$ as well as say that outside the bin from rank 1 to rank 1, the probability a person is friends with another is $1/10000$. Few observations are meaningless to statistics.

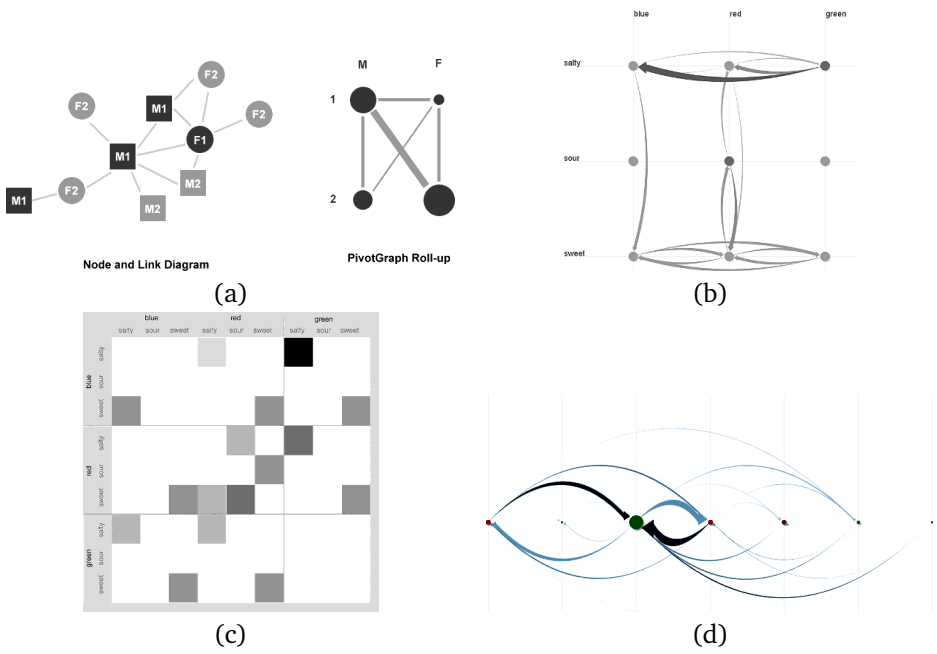


Figure 6.4: Visualizing multivariate graphs with PivotGraph. Insets: (a) summarizing graphs, (b) visualization of the summarized graph, (c) a correlation profile of the summarized graph, and (d) visualization using a linear layout.

So, we better include linking probability, number of arcs, and bin size, in our tool. This way, we prevent the problem of meaningless linking probabilities by showing the less problematic number of arcs and bin size.

One problem with arc correlation profiles is that they are useless to understand paths and structures like sociometric stars, communities, etc.; they only work at relational level. Otherwise, they are superior to network plot graphs because relational data is conveyed in a simpler way.

6.4 A surprisingly similar work

There is only one similar work: PivotGraph (Wattenberg, 2006). The insight behind PivotGraph is the creation of a *summarizing graph*. However, PivotGraph works with discrete information.

By viewing figure 6.4, we see that the figures are pretty similar to our designs. However, the central insight behind pivot graphs is the creation of a summarizing graph. In our case, we want to show detailed quantitative data.

Another difference is the aesthetical focus put on PivotGraph. As this program works with types instead of continuous data, vertices intentionally overlap, and arcs' width demonstrate the original number of arcs between vertices of these types (which heads' and tails' types match.). In our case, we want to show all of the vertices; in fact, with continuous data, vertex

overlapping should not happen often. (Unless we use ordinal data with repetitions.)

Also, arc correlation profiles of typological data have different statistical meaning than arc correlation profiles of continuous, quantitative data. For the first case, we should resort to neural networks, Bayesian networks, support vector machines, and similar technique for inference in discrete data. For the former case, which is our case, we should resort to traditional inferential statistics. (Continuous, quantitative data should be more important to quantitative sciences, like economics. Typological data should be more important to the rest. However, having too many types can be a problem, specially for theory making.)

6.5 Comments

We decided which empirical methods must be included in Network Observer, our application for network EDA. These visualization techniques should enable users to explore network data fast.

One additional requirement for our application might be a functionality for saving the created images. Another requirement might be that these images should be aesthetic and printer-friendly; if we use Network Observer to do research, we should include the created images in papers.

After developing empirical, exploratory methods, we should work on purely analytical methods, which is the topic of the next chapter.

References

Freeman, L. (2003). Finding social groups: A meta-analysis of the southern women data. In R. Breiger, C. Carley, P. P., editor, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*.

Henry, N., Fekete, J.-D., and McGuffin, M. J. (2007). Nodetrix: A hybrid visualization of social networks. In *InfoVis 2007*.

Hidalgo, C. A., Klinger, B., Barabasi, A.-L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317:482–487.

Lehmann, S., Jackson, A. D., and Lautrup, B. E. (2007). A quantitative analysis of measures of quality in science.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296:910.

Scott, J. (2000). *Introduction to Social Network Analysis: A Handbook*. Sage Publications, London, 2 edition.

Wattenberg, M. (2006). Visual exploration of multivariate graphs. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*.

Chapter 7

Analytical methods for studying networks

We now present an analytical method for studying network data with rich quantitative data. It is about the estimation of the function $f(\vec{x}, \vec{y})$ which indicates the likelihood a vertex is linked to another, based on their quantitative attributes $\vec{x}$ and $\vec{y}$.

7.1 An enlightening insight

When studying the shareholding networks of Japan as directed graphs, Garlaschelli et al defined a distribution function which modeled the probability that two vertices were connected by an arc (Garlaschelli et al., 2005). This function had two parameters: the two linked vertices. They observed that, with their data, the function could be more or less simplified by variable separation. In other words:

$$f(x, y) = \phi(x)\psi(y)$$

where $f(x, y)$ is the probability[1] that there exists an arc from x to y.

Garlaschelli et al sorted the vertices (matrix arrangement, like in figure 7.1), making $f(x, y)$ meaningful. However, they noted that three restrictions had to hold. First, $f(x, y)$ should not estimate a different number of arcs than the existing ones ($|A|$):

$$\sum_x \sum_y f(x, y) = |A|.$$

Two, by summing all of the firms up (y), one gets the outdegree distribution of shareholders:

$$k^{Out}(x) = \sum_y f(x, y) = \phi(x) \sum_y \psi(y).$$

[1] The original $f(x, y)$ modeled the probability a firm x had a shareholder y, the reverse of our function.

And three, by summing all of the shareholders up (x), one gets the indegree of firms:

$$k^{In}(y) = \sum_x f(x,y) = \psi(y) \sum_x \phi(x).$$

These two restrictions significantly simplified the problem of fitting the joint distribution $f(x,y)$ because the problem was reduced to fitting the in and outdegree distributions: $k^{Out}(x)$ and $k^{In}(y)$. For example, having computed the degree distributions, one could write

$$\begin{aligned} k^{Out}(x)k^{In}(y) &= \phi(x)\psi(y) \sum_x \phi(x) \sum_y \psi(y) \\ &= f(x,y) \sum_x \sum_y f(x,y) \\ &= |A| f(x,y), \end{aligned}$$

and immediately obtained $f(x,y)$.

Why separating variables? When studying topological data, researchers often resorted to analyzing adjacency matrices (block analysis) to discover communities (in social network analysis), clusters, relevant players, and to visualize the distribution of links. This is mainly true when graphs are not directed since their adjacency matrices are symmetrical, making variable separation easy to accomplish due to the rectangular patterns. We can see this in figure 7.1, where matrix A corresponds to the original network data and B is the reordered one. Note that after doing this, it is common that variable separation works well for estimating $f(x,y)$. (Nevertheless, non directed adjacency matrices become symmetrical with respect to the diagonal, which should make $f(y+x, y-x)$ easier to study than $f(x,y)$; in particular, the property $f(y+x, y-x) = f(y+x, x-y)$ should simplify things up.)

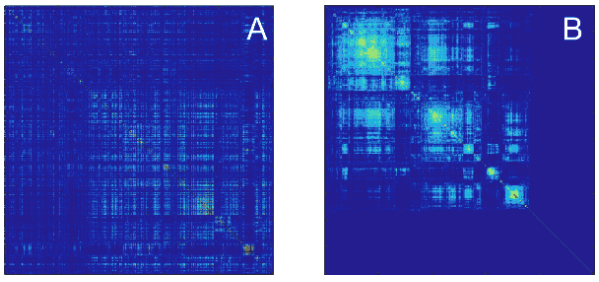


Figure 7.1: Reordering of rows and columns. These matrices correspond to the product space work from Hidalgo et al. (2007).

7.2 The continuous case

Far different from previous works, we are not interested in discrete distributions. We have vertices with many quantitative attributes whose distribution is best studied using continuous probability distributions (probability density functions).

The previous statement also implies that the parameters of $f(x, y)$ are not vertices but their attributes, and that we should avoid the density domain. Otherwise, we may have problems with the division of random variables of different kinds, which is not a trivial problem. Instead, we will recover the number of arcs and the distributions of x and y.

Analyzing our data, we found that the probability distributions are bell-shaped (like Gaussian distributions) but had strong skewness, and that X and Y random variables were in general correlated, specially when they were in the same domain (assets, debt, equity, etc.). The former observation implies that we need additional transformations to process the data, and the latter implies that we will not be able to separate variables in most cases.

7.2.1 The bivariate case

First, we study the bivariate case. We have two attributes, one for the tail of an arc (x) and one for the head (y). These values are distributed among the nodes following the frequency functions $X(x)$ and $Y(y)$. (These are frequencies, not distributions; for example, $\int_{-\infty}^{+\infty} X(x)dx = N$, where N is the number of nodes.) Also, the frequency of arcs if $A(x, y)$, and the joint pdf which says whether two nodes are linked through their attributes x and y is $\rho(x, y)$. In other words:

$$A(x, y) = X(x)Y(y)f(x, y)$$

$$\Rightarrow f(x, y) = \frac{A(x, y)}{X(x)Y(y)},$$

and note that we are interested in $f(x, y)$.

By having $A(x, y) = \phi(x, y)\psi(y)$, $f(x, y)$ becomes:

$$f(x, y) = \frac{\phi(x, y)\psi(y)}{X(x)Y(y)}$$

Given that $A(x, y)$ is a frequency function, it has already absorbed the distributions of x and y. So, the three restrictions presented in the previous section apply here as:

$$\int_X \int_Y A(x, y)dydx = |A|$$

$$\Rightarrow \int_X \int_Y f(x, y)Y(y)X(x)dydx = |A|$$

The outdegree constraint:

$$\int_Y A(x,y)dy = k^{Out}(x)X(x)$$

$$\Rightarrow k^{Out}(x) = \int_Y f(x,y)Y(y)dy$$

And the indegree constraint:

$$\int_X A(x,y)dx = k^{In}(y)Y(y)$$

$$\Rightarrow k^{In}(y) = \int_X f(x,y)X(x)dx$$

In the special case of variable separation, we can compute $k^{Out}(x)$ and $k^{In}(y)$ first, and multiply them:

$$\begin{aligned} k^{Out}(x)k^{In}(y) &= \frac{(\int_Y A(x,y)dy)(\int_X A(x,y)dx)}{X(x)Y(y)} \\ &= \frac{(\phi(x)\int_Y \psi(y)dy)(\psi(y)\int_X \phi(x)dx)}{X(x)Y(y)} \\ &= \frac{\phi(x)\psi(y)\int_X\int_Y \phi(x)\psi(y)dydx}{X(x)Y(y)} \\ &= |A|f(x,y) \end{aligned}$$

getting the desired $f(x,y)$ function, that represents the linking behavior of vertices (firms in our case) without taking into consideration the distribution of the chosen attributes. Recall that this level of isolation enables us to study this network at a local, actor level.

7.2.2 The multivariate case

First of all, when analyzing the arc likelihood between vertices, we are not restricted to study different attributes (one for the head, another for the tail). Now, we can focus on the relevant attributes, whenever vertices act as heads or tails in arcs. So, we now have to estimate only one joint pdf for vertices: $v(\vec{x})$. Therefore, the frequency of arcs is now:

$$a(\vec{x},\vec{y}) = v(\vec{x})v(\vec{y})f(\vec{x},\vec{y}).$$

Now, the problem of estimating $f(\vec{x},\vec{y})$ is somewhat *reduced* to estimating $v(\vec{x})$ (or $v(\vec{y})$) and $a(\vec{x},\vec{y})$. If we estime them as joint pdfs, then

$$a(\vec{x},\vec{y}) = Aa^e(\vec{x},\vec{y})$$

and

$$v(\vec{x}) = Nv^e(\vec{x}),$$

where v^e and a^e are estimated joint pdfs, and A and N are the number of arcs and vertices respectively. Note that the arc likelihood function becomes:

$$f(\vec{x}, \vec{y}) = \rho \frac{a^e(\vec{x}, \vec{y})}{v^e(\vec{x})v^e(\vec{y})},$$

where $\rho = A/N^2$ is the density of the network (allowing loops).

After doing this, **the problem of estimating $f(\vec{x}, \vec{y})$ has been reduced to the estimation of joint pdfs.**

7.3 Simplifying complex problems

The function $a(\vec{x}, \vec{y})$, as how we defined it, can be subject to additive separation by splitting the data in parts, simplifying the analysis of particular distributions. For example, if we separate the arcs of the digraph $G(V, A)$ so that $A = A_1 \cup A_2$ and $A_1 \cap A_2 = \emptyset$ (a partition), the frequency distributions of A_1 and A_2, namely $a_1(\vec{x}, \vec{y})$ and $a_2(\vec{x}, \vec{y})$, comply with

$$a(\vec{x}, \vec{y}) = a_1(\vec{x}, \vec{y}) + a_2(\vec{x}, \vec{y}).$$

It is easy to see that by summing up the number of arcs, indegrees and outdegrees of $G(V, A_1)$ and $G(V, A_2)$ we recover these of $G(V, A)$, since A_1 and A_2 are a partition of A. However, even though the counting properties hold, and that the sample frequency functions hold $a^*(\vec{x}, \vec{y}) = a_1^*(\vec{x}, \vec{y}) + a_2^*(\vec{x}, \vec{y})$, the separation $a(\vec{x}, \vec{y}) = a_1(\vec{x}, \vec{y}) + a_2(\vec{x}, \vec{y})$ will be as good as the goodness of fit of $a_1(\vec{x}, \vec{y})$ and $a_2(\vec{x}, \vec{y})$.

Note that, in general, we can write:

$$a(\vec{x}, \vec{y}) = \sum_i a_i(\vec{x}, \vec{y})$$

where $\{A_i\}_{i \in \mathbb{N}}$ is a partition of A, since we can perform an induction on the partitions. For example, the partition $\{A_1, A_2\}$ can be transformed to $\{A_1, A_2', A_3\}$ by splitting A_2 into $\{A_2', A_3\}$. We can repeat this process as many times as we want, being careful not to create meaningless A_i, like singletons. Each A_i has to have enough arcs to make $a_i(\vec{x}, \vec{y})$ statistically meaningful.

Partitioning the set of arcs A and using the property $a(\vec{x}, \vec{y}) = a_1(\vec{x}, \vec{y}) + a_2(\vec{x}, \vec{y}) + ...$ is specially useful when estimating a from a^* is hard while each a_i fits well to a_i^*; this way, it is possible to build a to fit well. (Basically, we are talking about estimating functions using a *divide and conquer* strategy.)

7.4 Modeling network dynamics

A first solution to this problem is the estimation of the arc functions at different times. This could gives us a general idea of how the network evolved, and if it changed rapidly or not.

A second solution is the estimation of the differences of the network, let us say $G_{t+1} - G_t$, which is also a network. The problem is that a network may change in several ways:

1. Vertex birth.
2. Vertex death. This destroy all of the arcs related to the vertex.
3. Arc birth.
4. Arc death, not caused by vertex death.
5. Vertex attribute change.
6. Arc weight change.

The problem of vertex birth and death is plain statistics. The same applies changes in vertex attributes and arc weights. However, birth death may be related to existing arcs.

Arc birth and death can be studied through the estimation of the arc functions for $G_{t+1}-G_t$ and G_t-G_{t+1}, respectively. By using this information, and vertex birth and death, it is possible to estimate G_{t+1} from G_t. Note that we modeled the evolution of the network to some extent with this. The last problem would be relating the change to the structure.

7.5 Comments

We defined a general framework for analyzing network data at relational level, by using joint pdfs. We defined how to estimate the arc likelihood function, $f(\vec{x}, \vec{y})$, in terms of other functions. However, how do we estimate the other functions? This problem is solved in the next chapter, where we present a novel technique for estimating joint pdfs from data. Of course, all of this is to be implemented in our application, Network Observer.

References

Garlaschelli, D., Battiston, S., Castri, M., Servedio, V. D. P., and Cardarelli, G. (2005). The scale-free topology of market investments. *Physica A*, 350:491–499.

Hidalgo, C. A., Klinger, B., Barabasi, A.-L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317:482–487.

Chapter 8

Estimating joint probability density functions

Previously, we identified the need of estimating joint pdfs from complicated data. However, estimating joint pdfs is a rather complicated task, and more research results are needed.

We now address the problem of estimating joint pdfs from data by developing an algorithm which estimates joint pdfs from several one-dimensional pdfs, under very simple assumptions.

8.1 Current methods

We now review current methods for estimating joint pdfs or cdfs. We also justify why we do not use them in our work.

8.1.1 Classic approach

There are several ways to estimate or approximate joint pdfs. The classic approach is concerned with the definition of joint pdfs with traditional distributions. In this line, we have the multivariate normal distribution, the multivariate Student distribution, the Wishart distribution (a generalization of chi square to many dimensions), etc. (Kotz et al., 2000).

8.1.2 Chow-Liu trees

Chow-Liu trees are a method for estimating joint probability distribution functions of discrete data, which consist in specifying which variables are related using a tree structure, much like Bayesian trees Chow and Liu (1968). The best property of Chow-Liu trees is that, if they are

built as maximum likelihood estimators, then they can be build using a maximum spanning tree algorithm, like Kruskal's.

Chow-Liu trees are not suitable to us because they are intended for discrete data. And if we discretized our data, the result might not be analytically useful. We want to obtain analytical, symbolic functions.

8.1.3 Machine learning techniques

The machine learning approach, which has been successful in approximating joint pdfs, uses discrete models to estimate any kind of functions (Mjolsness and deCoste, 2001). Typical examples are neural networks, Bayesian-belief networks, support vector machines, etc. which are designed to work with discrete data, but are extended to continuous data using discretization and fuzzy methods. In particular, fuzzy methods alone have been used to approximate joint pdfs as well (Buckley, 2005).

8.1.4 Copulas

In what is directly concerned with continuous data, we find an active research subject: copulas. A *copula* is a function defined on $[0,1]^n$ (n dimensions) which estimates joint cdfs from marginals cdfs (Schmidt, 2007), and can naturally be extended to estimate joint pdfs.

Research on copulas started after Sklar's theorem in 1959, when he demonstrated that always exists a copula $C : [0,1]^d \rightarrow [0,1]$ which estimates any given joint cdf. Each argument of the copula is marginal cdf evaluated on its parameter, i.e.

$$C(F_1(x_1), ..., F_n(x_n)) = F(x_1, ..., x_n),$$

where F_i, $i = 1..n$, are marginal cdfs, and $F(x_1, ..., x_n)$ is the joint cdf.

Copulas have several properties, like being increasing in each parameter, and that by setting parameter in 1, it is possible to build marginal cdfs and joint cdfs. For example,

$$F(x_1, x_2) = C(F(x_1), F(x_2), 1, ..., 1),$$

and also,

$$C(1, ..., 1, z, 1, ..., 1) = z,$$

because it recovers the marginal cdf.

Research on copulas has led to the development of several important copulas. Among these, there are Gaussian and Student-t copulas (fig 8.1), independence copulas ($C(u_1, ..., u_n) = u_1 \times ... \times u_n$), perfect positive dependence ($C(u_1, ..., u_n) = min\{u_1, ..., u_n\}$), archimedean copulas (Gumbel copulas, Frank copulas, Clayton copulas, ...), etc.

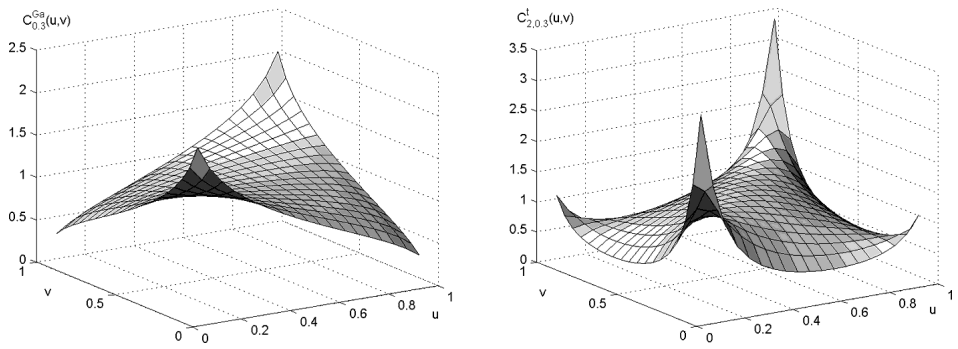


Figure 8.1: Two well studied copulas: Gaussian copula (left) and Student-t copula (right). Taken from Schmidt (2007).

Copulas have been actively used in various research fields, like finance, to model dependence. Note that variables are not only related through correlations (linear), but also at the tails of their distributions, any many more ways. (For example, $Z = X^Y$ is non linearly related to both X and Y.) To model dependence, sometimes functions like the following are used:

$$C(x, y, z) = A(x, y)B(y, z),$$

where the relations between x and y and y and z are explicitly modeled. Functions specified like this have certain appeal for modeling variable dependence.

Our problem with copulas is that there is no way to infer their shape from data, unless we try all the known formulas. Their analytical shapes are already predefined, and to model new cdfs, new copulas are needed. So, including them in Network Observer does not seem like a good idea.

8.2 A novel method

As we needed a more general method for computing joint pdfs, we had to develop our own method. Here, we reproduce how we deduced it.

Let us recall two particular cases of pdfs. First, if $\vec{X}$ is a vector of independent random variables, then its joint pdf is:

$$f(x_1, x_2, ..., x_n) = f_1(x_1)f_2(x_2) \cdots f_n(x_n),$$

and if not, one can generally write the joint pdf as:

$$f(x_1, x_2, ..., x_n) = f_1(x_1)f_2(x_2|x_1) \cdots f_n(x_n|x_1, ..., x_{n-1}),$$

where

$$f(a|b) =_{\text{def}} \frac{\partial}{\partial a} \lim_{\epsilon \to 0} \frac{\Pr(A \leq a \wedge B \in \mathcal{B}(b, \epsilon))}{\Pr(B \in \mathcal{B}(b, \epsilon))},$$

and $\mathcal{B}(b, \epsilon)$ is the ball centered in b with radius ϵ.

What if we could reduce a joint pdf of the second form to a pdf of the first form? If we could do that, then the problem of computing the pdf is reduced to reducing the problem and computing the marginals.

In the following, we suppose $X_1, ..., X_n$ have n degrees of freedom, so $Y_1, ..., Y_n$ can be created. If not, some variables may be discarded due to functional dependency.

8.2.1 Joint pdfs from marginals

Let us suppose we can define variables Y_k so that

$$Y_k =_{\text{def}} \phi_k(X_1, ..., X_k), \forall k \leq n,$$

so that $\{Y_k\}$ is a set of independent random variables and $\{\phi_k\}$ are increasing in their last parameter. Then, the joint pdf of $\vec{Y} = (Y_1, ..., Y_n)$ is

$$g(y_1, ..., y_n) = g_1(y_1) \cdots g_n(y_n).$$

In particular $g_k(y_k) = g_k(y_k|y_1...y_{k-1})$ because of their independence.

Now, let us focus on $f_k(x_k|x_1...x_{k-1})$. We first see that:

$$f_k(x_k|x_1...x_{k-1}) = \frac{\partial}{\partial x_k} \Pr(X_k \leq x_k|(\forall i < k)X_i = x_i).$$

Since ϕ_k is increasing in its last parameter:

$$\Pr(X_k \leq x_k|(\forall i < k)X_i = x_i) = \Pr(\phi(X_1, ..., X_k) \leq \phi(X_1, ..., X_{k-1}, x_k)|(\forall i < k)X_i = x_i).$$

Simplifying the latter term:

$$\begin{aligned}
\Pr(X_k \leq x_k|(\forall i < k)X_i = x_i) &= \Pr(\underbrace{\phi(X_1, ..., X_k)}_{Y_k} \leq \phi(X_1, ..., X_{k-1}, x_k)|(\forall i < k)X_i = x_i) \\
&= \Pr(Y_k \leq \underbrace{\phi(X_1, ..., X_{k-1}, x_k)}_{(\forall i<k)X_i=x_i}|(\forall i < k)X_i = x_i) \\
&= \Pr(Y_k \leq \phi_k(x_1, ..., x_k)| \underbrace{(\forall i < k)X_i = x_i}_{\text{equivalent to } (\forall i<k)Y_i=\phi_i}) \\
&= \Pr(\underbrace{Y_k \leq \phi_k(x_1, ..., x_k)|(\forall i < k)Y_i = \phi_i(x_1, ..., x_i)}_{Y_k \text{ is indepedent of } Y_{i<k}}) \\
&= \Pr(Y_k \leq \phi_k(x_1, ..., x_k)).
\end{aligned}$$

Going back to f_k:

$$\begin{aligned} f_k(x_k|x_1...x_{k-1}) &= \frac{\partial}{\partial x_k}\Pr(X_k \leq x_k|(\forall i<k)X_i = x_i) \\ &= \frac{\partial}{\partial x_k}\Pr(Y_k \leq \phi_k(x_1,...,x_k)) \\ &= \underbrace{g_k(\phi_k(x_1,...,x_k))}_{\text{The marginal pdf of } Y_k}\frac{\partial \phi_k}{\partial x_k}. \end{aligned}$$

Therefore, the joint pdf of $X_1,...,X_k$ is:

$$f(x_1,...,x_n) = \Pi_k\left(g_k(\phi_k(x_1,...,x_k))\frac{\partial \phi_k}{\partial x_k}\right).$$

Note that computing f was reduced to computing each g_k and ϕ_k. In particular, computing g_k corresponds to the classic problem of estimating univariate pdfs from samples. So, the difficulty of estimating f_k was really reduced to computing ϕ_k.

Note that we cannot use this approach to compute joint cdfs. If it holds that

$$F(x_1,...,x_n) = G(\phi_1(x_1),...,\phi_n(x_1,...,x_n)),$$

then it also holds that

$$\int_{\Omega_F} g(Y_1,...,Y_n)d\vec{Y} = \int_{\Omega_G} g(Y_1,...,Y_n)d\vec{Y},$$

where

$$\Omega_F = \{\phi_1(X_1), \phi_2(X_1,X_2),...,\phi_n(X_1,...,X_n) : (\forall i \leq n)X_i \leq x_i\}$$

and

$$\Omega_G = \{Y_1, Y_2,...,Y_n : (\forall i \leq n)Y_i \leq \phi_i(x_1,...,x_i)\}.$$

The above condition requres that $\Omega_F = \Omega_G$, a condition we cannot guarantee. For example, consider that $X_1 = Y_1$ and $X_2 = Y_2 - Y_1$; $X_1 \leq 0$ and $X_2 \leq 0$ is equivalent to $Y_1 \leq 0$ and $Y_2 \leq Y_1$ (Ω_F), which is different to $Y_1 \leq 0$ and $Y_2 \leq 0$ (Ω_G).

To address the problem of computing joint cdfs, we suggest using Monte Carlo methods (Robert and Casella, 2005), which are not affected by the *curse of dimensionality* as their convergence rate is independent of the number of dimensions (but instead are very slow) and are simple to implement.

8.2.2 Choosing ϕ_k

Which ϕ_k functions should we use? We propose reducing the problem of computing $\phi_k(x_1,...,x_k)$ to the problem of computing $\psi_k(y_1,...,y_{k-1},x_k)$, a function which creates

a new random variable by making x_k independent from $y_{i<k}$. Note that ϕ_k can be reduced to ψ_k,

$$\phi_k(x_1, ..., x_k) = \psi_k(\psi_1(x_1), \psi_2(\psi_1(x_1), x_2), \psi_3(\psi_1(x_1), \psi_2(\psi_1(x_1), x_2), x_3), ..., x_k),$$

and it holds that

$$\frac{\partial \phi_k}{\partial x_k} = \frac{\partial \psi_k}{\partial x_k}.$$

Dealing with such complicated form of ψ_k is not hard. We just have to keep a record of our transformations, to use the already computed variables in the next steps: $y_1 = \psi_1(x_1)$, $y_2 = \psi_2(y_1, x_2)$, $y_3 = \psi_3(y_1, y_2, x_3)$, ... $y_n = \psi_n(y_1, ..., y_{n-1}, x_n)$.

Now that the problem was reduced to using ψ_k, the next problem consists in estimating these functions.

For the case of linear relations among variables, we can use functions

$$y_k = \psi_k(y_1, ..., y_{k-1}, x_k) = x_k - \sum_{i<k} \alpha_i y_i$$

to build the set of independent random variable. Note that y_k can be obtained using a Gram-Schmidt method. Also, $\partial \psi_k / \partial x_k = 1$, which simplifies its use a bit.

8.3 Implementation

After we developed our method, we had to develop a software that computes joint pdfs using it.

The input is a single CSV file, which represents a spreadsheet. Each column represents a random variable, and headings (first row) were asummed. We did not bound the number of columns of the input file, so the number of dimensions is not an issue save for the available RAM.

The output is a textual description of the joint pdf computed. However, additional information and functionalities were added to augment the value for the users.

8.3.1 Design

Fortunately, our method for computing is suitable for object-oriented programming. We decided to code it in Java, which is an object-oriented language.

We decided to implement the software using two packages: one for computing joint pdfs, and other for the graphical user interface.

The package for computing joint pdf consists of eight classes: the class `JointPdf` for computing the joint pdf (an interface), the class `pdfType` which acts as the type for pdf objects, and six pdfs which extend `pdfType`.

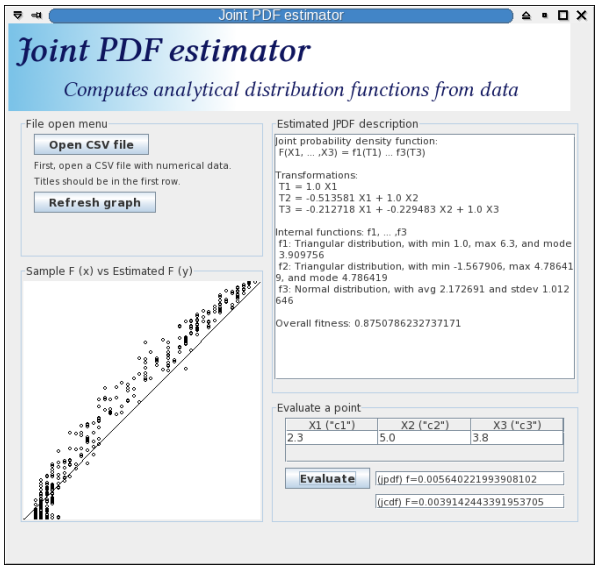


Figure 8.2: Joint PDF Estimator. Evaluated with cc42a dataset.

`JointPdf` manages all the computation. Its constructor receives a two-dimensional array of `doubles` (a `double` is a number type). Using this array, `JointPdf` executes a Gram-Schmidt routine to create a set of linearly independent random variables. Then, it creates a `pdfType` object for each variable. After the joint pdf is computed, `pdfType` can evaluate the built joint pdf and joint cdfs, and can generate a human readable version of the joint pdf.

`pdfType` computes the pdf for a random variable, which is an object that actually extends `pdfType` (a factory method). Each `pdfType` class can compute its pdf and cdf, and know its fitness score (which is simply an R^2 measure).

The graphical user interface features a text box where the readable version of the joint pdf is displayed, a graph that shows a scatter plot of the sample joint cdfs against the estimated joint cdfs, and an input for evaluating the joint pdf and joint cdfs.

8.3.2 Pdfs included

Due to the object-oriented nature of our software, adding new pdfs is fairly simple, and we included several well known pdfs in it:

Normal (Gaussian) It is well known that most random variables follow a Normal distribution. This is also supported by the *law of large numbers*, which explains why most averages follow Normal distributions.

We implemented this method using approximations. In particular, we approximated its cdf by using

$$F(x) \sim \frac{1}{1 + e^{-1.8138x}}$$

when $\mu = 0$ and $\sigma = 1$. For other μ and σ, we used $\overline{x} = (x - \mu)/\sigma$.

Box-Cox method Box-Cox method transforms random variables to Normal by powering them. We used the following transformation

$$z = \frac{x^\lambda - 1}{\lambda g^{\lambda-1}},$$

where g is the geometric mean of x. Note that z and x share the same units and that z becomes the logarithm of x ($z \to g\log(x)$) when $\lambda \to 0$, so this method support Lognormal distributions.

Box-Cox method is the most time-expensive one among the pdfs. Because of this, Box-Cox is used only when other pdfs de not achieve enough accuracy. Also, to improve the performance of our implementation, the iterated over λ until the skewness of the data was close to zero (symmetry) to avoid testing the normality of the data.

Power-law This pdf was included because it is quite common in social contexts, like in the distribution of wealth and degree distributions of several social networks.

This pdf has the form $f(x) = (\beta-1)x^{-\beta}$, where β is estimated using $\beta = 1+\frac{1}{\sum_k \log(x_k)/n}$.

Exponential This pdf was included because it is somewhat common. It often appears in memoryless processes, such as waiting times.

Triangular This pdf was included because it might fit data that Normals cannot fit.

Uniform This pdf was added "just in case" since it is most common in precision errors. Also, cdfs evaluated in their random variables follow the Uniform distribution.

Note that Box-Cox method, Power-Law and Exponential pdfs do not allow negative numbers. We patched this problem by using $\overline{x} = x - \min x$ or $\overline{x} = x - \min x + 1$ (Power-Law).

8.3.3 Quality of the estimations

To assess the quality of the estimation, we avoided using Pearson or similar goodness of fit tests which are based in histograms, as they need large samples to test goodness of fit. (If 10 bins are needed in one dimension, then 100 are needed in two, 1000 in three, etc. and we need enough observations per bin to perform goodness of fit tests.) And uniformity based tests which work with cdfs, like Kolmogorov-Smirnov or Cramer von Mises, are not suitable to test joint cdfs when variables are related (as this destroys any possibility of uniformity). Note that there exists a proposal for using a modified Kolmogorov-Smirnov that should work with multivariate data (Justel et al., 1997), and, in our case, could be translated to a goodness of fit using $\{G_i(Y_i)\}$ instead of $\{X_i\}$; note that $\{G_i(Y_i)\}$ is a set of indepedent uniform random variables, which simplifies the test a lot. However, we would like to include all the numerical and statistical errors in the evaluation of F_e. (But the work of Justel et al. (1997) might be used to test the quality of the ϕ_k functions chosen.) So, we decided

to compare the estimated joint cdfs to the sample joint cdf (which is computed by counting) expecting that:

$$F_e(x_1, ..., x_n) = F_s(x_1, ..., x_n) + \epsilon,$$

(where ϵ is a small error).

To measure the similarity of F_e (estimated joint cdf) and F_s (sample joint cdf), we generate a large random sample. For each observation, we compute both F_e and F_s, and then we perform a linear regression between them. In particular, we expect: high correlation ($r^2 \approx 1$), a null intercept ($b \approx 0$ in $F_e = aF_s + b$), and an inclination of one ($a \approx 0$ in $F_e = aF_s + b$). The above expectations led us design a simple score for assessing the goodness of fit of our data:

$$q = r^2 \left(\frac{1}{1 + |b|}\right) \left(\frac{1}{1 + |a - 1|}\right).$$

(Note that we need to perform a weighted linear regression in order to balance the uneven distribution of observations.)

8.4 Experimental results

We tested our software using several datasets. First, we tested it using four datasets which were random sampled by us. If X, Y, Z are independent random variables, then the columns of Ss1 are $A + 0.5B$, $C - A$, B; these of Ss2 are $A^{1.1} + B^{1.1}$, $B^{0.9} - 10C^{1.1}$, $A + C + AC/1000$; these of Ss3 are $A + AB/1000$, $B - \log C$, $20 \log(C + 5)$; and these of Ss4 are $(A + B)^2$, $(C - B)^2$, C^2. Ss2, Ss3 and Ss4 where designed to evaluate the tolerance of our estimations when slight nonlinear relations appeared.

Second, we used course data. These datasets correspond to the grades of students of Universidad de Chile in three different courses. We expect the presence of linear relations in these datasets.

Third, we used the voting intentions in the Chilean 1988 Plebiscite dataset (Fox, 2008). It has 2700 observations and involves 5 variables which relations are somewhat linear.

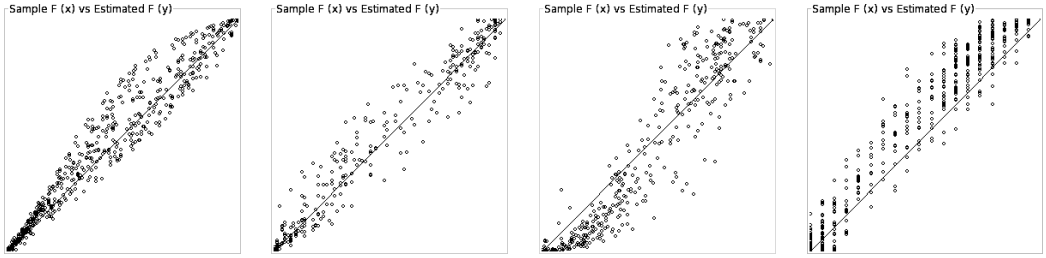
And fourth, we used several datasets from The Andrews & Herzberg archive at StatLib (Vlachos and Meyer, 1989). To use them, we removed their first columns and left only numerical sampled data. (We removed text data, row numbers, etc.) The sizes of these datasets range from 10 to 302.

Table 8.1 shows the results of the tests. Note that "Fitness" corresponds to the g measure we already defined, and its value is randomly generated (because we create a random sample to estimate it). Also note that F_e is randomly generated too (using Monte Carlo integration) and that F_s is estimated from the dataset by counting (another source of error). We cannot expect g to be close to 1 (for example, 0.97) under these conditions.

Dataset	Size	Variables	Fitness	Comments
Ss1	157	3	0.93	Fine, yet slightly biased around $F_s = 0$.
Ss2	157	3	0.88	Works fine, somewhat dispersed.
Ss3	157	3	0.88	Somewhat disperse and biased around $F_s = 0.75$. (Fig.8.3-a.)
Ss4	157	3	0.78	Small bias. Dispersion increases as F_s decreases.
cc10b	63	4	0.80	Works fine. Last variable ignored.
cc42a	41	3	0.87	Low dispersion, but slightly biased. (Fig.8.2.)
in50a	84	6	0.80	High dispersion. Seems unbiased.(Fig.8.3-b.)
voting	2700	5	0.65	High dispersion. Slight bias around $F_s = 0$. (Fig.8.3-c.)
T01.1	50	12	0.45	High dispersion.
T03.1	38	6	0.73	Small bias.
T07.1	184	4	0.30	Works bad.
T13.1	105	12	0.33	Works bad.
T17.1	127	12	0.28	Works bad.
T25.1	264	4	0.44	Works bad. Strong bias.
T30.1	19	6	0.81	Works well despite sample size. (Fig.8.3-d.)
T33.1	100	5	0.69	Disperse but overall fine.
T35.2	52	6	0.79	Work fine. Slight bias around $F_s = 1$.
T41.1	10	14	0.38	Works bad. Only 10 variables considered.
T53.1	302	9	0.12	Works bad.
T59.1	42	8	0.56	High dispersion. Seems unbiased.
T60.1	104	4	0.84	Works well. A bit disperse.

Table 8.1: Evaluation of Joint PDF Estimator with several datasets.

As we can see, our application worked well with several datasets. The results of datasets Ss2, Ss3 (fig.8.3-b) and Ss4 show that it is somewhat tolerant to nonlinear relations. The result of dataset T30.1 (fig.8.3-d) shows that it can approximate joint pdfs even when samples are small. (By taking relations into account, the *curse of dimensionality* should be reduced. For example, if $corr(X, Y) \approx 1$, we need as many observations as these needed to estimate X solely when estimating $f(x, y)$. Dimensions are relevant as the uncertainty increases.) Also, the results of the estimations give validity to the methodology proposed and the measure g included in the program.



(a) Ss3 (b) in50a (c) voting (d) T30.1

Figure 8.3: Scatter plots generated by Joint PDF Estimator.

8.5 Comments

Our novel method for estimating joint pdfs from sample data estimated precise functions for most of our datasets. We found that the estimations were good even though sample cumulative functions are rather discrete, imprecise functions, and we computed the cumulative joint pdf using random techniques.

Our method worked well when the relation between variables was more or less additive. However, when this was not the case, the precision of our method worsened.

We are aware that using correlations to build independent variables limits the generality of our method. To overcome this limitation, we thought of several solutions:

1. Using more advanced functions y_k to build the set of independent variables Y_k. For example, using a different operator instead of covariance. This is the most promising method as it is based on the developed theory.

2. Patching the bias of the estimated joint pdf with respect to the sample joint pdf. This can be done by applying an additional transformation after the evaluation of the estimated joint pdf. For example, several cases show that the bias of the joint pdf follows a sigmoid shape, and it is easy to transform a sigmoid function to a line.

3. Keeping a list of additional variables which are, in fact, transformations of the already existing ones. For example, from X_1, X_2 we can create $T_1 = X_1, T_2 = \log(X_1), T_3 = \exp(X_1), T_4 = X_2, T_5 = \log(X_2), T_6 = \exp(X_2)$, and then create Y_1, Y_2 using the most suitable variables.

4. Transforming all the variables to Normal. Normal random variables have several desirable properties, such as their addition remains Normal.

   Also, a software for estimating joint pdfs using this method may only need Box-Cox method to convert each variable.

These improvements are outside the scope of this work, so we do not develop these solutions any further. Nevertheless, improving this method may be a good idea.

References

Buckley, J. J. (2005). *Fuzzy Probabilities: New Approach and Applications*. Studies in Fuzziness and Soft Computing. Springer.

Chow, C. K. and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transaction on Information Theory*, 14(3):462–467.

Fox, J. (2008). Voting intentions of the chilean 1988 plebiscite dataset (communicate by flacso/chile). In: Fox, J. Applied Regression Analysis and Generalized Linear Models. 2nd ed. Sage Pubs. http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Chile.txt.

Justel, A., na, D. P., and Zamar, R. (1997). A multivariate kolmogorov-smirnov test of goodness of fit. *Statistics & Probability Letters*, 35:251–259.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions: Models and Applications*. John Wiley & Sons.

Mjolsness, E. and deCoste, D. (2001). Machine learning for science: State of the art and future prospects. *Science*, 293(14):2051–2055.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag New York, Inc.

Schmidt, T. (2007). Coping with copulas. In Rank, J., editor, *Copulas: From theory to application in finance*, pages 3–34. Risk Books.

Vlachos, P. and Meyer, M. (1989). The andrews and herzberg archive (from andrews, herzberg, 1985. data.). In: StatLib. http://lib.stat.cmu.edu/datasets/Andrews/.

Chapter 9

Network Observer: a network analysis tool

We now present *Network Observer*, the software application that we were designing. It includes the two empirical methods defined, and the analytical development which took us the two previous chapters to explain.

9.1 Overview of the requirements

Network Observer was designed to support exploratory data analysis of multivariate network data.

It should include the following functionalities:

1. Generation of **network plot graphs**. Regarding this item, the user should be able to pick up two attributes, one for the x-axis and one for the y-axis, and the graph should be displayed. Also, the user should be able to enable or disable filters for allowing arcs which point toward greater, equal or less x or y values. Also, the images built must be printer-friendly.

2. Generation of **arc correlation profiles**. Regarding this item, the user should be able to pick up two attributes, one for the tail and one for the head (of arcs), and an arc correlation profile should be displayed. The user might choose to display arc frequencies, bin sizes, or arc likelihood. Also, the images built must be printer-friendly.

3. Estimation of **arc likelihood functions**. Regarding this item, the user should be able to pick up the attributes which will be used to estimate the joint pdfs. The estimated functions should be displayed in a readable format.

We successfully implemented the above requirements. But we also extended the functionalities a bit.

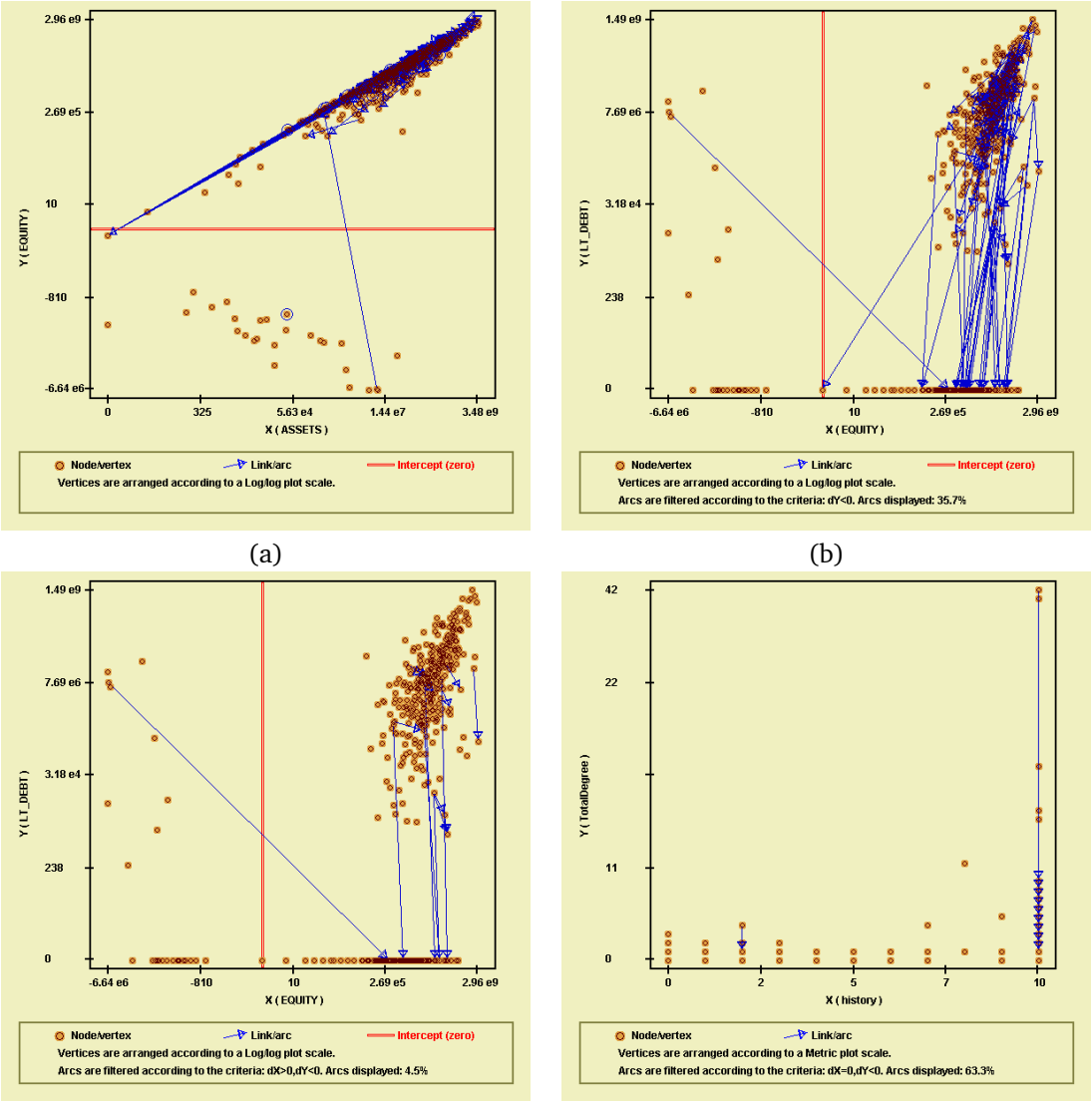


Figure 9.1: Network plot graphs generated by Network Observer. Interpretations: (a) Almost all firms invest in firms with positive equity; (b) 36% of the investments are to firms with less long-term debt, and practically all of the investor firms have positive equity; (c) less than 5% of the investments are to firms with less long-term debt but more equity; and (d) 63% of the investments are to old firms, which less have investors and investments.

9.2 Network plot graphs

The images generated by the network plot graph tool are colorful yet have good contrast. When converted to grayscale or printed in black and white printers, the images are still clear. Figure 9.1 shows a few images created by Network Observer.

In network plot graphs, vertices can be plotted like in scatter plots: using the tra-

ditional metric scale, the log/log scale (useful for handling attributes which follow long-tailed probability distributions), and a sorted scale which prevents vertex overlapping. In figure 9.1, insets (a), (b) and (c) follow a log/log scale (or layout). Only inset (d) follows a metric scale.

Note that each network plot graph created with Network Observed includes a box which contains all of the options used to create the graph. If filters were in use, it shows which filters were used and how many arcs are displayed. Also, it always shows which scale or layout was in use. This box also shows additional information, which shows how node or vertices look like, how arcs or links look like, and how the intercept looks like (this only appears when values range from negative to positive).

Images are saved in png format. Therefore, saved files are light as our tool uses only but a few colors. In figure 9.1, the file sizes of the insets are: (a) 9.3 kb, (b) 15.2 kb, (c) 11.7 kb, and (d) 5.7 kb. Image sizes are always 600x600 (this should change in the future.)

9.3 Arc correlation profiles

The images generated by the arc correlation profile tool can be colorful, in grayscale, or in black and white, according to the user's choice. Figure 9.2 shows a few images created by Network Observer.

Four options for displaying the measures were implemented: blocks, which sizes are proportional to what is measured (fig 9.2-(b,c)); grays, where darker cells or bins demonstrate greater amounts of what is measured (fig 9.2-(d)); heat graphs, where warmer tones reflect greater amounts of the measure with respect to cooler tones (fig 9.2-(a)); and an option for full information, which bins have text instead of blocks/colors.

Also, we added the possibility to apply transformations to the attributes involved in the correlation profiles: logarithm (actually $f(x) = \text{sign}(x)\log(|1+x|)$), fifth root ($f(x) = \sqrt[5]{x}$), third root ($f(x) = \sqrt[3]{x}$), square root ($f(x) = \sqrt{|x|}$), identity (default, $f(x) = x$), second power ($f(x) = x^2$), third power ($f(x) = x^3$), fifth power ($f(x) = x^5$), and exponential ($f(x) = e^x$). Fifth root, third root, third power, and fifth power were added because they preserve the sign ($f(-x) = -f(x)$) and are not as extreme as the logarithm and the exponential when it comes to transformations.

Again, images are saved in png format. In figure 9.2, file sizes are: (a) 7.3 kb, (b) 5 kb, (c) 5.4 kb and (d) 5.7 kb.

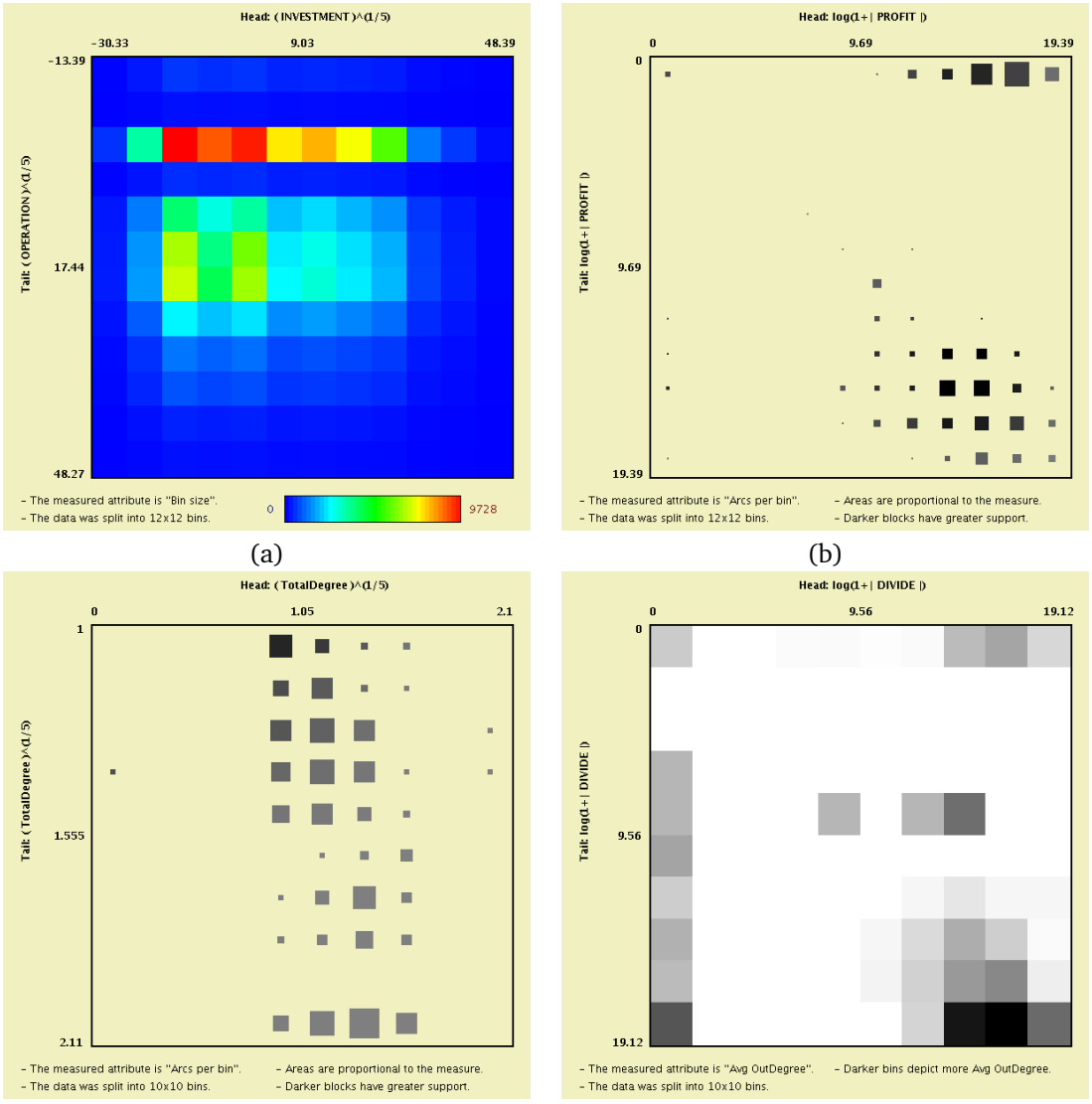


Figure 9.2: Arc correlation profiles generated by Network Observer. Remarks: (a) Heat graph of bin sizes; (b) block graph of arc frequencies (arcs per bin); (c) another block graph of arc frequencies; and (d) a grayscale graph of average outdegrees.

9.4 Arc likelihood estimator

The last tool is the arc likelihood estimator or, as we called it, the *Joint PDF Analyzer*. We called it that way because it explicitly computes the pdfs of the vertices and the frequency of arcs, and defines the arc likelihood as their quotient. Its user interface can be seen in figure 9.3.

This tool lets the user select some attributes (or all of them) from the list, and then lets him/her estimate the joint pdf. After the function is estimated, it can be saved

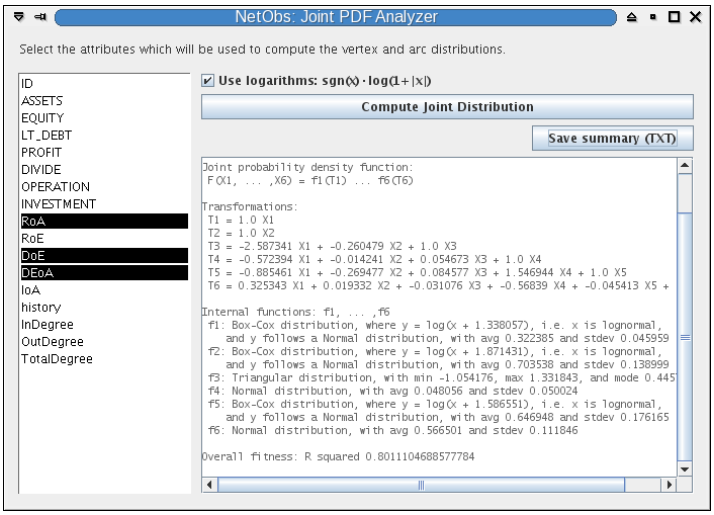


Figure 9.3: Interface of the *Joint PDF Analyzer* tool. Three attributes were chosen: RoA, DoE and DEoA. The logarithmic transformation was applied. Pressing the "Save summary (TXT)" button saves the result to a file.

(pressing the button). The text box displays information regarding the estimation; which functions have been estimated, and if the estimation process ended. (The overall process may take a while, so the activity in the text box is intended to ease the user who thinks that the application hanged up while it is still working.)

The output of this tool looks like:

```
This joint probability distribution was computed by NetObs.
The accuracy of the function is not necessarily high.

The probability that a vertex X links a vertex Y is:

              355 A(X1 ... X3 Y1 ... Y3)
F(X->Y) = ------------------------------------
           219024 V(X1 ... X3) V(Y1 ... Y3)

where Xi are the attributes of vertex X and Yi of vertex Y.
In object oriented notation:
  X1=X.RoA   Y1=Y.RoA
  X2=X.DoE   Y2=Y.DoE
  X3=X.DEoA   Y3=Y.DEoA
However, Xi and Yi were transformed using: sgn(z) log(1+|z|)
Unlike the logarithm, this function allows negatives.

The above function F(X->Y) is composed of two functions: A and V.
A represents the distribution of arcs, V the distribution of vertices.


The distribution of arcs is the following:
```

```
Joint probability density function:
 A(X1, ... ,X6) = a1(T1) ... a6(T6)

Transformations:
 T1 = 1.0 X1
 T2 = 1.0 X2
 T3 = -2.587341 X1 + -0.260479 X2 + 1.0 X3
 T4 = -0.572394 X1 + -0.014241 X2 + 0.054673 X3 + 1.0 X4
 T5 = -0.885461 X1 + -0.269477 X2 + 0.084577 X3 + 1.546944 X4 + 1.0 X5
 T6 = 0.325343 X1 + 0.019332 X2 + -0.031076 X3 + -0.56839 X4 + -0.045413 X5 + 1.0 X6

Internal functions: f1, ... ,f6
 a1: Box-Cox distribution, where y = log(x + 1.338057), i.e. x is lognormal,
    and y follows a Normal distribution, with avg 0.322385 and stdev 0.045959
 a2: Box-Cox distribution, where y = log(x + 1.871431), i.e. x is lognormal,
    and y follows a Normal distribution, with avg 0.703538 and stdev 0.138999
 a3: Triangular distribution, with min -1.054176, max 1.331843, and mode 0.445731
 a4: Normal distribution, with avg 0.048056 and stdev 0.050024
 a5: Box-Cox distribution, where y = log(x + 1.586551), i.e. x is lognormal,
    and y follows a Normal distribution, with avg 0.646948 and stdev 0.176165
 a6: Normal distribution, with avg 0.566501 and stdev 0.111846

Overall fitness: R squared 0.8011104688577784



The distribution of vertices is the following:

Joint probability density function:
 V(X1, ... ,X3) = v1(T1) ... v3(T3)

Transformations:
 T1 = 1.0 X1
 T2 = 1.0 X2
 T3 = -1.130663 X1 + 1.0 X3

Internal functions: f1, ... ,f3
 v1: Box-Cox distribution, where y = ( [x+2.970214]^3.999908 - 1 )*0.009501,
    and y follows a Normal distribution, with avg 0.759706 and stdev 0.118995
 v2: Triangular distribution, with min -4.95278, max 6.887259, and mode -0.87169
 v3: Normal distribution, with avg 0.481375 and stdev 0.385379

Overall fitness: R squared 0.9098768222898495
```

9.5 Other features

9.5.1 Input

To input data to Network Observer, the user can:

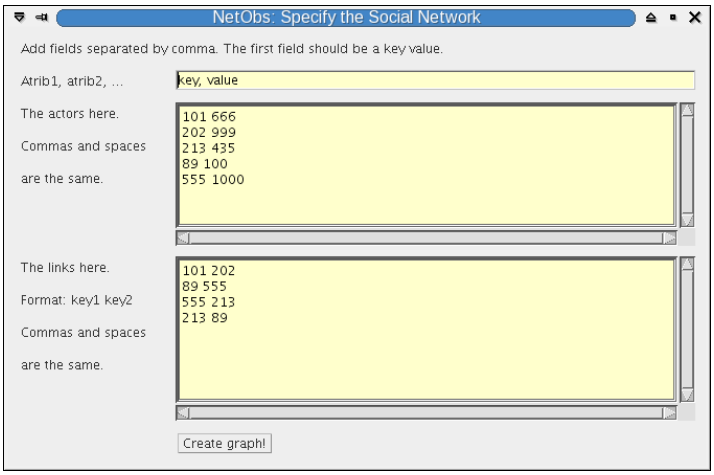


Figure 9.4: Entering the data manually.

Enter the data manually. To do this, a form has to be filled (figure 9.4). In the first field, the names of the attributes must be entered. In the second, the vector data. Note that in the first column, entries **must be unique**. In the third, the arc data in the form $(cod1, cod2)$. These $cods$ must appear in the first column of the second field.

Using a single file. By using a single file, the user can supply Network Observer with multivariate network data. This file is similar to the one used by Pajek, and has the form:

```
* atr1 atr2 atr3
10 200 300
15 250 150
...
*Arcs
10 15
15 25
...
```

Of course, in the above example, we could use more than three columns. Also, we could separate numbers by commas or semicolons and the file would still be valid.

Using one file for vertices and one for arcs. By using several files, the user can reuse vertex or arc data, while changing the other.

The file with vertex data must have the form:

```
atr1, atr2, atr3
100, 50, 123.5
```

```
111, 120, 110.3
....
```

The file with arc data must have the form:

```
100, 111
120, 100
...
```

Both files follow the csv format. This implies that can be edited by Microsoft Excel, OpenOffice Calc, Gnumeric, and other spreadsheet software.

We encourage the use of this input format with Network Observer.

9.5.2 Output

Network Observer lets the user save data in many ways:

As images. Arc correlation profiles and Network plot graphs let the user save the content in image (png) format.

Textual description. The Joint PDF Estimator tool lets the user save the description of the estimate.

Export to other formats. This utility enables users to save their data in several formats (currently Network Observer, Matlab/Octave, Pajek).

9.5.3 Object oriented design

We used the pattern *Model-View-Controller*, MVC, to design Network Observer. We believe that MVC is adequate for this type of development because we work with theory, and theory fits well within the *Model* part of the pattern. For example, when we developed Joint PDF Estimator (see the previous chapter), the joint pdf estimating engine was copied to Network Observer without any changes.

Once we developed the models, we focused in the *Controller* part. We created objects that could simplify the work of the interfaces, and could interact with the models. For example, selecting attributes was one of the several actions these objects could perform.

Finally, we developed the user interface (the *View* part). We created it using the GUI creator which comes with NetBeans. Note that both Network Observer and Joint PDF Estimator were developed with NetBeans.

9.6 Comments

We presented Network Observer, our application, which was designed to support the researcher who wishes to explore multivariate network data. Now that we explained how we designed and coded it, we are going to focus on exploring the data.

Part III

A study of the Chilean shareholding network

Chapter 10

Empirical analysis

We now study the Chilean shareholding network using empirical methods. We start by analyzing the dynamics of the network and then study the different views (graphs, as described in chapter 5) using Network Observer.

Disclaimer From a data analysis perspective, we are mostly concerned with the graph *firms invest in firms*, which is a subset of the original shareholding network *shareholders invest in firms*. The former is a rich multivariate network while the latter is not (for most shareholders, we only know their names). Also, we discarded shareholders with less than 1% of ownership.

Now, let us define some concepts.

Birth of an element We consider that an element (firm, arc) is born when it first appears in the database. In other words, an element is born in its *first period*. In a `SELECT` instruction, this would be `MIN(year+month/12.-0.25)`[1] given that we specified `GROUP BY id` (or a similar group). Note that vertices (firms - financial statements) start in March 2003 and that arcs (shareholding relations) start in December 2003. Thus, both periods cannot be used as reference for vertex or arc birth, since we know nothing prior to them.

Death of an element We consider that an element (firm, arc) dies when it last appears in the database. In other words, an element dies in its *last period*. In a `SELECT` instruction, this would be `MAX(year+month/12.-0.25)` given that we specified `GROUP BY id` (or a similar group). Note that the last period stored in the database is June

[1] We used `-0.25` because December 2004 would look like `2005` if we did not use it. Using `-0.25`, December 2004 becomes $2004 + 12/12 - 0.25 = 2004.75$. We believe it is easier to recognize 2004.75 as part of year 2004 when we look at charts.

2007, for both vertices and arcs. Thus, this period cannot be used as reference for vertex or arc death, since we know nothing after them.

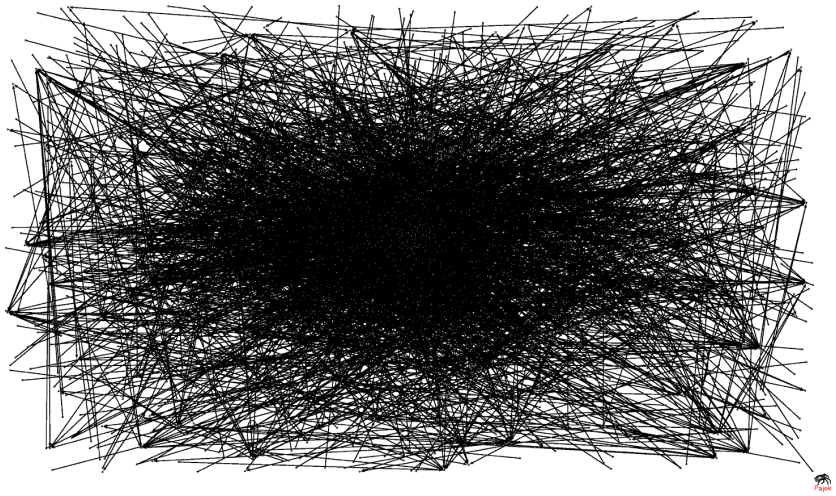


Figure 10.1: The Chilean shareholding network (June 2005).

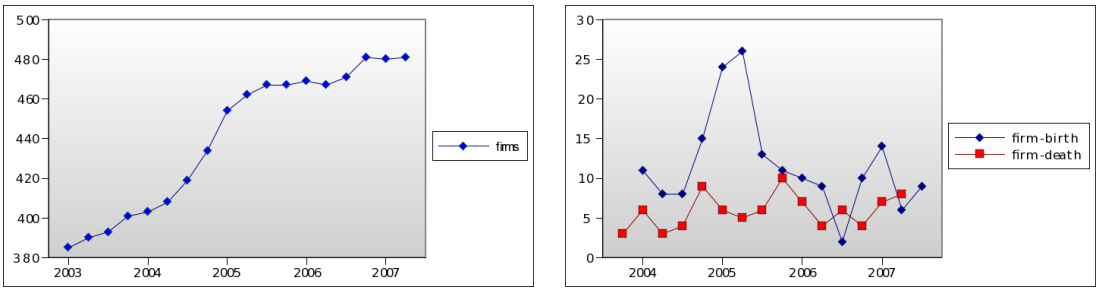
10.1 Dynamics

The Chilean shareholding network (figure 10.1) is a rather static network. We only see a few firms enter the network each period, and a few getting out from it. The same applies to the investment relationship (the arcs). In the following, we use the second approach to study network dynamics discussed in section 7.4 to justify that our network is rather static. However, we also discover interesting facts from this data.

Obs. 1. *Firm birth found its climax in March 2005.* Figure 10.2 shows that the number of firms quickly increased in the first periods, and then its growth became stable. By watching inset (b), we see that firm birth reached its peak in March 2005.

The number of firms change slightly, as can be seen in figure 10.2. Also note that several of the dying firms were born not too long ago; in particular, the mode of the lifespan of firms is one year (taking into account only firms that were born and died within the frame of study).

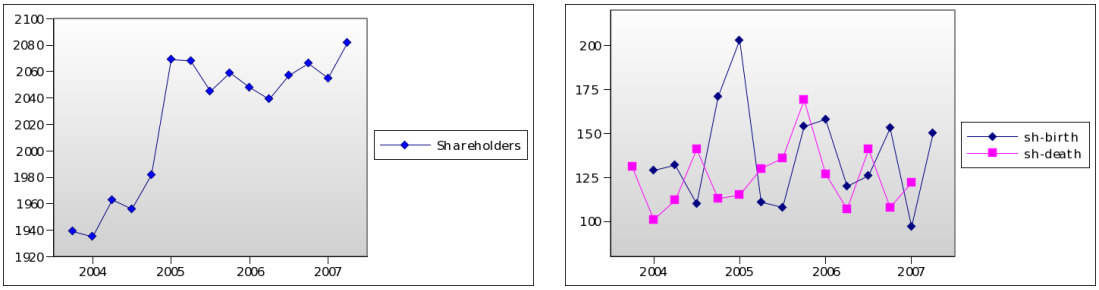
Obs. 2. *The dynamics of shareholders is similar to that of firms.* In particular, the number of shareholders seems proportional to that of firms. The same applies to birth and death processes. It is easy to verify this by comparing figure 10.3-(a) to figure 10.2-(a) and figure 10.3-(b) to figure 10.2-(b).



(a) (b)

Figure 10.2: Firm dynamics. Insets: (a) number of firms per period, (b) birth and death per period.

The conclusion we can draw from the above facts is that shareholders are as active as firms. As the number of shareholders is less smooth than the number of firms, we infer that the former adapt to the latter.



(a) (b)

Figure 10.3: Shareholder dynamics. Insets: (a) number of shareholders per period, (b) birth and death per period.

Obs. 3. *The number of firms and the number of shareholders are linearly related, specifically by* $V_{firms} \approx -635.56 + 0.535 V_{shareholders}$. The approximation is illustrated in figure 10.4. Note that all the firms were included in the shareholding network, but they represent approximately one fifth of the shareholders.

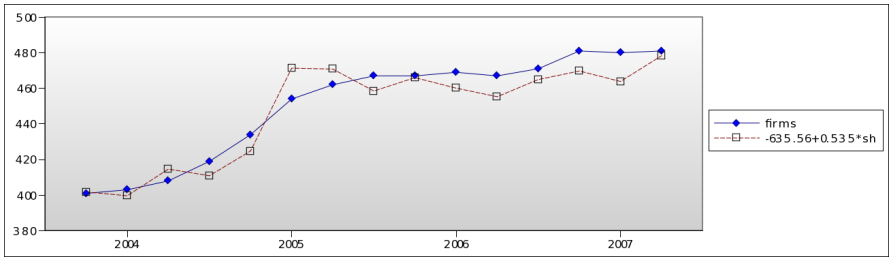
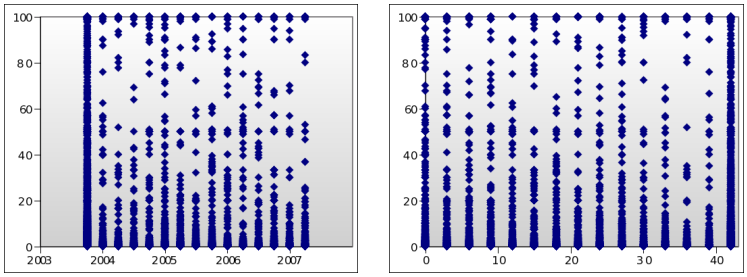


Figure 10.4: Relation of the number of firms and shareholders

Obs. 4. *Shareholder activity (investment) does not seem to change during the period under study.* The average amount invested in firms does not seem to depend on the period nor the time the shareholder exists in the database.



(a) (b)

Figure 10.5: Shareholder activity. Insets: (a) birth date vs average share, (b) lifespan vs average share.

The above facts suggest that shareholder behavior does not change over time. Even though the number of shareholders and the number firms change over time, investment does not seem to change at all. The next facts illustrate this property in detail.

Obs. 5. *In the network of investments between firms, arcs practically do not die.* We can see in figure 10.6 that arc death within the network of investment between firms remains small (between 0 and 8 per period, to be precise).

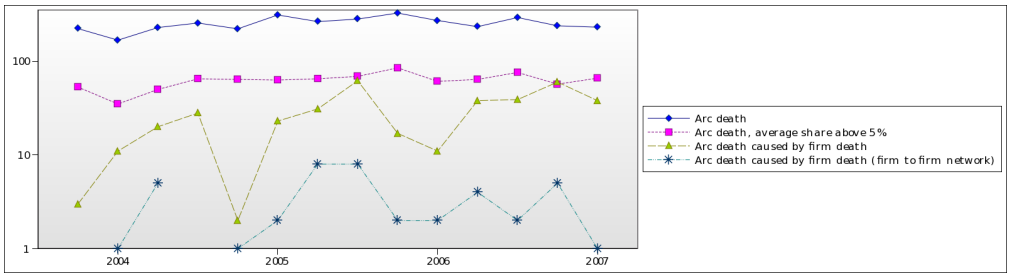
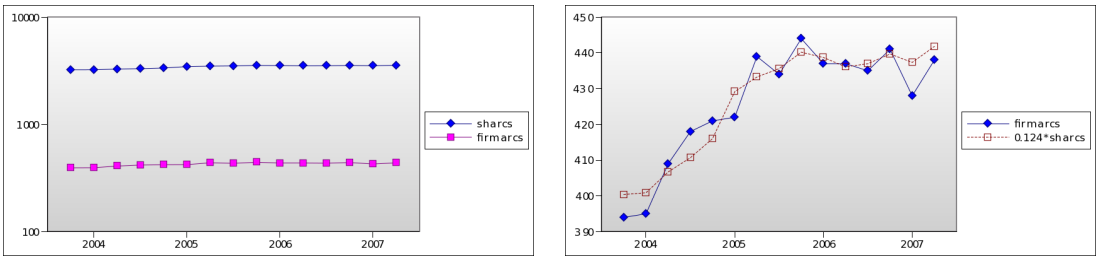


Figure 10.6: Arc death in detail

Also, from figure 10.6 we can see that arc death for all arcs and for those arcs which average share is greater than 5% are proportional. In fact, the correlation of both series is 0.83. Note that zeroes are not shown in the figure due to the scale of the graph (semilogarithmic).

Obs. 6. *The number of arcs in the shareholding network and the network of investments between firms is practically proportional, namely* $A_{firms} \approx 0.124 A_{shareholders}$*, where* A_x *is the number of arcs of network* x*.* This fact is illustrated in figure 10.7-(b), which shows how good the approximation is.



(a) (b)

Figure 10.7: Number of arcs. Insets: (a) all the arcs and the arcs between firms, (b) the arcs between firms and a estimation based on all the arcs.

Obs. 7. *The indegrees and outdegrees of the network of investment between firms follow a power law which parameter β do not change much over time; in particular, avg $(\beta_{outdegree}) = 5.83$ and avg $(\beta_{indegree}) = 6.14$.* Figure 10.8 shows that the distributions of indegrees and outdegrees do not change much over time. The degrees have variations in their parameters, but these variations are slight and show no clear tendency. (Note that we are modeling the power laws as $f(k) = (\beta - 1)k^{-\beta}$.)

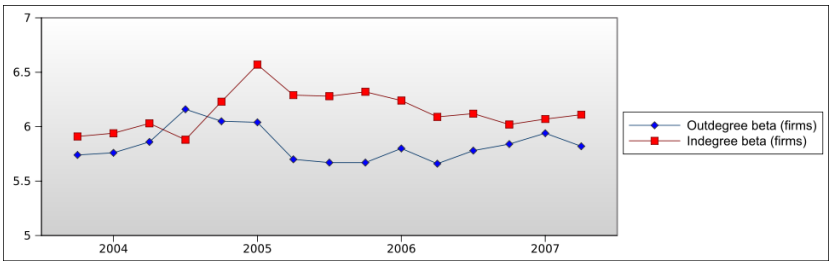


Figure 10.8: Indegree and outdegree of firms.

Obs. 8. *The correlation between firms' attributes slightly changes over time.* Table 10.1 illustrates the extent of this fact. Correlations do change over time, but not too much, i.e. ± 0.1. However, the way correlations change partially explain the evolution of the Chilean economy. For example, note how profits become less correlated with equity, debt (and assets), and investments, and more with operational income.

Over time some variables keep their correlations while others show tendencies. For example, the correlation between assets and debt is almost constant while profits and dividends become more correlated as time goes on (figure 10.9). But it is also important to note that correlations tend to become stable since June 2005.

The above facts are important to us because they let us reduce the study of the Chilean shareholding network (figure 10.1) to the study of the shareholding network of firms (figure 10.10).

Correlations	ASST	EQUT	LDEB	PROF	DIVD	OPER	INVS	RoA	RoE	DoE	DEoA	IoA
ASSETS	1											
EQUITY	0.95	1										
LT_DEBT	0.85	0.64	1									
PROFIT	0.75	0.77	0.55	1								
DIVIDE	0.49	0.47	0.4	0.73	1							
OPERATION	0.54	0.38	0.59	0.63	0.6	1						
INVESTMENT	0.53	0.69	0.17	0.72	0.39	0.05	1					
RoA	0.06	0.06	0.04	0.12	0.08	0.08	0.09	1				
RoE	0.03	0.02	0.02	0.03	0.04	0.04	0.01	0.38	1			
DoE	-0.02	-0.03	0.01	-0.02	-0.02	-0.01	-0.03	0	-0.28	1		
DEoA	0.08	0.08	0.07	0.07	0.05	0.04	0.05	0.04	0.03	0.02	1	
IoA	0.03	0.04	0.01	0.07	0.03	0	0.1	0.74	0.32	-0.02	0.04	1

(a) December 2003

Correlations	ASST	EQUT	LDEB	PROF	DIVD	OPER	INVS	RoA	RoE	DoE	DEoA	IoA
ASSETS	1											
EQUITY	0.95	1										
LT_DEBT	0.82	0.61	1									
PROFIT	0.68	0.64	0.51	1								
DIVIDE	0.49	0.4	0.44	0.84	1							
OPERATION	0.45	0.3	0.5	0.81	0.74	1						
INVESTMENT	0.62	0.75	0.22	0.53	0.41	0.02	1					
RoA	0.03	0.03	0.02	0.06	0.05	0.03	0.06	1				
RoE	0.01	0.01	0.01	0.03	0.03	0.02	0.01	0.09	1			
DoE	0.01	-0.02	0.07	-0.01	0	0.01	-0.02	0.01	-0.02	1		
DEoA	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0	0.01	1	
IoA	0.01	0.02	0	0.03	0.03	0	0.06	0.98	0.09	0	0.01	1

(b) December 2004

Correlations	ASST	EQUT	LDEB	PROF	DIVD	OPER	INVS	RoA	RoE	DoE	DEoA	IoA
ASSETS	1											
EQUITY	0.96	1										
LT_DEBT	0.79	0.6	1									
PROFIT	0.59	0.54	0.48	1								
DIVIDE	0.51	0.45	0.44	0.95	1							
OPERATION	0.43	0.31	0.46	0.9	0.89	1						
INVESTMENT	0.61	0.75	0.21	0.29	0.18	-0.06	1					
RoA	0.05	0.04	0.03	0.06	0.06	0.05	0.04	1				
RoE	0.05	0.04	0.04	0.07	0.06	0.06	0.03	0.08	1			
DoE	0	-0.01	0.02	-0.01	-0.01	0	0	0	0.19	1		
DEoA	0.05	0.05	0.05	0.03	0.02	0.02	0.04	0.81	0	0.01	1	
IoA	0.03	0.03	0.01	0.02	0.01	0	0.05	0.96	0.03	0	0.82	1

(c) December 2005

Correlations	ASST	EQUT	LDEB	PROF	DIVD	OPER	INVS	RoA	RoE	DoE	DEoA	IoA
ASSETS	1											
EQUITY	0.96	1										
LT_DEBT	0.82	0.65	1									
PROFIT	0.49	0.43	0.44	1								
DIVIDE	0.34	0.27	0.35	0.98	1							
OPERATION	0.36	0.25	0.4	0.96	0.98	1						
INVESTMENT	0.7	0.82	0.32	0.18	0	-0.05	1					
RoA	0.12	0.11	0.09	0.24	0.22	0.22	0.09	1				
RoE	0.04	0.03	0.04	0.11	0.1	0.11	0.02	0.2	1			
DoE	-0.02	-0.03	0.01	-0.01	-0.01	-0.01	-0.03	-0.01	-0.15	1		
DEoA	0.09	0.1	0.08	0.03	0.02	0.01	0.08	0.49	-0.01	-0.01	1	
IoA	0.06	0.08	0	0.03	0	-0.02	0.19	0.51	0.07	-0.03	0.05	1

(d) December 2006

Table 10.1: Correlation between firms' attributes.

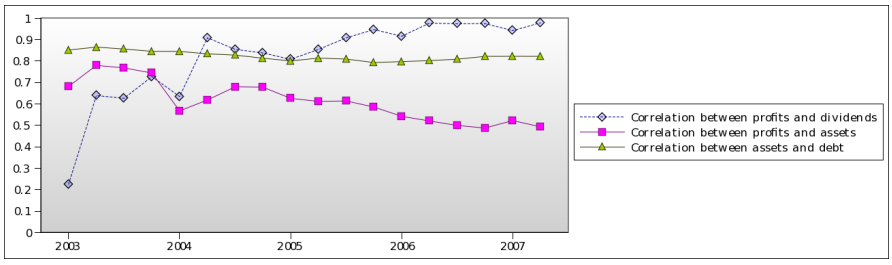


Figure 10.9: Correlations over time.

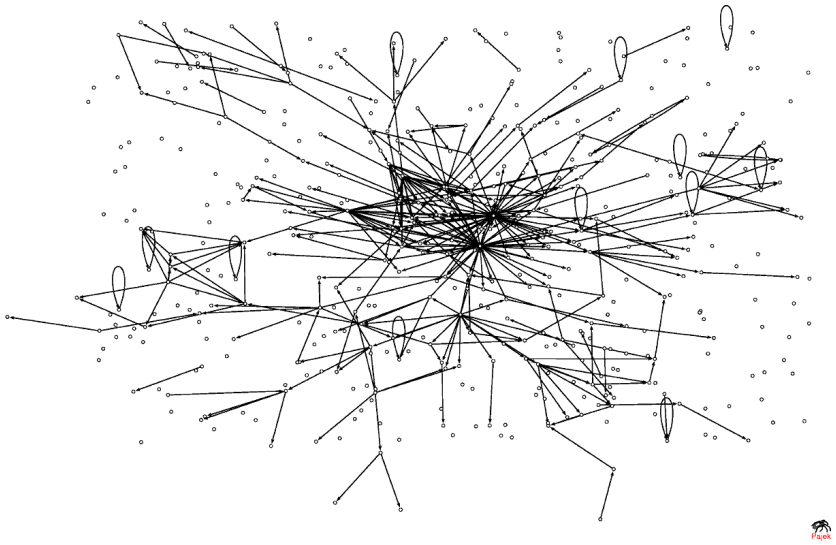


Figure 10.10: The network of investments between firms (June 2005).

10.2 Behavior

We call *behavior* to the rationale behind vertex linking or "in which firm another firm is going to invest". To understand this, we study the result of this behavior, i.e. the relation between connected vertices.

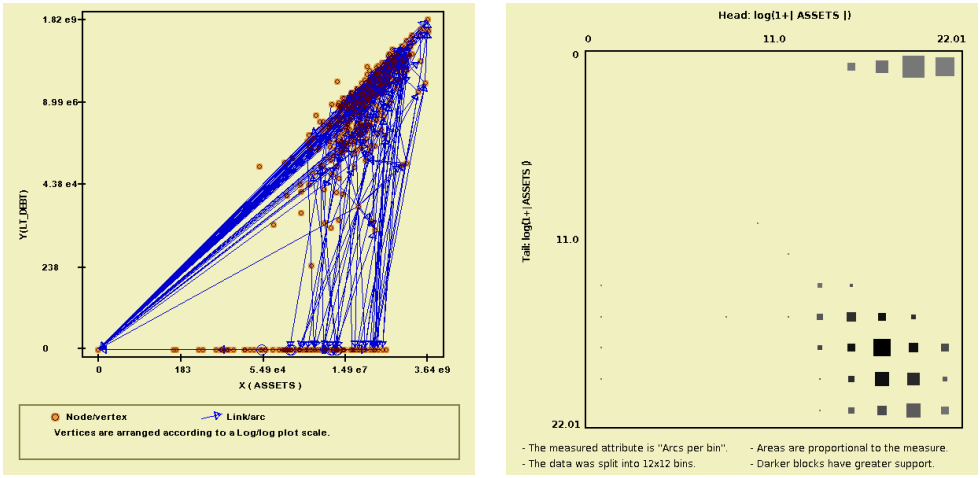
10.2.1 Financial structure

We now study how financial structures condition this shareholding network. To do this, we study the arc likelihood between two firms based on their assets, equity, and debt, and derived ratios such as D/E (debt to equity) and $(D+E)/A$ (proportion of non current assets, as D is long term debt only). Note that ratios are illustrative because they do not depend on the size of the firms (are adimensional; do not even have units). However, they do not add additional information to that contained in assets, debt and equity.

Obs. 9. *A large number of investments is made by financial firms, which have no assets.* Figure 10.11 shows that financial firms[2] participate in a large number of investments ($130/332 = 39\%$ in Aug. 2004, $133/355 = 37\%$ in Aug. 2005, $124/354 = 35\%$ in Aug. 2006). Note that practically no firms invest in financial firms (less than five do that).

Obs. 10. *Financial firms work with several, small investments.* Figure 10.12-(a) shows that financial firms have shares below 5%. This is consistent with the rational behavior of agents who use stocks as financial assets rather than instruments of

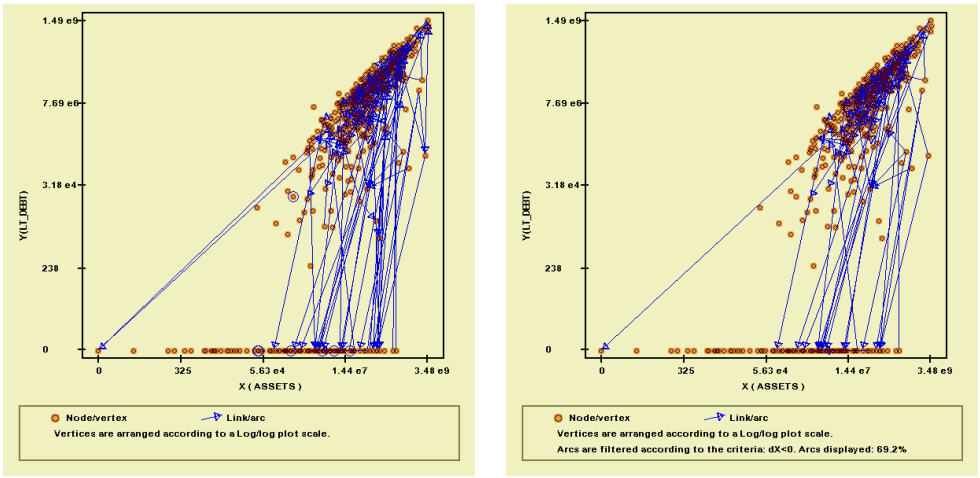
[2] Recall chapter 5, section *Integrity Problems* where we left some financial firms with no assets



(a) (b)

Figure 10.11: The role of financial firms. Insets: (a) Network plot graph, note the large number of arcs in which participate vertices at $(0, 0)$, (b) note that several arcs start in firms without assets. (Both figures depict data from August 2004)

power. (We can discard hypotheses such as *financial firms use their customers' money to acquire other firms*.)



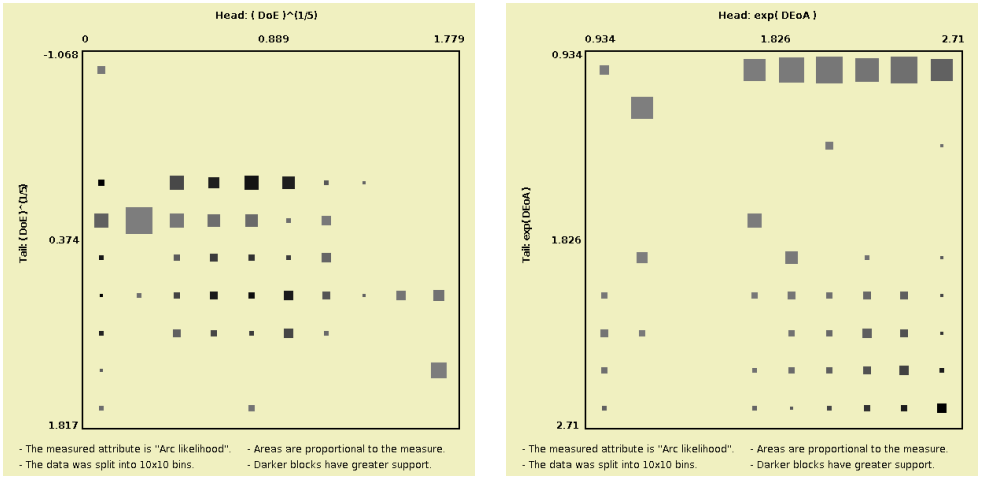
(a) (b)

Figure 10.12: The direction of greater investments. Insets: (a) Financial firms play no role in the network of large investments (shares above 5%), where (b) larger firms own smaller firms. (August 2005)

Obs. 11. *Ownership above 5% is done by firms with greater assets.* Figure 10.12-(b) shows that about the 70% of firms invested in firms with lower assets, owning more than the 5% of them, for August 2005. The same applies to other periods. (Note that the same applies to equity (72%) and debt (61%), since they are highly correlated to assets.)

Obs. 12. *Ownership below 5% is mostly done by financial firms.* Actually, financial firms are responsible (the tail) of 60% to 70% of arcs with shares below 5%. This implies that most firms with positive assets invest in other firms in order to gain power over them rather than to earn money from dividends.

Obs. 13. *Most firms are not concerned with D/E and $(D+E)/A$ ratios.* Figure 10.13 illustrates this. Inset (a) shows that tails' and heads' D/E are not related, and inset (b) shows that no firm is concerned with $(D+E)/A$ save financial firms. However, financial firms do not invest in themselves, which may explain inset (b).



(a) (b)

Figure 10.13: Topology and structural ratios. Insets: (a) Most firms do not seem to be concerned with debt to equity ratio, and (b) most firms do not seem to be concerned with $(D+E)/A$, save financial firms. (August 2005)

To this moment, we have found two types of firms with rather different behaviors: financial firms, which tend to invest in any kind of firms, and non financial firms which tend to invest in smaller firms, seeking ownership. This essential difference is used in the analysis of the relations between firms.

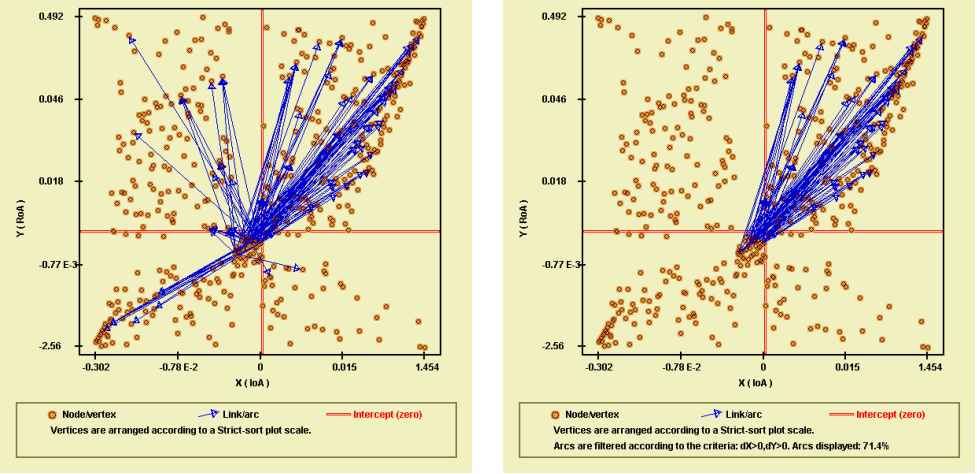
10.2.2 Earnings

There are four main variables related to earnings: [net] profits, dividends (what shareholders earn), operational flow (due to production, sales), and investment flow (due to investments). From these, two ratios were built: RoA (return on assets, or profits/assets), and IoA (investment on assets, or investment/assets). Of course, the latter do not add new information to the already added by the main variables.

Obs. 14. *Non financial firms tend to invest in firms with less profits and dividends.* We did not illustrate this observation using figures because it is enough to recall that

these firms invest in firms with less assets and that the greater the assets, the greater the profits and dividends.

Obs. 15. *Financial firms invest mostly in firms with both positive RoA and IoA (in average).* Financial firms, as can be seen in figure 10.14, mainly invest in firms with good performance. This has sense if we consider financial firms as agents who seek economic performance from their investments.



(a) (b)

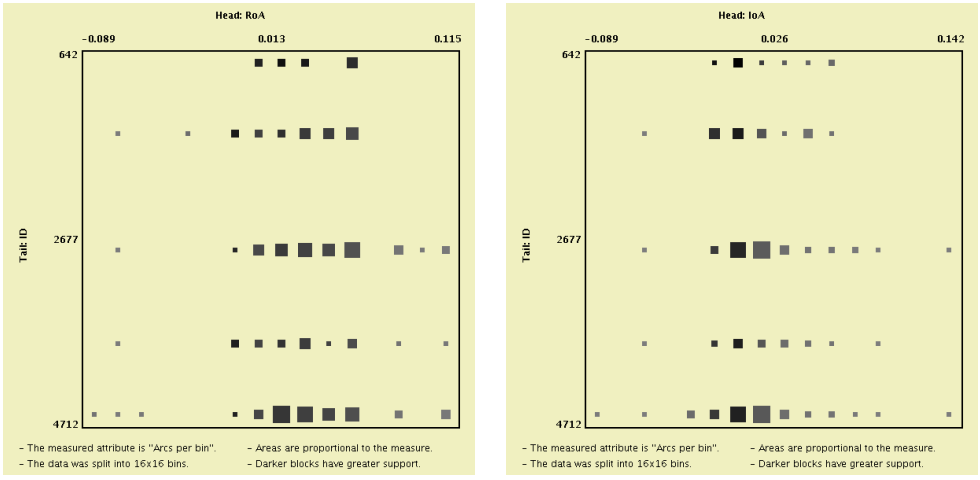
Figure 10.14: Financial firms invest in performance. Insets: (a) In which firms financial firms do invest, and (b) they mostly invest in firms with positive RoA and IoA (June 2006)

Obs. 16. *Financial firms share similar investing patterns.* Figure 10.15 illustrates this well. We could think of a rational explanation for this: if one firm found a better portfolio than the rest, then the rest would copy it (since it is public information). Eventually, portfolios should converge to a stable solution. (This solution should be optimal, the so-called market portfolio.)

10.2.3 Topological attributes

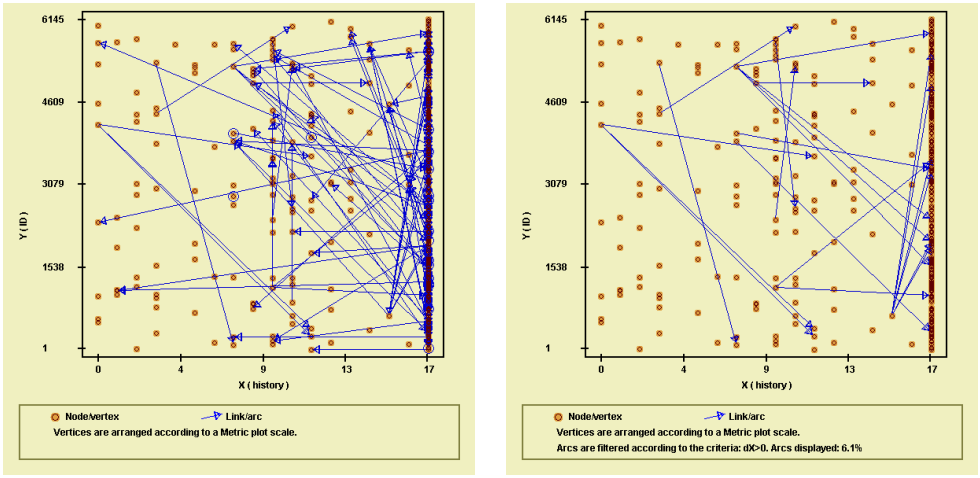
We now study how topological variables condition the behavior of this network. In particular, we study the following variables: the number of periods a firm existed, its indegree, its outdegree, and its total degree (indegree + outdegree).

Obs. 17. *Most investments were made by firms which were born before the period under study, but it is more likely that a new firm invests in an older one.* To June 2007 (the last period of our data), 12.5% of firms invested in new ones while 6.1% invested in older ones. Note that old firms represent about 87% of firms in this period, so it is more probable that a new firm invested in an old firm than the opposite.



(a) (b)

Figure 10.15: Financial firms invest in performance. Insets: (a) In which firms financial firms do invest, and (b) they mostly invest in firms with positive RoA and IoA (June 2006)

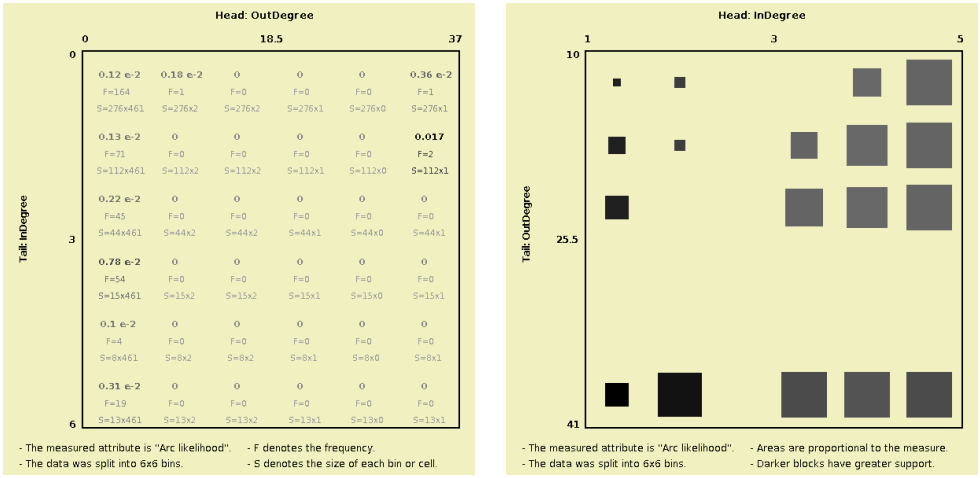


(a) (b)

Figure 10.16: Age of firms and investments. Insets: (a) Investments arranged by the age of the parts (17 is the oldest in our time horizon), and (b) 6.1% of the investments was made by younger to older firms.(June 2007)

Obs. 18. *Indegree follows a power law, but is bounded to 6 (depending on the period).* This is natural since we are only taking into account shares over 1%. Also note that, the higher the indegree of a firm, the more likely it will invest in other firms (first column of figure 10.17-(a)). (*May this imply paths play a role in this network?*)

Obs. 19. *Financial firms not only have similar investing patterns: they do invest in common firms.* Figure 10.17-(b) illustrates this fact pretty well. This inset was built using investments made by financial firms alone. If a firm has an indegree above one, we can say that more than one financial firm invested in it. Note that it is less likely



(a) (b)

Figure 10.17: Degree and types of firm. Insets: (a) Firms with higher indegree tend to invest more, but not necessarily in firms with given outdegrees, and (b) financial firms have high outdegrees and invest in somewhat *famous* firms, with indegrees above 2. (June 2006)

that financial firms invest in firms with indegree one. (*Does this imply prefferential attachment from few vertices?*)

10.3 Comments

We basically discovered one key fact of the Chilean shareholding network: big investments (share above 5%) are made by bigger firms on smaller ones. This can be explained by the fact that shares are part of the assets, so firms need to have enough assets to be able to own other firms. This fact explains why the richer own the poorer in this network, and why the opposite is often unfeasible.

Note that we only studied the dynamics of this network at a relational level; we did not study paths. If investments only depend on the attributes of the two related firms, paths will be meaningless to arc birth. However, if paths play a role in investment, we are facing the problem of analyzing [non relational] *emergence*, a subject of interest in nonlinear and complexity sciences. Unfortunately, we did not develop any empirical method to evaluate emergence. (It is outside our scope, but we could have used a histogram of path lengths to assess emergence, for example.)

Paths should play a role in investment. As discussed in chapter 3, ownership structures play a role in economies. However, the structure may not be observable; natural people may be shareholders of smaller firms on behalf of larger corporations, and we do not have enough information to detect if a person acts on behalf of a firm or not since people are only *names* to us (nothing else was available).

Chapter 11

Modeling the network of investments between firms

After studying the Chilean shareholding network, putting special emphasis to the investment between firms, we now model the latter subnetwork. We will particularly model the network which vertices are firms which financial statements are known by us. We will be discarding those firms without financial statements and without assets.

11.1 Model design

To model the **network of investments between firms which financial statements are known**, we prepare the simulations and estimate some basic parameters. But first, let us study some important properties of this network.

Obs. 20. *In this network, the quantity A/V remains practically invariant, where A and V are the number of arcs and firms. In particular, $A/V \approx 0.77909$.* This is illustrated in figure 11.1. Note that this property implies that the density of the network is approximately $\rho = A/V^2 = 0.48255/V$, and that the average indegree and outdegree of firms is 0.48255.

Obs. 21. *In this network, for each firm that dies, two new firms are born.* Between December 2003 and June 2007, 176 firms were born while 88 died. Thus, the ratio *firm birth* to *firm death* is 2.

Thanks to these invariant properties, we can model the growth (and decay) of this network without worrying about the level of activity of the Chilean economy.

Model We will use a discrete time dynamical model. Each step, the following will happen:

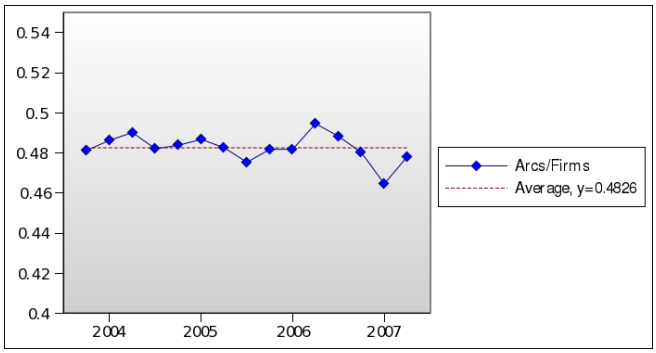


Figure 11.1: Ratio Arcs to Firms over time.

- A new firm will be born. Its financial statements will be sampled and remain fixed. Note that we are mainly concerned with its assets.
- The firm is then wired to the older firms, and viceversa. An arc from firm a to firm b will be created according to the probability:

$$f(\vec{x}_a, \vec{x}_b) = \rho_B(t) \frac{A(\vec{x}_a, \vec{x}_b)}{V(\vec{x}_a)V(\vec{x}_b)},$$

  where t is the time or iteration number, $\vec{x}_a$ and $\vec{x}_b$ are the financial statements of firms a and b, and A and V are pdfs sampled from the original network.
- A randomly chosen firm will die (with probability 0.5). Naturally, its related arcs will die too.
- Existing arcs might individually die, following the arc death rate.

In this model, the expected number of firms per time is:

$$v(t) = 0.5t,$$

and the expected number of arcs is:

$$a(t) = 0.48255v(t),$$

due to the invariance of the ratio $k = a/v = a(t)/v(t) = 0.48255$.

Now, let us review how the number of arcs varies. Each step, the following happens:

- When a firm is born, it links toward older firms creating $\rho_B(t)v(t)$ arcs in average, and viceversa, creating another $\rho_B(t)v(t)$ arcs in average. Replacing $v(t)$ by $0.5t$, $\rho_B(t)t$ arcs are expected to be born.

- When a firm dies, which happens with probability 0.5, all of its arcs disappear, destroying $2\rho v(t) = 2k$ arcs in average. Since this happens with probability 0.5, the expected death of arcs due to this reason per step is just k.
- Additional arc death destroys $\lambda a(t) = 0.5k\lambda t$ arcs.

From the previous processes, we need to define $\rho_B(t)$ and λ. First, let us estimate $\rho_B(t)$ from the expected value of $a(t+1) - a(t)$:

$$\begin{aligned} a(t+1) - a(t) &= \rho_B(t)t - k - 0.5k\lambda t \\ \rho_B(t)t &= k(v(t+1) - v(t)) + k + 0.5k\lambda t \\ \rho_B(t)t &= 0.5k + \underbrace{k + 0.5k\lambda t}_{\text{Arc death}} \\ \rho_B(t) &= 1.5k/t + 0.5k\lambda. \end{aligned}$$

We defined $\rho_B(t)$ in function of λ, so we need to estimate the latter. However, we can assume $\lambda \approx 0$ since arc death by firm death is enough to completely model arc death; first let us recall that **153 firms were born** and that **74 arcs died** from December 2003 to March 2007[1]. Now, computing the number of arcs killed due to firm death, we get $\mathbf{k\Delta t = 0.48255 \cdot 153 \approx 73.8}$, which very close to 74. Therefore, we will assume $\lambda = 0$ and $\rho_B(t) = 1.5k/t$.

Implementation decisions To have better control of the number of vertices, one vertex will die each even step. Steps start in zero, when a vertex is born and nothing else happen (we cannot kill it yet). But we found a problem with $f(x, y)$: it creates too many arcs during the first iterations. This has sence since this function was designed for large graphs (and works well with them).

To address this problem, we will recompute ρ_B not to solve $\Delta a = 0$ but $a + \Delta a/\Delta v = k(v+1)$, and we will use a and b as parameters for ρ_B. Note that v increases in one after two steps. This way, we expect ρ_B to provide stability to the number of arcs of the simulation. So, let us solve this problem:

$$\begin{aligned} a + \frac{\Delta a}{\Delta v} &= k(v+1) \\ \frac{\Delta a}{\Delta v} &= k(v+1) - a \\ \underbrace{2v\rho_B - 2k}_{\text{Even step}} + \underbrace{2v\rho_B}_{\text{Odd step}} &= k(v+1) - a \\ 4v\rho_B - 2k &= k(v+1) - a \\ 4v\rho_B &= k(v+3) - a \\ \rho_B &= \frac{3k + kv - a}{4v}. \end{aligned}$$

[1] we define the moment of death of anything as its last period, and June 2007 is the last period of everything in our database, so we cannot use it as reference for death.

Note that, if $a \approx kv$ as it should be, $\rho_B \approx 3k/4v = 1.5k/v$, which was the previous form of ρ_B. Fortunately, this solution effectively solved the problem.

The simulations In the next sections, we present different simulations which are based on different probabilities:

$$f(\vec{x}_a, \vec{x}_b) = \rho_B(t)\frac{A(\vec{x}_a, \vec{x}_b)}{V(\vec{x}_a)V(\vec{x}_b)},$$

where A and V are estimated to take into account specific attributes per simulation. Note that we will not specify the constants in detail within these functions to avoid writing large formulas. Also, note that these pdfs were sampled from the last period of the network, to replicate the properties of the more mature network.

We will assess the quality of the models according to their capacity to reproduce some basic properties of the original network. In particular, we will compare their arc correlation profiles for the following pairs of attributes: assets to assets, indegree to outdegree (which states if vertices participate in paths), and degree to degree.

11.2 Wiring based on assets

11.2.1 Arc likelihood function

We estimated A and V to only consider the assets of the firms. Or better stated, they only work with $\log(1 + \text{assets})$, which we will simple call *log-assets*.

Function V was estimated to be the following:

$$V(x) = \frac{1}{2.69\sqrt{2\pi}} \exp\left(\frac{(x - 16.96)^2}{2 \cdot 2.69^2}\right),$$

and function A was estimated to be:

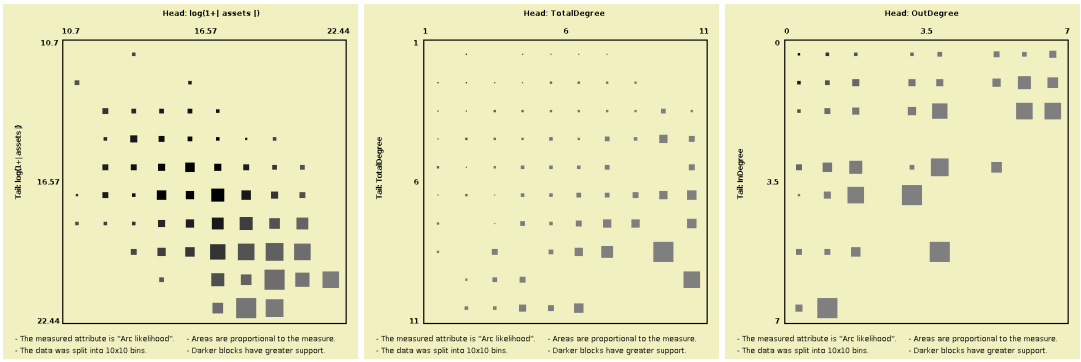
$$A(x, y) = \frac{1}{1.99\sqrt{2\pi}} \exp\left(\frac{(x - 18.3)^2}{2 \cdot 1.99^2}\right) \frac{1}{1.75\sqrt{2\pi}} \exp\left(\frac{(y - 0.53x - 7.93)^2}{2 \cdot 1.75^2}\right),$$

where x and y are the log-assets of the tail and the head of the potential arc.

11.2.2 Results

We simulated several networks getting always similar results. Reference results are illustrated in figure 11.2.

Arc correlation profiles - original network



Arc correlation profiles - simulated network

Figure 11.2: Results of the simulation based on assets.

Assets to assets We see that the correlation profile of the simulated network is more condensed than the original. In particular, the original becomes more sparse the greater the assets of both head and tail are, while the simulated remains centered around the diagonal.

Degree to degree Both correlation profiles are somewhat similar in that the likelihood increases toward the bottom right corner, when the totaldegrees of both head and tail are greater. However, the distribution of the degrees in the simulated network seems more continous, without empty spaces like in the original.

An essential difference between both correlation profiles is that the simulated one is symmetrical while the original is not; in the original, firms with higher degree tend to invest in firms with lower degree.

Indegree to outdegree Both correlation profiles seem similar in that they look symmetrical. However, note that the original is not symmetrical.

11.3 Wiring based on assets and indegrees

11.3.1 Arc likelihood function

Now, we estimated A and V to take into account the assets and indegrees of firms. Again, we only work with log-assets and *log-indegrees* ($\log(1 + \text{indegree})$).

Function V was estimated to be:

$$\begin{aligned} V(x_1, x_2) &= \frac{1}{2.69\sqrt{2\pi}} \exp\left(\frac{(x_1 - 16.96)^2}{2 \cdot 2.69^2}\right) \times \\ &\quad \frac{1}{x_2 - 0.038x_1 + 1.85} \frac{1}{0.25\sqrt{2\pi}} \exp\left(\frac{(\log(x_2 - 0.038x_1 + 1.85) - 0.35)^2}{2 \cdot 0.25^2}\right), \end{aligned}$$

where x_1 and x_2 are the log-assets and log-indegrees of the firm, respectively. (Note that more precise values were used in the code.)

Function A was estimated to be:

$$\begin{aligned} A(x_1, x_2, y_1, y_2) &= \frac{1}{1.99\sqrt{2\pi}} \exp\left(\frac{(x_1 - 18.3)^2}{2 \cdot 1.99^2}\right) \times \\ &\quad \frac{1}{0.443\sqrt{2\pi}} \exp\left(\frac{(x_2 - 0.393)^2}{2 \cdot 0.443^2}\right) \times \\ &\quad \frac{1}{1.75\sqrt{2\pi}} \exp\left(\frac{(y_1 - 0.53x_1 - 7.94)^2}{2 \cdot 1.75^2}\right) \times \\ &\quad \frac{1}{0.35\sqrt{2\pi}} \exp\left(\frac{(y_2 - 0.029x1 - 0.24x_2 - 0.27)^2}{2 \cdot 0.35^2}\right), \end{aligned}$$

where x_1 and x_2 are the log-assets and log-indegree of the tail, and y_1 and y_2 are the log-assets and log-indegree of head of the potential arc.

Unfortunately, we find ourselves with a new continuos approximation to a *very* discrete distribution: indegrees. In the data, indegrees range from 0 to 6. No less, no more. Naturally, this is wrong. Our patch consisted in using:

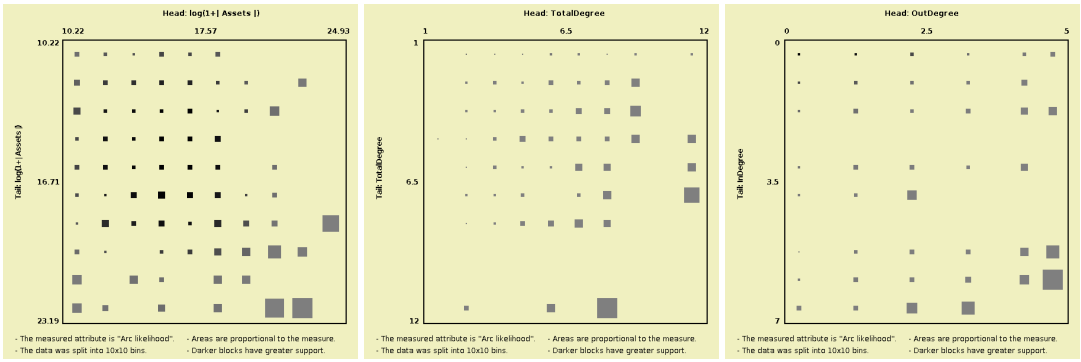
$$f(x_1, x_2, y_1, y_2) = \rho_B \left(0.4 \frac{A(x_1, x_2, y_1, y_2)}{V(x_1, x_2)V(y_1, y_2)} + 0.6\right),$$

to soften the function (avoiding over concentrated hubs).

11.3.2 Results

We simulated several networks getting always similar results, like these illustrated in figure 11.3.

Arc correlation profiles - original network



Arc correlation profiles - simulated network

Figure 11.3: Results of the simulation based on assets and indegrees.

Assets to assets Due to the patch, the simulated network presents a sparce correlation profile from assets to assets. That wealthier firms invest in poorer firms is still visible, tough.

Degree to degree Both correlation profiles seem similar. However, the simulated network presents a profile which looks more like the transposed of the original than the original itself; while in the original, there is a slight tendency that firms with higher degree invest in firms with lower degree, in the simulated one, this tendency is reversed.

Indegree to outdegree Both correlation profiles seem quite different. The simulated network exhibits a uniform profile, which is quite different than the original. Also, note that in the simulated network, indegrees reach 7 and outdegrees 5, while in the original indegrees reach 4 and outdegrees 7.

11.4 Wiring based on assets and degrees

11.4.1 Arc likelihood function

Now, we estimated A and V to take into account the assets, indegrees and outdegrees of firms. Again, we only work with log-assets, log-indegree and log-outdegree.

Function V was estimated to be:

$$\begin{aligned} V(x_1,x_2,x_3) &= \frac{1}{2.69\sqrt{2\pi}}\exp\left(\frac{(x_1-16.96)^2}{2\cdot 2.69^2}\right)\times \\ &\quad \frac{1}{x_2-0.038x_1+1.85}\,\frac{1}{0.25\sqrt{2\pi}}\exp\left(\frac{(\log(x_2-0.038x_1+1.85)-0.35)^2}{2\cdot 0.25^2}\right)\times \\ &\quad \frac{1}{x_3-0.056x_1+2.25}\,\frac{1}{0.25\sqrt{2\pi}}\exp\left(\frac{(\log(x_3-0.056x_1+2.25)-0.39)^2}{2\cdot 0.25^2}\right), \end{aligned}$$

where x_1, x_2 and x_3 are the log-assets, log-indegree and log-outdegree of the firm, respectively.

Function A was estimated to be:

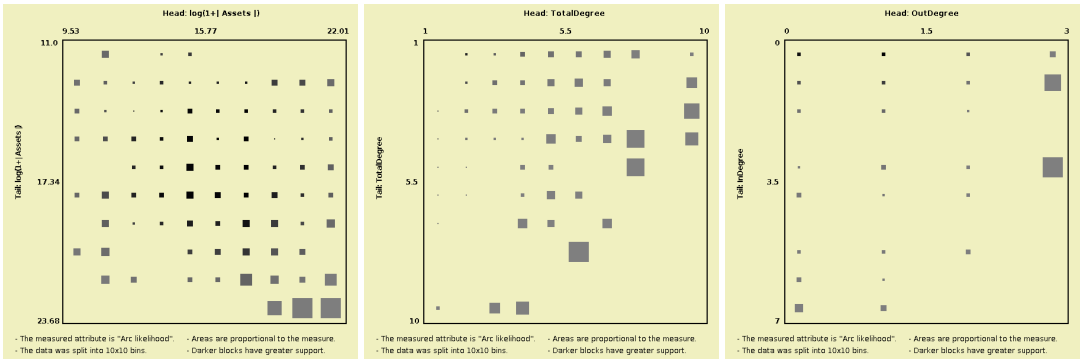
$$\begin{aligned} A(x_1,x_2,x_3,y_1,y_2,y_3) &= \frac{1}{1.99\sqrt{2\pi}}\exp\left(\frac{(x_1-18.7)^2}{2\cdot 1.99^2}\right)\times \\ &\quad \frac{1}{0.44\sqrt{2\pi}}\exp\left(\frac{(x_2-0.39)^2}{2\cdot 0.44^2}\right)\times \\ &\quad \text{Tri}(x_3-0.053x_1+0.245x_2,-0.44,-0.15,1.84)\times \\ &\quad \frac{1}{1.73\sqrt{2\pi}}\exp\left(\frac{(y_1-0.5x_1-0.12x_2-0.47x_3-7.74)^2}{2\cdot 1.73^2}\right)\times \\ &\quad \frac{1}{0.25\sqrt{2\pi}}\exp\left(\frac{(y_2-0.25x_1-0.27x_2-0.09x_3-0.23)^2}{2\cdot 0.25^2}\right)\times \\ &\quad \frac{1}{0.45\sqrt{2\pi}}\exp\left(\frac{(y_3+0.06x_1-0.07x_2-0.29x_3+0.43-0.09y_1)^2}{2\cdot 0.45^2}\right), \end{aligned}$$

where x_1, x_2, and x_3 are the log-assets, log-indegree, and log-outdegree of the potential tail; y_1, y_2, and y_3 are the log-assets, log-indegree, and log-outdegree of the potential head; and $\text{Tri}(z, min, mode, max)$ is the pdf of a triangular distribution with given min, $mode$, and max, evaluated in z.

Unfortunately, we found ourselves with a new continuos approximation to *very* discrete distribution: in and outdegrees. Moreover, A requires positive indegrees and outdegrees to work, so it cannot start creating arcs in a disconnected graph (but can continue creating new arcs when arcs exist). Therefore, we used the following patch for this situation:

$$f(x_1,x_2,x_3,y_1,y_2,y_3) = \rho_B\left(0.8\frac{A(x_1,x_2,x_3,y_1,y_2,y_3)}{V(x_1,x_2,x_3)V(y_1,y_2,y_3)}+0.2\right).$$

Arc correlation profiles - original network



Arc correlation profiles - simulated network

Figure 11.4: Results of the simulation based on assets, indegrees and outdegrees.

11.4.2 Results

We simulated several networks getting always similar results. Reference results are illustrated in figure 11.4.

Assets to assets The arc correlation profile of the simulated network seems too sparse, almost uniform, in comparison to that of the original. Both profiles are not alike at all.

Degree to degree Again, both correlation profiles are not similar. In the simulated network, is it likely that a firm with low degree is linked to a firm with high degree, which is the opposite of the original.

Indegree to outdegree In the simulated network, it is probable that a firm with high indegree is linked to a firm with high outdegree, which is unlikely in the original.

Also, like in the previous simulation, outdegrees are small compared to indegrees, which is the inverse of the original network.

11.5 Constrained wiring based on assets and degrees

11.5.1 Arc likelihood function

We estimated A and V to only consider the assets of the firms, using log-assets, like in the first simulation. Also, we will use the same A and V. However, now we will only accept arcs within:

$$\frac{x_2}{3} + \frac{y_3}{6} \leq 1,$$

where x_2 is the indegree of the potential tail, and y_3 is the outdegree of the potential head.

Using the property that ρ_B adapts itself to keep $a \approx kv$, we write the arc likelihood function for this case:

$$f(x_1, x_2, x_3, y_1, y_2, y_3) = \begin{cases} \rho_B \frac{A(x1,y1)}{V(x_1)V(y_1)}, & \text{if } \frac{x_2}{3} + \frac{y_3}{6} \leq 1 \\ 0, & \text{else.} \end{cases}$$

11.5.2 Results

We simulated several networks getting always similar results. Reference results are illustrated in figure 11.5.

Assets to assets Like in the first simulation, both arc correlation profiles seem similar, yet the original becomes more sparse as assets get greater.

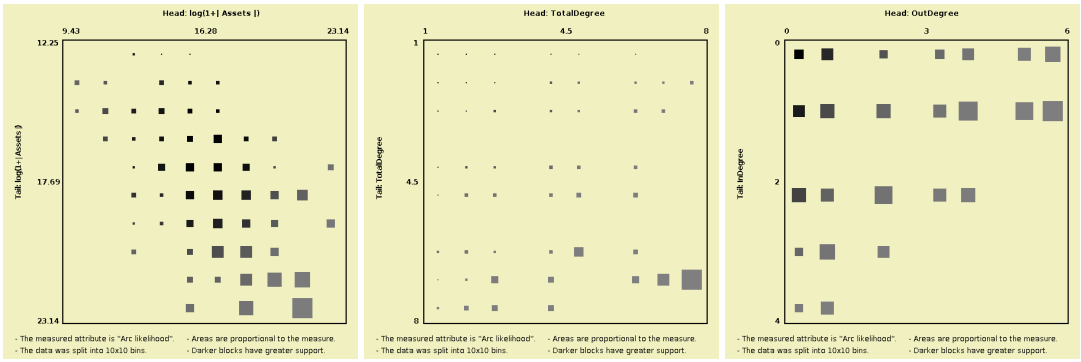
Degree to degree Both profiles seem similar. However, the slight tendency that firms with greater degrees are linked to firms with lower degrees is more significative in the original.

Indegree to outdegree Note that, so far, this simulation is the closest to the original in the profile from indegree to outdegree. In particular, the relation outdegree to indegree in the simulation is $6:4$ while in the original is $7:4$. And this was obtained only by applying a simple constraint to the arc likelihood function. To yield better results, a more precise constraint might be needed.

11.6 Comments

We studied, modeled and simulated the network of investment between firms. Without resorting to economic hypothesis but to statistical facts, we reproduced the network combining a discrete time model with continuous probabilities. Anyway, an

Arc correlation profiles - original network



Arc correlation profiles - simulated network

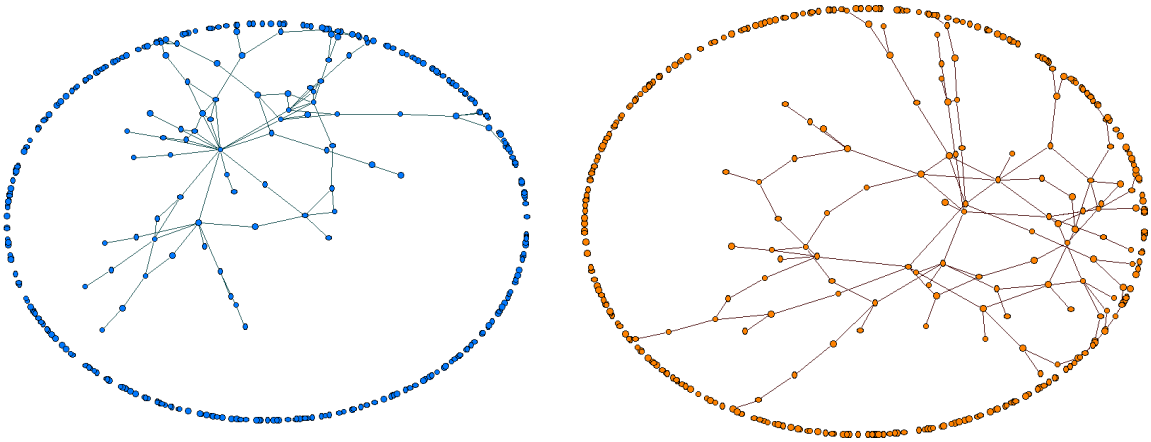
Figure 11.5: Results of the simulation based on assets.

ideal model might not be successful to explain the multivariate nature of the data; it would require a large amount of knowledge to do this and, due to the number of variables, the model might end up being complex.

To our surprise, the simplest simulations were the most successful. It was not necessary to include more variables than the assets in the pdfs; we only had to add the constraint on the degrees to reproduce the profiles of the degrees. (And that constraint was just a simple approximation.) However, that constraint would get outdated if more vertices were born. So, the last simulation model was not extensible, unless we modeled the constraint in function of the number of vertices. Therefore, the most predictive simulation was the first one.

Anyhow, it seems that shareholdings are more concentrated in the original network, as can be seen in figure 11.6. Probably, a better simulation model might be able to reproduce this fact as well. In particular, the arc likelihood function should not be estimated to take the topology into account in the pdfs. The topology should be included in a different way in these models. Note that the pdfs are somewhat static while the topological attributes (degrees, centralities) are dynamic.

Regarding centralities, the original network and the simulated network had some differences. If we sort their eigenvector centralities, their correlation is 0.627 (not so



(a) (b)

Figure 11.6: The (a) original and the (b) first simulated network.

bad), but if we raise each element of the simulated eigencentrality by 8, their correlation becomes 0.999. In other words, **ownership is much more concentrated in the original network**, since our eigenvectors are unitary ($||\vec{c}_{\text{net}}|| = 1$), thus powering by 8 the elements of a vector like this only makes its smaller elements much smaller. The distribution of indegrees are quite similar, with correlation of 0.98, but outdegrees with 0.91. Concerning Katz/Bonacich's centralities, their correlation is 0.956 when $\beta = 0.5$, 0.934 when $\beta = 0.7$, and 0.92 when $\beta = 0.9$.

Chapter 12

Conclusion

12.1 Summary

We studied the Chilean shareholding network starting practically from scratch. After downloading and preprocessing the data, we developed several methodologies to analyze it. We focused in the multivariate aspect of the network data, which is not traditionally studied within the social networks and cross shareholding disciplines. (In general, multivariate data is studied using linear regression and ANOVA.) But we went further in analyzing multivariate data; we literally reduced the topology to the distribution of arcs and vertices. Using the right attributes, we can determine the probability that two firms are related knowing only their characteristics. We even developed a software application to perform these tasks.

Once we had developed the methodologies, we applied them to study the Chilean shareholding network. In particular, we found that wealthier firms become the owner of the poorer firms (seeking ownership or participation at least), and that financial firms tend to invest in economic performance (which was expected).

Finally, we reproduced the topology and dynamics of a subset of this network, the network of investment between firms (restricting ourselves to those firms which financial statements are in the data). We developed a general simulation model and ran several cases, reproducing the original network in the simplest ones. In particular, we found that ownership is much more concentrated in the original network than in the best simulation (using eigenvector centrality to compare them, recall chapter 2). We consider this to be emergent behavior, since it was not explained by simple relational interaction. We may think of limitations of the methodologies used, but we believe this happens because some shareholders desire to have control of portions of the market. Their *awareness* may be at network level, while our concern was at relational level. Naturally, behavior due to network level knowledge cannot be explained with relational level analysis (*emergence*).

12.2 Discussion

In spite of the complexity of this thesis, we could have extended the analytical methodologies and the software. Time is not infinite, so this is left for future research.

First of all, we did not develop analytical methodologies to study the dynamics of networks. We put much more emphasis to the multivariate aspect of the data. Anyway, our efforts in the study of the dynamics are the characterization of its processes (birth and death of firms and arcs), and the simulation models based on them.

Second, the use of joint pdfs has two drawbacks. First, it does not work well with discrete data. Second, *pdfs say nothing*; the researcher will have to find out the expected values, etc. of the pdfs, using additional effort. In many situations, knowing $\mathbb{E}(y|\vec{x})$ and Var$(y|\vec{x})$ might be more practical than knowing $f(y|\vec{x})$. (Note that we computed $f(x_1, ..., x_n)$ by computing $f(x_k|x_1, ..., x_{k-1})$, so we already worked in the conditional.)

Third, arc likelihood functions have to be compatible with discrete data. We should be aware that not all networks are composed of continuous data, but all of them have discrete data (degrees, for example). If that data is ordinal, discrete distributions could be used.

Fourth, Network Observer should let the user rearrange the parameters of the joint pdfs. Let us suppose that we want to estimate the joint distribution of θ and κ, where θ is discrete and κ is continuous. If we estimated the joint pdf of those variables, assuming that they were continuous, having $f(\kappa)$ and $f(\theta|\kappa)$ could be less convenient than $f(\theta)$ and $f(\kappa|\theta)$. In particular, we could have replaced the continuous distribution $f(\theta)$ by the discrete $\Pr(\Theta = \theta)$. Then, $f(\kappa|\theta) = \Pr(\Theta = \theta)f(\kappa|\theta)$ could be a better estimation than the original.

Fifth, preprocessing the data was quite an issue. It has small theoretical value but high technical value. It could be great if a better methodology for matching names were developed. Moreover, it would be better if it were designed to handle firm renaming too; for example, CB S.A., a firm with many arcs, changed its name to Curauma S.A. during 2008. If we had control of this issue, arc and firm death would have been probably smaller.

Part IV

Appendix

Appendix A

Database

We stored the processed data in a relational database. We now describe the details regarding how the data was stored.

A.1 Schemas

The following are the schemas of the original tables. They are presented in the way they were created. Note: `key1` and `key2` are the words used to identify a firm or person. These keys are the two most unique words in their names.

```
CREATE TABLE accionistas(
  rut integer,
  ano integer,
  mes integer,
  accionista varchar,
  share decimal,
  key1 varchar,
  key2 varchar
);

CREATE TABLE empresas(
  rut integer,
  nombre varchar,
  vigencia integer,
  key1 varchar,
  key2 varchar
);

CREATE TABLE financieros(
  rut integer,
  ano integer,
  mes integer,
  tipo varchar,
  activos bigint,
  patrimonio bigint,
  deudalp bigint,
  dividendos bigint,
  utilidad bigint,
  operacion bigint,
  inversion bigint
);
```

```
CREATE TABLE tipocambio(
  tipo varchar,
  ano integer,
  mes integer,
  cambio decimal
);
```

The following are the intermediate tables used to transform the keys to ids.

```
CREATE TABLE keys(
  key1 varchar,
  key2 varchar
);

CREATE TABLE keysnum(
  id integer,
  key1 varchar,
  key2 varchar
);
```

Finally, the following are the derived tables, which resort to `ids` for identification. Also, the currencies were all converted to CLP.

```
CREATE TABLE zaccounts(
  id integer,
  year integer,
  month integer,
  assets bigint,
  equity bigint,
  debt bigint,
  dividends bigint,
  profits bigint,
  operations bigint,
  investment bigint
);

CREATE TABLE zfirms(
  id integer,
  rut integer,
  name varchar
);

CREATE TABLE zshareholders(
  id1 integer,
  id2 integer,
  year integer,
  month integer,
  share decimal
);
```

A.2 Network creating script

We created a large SQL script to create the files to be opened by Network Observed. But we created that script with another script, which was written in Perl:

```
#!/usr/bin/perl

print "
SELECT load_extension('./libsqlitefunctions.so');
.read dump2.sql
```

```
.separator ,
create table tmp(a integer,b integer,c bigint,d bigint);
create index itmpa on tmp(a);
create index itmpb on tmp(b);
create index itmpcd on tmp(c,d);
";

for($ano=2003;$ano<=2007;$ano++) {
 for($mes=3;$mes<=12;$mes+=3) {
  grafgral($ano,$mes);
 }
}
print ".quit\n";


sub grafgral {

 ($a,$m)=@_;
 if ($a==2003 and $m<12) {
  return;
 }
 if ($a==2007 and $m>6) {
  return;
 }
 system("rm -rf $a-$m");
 system("mkdir $a-$m");


# The following instructions create the arcs of the relation
# A invests in B, and its variants according to the amount of
# ownership (less than 5%, more than 5%, than 10% and 20%)

 print "
.header OFF

.output $a-$m/GRall.csv
select distinct id1,id2 from zshareholders
where year=$a and month=$m;

.output $a-$m/GRfirm.csv
select distinct a.id1,a.id2 from zshareholders a,zfirms b
where a.year=$a and a.month=$m and a.id1=b.id;

.output $a-$m/GRfirm5-.csv
select distinct a.id1,a.id2 from zshareholders a,zfirms b
where a.year=$a and a.month=$m and a.id1=b.id and a.share<5;

.output $a-$m/GRfirm5+.csv
select distinct a.id1,a.id2 from zshareholders a,zfirms b
where a.year=$a and a.month=$m and a.id1=b.id and a.share>=5;

.output $a-$m/GRfirm10+.csv
select distinct a.id1,a.id2 from zshareholders a,zfirms b
where a.year=$a and a.month=$m and a.id1=b.id and a.share>=10;

.output $a-$m/GRfirm20+.csv
select distinct a.id1,a.id2 from zshareholders a,zfirms b
where a.year=$a and a.month=$m and a.id1=b.id and a.share>=20;
";


# The following instructions separates the above arcs per
# tail and head. Basically, the four combinations
# {Eq0,EqX}->{Eq0,EqX}.
```

```
print "
.output /dev/null
delete from tmp;
insert into tmp select a.id1,a.id2,b.equity,c.equity
from zshareholders a,zaccounts b,zaccounts c
where a.year=$a and a.month=$m and b.year=$a and b.month=$m
and c.year=$a and c.month=$m and a.id1=b.id and a.id2=c.id;

.output $a-$m/GR-Eq0-Eq0.csv
select distinct a,b from tmp where c=0 and d=0;

.output $a-$m/GR-Eq0-EqX.csv
select distinct a,b from tmp where c=0 and d<>0;

.output $a-$m/GR-EqX-Eq0.csv
select distinct a,b from tmp where c<>0 and d=0;

.output $a-$m/GR-EqX-EqX.csv
select distinct a,b from tmp where c<>0 and d<>0;
";


# The following instructions create the arcs of the relation
# "shareholders with a common firm". As the relation is symmetrical,
# (a,b) and (b,a) arcs are present. Removed: "a.id2>b.id2", which
# made the arcs unique ( (a,b) => (b,a) doesn't exist -- this
# was removed). Note that (a,a) is not allowed.
# Flavors: both shareholders own more than X% of the same firm.

 print "
.output /dev/null
delete from tmp;
insert into tmp select a.id2,b.id2,a.share,b.share
from zshareholders a, zshareholders b
where a.id1=b.id1 and a.year=$a and a.month=$m and b.year=$a and b.month=$m
and a.id2<>b.id2;

.output $a-$m/GR-ShComFirm.csv
select distinct a,b from tmp;

.output $a-$m/GR-ShComFirm5-.csv
select distinct a,b from tmp where c<5 and d<5;

.output $a-$m/GR-ShComFirm5+.csv
select distinct a,b from tmp where c>=5 and d>=5;

.output $a-$m/GR-ShComFirm10+.csv
select distinct a,b from tmp where c>=10 and d>=10;
";


# The following insts create the relation "firms with a
# common shareholder". Flavors: that the shareholder own more
# than the X% of the firm. Note: symmetrical; (a,b) <=> (b,a)
# (a,a) is not allowed.

print "
.output /dev/null
delete from tmp;
insert into tmp select a.id1,b.id1,a.share,b.share
from zshareholders a, zshareholders b
where a.id2=b.id2 and a.year=$a and a.month=$m and b.year=$a and b.month=$m
and a.id1<>b.id1;

.output $a-$m/GR-FirmComSh.csv
```

```
select distinct a,b from tmp;

.output $a-$m/GR-FirmComSh5-.csv
select distinct a,b from tmp where c<5 and d<5;

.output $a-$m/GR-FirmComSh5+.csv
select distinct a,b from tmp where c>5 and d>5;

.output $a-$m/GR-FirmComSh10+.csv
select distinct a,b from tmp where c>10 and d>10;
";


# NODES/VERTICES: Shareholders

 print "
.header ON\n";
 print "
.output $a-$m/NDall.csv
select id1 from zshareholders where year=$a and month=$m
union
select id2 from zshareholders where year=$a and month=$m;

.output $a-$m/NDshare.csv
select id1, avg(share) as AvgShare, count(distinct id2) as NumInvestmnt,
max(share) as MaxShare, min(share) as MinShare,
avg(share*share) as AvgShare2, sum(share) as TotalShare
from zshareholders
where year=$a and month=$m
group by id1;

.output $a-$m/NDshare5+.csv
select id1, avg(share) as AvgShare, count(distinct id2) as NumInvestmnt,
max(share) as MaxShare, min(share) as MinShare,
avg(share*share) as AvgShare2, sum(share) as TotalShare
from zshareholders
where year=$a and month=$m
group by id1
having max(share)>=5;
";


# NODES/VERTICES: Firms

 print "
.output $a-$m/NDfirm.csv\n";
 print "select DISTINCT A.id as ID,A.assets as ASSETS,A.equity as EQUITY,A.debt as LT_DEBT,
A.profits as PROFIT,A.dividends as DIVIDE,A.operations as OPERATION,A.investment as INVESTMENT,
max( 1.0*A.profits*A.assets/(1+A.assets)/(1+A.assets) ) as RoA,
max( 1.0*A.profits/(1+A.equity)*A.equity/(1+A.equity) ) as RoE,
max( 1.0*A.debt/(1+A.equity)*A.equity/(1+A.equity) ) as DoE,
max( 1.0*(A.debt+A.equity)/(1+A.assets)*A.assets/(1+A.assets) ) as DEoA,
max( 1.0*A.investment/(1+A.assets)*A.assets/(1+A.assets) ) as IoA,
$a*4 + $m/3 - min( B.year*4 + B.month/3 ) as history ";
 print "from zaccounts A, zaccounts B where A.year=$a and A.month=$m and B.id=A.id
group by A.id,A.assets,A.equity,A.debt,A.profits,A.dividends,A.operations,A.investment;\n";

 print "
.output $a-$m/NDfirm-Eq+.csv\n";
 print "select DISTINCT A.id as ID,A.assets as ASSETS,A.equity as EQUITY,A.debt as LT_DEBT,
A.profits as PROFIT,A.dividends as DIVIDE,A.operations as OPERATION,A.investment as INVESTMENT,
max( 1.0*A.profits*A.assets/(1+A.assets)/(1+A.assets) ) as RoA,
max( 1.0*A.debt/(1+A.equity)*A.equity/(1+A.equity) ) as DoE,
max( 1.0*(A.debt+A.equity)/(1+A.assets)*A.assets/(1+A.assets) ) as DEoA,
max( 1.0*A.investment/(1+A.assets)*A.assets/(1+A.assets) ) as IoA,
```

```
$a*4 + $m/3 - min( B.year*4 + B.month/3 ) as history ";
 print "from zaccounts A, zaccounts B where A.year=$a and A.month=$m and B.id=A.id and A.equity>0
group by A.id,A.assets,A.equity,A.debt,A.profits,A.dividends,A.operations,A.investment;\n";

 print "
.output $a-$m/NDfirm-Eq+As+.csv\n";
 print "select DISTINCT A.id as ID,A.assets as ASSETS,A.equity as EQUITY,A.debt as LT_DEBT,
A.profits as PROFIT,A.dividends as DIVIDE,A.operations as OPERATION,A.investment as INVESTMENT,
max( 1.0*A.profits*A.assets/(1+A.assets)/(1+A.assets) ) as RoA,
max( 1.0*A.debt/(1+A.equity)*A.equity/(1+A.equity) ) as DoE,
max( 1.0*(A.debt+A.equity)/(1+A.assets)*A.assets/(1+A.assets) ) as DEoA,
max( 1.0*A.investment/(1+A.assets)*A.assets/(1+A.assets) ) as IoA,
$a*4 + $m/3 - min( B.year*4 + B.month/3 ) as history ";
 print "from zaccounts A, zaccounts B where A.year=$a and A.month=$m and B.id=A.id and A.equity>0 and A.assets>0
group by A.id,A.assets,A.equity,A.debt,A.profits,A.dividends,A.operations,A.investment;\n";
}
```

This script generated a SQL script of 2078 lines. It created the following structure of files per period:

```
NDshare.csv NDshare5+.csv NDfirm-Eq+.csv NDfirm-Eq+As+.csv NDfirm.csv NDall.csv
GR-ShComFirm.csv GR-ShComFirm5+.csv GR-ShComFirm5-.csv GR-ShComFirm10+.csv
GRfirm.csv GR-FirmComSh.csv GR-FirmComSh5+.csv GR-FirmComSh5-.csv GR-FirmComSh10+.csv
GRfirm5+.csv GRfirm5-.csv GRfirm20+.csv GRfirm10+.csv GR-EqX-EqX.csv GR-EqX-Eq0.csv
GR-Eq0-EqX.csv GR-Eq0-Eq0.csv GRall.csv
```

Appendix B

Simulators

The following is the source code of the first simulator, explained in sections 10.1 and 10.2.

```
#!/usr/bin/perl

# SIM2x:
# This program simulates a shareholding network
# which only uses ASSETS to motivate wiring
# among vertices.


# Simulation vars
$N=500;
@alive=();
@lassets=();
@ea=();
@arcs=(());

# First vertices
$alive[0]=1;
$lassets[0]=norm(16.9573,2.6915);
$ea[0]=1.5;
$alive[1]=1;
$lassets[1]=norm(16.9573,2.6915);
$ea[1]=0.5;

$num_f=2;
$sub_b=0;


$i=2;
for($i=2;$num_f<=$N;$i++) {
 $arc_b=0;
 $arc_d=0;

 # Node birth
 $alive[$i]=1;
 $lassets[$i]=norm(16.9573,2.6915);   #mean 16.3, stdev 2.68
 $ea[$i]=1/(1+exp(norm(0.4,0.7)));
 $num_f++;

 # Arc birth
 for($j=0;$j<$i;$j++) {
   if ($alive[$j]==0) {
     next;
   }
```

```
   if (rand()< attract($lassets[$i],$lassets[$j])*rho($num_f,$sub_b))
   { $arcs[$i][$j]=1; $arc_b++; }

   if (rand()< attract($lassets[$j],$lassets[$i])*rho($num_f,$sub_b))
   { $arcs[$j][$i]=1; $arc_b++; }

 }

 # Vertex death
 if ( ($i%2)==0 ) {
   # We kill the i-1 vertex
   $k=int($i-1);
   $alive[$k]=0;
   for($l=0;$l<=$i;$l++) {
     if ($arcs[$l][$k]==1) { $arcs[$l][$k]=0; $arc_d++; }
     if ($arcs[$k][$l]==1) { $arcs[$k][$l]=0; $arc_d++; }
   }
   $num_f--;
 }

 # Updated number of arcs
 $sub_b+=$arc_b-$arc_d;

 # Writing the output
 if ($num_f%50==0) {
   saveto();
 }

}


# ATTRACT
# Vertex attraction - wiring function - A(x,y)/V(x)V(y)
sub attract{
 local $a1,$a2,$arc,$ver;
 ($a1,$a2)=@_;
 $ver=norm_pdf($a1,16.9573,2.6915)*norm_pdf($a2,16.9573,2.6915);
 $arc=norm_pdf($a1,18.729,1.99)*norm_pdf($a2-0.52646*$a1,7.9347,1.75365);
 return $arc/$ver;
}


# RHO
# Adaptive density.
sub rho {
 local $v,$a,$k;
 $k=0.48255; # Key constant
 ($v,$a)=@_;
 return (3*$k+$k*$v-$a)/4/$v;
}


# SAVETO
# It saves the network to a file
sub saveto{
 open(OUT,">output/sim2-$num_f.netobs");
 print OUT "* id, assets, equity, debt\n";
 for($_i=0;$_i<$i;$_i++) {
   if ( $alive[$_i]==0 ) {next;}
   print OUT "$_i,".exp($lassets[$_i]-1);
   print OUT ",".( exp($lassets[$_i]-1)*$ea[$_i] );
   print OUT ",".( exp($lassets[$_i]-1)*(1-$ea[$_i]) )."\n";
 }
 print OUT "*Arcs\n";
```

```
 for($_i=0;$_i<$i;$_i++) {
   if ($alive[$_i]==0) {next;}
   for($_j=0;$_j<$i;$_j++) {
     if ($alive[$_j]==0) {next;}
     if ($arcs[$_i][$_j]==1) {
       print OUT "$_i,$_j\n";
     }
   }
 }
 close(OUT);
}


# NORM
# It samples a random number which follows a normal distro
# Parameter: mean and standard deviation
sub norm{
 local $mean,$stdev;
 ($mean,$stdev)=@_;
 return $mean+$stdev*sqrt(-2*log(rand()))*cos(2*355/113.*rand());
}


# NORM_PDF
# Evaluation of a number in a normal pdf
sub norm_pdf{
 local $var,$mean,$stdev,$pi;
 $pi=355.0/113.0;
 ($var,$mean,$stdev)=@_;
 return exp( -($var-$mean)*($var-$mean)/2/$stdev/$stdev )/$stdev/sqrt(2*$pi);
}
```

The rest of the simulators are variations of this one. The main difference between those scripts is their `attract` function and the variables needed to keep a registry of the in and outdegrees of vertices.

Appendix C

Network files

C.1 Firms

The following corresponds to the first 20 lines of the `NDfirms.csv` file, for June 2006:

```
ID,ASSETS,EQUITY,LT_DEBT,PROFIT,DIVIDE,OPERATION,INVESTMENT,RoA,RoE,DoE,DEoA,IoA,history
1,247188684,205926215,35419226,14508310,14772168,0,14536967,0.0586932607425244,
   0.070453923795434,0.171999595369637,0.976361195592908,0.0588091924239607,13
5,175090,69612,0,33533,0,51451,-342,0.191516459911967,0.481699083418281,0.0,
   0.397573846878951,-0.00195325885813655,7
6,4283643,115966,2620166,843729,0,1149022,1214,0.196965201364715,7.27553290508814,
   22.593870721278,0.638739204579243,0.00028340350332484,13
36,1475788,1431146,10806,-18007,0,0,30854,-0.0122016004987103,-0.0125822067319321,
   0.00755058177071465,0.977071263526205,0.0209067685781756,13
53,1210984,877188,0,-98743,0,-40746,8332,-0.0815393406694484,-0.112567402729023,0.0,
   0.724358497950762,0.00688034378596806,13
93,42454910,6917811,1806771,522632,0,1557869,-200466,0.0123102834366963,0.0755487319475084,
   0.261176617523481,0.205502298532617,-0.00472185644855417,3
105,9487916,2040656,0,0,0,0,0,0.0,0.0,0.0,0.215079430492546,0.0,5
106,4826698,2495423,1358705,192972,0,389778,-16367,0.0399801106345933,0.0773303144754777,
   0.544478395463611,0.798501667806127,-0.00339092962065165,13
107,16962073,4680791,11260655,924872,0,-343809,1260705,0.0545258761089075,0.197588741907612,
   2.40571522816742,0.939828765053801,0.0743249278169089,12
110,15440836,7549833,6667292,235668,0,405820,-96733,0.0152626431285659,0.0312149868175922,
   0.88310433274368,0.920748278040346,-0.0062647506566677,13
112,5063971,3516811,565423,138214,203000,380986,-2682,0.027293589440543,0.039300923876245,
   0.160777101313022,0.806132655130634,-0.00052962367690347,13
113,1354387231,896581021,397798371,42767977,20942777,38899235,16529011,0.0315773627792311,
   0.0477011847260568,0.443683683677521,0.955693734009083,0.0122040510994687,13
142,40938311,10895314,4329414,184829,436645,3655727,-587749,0.00451481720802711,
   0.0169640788757296,0.397364702409729,0.37189436701996,-0.0143569423586165,8
147,4559923,1314999,3130800,-30616,0,104388,-69241,-0.00671414551774394,-0.0232821115725253,
   2.38083469137909,0.974971956775558,-0.01518467957258,2
148,11810636,7170793,4519121,-80683,0,268867,-13818,-0.00683138370679039,
   -0.0112516115716595,0.630211991836318,0.989778367603843,-0.0011699621984858,13
154,145481,126953,3662,5165,0,3276,2127,0.0355024298354857,0.0406837068227288,
   0.0288448662894159,0.897802492345008,0.0146202649099861,13
164,1114432536.3,323844822.78,449660282.88,52142882.58,19985853.78,218605412.94,17466669.54,
   0.046788729499536,0.161011937169052,1.38850538428542,0.694079793147404,0.0156731510788839,13
165,49418255,42926656,2085229,1719126,866261,2456774,-65345,0.0347872649575642,
   0.0400479813732529,0.0485765511957628,0.910835139329175,-0.00132228459615644,5
225,138798880,103035017,31411626,3654907,955,0,3656214,0.0263323950981104,0.0354724736839229,
   0.304863592056988,0.968643558670739,0.026341811600471,13
```

C.2 Shareholders

The following corresponds to the first 20 lines of the `NDshare.csv` file, for June 2006:

```
id1,AvgShare,NumInvestmnt,MaxShare,MinShare,AvgShare2,TotalShare
1,62.82,2,63.97,61.67,3947.6749,125.64
6,33.33,1,33.33,33.33,1110.8889,33.33
8,7.52,2,10.23,4.81,63.8945,15.04
9,3.44,1,3.44,3.44,11.8336,3.44
10,0.1,1,0.1,0.1,0.01,0.1
11,50.01,1,50.01,50.01,2501.0001,50.01
13,4.65,1,4.65,4.65,21.6225,4.65
27,1.69,1,1.69,1.69,2.8561,1.69
29,11.0,1,11,11,121.0,11
30,0.05,1,0.05,0.05,0.0025,0.05
36,1.69,1,1.69,1.69,2.8561,1.69
38,13.36,1,13.36,13.36,178.4896,13.36
41,25.0,1,25,25,625.0,25
42,50.0,1,100,0,5000.0,100
43,0.01,1,0.01,0.01,0.0001,0.01
46,0.11,1,0.11,0.11,0.0121,0.11
47,0.01,1,0.01,0.01,0.0001,0.01
50,0.02,1,0.02,0.02,0.0004,0.02
52,3.29,1,3.29,3.29,10.8241,3.29
```

C.3 Investments

The following corresponds to the first 20 lines of the `GRall.csv` file, for June 2006:

```
1,1119
1,3127
6,5
8,1937
8,5548
9,1096
10,1256
11,5460
13,4458
27,5097
29,5191
30,692
36,36
38,5460
41,1970
42,5206
43,6
46,1506
47,1152
50,558
```