Inferring Companies Similarities from Brazilian Government Expenditure Data*

Marcelo Pita
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Coordenação Estratégica de Tecnologia
Serviço Federal de Processamento de Dados
Belo Horizonte, Brazil
marcelo.pita@serpro.gov.br

Gustavo da Gama Torres
Coordenação Estratégica de Tecnologia
Serviço Federal de Processamento de Dados
Belo Horizonte, Brazil
gustavo.gamatorres@serpro.gov.br

ABSTRACT

A graph-based method is proposed for inferring similarities among companies from their affiliations in the context of expenditure financial transactions in the Brazilian Federal Government. There are trusted and untrusted companies. We performed a basic cluster analysis in the companies network to verify whether clusters (connected components) are discriminative concerning companies trustworthiness. Results show evidences that this is true, reinforcing the following hypotheses: (1) there are suppliers associations, which evidences the formation of cartels; and (2) public agencies and agents play an important role in the legality of financial transactions.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*Network problems*; J.1 [**Computer Applications**]: Administrative Data Processing—*Government*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economy*

General Terms

Algorithms, Economics, Experimentation, Human Factors

Keywords

e-Government, open government data, data science, social network analysis, economics, clustering, classification model

*This work was sponsored by Universidade Federal de Minas Gerais (UFMG) and Serviço Federal de Processamento de Dados (SERPRO), and performed at the Laboratório de Computação de Alto Desempenho, Coordenação Estratégica de Tecnologia at Belo Horizonte (LCAD/CTBHE/CETEC/SERPRO).

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil

1. INTRODUCTION

Corporate scandals and the corruption of public officers are recurrent issues in the Brazilian media. These situations have been described as very negative things, unwanted, but difficult to avoid.

There are groups that advocate the existence of capacity within public governance and regulation systems to identify and correct problematic situations. Although the institutions are basically well established, there is a growing perception, reinforced by recurrent news, that the scandals reveal the existence of a structural deficit in public governance, which would represent a threat to the political system as a whole. In fact, the perception that the procurement is dominated by supplier associations acting under economic, fiscal, criminal and political infringing conduct is common sense understanding. There is no numeric confirmation to measure the real size of the problem. Nevertheless, it is the lack of awareness of the extent of socio-political and administrative phenomena that allow us to establish the existence of accountability deficit, which inhibits government responsiveness.

The "bureaucratic machine" has difficulty to see the accountability mechanisms as something that help the government, and not as an additional burden inside its routines. However, corruption is contingent and contexts change. The adequacy of the state public functions presupposes not only the "rule of law", but, above all, the recognition of the dynamics of changes, which define an underlying evolutionary process. Typically, this process is supplied by unexpected events which are sometimes difficult to understand and explain [1]. It leads to the necessity of supplying management accountability means. Some of those means are data discovery techniques.

This text proposes the initial examination of the cartel formation issue by Brazilian government suppliers as a case study in experimental computing [2]. The study has an economic interpretation, but this work is mainly an exercise of data science, more specifically about data science process in this emerging subject which is related to the problem of hypotheses formulation on regulatory activities. A plausible hypothesis must be known in order to support the decision whether to continue or to terminate a research line. Also, the study must be agile enough so that the effort that pre-

cedes the decision does not imply in a waste of resources.

The presupposition assumed, based on the "theory of evolutionary governance" [3], is that the critical phenomena for public governance are emergent, with probabilistic behavior with an underlying structure, typical of complex systems. Unveiling this structure is a prerequisite for the understanding of the problem that involves relationships between agents. In this case, the analysis, based on complex networks, is possible [4], and can be imposed as necessary, according to the relationships that are established in such problems. The case study initially identifies the validity of spending resources to deepen the understanding of government procurement and its "pathologies". Additionally, it allows giving insights about the generalization of complex networks approach for studies on regulation.

2. HYPOTHESES FORMULATION PROBLEM

Investigations regarding the regularity of transactions between business and government are an old and recurring theme. George Stigler, co-authored with Claire Friedland, coined the term "regulatory capture" [5] as he started a series of studies about the ineffectiveness of State activity based on regulation. These studies showed how the corporations would act to reverse the State regulatory bodies control interface. Although it was not his first goal, his studies raised important issues concerning public transparency, violated due to the influence of the external agent to give rise to regulatory sharing of public interest information formatted in selected information from publications and not representative of the control action. He described, in this way, the most basic way to produce gaps in accountability mechanisms. Although the studies were focused on regulatory agencies, the control problem is similar when transposed to the government procurement context.

The Society is developing a new perception towards the necessity of transparency as a requirement to proceeding on the public sphere, whether governments, companies or individuals. Thus, it seeks to improve the definition of the boundaries between the public and the private area, as a means of preserving privacy in the private sphere. A wide regulation has been built over the years, defining criteria for public and corporate governance. The goal has been to define ways to influence the behavior of organizational systems for the benefit of public governance.

Human societies, organizations and individuals are typically organized into networks. There are in such networks induction control structures which are not random and reveal models of fundamental structures [6]. According to the authors, real networking profiles provide information about the high-level organization and show a strong correlation with certain control properties. This type of observation leads to the fact that effective regulation requires the identification of the actual network structure. The problem, in case it happens, comes from the fact that the institutional framework of each country is unique, unfolding as following: (a) the quality of the data depends on the informational potential of accountability mechanisms; (b) analysis methods of complex network depend on the information structure [7].

For some situations it is necessary to define a strategy of analysis, and identifying if the study under such strategy is viable, as soon as possible, before spending unnecessary effort, and only then to conclude about the feasibility of the task. The hypotheses formulation problem is a creative process, rather than the art with digital investigations [9]. It is necessary to formulate multiple "occurrence hypotheses". Some occurrence hypotheses are supported on preliminary experimental evidences, which justify the choice of assumptions and established correlations made with the theories, not requiring logical consistency.

This work aims at developing the research from first occurrence hypotheses. This is established as the first step of the identification of the complex structure of the government purchases network, from the available data. The first issue of the hypothesis formulation process is the identification of topological elements of the control structure. It refers to these questions: (a) do suppliers act alone? (b) are there associations of suppliers? (c) is there a large syndicate of crime? (d) are they hierarchical? Thus, the formulation of hypotheses need to be preceded by a kind of data validation, a situation frequently found in data mining. Such validation requires obtaining experimental evidence, which usually is not a trivial task.

3. TECHNICAL FRAMEWORK

A graph-based method is proposed for inferring similarities among companies from their affiliations in the context of expenditure financial transactions in the Brazilian Federal Government. Affiliations are interpreted as direct payments (from government offices to companies) and payment cards (from authorized people to companies) transactions. The proposed generative method aims to produce an acceptable approximation of the real world companies similarity network with edges representing probabilities of similar behavior.

As one can see, non-tecnhical motivation of this study is closely related to investigation of financial fraud (*e.g.* misappropriation of money, money laundering) involving public funds. In such scenarios, similarity among untrusted agents is important. Although this approach is far from being a potential source of evidence of companies involvement in financial crime, it could provide a strongly reduced sample to start investigations by identifying behaviors that are similar to those of admittedly untrusted agents.

Expenditure data are available in the Federal Government transparency Web portal[1]. This work is particularly aimed at direct expenditures and financial transfers. Direct expenditures include government-to-companies payments and federal government payment card (FGPC) usage. There are kinds of transfers which are resources moved from federal government to states, cities or companies, but for the purpose of this work, only the usage of civil defense payment card (CDPC) is included. Untrusted agents data are also available in the portal, and include suspended untrusted (for-profit and nonprofit) companies and expelled public employees, from which we used only the database of untrusted companies.

Expenditure data allow to interpret the following possible connections:

- **Government office to company:** connection weighted with the total amount of money paid by the govern-

[1]Web address: http://www.transparencia.gov.br

ment office to the company along the following period of time: 2011–2014.

- **Individual to company:** connection weighted with the total amount of money paid by the individual (FGPC or CDPC) to the company along the following period of time: 2010–2014.

Government offices and individuals are interpreted as affiliations. The research hypothesis is that *a similarity network of companies can be inferred from their shared affiliations and connections weights, and actual similar companies are in the same connected component in the graph.* An initial graph that mixes companies, people and government offices is generated and then a final graph of companies connections is inferred.

To validate the mentioned hypothesis, we investigated whether clusters formed by companies connections are sensitive to trustworthiness of companies. Results show evidences that trusted companies behave similarly, as well as untrusted ones, since homogeneity of connected components is high.

4. METHOD AND COMPANIES NETWORK GENERATIVE MODEL

Target network is formed by companies as nodes and similarity probabilities among companies (connections) as weighted edges. A variation of Jaccard similarity coefficient [8] was used. The original Jaccard index applied to nodes a and b of a graph could be defined according to equation 1.

$$J(a,b) = \frac{|N_a \cap N_b|}{|N_a \cup N_b|} \quad (1)$$

Where N_a and N_b are neighbors of nodes a and b, respectively.

The proposed variation states that the intersection between a and b is weighted by the similarity of connection values of nodes a and b with their shared neighbors. For example, if nodes a and b share a neighbor c, then the similarity between connections (a,c) and (b,c) is given by equation 2.

$$s((a,c),(b,c)) = 1 - \frac{|w_{(a,c)} - w_{(b,c)}|}{max(w_{(a,c)}, w_{(b,c)})} \quad (2)$$

Where $w_{(a,c)}$ and $w_{(b,c)}$ are weights of (a,c) and (b,c), respectively.

When (a,c) and (b,c) are equal and positive, the sub-expression $|w_{(a,c)} - w_{(b,c)}|$ is equal to 0. In this case, similarity between connections is 1. On the other hand when $w_{(a,c)}$ and $w_{(b,c)}$ are different and positive, similarity is proportional to difference, considering values magnitude, since this difference is normalized by the higher value $(max(w_{(a,c)}, w_{(b,c)}))$.

The complete intersection between nodes a and b, $S(a,b)$, is therefore the sum of connections similarities with shared neighbors. This is captured by equation 3.

$$S(a,b) = \frac{\sum_{i \in N_a \cap N_b} s((a,i),(b,i))}{|N_a \cup N_b|} \quad (3)$$

The maximum value of $\sum_{i \in N_a \cap N_b} s((a,i),(b,i))$ is $|N_a \cap N_b|$. This happens when weights of all shared connections of a and b are equal, so in this case equation 3 behaves as the original Jaccard similarity (equation 1).

Using equation 3 a network of companies is built from their affiliations, that is, valued relations with government offices and employees.

5. EXPERIMENTS AND RESULTS

Connections among companies indicate degrees of similarity ($]0.0, 1.0]$) regarding valued services provided to government offices and employees. Connections with weight 0.0 and companies with no connections were not included in the initial graph. We defined the σ parameter as an inferior threshold for removal of edges in the initial graph to produce the final one. This means that only nodes with weights greater than or equal to σ were kept in the final companies' network. No social network analysis [7] centrality metric, such as nodes and edges betweenness and nodes degrees, were used to increase clustering coefficient.

There are trusted and untrusted companies. We hypothesize that trusted companies have similar behavior among themselves, as well as the untrusted ones. In terms of the generated graphs and according to the mentioned hypothesis, clusters (here connected components) must be uniform regarding the presence of trusted and untrusted companies. That is, it is expected the existence of clusters formed almost entirely by trusted companies and clusters formed almost entirely by untrusted companies. This hypothesis is here called the **homogeneity hypothesis**.

We experimented four values of σ: 0.25, 0.5, 0.75 and 1.0. For information, a connection between two companies with weight 1.0 means that they behave perfectly equal, while weights more and more close to 0.0 imply to less similarity between companies.

Below there is some information about the graph for $\sigma = 0.25$:

- Number of companies: 7377
- Number of connections: 409796
- Average degree: 111.1010
- Number of connected components: 524
- Number of untrusted companies: 324

Figure 1 shows companies and their connections for $\sigma = 0.25$. Companies are nodes and connections are edges. Trusted companies are blue nodes, while untrusted companies are red. It turns out that, as stated by the homogeneity hypothesis, clusters are almost entirely uniform regarding trustworthiness of companies.

Below there is some information about the graph for $\sigma = 0.5$:

- Number of companies: 6576
- Number of connections: 304773
- Average degree: 92.6925
- Number of connected components: 716
- Number of untrusted companies: 138

Figure 2 shows companies and their connections for $\sigma = 0.5$. Homogeneity hypothesis is again visually evidenced, besides strong classes (trusted and untrusted) unbalancing.

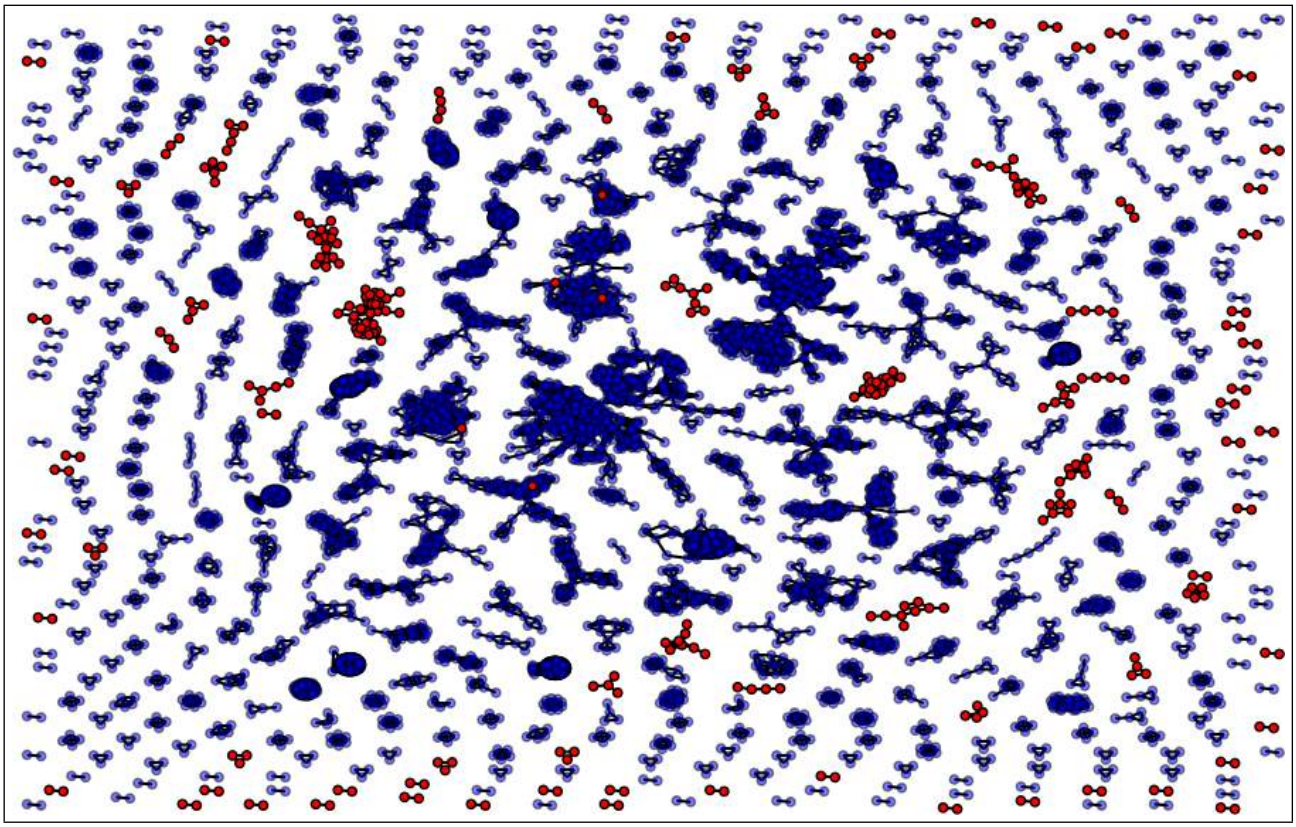

Figure 1: Network of trusted (blue) and untrusted (red) companies for $\sigma = 0.25$.

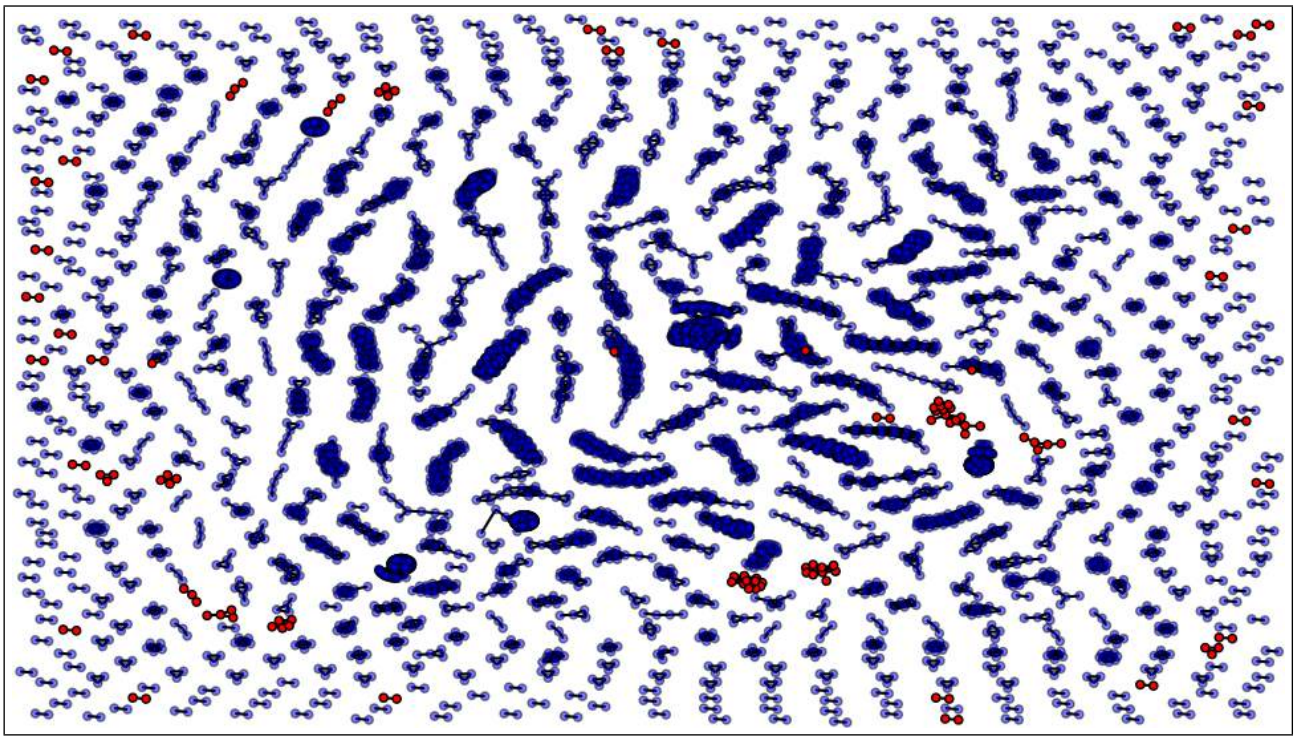

Figure 2: Network of trusted (blue) and untrusted (red) companies for $\sigma = 0.5$.

Same visual results for $\sigma = 0.75$ (figure 3) and $\sigma = 1.0$ (figure 4) in more extreme class balancing conditions.

Below there is some information about the graph for $\sigma = 0.75$:

- Number of companies: 5773
- Number of connections: 215294
- Average degree: 74.5865
- Number of connected components: 780
- Number of untrusted companies: 82

Below there is some information about the graph for $\sigma = 1.0$:

- Number of companies: 2190
- Number of connections: 121464
- Average degree: 110.9260
- Number of connected components: 386
- Number of untrusted companies: 14

We measured the classification performance for these four graphs according to the following metric.

Be C the connected components set. Each connected component C^i is a set of nodes. C^i_t is the subset of C^i that contains only trusted companies, while C^i_u is the subset of untrusted companies. Classification performance of connected component C^i, π^i, is defined according to equation 4. The max function in the numerator indicates that the connected component represents the class with more elements.

$$\pi^i = \frac{max(|C^i_t|, |C^i_u|)}{|C^i_t| + |C^i_u|} \quad (4)$$

For calculation of the global graph classification performance, Π, we summed each connected component performance weighted by its size. So, errors in larger connected components are more penalized than in smaller components. This is shown in equation 5.

$$\Pi = \frac{\sum_{i=1}^{|C|} |C^i| \ \pi^i}{\sum_i^{|C|} |C^i|} \quad (5)$$

Results of performance for different σ are shown in table 1. Classes unbalancing issues and statistical relevance of results were not addressed. Nevertheless, we argue that numerical results are consistent with visual results and encourages further investigation.

	σ			
	0.25	0.5	0.75	1.0
Π	**99.932%**	**99.939%**	**99.948%**	**99.954%**

Table 1: Classification performance results.

6. CONCLUSIONS AND FUTURE WORK

This is a preliminary work that proposes a generative model for inferring companies similarity networks based on Government expenditures data.

Despite the lack of strong statistical rigor, preliminary results and visual inspection show evidences that the generated models produced clusters (connected components) that are discriminative to trustworthiness of companies. Such discriminative clusters topology reinforces the hypothesis that there are suppliers associations, which evidences the formation of cartels. Furthermore, affiliation-based clustering reinforces the hypothesis that public agencies and agents could play an important role in establishing the legality of financial transactions. The study leads to the next steps of investigation.

Finally, since the interpretation of affiliation and derived probability of "friendship" are domain specific – in this case, the Brazilian Federal Government expenditure financial transactions –, this work could be generalized with adaptations. The same research methodology can be reused to produce valid methods for other domains.

7. REFERENCES

[1] Duit, A .: Galaz V. *Governance and Complexity: Emerging issues for governance theory.* In: Governance: An International Journal of Policy, Administration and Institutions, v. 21(3), 2008, pp. 311–335.

[2] Feitelson, D. *Experimental Computer Science: The need for a cultural change.* School of Computer Science and Engineering, Hebrew University, Jerusalem, Technical report, 2005.

[3] Van Assche, K., Beunen, R. and Duineveld, M. *Evolutionary Governance Theory: an introduction.* Springer, 2014.

[4] Barabási, A. and Albert, R. *Emergence of scaling in random networks.* Science, v. 286, 1999, p. 509–512.

[5] Stigler, G. and Friedland, C. *What Can Regulators Regulate? The Case of Electricity.* Journal of Law & Economics, v. 5(1), 1962.

[6] Ruths, J. and Ruths, D. *Control Profiles of Complex Networks.* Science, v. 343, 2014, pg. 1373–1375.

[7] KLEINBERG, J. and EASLEY, D. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, 2010.

[8] JACCARD, P. *Étude comparative de la distribution florale dans une portion des Alpes et des Jura.* In Bulletin de la Société Vaudoise des Sciences Naturelles, v. 37, pp. 547–579, 1901.

[9] Carrier, B. *A Hypothesis-based Approach to Digital Forensic Investigations.* Purdue University, CERIAS Tech Report 6, PhD Thesis, 2006.

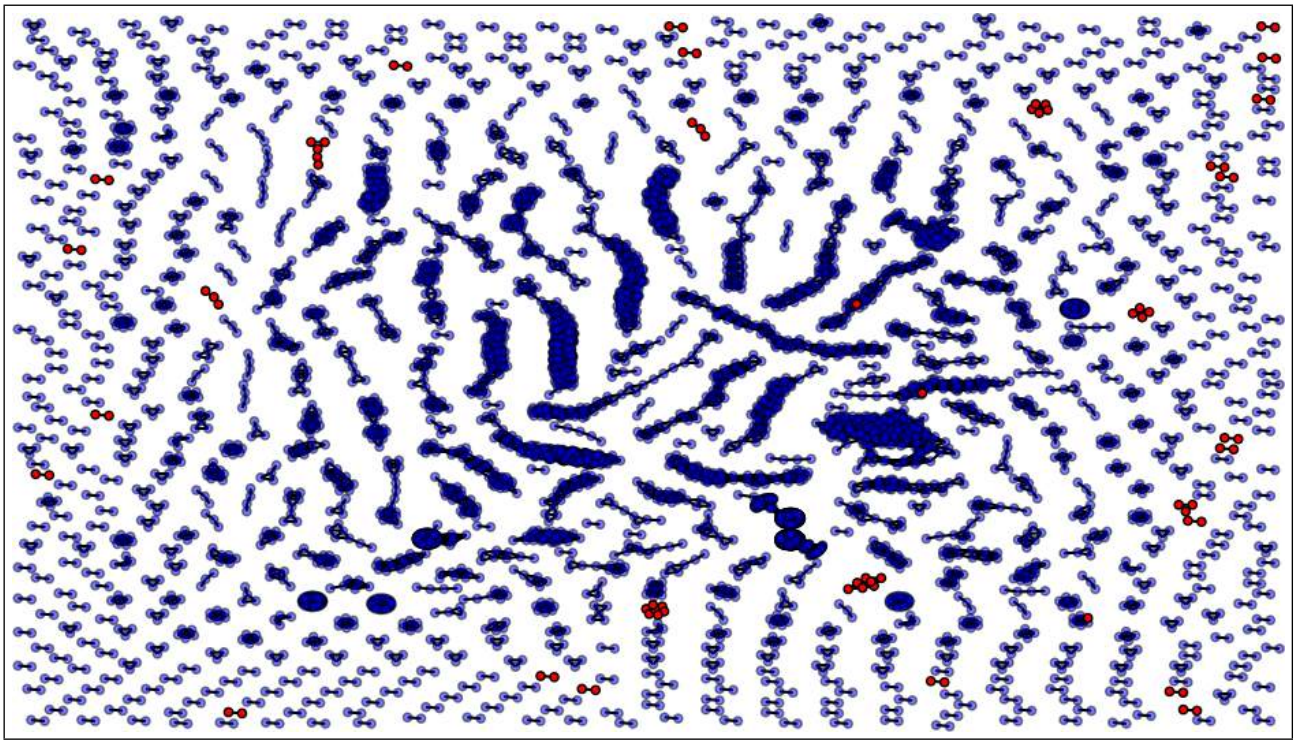

Figure 3: Network of trusted (blue) and untrusted (red) companies for $\sigma = 0.75$.

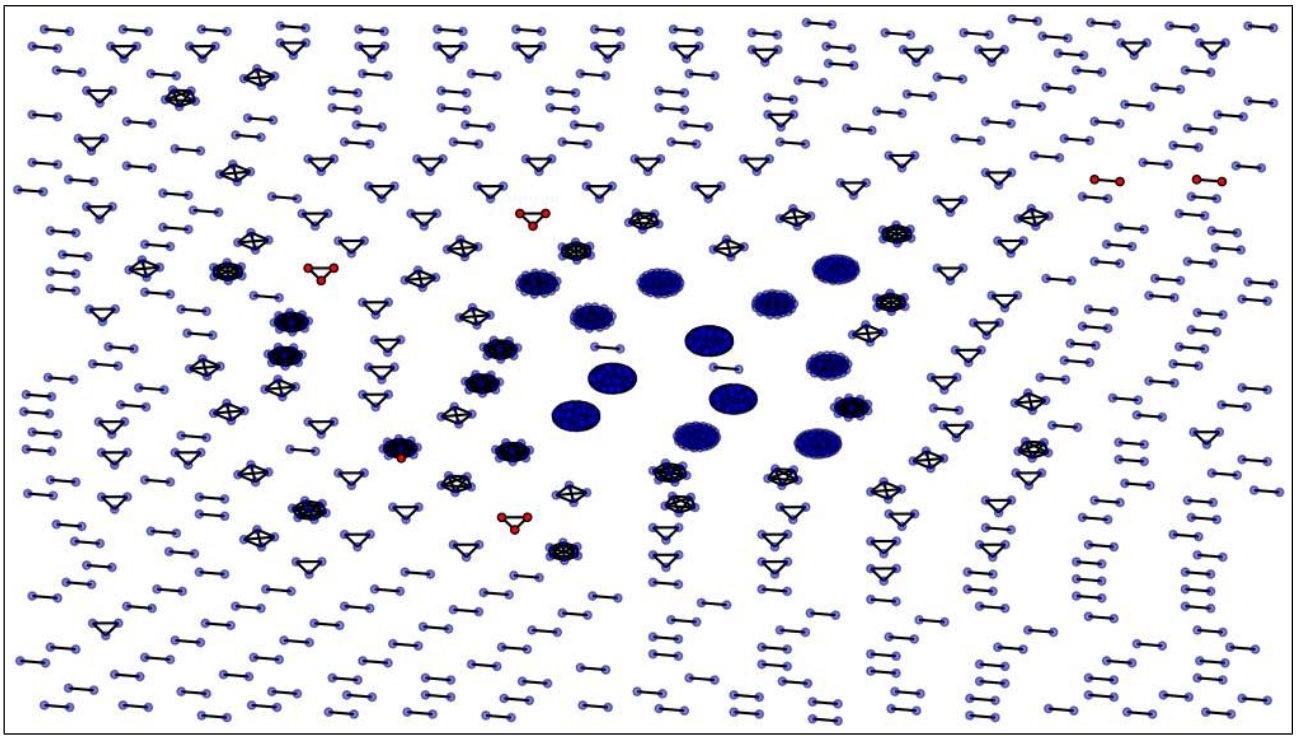

Figure 4: Network of trusted (blue) and untrusted (red) companies for $\sigma = 1.0$.